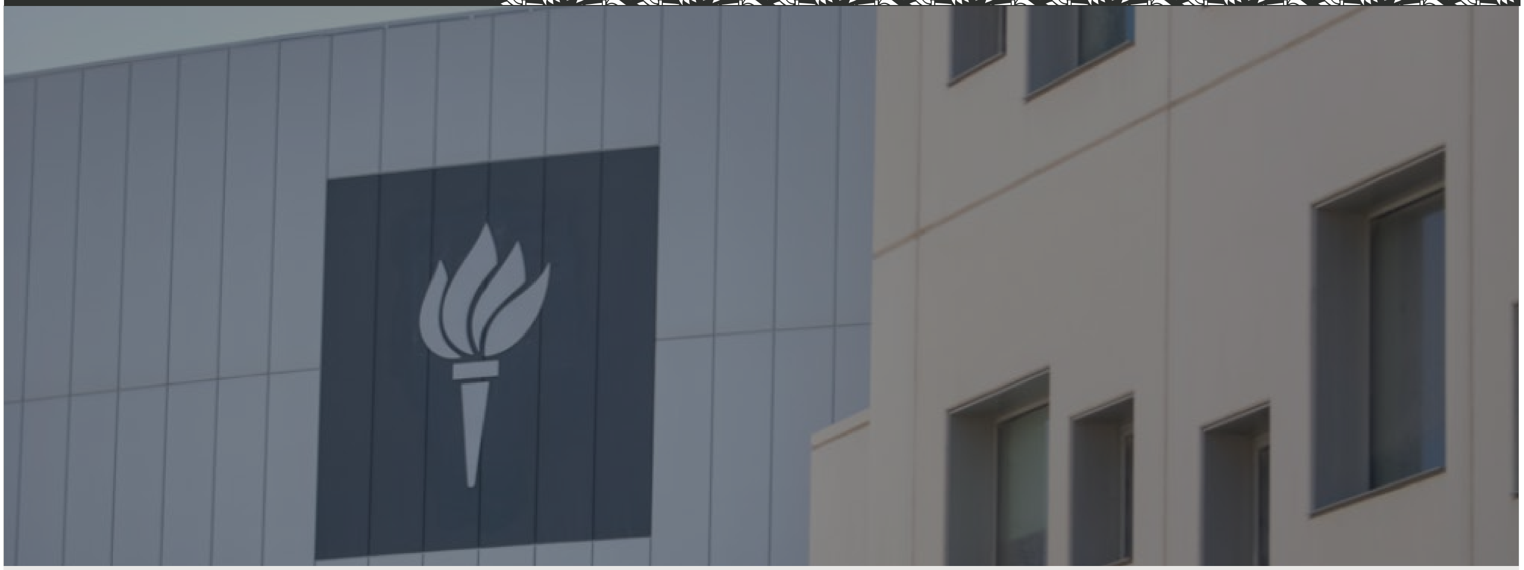
جامعة نيويورك أبوظبي
NYU | ABU DHABI



The Correct Use of Hypothesis Testing and Choosing Appropriate Comparison Groups When Estimating the Impacts of Location Based Policies: A Response to Neumark and Young

John C. Ham, Ayşe İmrohoroğlu, Heonjae Song, Charles Swenson

Working Paper # 0022
October 2018

Division of Social Science Working Paper Series
New York University Abu Dhabi, Saadiyat Island P.O Box 129188, Abu Dhabi, UAE
https://nyuad.nyu.edu/en/academics/divisions/social-science.html

The Correct Use of Hypothesis Testing and Choosing Appropriate Comparison Groups When Estimating the Impacts of Location Based Policies: A Response to Neumark and Young

John C. Ham*
NYU Abu Dhabi and NYU Wagner School
IFAU, IPR and IZA

Ayşe İmrohoroğlu
University of Southern California

Heonjae Song
University of Seoul

Charles Swenson
University of Southern California

First Draft: March 2018

This Draft: October 2018

* We would like to thank Jean Imbs for helpful discussions.

