STAT 156 Final Project

David Zhang

December 2023

1 Paper Summary and Summary Statistics

We replicate "Early Childhood Education by Television: Lessons from Sesame Street" by Kearney and Levine (2019).

1.1 Research Question

Broadly speaking, the paper's purpose and results aim to contribute towards providing evidence of the causal effect of early childhood interventions on children's long-term outcomes. In particular, Kearney and Levine choose to analyze the effect of the popular kids TV show *Sesame Street.*

Sesame Street initially aired in 1969, and its main purpose was actually to reduce educational inequality experienced by disadvantaged children from differences in preschool experience. Thus, it is natural to consider the causal effect of *Sesame Street* on children's outcomes. Specifically, Kearney and Levine seek to understand and quantify the intent-to-treat effect of *Sesame Street* exposure in preschool on various educational and labor market outcomes, as data is not available with regards to the amount of individual exposure to *Sesame Street.*

1.2 Data Overview

To obtain data on *Sesame Street* broadcast coverage rates, the authors rely on data reported in the 1968-1969 edition of TV Factbook, which provides a listing of every TV station, along with technical specifications such as UHF/VHF channel[1], geographic location, height of the tower, and transmission signal power. For commercial stations in the data, TV Factbook also lists surrouding counties and coverage reates, defined as the fraction of TV households who can receive the signal. Simulated coverage rates are obtained by regressing country-level coverage rates on the technical specifications. The model is also then applied on the noncommercial stations to simulate their coverage rate.

[1] UHF - ultra-high frequency, VHF - very high frequency.

For the results that we will be replicating, the data used for the regressions are microdata from the 1980, 1990, and 2000 US Census of the Population. These are available from IPUMS-USA. The census samples are restricted to individuals born between 1959 and 1968. Additionally, observations are restricted only to those whose recorded state of residence in the census is the same as their state of birth. The authors assign county of residence in the census year to be the county of birth.

1.3 Summary Statistics

Suprisingly, the original paper does not have any summary statistic tables showing distributional characteristics (mean, median, etc.) of the census data and variables of interest. As such, we present some that we generated here for the 1980, 1990, and 2000 census data and the relevant covariates. Variables with a (Y) are outcomes of interest; the others are the treatment variables. More details on the covariates are in the models section.

Table 1: 1980 Census Summary Statistics

Statistic	Mean	St. Dev.	25th Perc.	Median	75th Perc.	Max	N
gradeage (Y)	0.798	0.402	1	1	1	1	715,458
preschlcov	0.342	0.366	0.000	0.205	0.733	0.940	715,458
covst6768	0.142	0.289	0.000	0.000	0.000	0.940	715,458
covst69	0.074	0.221	0.000	0.000	0.000	0.940	715,458
covst7072	0.212	0.333	0.000	0.000	0.479	0.940	715,458
covst7374	0.131	0.280	0.000	0.000	0.000	0.940	715,458

Table 2: 1990 Census Summary Statistics

Statistic	Mean	St. Dev.	25th Perc.	Median	75th Perc.	Max	N
hsdrop (Y)	0.143	0.350	0	0	0	1	667,530
hsgrad (Y)	0.325	0.468	0	0	1	1	667,530
anycol (Y)	0.532	0.499	0	1	1	1	667,530
preschlcov	0.323	0.360	0.000	0.000	0.687	0.940	667,530
covst6768	0.140	0.286	0.000	0.000	0.000	0.940	667,530
covst69	0.069	0.212	0.000	0.000	0.000	0.940	667,530
covst7072	0.199	0.324	0.000	0.000	0.457	0.940	667,530
covst7374	0.124	0.272	0.000	0.000	0.000	0.940	667,530

Table 3: 2000 Census Summary Statistics

Statistic	Mean	St. Dev.	25th Perc.	Median	75th Perc.	Max	N
lnhrwage (Y)	2.714	0.678	2.312	2.724	3.101	6.908	458,043
working (Y)	0.887	0.316	1	1	1	1	458,043
inpov (Y)	0.053	0.224	0	0	0	1	458,043
preschlcov	0.324	0.358	0.000	0.000	0.687	0.949	458,043
covst6768	0.140	0.284	0.000	0.000	0.000	0.949	458,043
covst69	0.067	0.209	0.000	0.000	0.000	0.949	458,043
covst7072	0.201	0.322	0.000	0.000	0.463	0.949	458,043
covst7374	0.123	0.271	0.000	0.000	0.000	0.949	458,043

Explaining the names of the outcomes:

1. gradeforage: Indicator for grade-for-age status. This is the primary outcome of interest of the paper, and is investigated in Table 4.

2. The following are investigated in Table 6:

   (a) hsdrop: Indicator if the individual dropped out of high school.

   (b) hsgrad: Indicator if the individual graduated high school.

   (c) anycol: Indicator if the individual went to any college.

   (d) lnhrwage: Logarithm of the individual's wage.

   (e) working: Indicator if the individual is working

   (f) inpov: Indicator if the individual is in poverty.