** Corresponding author, jch18@nyu.edu. Mailing Address: Division of Social Sciences, Social Science Building (A5), NYU Abu Dhabi, Saadiyat Campus, P.O. Box 129188, Abu Dhabi, United Arab Emirates.

ABSTRACT

We are grateful to Neumark and Young (2017, hereafter NY hereafter) for spotting an error in Ham, Swenson, İmrohoroğlu, and Song's (2011, hereafter HSIS) 1990 poverty rate data. Our corrected estimates reported here of the impacts on the Poverty Rate of Enterprise Zones (ENTZs), Enterprise Communities (ENTCs), and Empowerment Zones (EMPZs) are smaller than those in HSIS. However, they are still quite sizable and statistically significant. We show here that NY obtained similar results with the NCBD data to our new estimates reported here. The NCBD data uses a different approach than that used by HSIS to deal with the changing borders of some Census tracts over time.

However, we find NY's criticisms of the HSIS results for ENTZs, ENTCs and EMPZs to be deeply flawed, and suffer from several important errors. First, their criticisms arise from their making a fundamental error in hypothesis testing. Second, they use an incorrect approach for comparing parameter estimates from different studies. Third they use a comparison group for EMPZs where about half the comparison tracts are impacted by other labor market programs, leading to downward biased estimates of the EMPZ impacts. We argue that this bias is over 50 percent of their (under)estimated treatment effect for EMPZs.

1. Introduction

Neumark and Young (2017; hereafter NY) perform an important service by documenting an error in the 1990 poverty rates in Ham, Swenson, İmrohoroğlu, and Song (2011; hereafter HSIS). The error led to overly high estimates of the impacts of three programs: Enterprise Zones (ENTZs), Enterprise Communities (ENTCs), and Empowerment Zones (EMPZs) on the Poverty Rate. The reduction in the poverty rate (standard error) in EMPZs using the new data is **8.21** (1.51) percentage points (pp) as opposed to **8.8** (2.78) pp; the reduction in the poverty rate in ENTCs using the new data is **11.67** (1.51) pp as opposed to **20.28** (2.78) pp; and the reduction in the poverty rate in ENTZs using the new data is **1.67** (0.37) pp as opposed to **6.10** (1.21) pp.[1] While the new estimates are smaller in magnitude, they are still quite sizable and statistically significant.[2] The corrected tables, data, and our original programs are available at http://www-bcf.usc.edu/~aimrohor/research.htm.[3]

NY also make a positive contribution by replicating our results (with the corrected Poverty Rate data) for ENTZs on all the outcomes we consider, and for the impact of ENTCs and EMPZs on the poverty rate, using the Neighborhood Change Database (NCDB). The potential advantage of using the NCDB is that a private firm has dealt with the issue of the changes in the boundaries Census tracts over time; an issue that we dealt with ourselves in HSIS. When NY use our

[1] Numbers given in parentheses after point estimates are standard errors.

[2] The finding that the ENTC effect is bigger than the EMPZ effect might be somewhat surprising since the EMPZ is a more generous program. However, there is about a 50 percent overlap in their confidence intervals. As we note below, comparing confidence intervals is the preferred means of comparing two estimates when we do not know their covariance.

[3] As we discuss below, we did not update the IV estimates, or the ENTZ estimates by state, since we now believe there is insufficient variation in the data for these estimators to have a well-behaved asymptotic distribution. In HSIS we did not use the Hausman test to choose the model and instead went with the model chosen when we did not use IV because we were worried that using both the Hausman test and IV was asking too much of the data. However, it is worth noting that using the Hausman test and IV approach together makes the IV estimates much closer to the Non-IV estimates.

estimation strategy, programs and comparison groups with the NCBD, they obtain estimated treatment effects on many of the outcome variables that are very similar to ours. The results for the impacts of ENTZs on all our outcome variables are summarized in Table 1, while the results for the poverty rate for ENTCs and EMPZs are shown in Table 2. In both tables, NCDB refers to the results in NY using the NCDB data. For the poverty rate, HSIS denotes our estimates using the corrected poverty rate data, and for other variables, HSIS denotes our previous estimates which were unaffected by the error.[4]

While we are pleased to see that NY could replicate our results using a different dataset, we believe that there are serious problems with the rest of their comments. First, they claim that the HSIS results for ENTZs are outliers compared to the rest of the ENTZ literature. However, here they make a fundamental error in hypothesis testing by acting as if not being able to reject a null hypothesis allow one to accept the null hypothesis. Specifically, they incorrectly act as if one can treat insignificant coefficients with large confidence intervals as precisely estimated zero coefficients, which they then can use to reject other researchers' precisely estimated effects. This error is also present in Neumark and Simpson (2015, hereafter NS) and Neumark and Kolko (2010, hereafter NK).