We also replicate parts of Table 3 in the original paper, which describes some statistics from the 1970 census based on if the observations are not in the relevant census data, in weak reception, or in strong reception areas[2].

	Perc. FM Household	Perc. Low Income	Perc. No HS Degree	Perc. Black	Mean Family Income	Mean Unemp. Rate	n
Not in census	8.922%	32.841%	56.918%	9.131	$7174.527	4.526%	73,442,763
Low Coverage	10.255%	18.114%	46.037%	9.531	$9852.780	4.310%	66,915,590
High Coverage	10.101%	17.962%	43.511%	8.788	$9927.575	4.522%	59,182,563

Table 4: Demographic Characteristics of 1970 Census

1.4 Results

From a graphical analysis, Kearney and Levine find that those living in areas with more *Sesame Street* coverage are 1.5 to 2 percentage points more likely for children to be at the appropriate grade level for their age as a whole (referred to as grade-for-age).

For their econometric analysis, they estimate two equations. Estimates from equation 1 suggest that an increase in coverage rates (of *Sesame Street*), similar to moving from an area with no coverage to an area with coverage, would cause a 3.2 percentage point increase in the rate of grade-for-age status. Across the whole data, this result is statistically significant, and is larger for boys and largest for black, non-Hispanic children. However, the results are not statistically significant across ethnic groups. Simulation results suggest that the introduction of *Sesame Street* for white and black, non-Hispanic children would have similar effects on elementary school performance as Head Start participation[3].

[2] Perc. - percent. FM Household - Female-headed household. HS - high school.

[3] Head Start is a federally funded, nationwide preschool program for poor children in the U.S.

The results for equation 2 suggest that those who started school before the introduction of *Sesame Street* have no significant difference in outcomes. However, those that started school after the introduction of *Sesame Street* show a strong increase in grade-for-age status, providing evidence for a causal effect.

Using the same equation forms as earlier, Kearney and Levine also consider the effect on later life outcomes, particularly high school graduation, college attendance, wage/socioeconomic status, and employment. Generally, they find small effects that are not statistically significant, suggesting little to no causal relationship.

2 Empirical Specifications

We now go over the method of identifying the intent-to-treat effect of *Sesame Street.*

2.1 Calculating *Sesame Street* Coverage Rates

The first step of Kearney's and Levine's method of identifying the causal effect lies in their calculation of the coverage rates of *Sesame Street.* As alluded to earlier, the coverage rates are simulated using the TV Factbook data. Although coverage rates alone are also a function of the share of households with a UHF tuner and other confounding factors, simulating the coverage rates by regressing them on only technological factors effectively acts as an instrumental variable, making the use of the simulated coverage rates robust to confounding. The simulation is done with OLS regression, with the following regression equation:

$$\begin{aligned} CoverageRate = \beta_0 + \beta_1 Distance + \beta_2 UHF + \beta_3 Distance \times UHF \\ + \beta_4 HeightAboveGround + \beta_5 VisualPower. \end{aligned}$$

2.2 Estimated Econometric Models

Kearney and Levine estimate two models:

$$\begin{aligned} Outcome_{ijc} &= \beta_0 + \beta_1(preschool69_{ic} \times SSCov_j) + \beta_2 Policy_{jc} \qquad (1)\\ &+ \beta_3 X_{ijc} + \gamma_c \times \gamma_s + \delta_j + \epsilon_{ijc} \\ Outcome_{ijc} &= \beta_0 + \sum_{c=1967/68}^{1973/74} \beta_c \times (\gamma_c \times SSCov_j) + \beta_2 Policy_{jc} \qquad (2)\\ &+ \beta_3 X_{ijc} + \gamma_c \times \gamma_s + \delta_j + \epsilon_{ijc}. \end{aligned}$$

$Outcome_{ijc}$ represents the various educational/labor market outcomes referenced earlier for individual i, in county j, in cohort c. Subscript s indexes states. $preschool69_{ic}$ is an indicator variable for being preschool age when *Sesame Street* began in the fall of 1969. $SSCov_j$ represents *Sesame Street* coverage rates. $Policy_{jc}$ includes controls for policy changes that could result in omitted variable bias: they include an indicator for the introduction of the Food Stamp Program, and expenditures for Head Start in 1968 and 1972.

δ_j represents county fixed effects, which controls for time-invariant differences across counties. $\gamma_c \times \gamma_s$ represent state × birth cohort fixed effects controlling for time-varying changes in outcomes across states. X_{ijc} represents additional covariates used as controls, including but not limited to race/ethnicity, age, and socioeconomic status (when available). Standard errors are clustered at the county level, and are estimated via clustered bootstrap.

Equation (1) estimates a difference-in-difference specification between those who start preschool in 1969 and those who don't: β_1 is the causal effect of interest.

Equation (2) relaxes pre-assignment of treatment and control groups by cohort. The effect of *Sesame Street* coverage is allowed to vary by grouped birth cohort, providing a specification check to see if the "treatment effect" really starts with the appropriate groups of birth cohorts. The choice of groupings for cohorts is not arbitrary: Kearney and Levine group them into two year intervals to increase power[4], and leave 1969 as its own cohort. Thus, c takes on the possible values of $c \in \{1967/68, 1969, 1970/72, 1973, 74\}$.

2.3 Assumptions for Identification and Criticisms

The primary assumption for identification of the intent-to-treat effect is no omitted variable bias. Mathematically, no omitted variable bias implies that

$$\mathbb{E}[\epsilon_{ijc} \mid Z_{ijc}] = 0, \forall i, j, c$$

[4] Kearney and Levine do not explicitly explain how this increases power.

for each of the model specifications, where Z_{ijc} is short-hand for all of the covariates included for each of regressions. In plain English, this is saying that knowing the covariates in the model removes all outside effects from the relationship of interest other than random noise, implying that the coefficients of the linear model are unbiased for the true values. This is similar to the ignorability assumption. This is a strong assumption which is untestable. However, Kearney and Levine collect a rich set of covariates that control for many of the intuitive and likely most significant confounding factors, and the fixed effects specification controls for many other things not collected in the data.

The above assumption is implied by the models. Additionally, Kearney and Levine explicitly/implicitly impose some other assumptions about the population and methods they are considering, and has some issues regarding inference:

1. Using the simulated coverage rates aims to isolate the effect of coverage rates only due to technology, and implicitly assumes that technological specifications of the towers are uncorrelated with confounders. However, it could be the case that the technological specifications of the towers is correlated with wealth of the county, which could be possible. In this case, the estimates of the causal effect are biased, in the cases where socioeconomic status cannot be controlled for.

2. Rural locations are not included in the analysis, so the comparisons are only among those living in metropolitan areas. This is driven by the fact that distance from the towers matters, and rural areas tend to be farther away from the towers. These observations are dropped from the analysis, and so there are problems of external validity.

3. This paper suffers from multiple inference, and the authors do not adjust for it. For each table we replicate in this paper, we look at 5 coefficients, representing the combined number of coefficients from each model. For one set of regressions with both models, the probability that we make at least one type 1 error (falsely reject the null) is $1-(0.95)^5 \approx 0.226$ with a significance level $\alpha = 0.05$. The Bonferroni correction yields a family-wise significance level of $\alpha_f = 0.01$. At this significance level, several results presented are no longer statistically significant. Particularly in Table 4, 30 comparisons are done, yielding a family-wise $\alpha_f = \frac{0.05}{30} \approx 0.0016$. Under this Bonferroni correction, many of the results are not significant at the family-wise level.

4. The original paper has a limitation in that it cannot capture specific mechanisms of the long-term effects of educational interventions. Educational interventions may be effective in improving narrower measures of academic achievement and/or enhancing a child's socio-emotional development. Unfortunately, the Census data does not allow for the authors to distinguish between these types of intermediate outcomes, as there are no measures of noncognitive outcomes available.

5. Many of the outcomes are binary, but the authors estimate them using linear models instead of something like logistic regression. Although this is likely for interpretation and inference purposes, this likely results in some inaccurate interpretation of coefficients up to a certain point, since the linear model is not bounded between 0 and 1.

3 Result Replication

We first discuss the bootstrap methodology used throughout the paper.

3.1 Bootstrap Methodology

Due to the use of simulated coverage as one of the key covariates and the clustered nature of the data, Kearney and Levine use a clustered bootstrapping procedure to estimate the standard error of each coefficient. At a high level, their bootstrapping algorithm for a single repetition is as follows:

1. Randomly sample clusters from the original data used to simulate coverage, where the number of clusters sampled is the number of unique clusters. The cluster is the station ID (referred to as `stationid`). Denote this as the stage 1 sample.
2. Simulate the coverage on the stage 1 sample using OLS. Then,
   (a) If the simulated coverage (`covratehat`) is negative or the distance is greater than 200 miles (`distance > 20`), set it to 0.
   (b) For each unique identifier (a combination of state and city code referred to as `statecty`), set the `covratehat` to the maximum value for that identifier.
3. For each group or response (as seen in Table 6) variable:
   (a) Filter the grouped census data only for observations without NA values in relevant covariates and responses. Drop all of the coverage variables already recorded in the census data
   (b) Take a clustered bootstrap sample from the census data in the same way as the stage 1 sample, with the cluster being an identifier variable referred to as `dmaindex`.
   (c) Perform a many-to-one merge with the stage 1 sample on `statecty`, and only keep rows where there is a match.
   (d) Recreate the coverage variables used as covariates, and re-estimate Equation (1) and Equation (2).
   (e) Store the coefficients in a matrix.