Their very basic error is identified by many econometric texts and review articles. For example, in his widely used undergraduate text Wooldridge (2016, p. 120) states

> *When H_0 is not rejected, we prefer to use the language "we fail to reject H_0 at the x% level rather than H_0 is accepted at the x% level."*

[4] NY attempt to 'replicate' the HSIS estimated impacts on the HSIS data (absent the poverty data) and find 'small' differences. This is surprising given that they are using the HSIS' estimation programs and data, but they make no attempt to explain these differences. The HSIS approach is somewhat complicated in that it uses Hausman tests to choose the comparison group, and there is nothing in NY to suggest that they understand the HSIS estimation approach. Hence we find their replication uninformative.

Wooldridge continues

> *…The estimated elasticity with respect to price is -.954 and the t-statistic for testing $H_0 : \beta = -1$ is .393; therefore we cannot reject H_0 but there are many other values for β (more than we can count). For example, the t-statistic for $H_0 : \beta = -.9$ is -.462, so this null is not rejected either. Clearly $H_0 : \beta = -1$ and $H_0 : \beta = -.9$ cannot be true, so it makes no sense to say that we "accept" either of these.*

Further, Romano, Shaikh and Wolfe (2009, p.77) in their review article on hypothesis testing state that

> *Importantly, acceptance of H_0 does not necessarily demonstrate that H_0 is indeed true; there simply may be insufficient data to show inconsistency of the data with the null hypothesis. Therefore, the decision that "accepts" H_0 should be interpreted as a failure to reject H_0.*

Finally, in his chapter in the Handbook of Econometrics (1984, pp. 776-777) on hypothesis testing, Engle writes

> *Testing is inherently concerned with one particular hypothesis which will be called the null hypothesis. If the data fall into a particular region of the sample space called the critical region then the test is said to reject the null hypothesis, otherwise it accepts. As there are only two particular outcomes, an hypothesis testing problem is inherently much simpler than an estimation problem where there are a continuum of possible outcomes. It is important to notice that both of these outcomes refers only to the null hypothesis - we either reject or accept it. To be even more careful in terminology, we either reject or fail to reject the null hypothesis. This makes it clear that the data may not contain evidence against the null simply because they contain very little information at all concerning the question being asked.*

The Romano et al and Engle quotes raise the question of power, i.e. the probability of rejecting the null hypothesis when it if false. Our argument below is that Neumark and Young ignore power considerations and make statements based on uninformative estimates.

Another way to view this issue is as follows. Using data to only test the null hypothesis that a parameter is zero ignores the important information in the confidence interval (CI) for this parameter. If the CI contains zero but is quite wide, the data are basically uninformative about the parameter of interest, including whether it equals zero; this is sometimes described as obtaining a 'big zero' estimate. On the other hand, if a CI contains zero but is narrow, the data are informative about the parameter of interest, and this is sometimes described as obtaining a 'small zero' estimate. NY act as if a big zero estimate is an informative small zero estimate.