Kearney and Levine repeat this 400 times, and then calculate the standard deviation with respect to the coefficients in each group. Unfortunately, due to the size of the data and complexity of the algorithm, it is too computationally intense to replicate on our computers, even with parallelization and Kaggle notebooks, which provide significantly more memory. Thus, we implement a smaller and more compromised version of their bootstrapping algorithm to estimate standard errors, although they are likely not valid to use for specific inference. The alternative algorithm for one repetition is as follows:

1. Instead of taking a clustered bootstrap sample, we proceed by proportion subsampling[5].
    (a) We first choose a subsample size n. For these results, we use $n = 30,000$.
    (b) For the kth cluster that we consider in the data, we compute the proportion of the total data that is in it, denoted as p_k.
    (c) For each cluster k, randomly sample with replacement $p_k \times n$ data points, rounding to the nearest integer. This ensures that the subsample maintains a similar cluster structure as the original data.
2. For each group:
    (a) Instead of re-simulating `covratehat`, we compromise by randomly permuting `covratehat` within the group. When we consider different response variables as in Table 6, we randomly permute `covratehat` before drawing a sample, since there are no covariates (race, sex) that we care about that may possibly negatively affect the permutation. We do this to preserve the randomness that is inherent in the generation of `covratehat`, although this overestimates the variance, since permuting `covratehat` is a significantly more heavy-handed method of resimulating the coverage.
    (b) With the permuted `covratehat`, we recalculate the coverage variables.
    (c) We then estimate the regressions, and store the coefficients in a vector. This vector is then appended to an empty list.

We repeat this process 200 times, generating 200 lists consisting of vectors of coefficients. We then ”squash” these lists by row-binding the vectors into matrices index-wise, creating a list of matrices with as many matrices as groups. We then calculate standard errors of the coefficients column-wise. This method is computationally feasible, but is clearly not equivalent to the authors' bootstrap method. As such, we do not interpret these standard errors for inference, but rather use them to get an idea of the relative variance in the estimates across groups, and also compare their relative magnitude to the relative magnitude of the standard errors in the original paper.

[5] When we consider different response variables as in Table 6, we randomly permute `covratehat` before drawing a sample.

3.2 Replicated Tables

We replicate Table 4, Table 5, and Table 6 of the paper. We first discuss some terminology used in the tables:

1. First, the reports for the coefficients are for the entire dataset. We run regressions across the entire data, and only bootstrap the standard errors.
2. For each table, there are two tables to represent each equation estimated. For example, Table 4a refers to the results from estimating Equation (1), and Table 4b refers to the results from estimating Equation (2).
3. β_1 refers to the coefficient on $preschool69 \times SSCov$ from Equation (1). β_c, where $c = 1967/68 \dots 1973/74$, refers to the coefficients of interest for the cohorts in Equation (2).
4. GFA - grade-for-age (indicator on whether a student is in their grade for age or not)
5. Reported sample size is the one used to calculate the coefficients of interest. They **are not** used to calculate the standard errors (S.E.) due to computational restraints.

3.3 Table 4

Table 4 is the central result of the paper: it uses 1970 census data.

Table 5: Table 4a (Aggregate Effect)

	All	Boys	Girls	White	Black	Hispanic
β_1	0.103	0.102	0.102	0.060	0.186	0.120
$S.E.$	0.092	0.142	0.126	0.117	0.183	0.298
Sample Size	715,458	359,548	355,910	512,178	132,828	61,283
Mean GFA	0.798	0.761	0.835	0.832	0.703	0.711

Table 6: Table 4b (Event Study)

	All	Boys	Girls	White	Black	Hispanic
$\beta_{1967/68}$	-0.051	-0.031	-0.063	-0.034	-0.045	-0.097
$S.E._{1967/68}$	0.078	0.127	0.100	0.090	0.234	0.327
β_{1969}	0.029	-0.012	0.069	0.058	0.027	0.004
$S.E._{1969}$	0.157	0.242	0.208	0.201	0.309	0.527
$\beta_{1970/72}$	0.074	0.078	0.069	0.029	0.183	0.058
$S.E._{1970/72}$	0.103	0.164	0.144	0.139	0.205	0.372
$\beta_{1973/74}$	0.114	0.114	0.116	0.103	0.153	0.133
$S.E._{1973/74}$	0.136	0.205	0.189	0.186	0.259	0.442
Sample Size	715,458	359,548	355,910	512,178	132,828	61,283

Table 4a and Table 4b correspond to the top and bottom half of Table 4, respectively. The first thing that we can notice is that the sample size for each of the groups is exactly the same as those in the paper, as well as the mean GFA. Interestingly, although the sample size suggests that we have cleaned the data exactly the same way as the authors have, we have ended up with slightly different estimates for β_1, and drastically different estimates for coefficients in Table 4b.

Table 4a We first start with the interpretation of the coefficients. For a 10 point increase in coverage rates of *Sesame Street*, the coefficient implies that the rate of occurrence for grade-for-age status increases by $0.1 \times 0.103 \times 100 = 1.03$ percentage points on average, for those who were preschool age in 1969. The interpretation of the other coefficients is the same, except conditioning on the particular group.

This is very close to Kearney and Levine's estimate of $\beta_1 = 0.105$. Differing from Kearney and Levine, we find that boys and girls experience the same effect from *Sesame Street* coverage, since the coefficient is the same for both groups ($\beta_1^{Boys} = \beta_1^{Girls} = 0.102$). The coefficient for white, non-Hispanic children is roughly the same as in the paper, which is 0.068. For black and Hispanic children, we note that our estimates for the effect of *Sesame Street* are noticeably higher. In particular, based on the standard errors reported by Kearney and Levine, our estimate for β_1^{Black} is around 1.7 standard errors higher than the one estimated by Kearney and Levine. The estimated effect for Hispanics is around 0.6 standard errors higher than the one estimated by Kearney and Levine. Note that the mean GFA calculated per group is exactly the same as in the original paper, indicating that potential differences in estimates are very likely not a result of a mix-up of groups.

Although we can't really interpret the standard errors for inference, we can try to interpret them in terms of relative magnitude for the variance of the estimate. The $S.E.$ for all, boys, and girls are roughly the same in terms of magnitude when also considering the variance of the alternative bootstrap procedure, which is similar to Kearney and Levine, suggesting that the variance is relatively similar. While the $S.E._{White}$ from Kearney and Levine is roughly half of the ones from All, Boys, and Girls; we find that it is relatively similar. For black and hispanic children, the $S.E.$ is quite high. This is likely due to the low proportion of black and hispanic children in the dataset, which is magnified even more since we only sample $30,000$ total observations. Similar to Kearney and Levine, the $S.E.$ for Hispanics is the highest, more than double those of All, Boys, and Girls.

With regards to the coefficient estimates in Table 4a, we have thought of some possibilities as to why the coefficients differ. Perhaps one of the more likely reasons is that there is some difference in the algorithms for fixed effects regression in R and in Stata. In R, we estimate the fixed effects regressions using the `plm` function within the `plm` package. `plm` uses the within/demeaning transformation to the data, and then runs the standard `lm` in `R`. On the other hand, `reghdfe` in `Stata` performs the within/demeaning transformation, and uses an iterative method known as FGLS[6] to estimate the

[6] Feasible Generalized Least Squares

coefficients. The details are out of scope for this paper, but on first impression seem significantly different enough to cause the small differences that we observe. Experiments trying the same type of regression using the `R` package `fixest` also provide different estimates, aligning with this thought. Another likely reason could be an error in the data cleaning. We experienced several significant bugs during the procedure of replicating the results involving the calculation of relevant covariates, but managed to fix them. It is highly likely that fixing these bugs introduced other, more difficult to notice bugs in the data that are influencing the estimation of the coefficients. However, we are unaware of any possible errors, if there are any at all.

Table 4b Upon first inspection, Table 4b shows significantly different results across the board, compared to Kearney and Levine. We find that for the first row, the coefficients for the first row are much further from the results of Kearney and Levine. First, all of the coefficients are negative in the first row. $\beta_{1967/68} = -0.051$ is much different from the estimate of -0.002 from Kearney and Levine. Although the sign is the same, the magnitude of ours is much higher. For the other groups in the first row, the magnitude is somewhat close, although the sign is different. However, we argue that this is not very concerning, since the coefficients are still very close to 0, which is the expected result. For the second row, the coefficients differ significantly as well: β_{1969} suggests that boys are negatively affected by *Sesame Street* exposure, although this coefficient is small and close to 0. This does not align with the results of Kearney and Levine, since they estimate $\beta_{1969} = 0.085$, which is much higher than ours. However, their estimate is not significant at the 5% level: it is approximately 1.89 standard errors above 0. The last two rows reflect similar results as Kearney and Levine and similar results to Table 4a, in that the effect of *Sesame Street* is higher for those that have more opportunity to be exposed, compared to those that are preschool age right when it starts and those who were preschool age before its release. Generally speaking, although the numbers do not match up exactly, the estimates still hold the same general pattern in that the effect of *Sesame Street* is close to 0 for the first cohort, and then becomes positive and increases for future cohorts, passing the specification check.

Our results suggest that the effect observed in Table 4a is mostly a result of the effect of *Sesame Street* on future cohorts. This makes sense. A child that goes to preschool in 1969 right when *Sesame Street* released will have much less exposure compared to future cohorts that will grow up with *Sesame Street*. We therefore expect the effect of *Sesame Street* exposure to be much stronger on those future cohorts.

When also considering the variance of the alternative bootstrap procedure, the relative magnitude of the standard errors matches up with the relative magnitude of the standard errors in the paper. The standard errors of All, Boys, Girls, and White stays relatively the same in magnitude, and we see that the standard errors significantly increase for the Black and Hispanic groups, likely due to the smaller proportion of observations.

There are several things wrong with the replication of Table 4b. The best explanation is likely to be something wrong with the data cleaning procedure. We believe that the differences between estimates is too high to be explainable by something like the difference in estimation algorithms, unless either of the algorithms happens to be somewhat more unstable for coefficients close to 0, as the estimates from Kearney and Levine are quite close to 0 for the first two rows. The primary issue has to do with the 1967/68 and 1969 cohorts, since the estimates for these cohorts are off the most. It is likely that there is an issue with the data cleaning procedure that incorrectly modifies the covariates of these cohorts, since we had many data cleaning bugs specifically around the covariates for the event study approach.