Further, there is related issue whether a parameter estimate from one study can lead one to reject an estimate from another study based on a different approach or on different data, in the absence of a compelling *a priori* reason. (Such a reason may be that the estimate in one study is much more likely to be consistent than in the other study, or that the data in one study are much more likely to be informative than in the other study.) Usually such an argument boils down to whether one of the estimates is precisely estimated, or whether the coefficients are equal (which is again an issue of hypothesis testing). Considering the equality of the coefficients, suppose study one provides an estimate $\hat{\theta}_1$ of the parameter of interest, and another study produces an estimate $\hat{\theta}_2$. Before one argues that $\hat{\theta}_1$ can be used to reject $\hat{\theta}_2$, one should consider the null hypothesis $H_0 : \theta_1 = \theta_2$. The problem that arises here is that one generally cannot calculate $Var(\hat{\theta}_1 - \hat{\theta}_2)$ because we cannot calculate $Covar(\hat{\theta}_1, \hat{\theta}_2)$, if the estimates come from different studies. In this case one cannot test $H_0 : \theta_1 = \theta_2$. In such a situation, researchers often compare the CIs for $\hat{\theta}_1$ and $\hat{\theta}_2$. If there is substantial overlap in the CIs, most researchers would conclude here that the estimates are not different, and would choose between for $\hat{\theta}_1$ and $\hat{\theta}_2$ on other grounds, e.g. efficiency.[5]

To see how this affects the argument in NY, as well as in NS and NK, that NK's 95 percent CI for California of [-10.91, 7.51] for the ENTZ effects rules out the HSIS estimated 95 percent CI of [0.368, 7.032]. But all the (random) NK CI tells us is that with 95% probability, this CI contains the (nonrandom) true parameter value; it does not rule out any estimates of the true parameter within their CI. Since HSIS' 95% CI lies completely in the NY CI, the NY CI cannot tell us anything about the validity of the HSIS CI. In fact the NY CI is essentially uninformative. We believe this occurs because HSIS focus on an average across states of the impacts of the individual ENTZ programs. They argue that previous results at the state level have not been useful because they are so noisy, while the average effect across states should be estimated with more precision. Thus it is not surprising that the HSIS estimated impact is an outlier in terms of precision. Note that HSIS' estimates for the effect of individual state programs are just as noisy as those in the literature, driving home this point.

We have other substantial concerns with the analysis in NY. Following Busso, Gregory, and Kline (2013; hereafter BGK), NY's *comparison group for EMPZs consists of Census tracts that were unsuccessful EMPZ applicants, regardless of whether these unsuccessful EMPZ applicants were covered by another program*. Unfortunately, about half of their comparison group are likely to be (positively) affected by other labor market programs. Specifically, about 34 percent of their comparison tracts were ENTCs and about 16 percent of their comparison group were ENTZs; the

[5] This reasoning is used repeatedly in the specification tests in Hausman (1978).

two programs were mutually exclusive. Moreover, HSIS found significant positive effects for both programs. In this case the EMPZ treatment estimates based on the NY comparison group will be downward biased. Not surprisingly, they find that the estimated impact of EMPZs is diminished when they change to this comparison group from HSIS' fairly standard comparison group (untreated tracts that are geographically near the treatments). In Section 2.3 we estimate this bias, and find it to be about 50% of the size of the NY estimated treatment effect.

To evaluate *the impact of being assigned to an ENTC*, NY move to a comparison group consisting of unsuccessful EMPZ applicants that were not treated by another program. Again this differs from HSIS' more standard comparison group for ENTCs consisting of *untreated tracts* that are geographically near the ENTC tracts. However, it is very unusual in the program evaluation literature to use unsuccessful applicants for one program (EMPZs) as a comparison group for recipients of another program (ENTCs). NY find that using their comparison group, their estimate of the impact of the ENTC program is not statistically significant, and argue that this negates the HSIS' statistically significant estimate of the ENTC. However, this is another example of their error of first treating insignificant coefficients with relatively large confidence intervals as precisely estimated zero coefficients, and second acting as if these 'big zeros' can be used to reject considerably more precise and significant estimates.