3.4 Robustness Check: Table 5

Table 7: Table 5a (Aggregate Effect)

	All	Boys	Girls	White	Black	Hispanic
β_1	0.049	0.143	-0.055	0.030	0.251	-0.164
$S.E.$	0.093	0.131	0.125	0.109	0.198	0.376
Sample Size	199,102	10,722	14,905	18,937	4907	1524

Table 8: Table 5b (Event Study)

	All	Boys	Girls	White	Black	Hispanic
$\beta_{1967/68}$	0.023	0.014	0.029	0.039	0.027	-0.106
$S.E._{1967/68}$	0.097	0.170	0.131	0.110	0.281	0.427
β_{1969}	0.044	0.131	-0.044	0.042	0.205	-0.373
$S.E._{1969}$	0.189	0.280	0.248	0.223	0.380	0.750
$\beta_{1970/72}$	0.056	0.150	-0.052	0.073	0.318	-0.426
$S.E._{1970/72}$	0.127	0.178	0.164	0.147	0.265	0.436
$\beta_{1973/74}$	0.072	0.181	-0.039	0.000	0.235	0.062
$S.E._{1973/74}$	0.156	0.228	0.193	0.199	0.345	0.582
Sample Size	199,102	10,722	14,905	18,937	4907	1524

The data used for Table 5 uses census data from 1970, and looks at individuals who were preschool age in 1959 to establish a sort of "placebo group" to compare to. This implies that there should not have been a causal effect on school performance for these individuals, since *Sesame Street* first broadcasted in 1969. As such, we should expect to see many coefficients very close to 0, and Kearney and Levine find this exact result. This serves as a robustness check in order to see if there are any differential county trends that could bias the estimates in table 4. If there are any noticeable differences in outcomes for those in the 1970 census, it suggests a bias on the estimates in table 4, caused by county trends. Passing this robustness check provides strong evidence of a causal relationship in table 4.

However, looking at Table 5a and Table 5b, we can immediately notice that the results differ significantly from the original paper. However, the results in the table suggest that there is in fact a causal effect of coverage on the children in the 1970 census.

Let's first discuss the implications under the assumption that we have done everything else correctly. Note that the sample sizes in each group match the numbers in the original paper as a slight robustness check. We have that $\beta_1^{All} = 0.049$, which although quite small, is different enough from 0 to warrant some attention. Additionally, the standard error is quite low in magnitude even with the alternative bootstrap procedure, suggesting more that this is a significant result. The same holds for other groups as well. The fact that these coefficients are far from 0 suggests that there are trends within counties that are biasing the estimates, i.e. there are unmeasured confounders and/or omitted variable bias, violating the primary assumption for identification. Thus, this implies that the results in Table 4a and Table 4b are biased, although the direction of the bias is unclear. In this case, we fail the robustness check, and depart from the results of Kearney and Levine, suggesting biased estimates in Table 4.

However, it is more likely that there is something wrong with the data cleaning/estimation process. The biggest indicator is that within each row, the coefficient estimates vary quite wildly. We can see that in Table 5a, the coefficient for Girls and Hispanic is negative while the others are positive. Additionally, we can see that conditioning on ethnicity yields wildly different estimates. If we only consider White, we get something that seems to make sense, i.e. something close to 0. For Black, we get a large positive causal effect, and a large negative effect for Hispanic. This pattern also holds for Table 5b in an even more pronounced manner for some cohorts, particularly for cohort $c = 1970/72$. This variance also suggests a presence of a lot of noise, which is suggestive of a non-statistically significant result. Unfortunately, we cannot rely on the standard errors to give us an idea as to whether or not these coefficients are significant or not, but we at least have more of an indication that there is something wrong with the data rather than an actual confounding problem. Additionally, further inspection of Table 5 in the original paper shows similar behavior with the coefficients in regards to the variance of the estimates, but again we cannot know for sure due to the standard errors. We pass through the robustness check uncertain.

3.5 Table 6

Table 6 looks at census data from 1990 and 2000, and tries to see if *Sesame Street* exposure has any positive effect on longer term outcomes. Since those that start preschool in 1969 or later will be around high school or college age, the authors examine the effect of *Sesame Street* exposure on the likelihood of dropping out or graduating high school, and attending college.

Although not completely exact, our results for Table 6a closely match with the results in the original paper. Kearney and Levine do not find large or statistically significant effects at all. The only difference we have is that they find that the effect on employment is statistically significant: their estimate is 2.25 standard errors away from 0. Their estimate of $\beta_1^{Employment} = 0.027$ implies that a 30 point increase in coverage rates would result in a $0.3 \times 0.027 \times 100 = 0.81$ percentage point increase in employment,

Table 9: Table 6a (Aggregate Effect)

	1990 Census			2000 Census		
	HS Dropout	HS Grad	Any College	log hourly wage	Employed	In Poverty
β_1	0.005	0.002	-0.007	0.023	0.001	-0.005
$S.E.$	0.021	0.029	0.028	0.039	0.020	0.015
Sample Size	667,530	667,530	667,530	458,043	458,043	458,043
Mean rate	0.143	0.325	0.532	2.714	0.887	0.053

Table 10: Table 6b (Event Study)

	1990 Census			2000 Census		
	HS Dropout	HS Grad	Any College	log hourly wage	Employed	In Poverty
$\beta_{1967/68}$	-0.014	-0.003	0.017	0.024	-0.018	-0.001
$S.E._{1967/68}$	0.032	0.047	0.049	0.067	0.033	0.022
β_{1969}	-0.019	0.005	0.015	0.023	-0.014	0.004
$S.E._{1969}$	0.044	0.059	0.062	0.079	0.037	0.027
$\beta_{1970/72}$	-0.012	-0.002	0.014	0.039	-0.010	-0.001
$S.E._{1970/72}$	0.029	0.042	0.046	0.060	0.030	0.020
$\beta_{1973/74}$	0.008	0.007	-0.015	0.034	-0.006	-0.011
$S.E._{1973/74}$	0.036	0.047	0.051	0.066	0.032	0.023
Sample Size	667,530	667,530	667,530	458,043	458,043	458,043

which is negligible. Our result implies an essentially ”infinitesimal” increase in employment. One thing to note is that the mean rate is not the same as the original paper, specifically in the Employed and In Poverty columns. Since our results differ the most in these columns, this is likely the reason why.

For this table, we also note that the standard error estimates are within the same order of magnitude as the ones reported in the original paper, although they are significantly higher than those in the original paper. They follow the same general pattern as the standard errors in the original paper, when also considering the additional variance from the alternative bootstrapping. In general, we have the same qualitative results for Table 6a, and slight differences could be chalked up to differences in estimation algorithms as well.

The same holds generally for Table 6b. All coefficients are small and quite close to 0, and we again note that the standard errors are in the same order of magnitude. This suggests that the effect of *Sesame Street* does not carry on into later life outcomes, aligning with the results of Kearney and Levine.

4 Re-analysis of Results

For the re-analysis, we take another look at Equation (1) and try to estimate the coefficient of interest using the AIPW/Doubly Robust estimator, using the function made in Chapter 12. We make some slight changes. Since we are unsure how to generalize methods learned to continuous treatments, we binarize the continuous treatment. Specifically, we define the treatment variable to be an indicator which equals 1 if $preschool_{69} \times SSCov$ is above the average value for an observation, and 0 otherwise. This will also slightly touch on Figure 5 in the original paper, which we do not replicate.

We argue that this method can be applied in this setting, since we have a rich set of covariates. The justification of the use of the AIPW estimator aligns with the authors' own justifications of their own model specifications. In particular, we assume and believe that conditional on the covariates X, we have that ignorability holds for a treatment Z:

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} \mid X.$$

In other words, conditional on the covariates, the treatment is independent of the potential outcomes, and is as good as randomly assigned.
To ensure the strong overlap condition,

$$0 < \alpha_L \leq e(X) \leq \alpha_U < 1,$$

where $e(X) = Pr(Z = 1 \mid X)$ denotes the propensity score and α_L, α_H denote lower and upper bounds, we truncate our observations for $e(X) \in [0.05, 0.95]$, as suggested by Kurth et al. (2005). To estimate the standard error, we run the same alternative bootstrapping procedure, except we randomly permute the treatment instead of permuting `covratehat`, since this is equivalent to randomly permuting `covratehat` and recalculating the treatment assignment variable.

4.1 Table 4a (DR)

Table 11: Table 4a (DR)

	All	Boys	Girls	White	Black	Hispanic
$\hat{\tau}_{DR}$	0.169	0.231	0.197	0.224	0.054	0.403
$S.E.$	0.006	0.010	0.009	0.006	0.017	0.032
Sample Size	715,458	359,548	355,910	512,178	132,828	61,283
$Pr(Z = 1)$	0.491	0.498	0.483	0.478	0.520	0.526

We compare the results from the original paper to Table 4a (DR). An initial look at the last row shows that the treatment/control groups are well-balanced by cutting at the mean, indicating that the distribution is roughly symmetric across groups.

The first thing to notice is that the standard errors have decreased significantly and are extremely small, despite the high variance of the alternative bootstrapping procedure. This is a strength of the AIPW estimator, since it residualizes the outcomes by the outcome estimator, reducing the variance. However, these standard errors are still likely not valid for inference, since the bootstrap procedure does not perfectly preserve the clusters. However, we can still note that the variance has significantly reduced compared to the variances in Table 4a. Assuming that the standard errors are still conservative and valid for inference, all of the estimates for the treatment effect are highly significant and robust to multiple inference compared to the original paper.

With regards to the estimates of the causal effect, we recall that at its core, $\tau_{DR} = \mathbb{E}[Y(1) - Y(0) \mid X]$, and $\hat{\tau}_{DR}$ is an estimator for this. In this case, since the potential outcomes are binary (grade-for-age), we interpret the causal effect as $\tau_{DR} = Pr(GFA = 1 \mid Z = 1, X) - Pr(GFA = 1 \mid Z = 0, X)$, which can be interpreted as a risk difference. We can see that for across the entire data, a child exposed to *Sesame Street* with above average coverage is 16.9 percentage points more likely to have grade-for-age status, compared to children in below average coverage, on average. This is significantly higher than the result reported in the original paper: for a 30 point increase in coverage rates, which is tantamount to moving from a below average coverage area to an above average coverage area for a typical person as claimed by Kearney and Levine, they only report a 3.2 percentage point difference. We find that the results are significantly more pronounced like this across the board, particularly for Hispanics. Interestingly, the causal effect only changes marginally for black, non-Hispanics: the original paper's results suggest that moving from low (below average) to high (above average) coverage only results in a 3.15 percentage point increase in the rate of grade-for-age for black, non-Hispanics: the AIPW estimator finds it to be 5.4.

4.2 Table 5a (DR)

For Table 5a (DR), we first note that the estimated treatment effects are far from 0, which is what we do not expect, and is what Kearney and Levine do not find. Additionally, the AIPW shrinks the variance such that all of these are highly significant. Additionally, the effect on Hispanics is negative and larger than 1, which makes no sense in this context. These results further corroborate with the results from Table 5, indicating that there is likely a problem with the data cleaning procedure. We again fail the robustness check.

Table 12: Table 5a (DR)

	All	Boys	Girls	White	Black	Hispanic
$\hat{\tau_{DR}}$	0.066	0.045	0.244	0.326	0.397	-1.672
$S.E.$	0.007	0.010	0.010	0.006	0.021	0.062
Sample Size	199,102	98,977	100,125	157,779	28,712	10,917
$Pr(Z = 1)$	0.566	0.578	0.554	0.556	0.610	0.594

Table 13: Table 6a (DR)

	1990 Census			2000 Census		
	HS Dropout	HS Grad	Any College	log hourly wage	Employed	In Poverty
$\hat{\tau_{DR}}$	-0.002	-0.007	0.009	0.072	0.007	-0.020
$S.E.$	0.004	0.006	0.005	0.007	0.004	0.003
Sample Size	667,530	667,530	667,530	458,043	458,043	458,043
$Pr(Z = 1)$	0.473	0.473	0.473	0.476	0.476	0.476

4.3 Table 6a (DR)

As a first note, Table 6a (DR) does not change very significantly from the results in Table 6a. All of the estimated effects are small and close to 0. In fact, some of them are even closer to 0 than the original paper. One interesting thing to note is that although the coefficients are not significant, the sign on HS Dropout is negative while the sign is positive in the original paper, giving a tiny indication of a little more sensitivity.

There are two significant changes in the results. For log hourly wage, the causal effect has increased and the standard error has shrunk. For In Poverty, the causal effect slightly increases in magnitude, and the standard error shrinks significantly. Even so, the values for the coefficients are still quite small and close to 0.

4.4 Re-analysis Caveats

The results from the AIPW estimator are quite promising. However, using the AIPW estimator instead of Equation (1) does not allow us to control for some possible confounders. In particular, we are unable to control for fixed effects. For counties (δ_j), the time-invariant differences across counties can bias the estimated causal effect through things like differences in cultural attitudes across counties, government policies in specific counties, county infrastructure: the list goes on. We are also unable to control for state × birth cohort fixed effects ($\gamma_c \times \gamma_2$), which control for time-varying changes in outcomes across states. Going hand in hand with this, the AIPW estimator does not consider the clustered nature of this dataset. It is likely that the AIPW estimator can be adapted to account for clustering, but it is unclear how to do this. Also, we truncate the propensity score in order to maintain strong ignorability: we potentially bias the estimates through this, and this injects more arbitrariness into the estimation procedure. Finally, we inherently lose some interpretational power by binarizing the treatment variable.

As such, the results from AIPW are likely biased, since it is highly unlikely that the effect of including fixed effects is negligible. The initial guess is that the fixed effects bias the estimates upward, especially in Table 4a (DR). This makes sense intuitively especially when considering county infrastructure as a fixed effect: the grade-for-age rate likely increases along with county infrastructure, which would be one source of upward bias.

5 Conclusion

While our results do not match up exactly, we find that the high level results are replicated quite well. We find evidence of a positive intent-to-treat effect of *Sesame Street* coverage on grade-for-age rates, as well as on other response variables, which aligns with what Kearney and Levine find. Additionally, we show that the AIPW estimator has a lot of potential in more precisely estimating the treatment effect through significantly reducing the variance. There is much opportunity in revisiting this paper to examine an AIPW estimator for continuous treatment, and an AIPW estimator that is robust to clustered data.

6 Final Remarks

We want to thank Professor Ding and Sizhu for a great semester. Thank you for the great teaching, and all the help you've given in office hours and sections.