2. A More Detailed Discussion

2.1 ENTZ Results

As noted above, in their introduction NY claim that the HSIS estimates are outliers; this claim is repeated from NS and NK. However, NY's claim seems to be a function of comparing point estimates as opposed to comparing their confidence intervals. In Figure 1, we contrast estimates across 12 different studies of the most commonly used measure of the ENTZ impact—its effect on log employment. The studies summarized are: NK; Elvery (2009) for California and Florida; Bondonio and Engberg (2000; hereafter BE) for California and New York; Greenbaum and Engberg (2000; hereafter GE) for California, Florida, and New York; O'Keefe (2004) for California; Freedman (2013) results for Texas; the Bondonio and Greenbaum (2007; hereafter BG) national study and HSIS.[6]

From Figure 1, we see that HSIS' CI (for the average state impact) is completely contained in the Cs for five of the other 11 studies: The BG national results; the Freedman Texas results; the GE California results; the O'Keefe California results, and the NK California results. About 95 percent of HSIS' CI is contained in the BE New York CI, about 80 percent of HSIS' CI is contained in the BE California CI, and about 75 percent of HSIS' CI is contained in the GE New York CI. HSIS' CI has less than 50 percent overlap with the GE Florida CI and the Elvery California CI, and no overlap whatsoever with the Elvery Florida CI. In summary, the HSIS CI has complete or very substantial overlap with eight of the 11 CIs from the literature. Figure 1 is inconsistent with NY's claim that the HSIS' estimates being outliers in terms of parameter estimates, but HSIS' estimates clearly are an outlier in terms of precision. But again, the precise nature of the HSIS estimated average ENTZs impacts is not a surprise given that they estimate an average effect of ENTZ designation across several state programs, while all but one of the other papers consider a single the effect of a state program. Further, we would expect the state effect estimators to be poorly identified given the analysis in Conley and Taber (2011), showing that limited program variation can lead to inconsistent results. For the above reasons we do not discuss the HSIS

[6] We show in Table A how we calculated each CI in Figure 1.

estimated individual state impacts here, except to note that they are similar in terms of precision to the estimated individual state programs in the literature.[7]

It is interesting to compare our Figure 1 with the figure in NS, which we have repeated as Figure 2 here. NS' figure does not incorporate standard errors or CIs, and instead takes the unconventional step of reporting the lower and upper point estimates in each paper.[8] As a result, the NS figure (implicitly) vastly overstates the precision of previous results, as one would expect.

One interesting question for future research is why the confidence intervals are so large for the BE New York results, the BE California results, and the NK California results. In the case of the NK California results, we speculate that their extremely large CI arises because they use the National Establishment Time-Series (NETS) data in which firm employment is imputed for a large number of establishments.[9]

Note that we have not re-estimated the IV results in HSIS to correct for the error in the 1990 poverty rates for two reasons. First, as noted by HSIS the IV estimates are local average treatment effects, and since no other study has valid IV estimates, we have nothing to compare them to.[10] In spite of this, NY argue that they are skeptical of our work because of the large IV estimates, ignoring generally accepted argument that local average treatment effects are not generally

[7] HSIS write regarding the individual state impacts "As expected, many of these effects are imprecisely estimated and thus statistically insignificant, and thus we do not discuss them in detail.

[8] It is also difficult to interpret the NS figure with regard to the HSIS results. They report a combination of 'various estimates' for HSIS' individual state impacts and the average state impacts, without stating which results they used.

[9] For example, we looked at the California enterprises in 1990 since NK used a dataset containing such firms. We found that employment was imputed in approximately 50 percent of establishments in the NETS data.

[10] Hanson presents an IV estimate of the impact of EMPZ designation using the same control group as NY. We do not compare his IV estimates with ours, because his first-stage equation generates a very small F-statistic for his excluded instruments; this F-statistic is well below the suggested critical values in Staiger and Stock (1997) and Stock and Yogo (2005).

comparable to OLS results. Moreover, given the importance that NY place on IV estimates for the credibility of results (which we disagree with), it is quite surprising that NY do not provide readers with any IV estimates based on their comparison groups.

A second reason for not using IV procedures in our revised estimates is that, in retrospect, we believe that triple-difference IV estimation is likely to demand too much of the data, and that such estimates are likely to have poor asymptotic properties along the lines raised by Hahn, Ham, and Moon (2011). They note that fixed-effects estimates will have a non-standard asymptotic distribution if there is insufficient within-state variation in the data. Note that with only 2 periods of data (as used in the above studies), fixed effect and first difference estimates will generally be numerically equal, so that the Hahn et al theoretical results apply to them and help explain the variation across studies. This problem will be accentuated in HSIS when we use IV estimation, because such estimation further reduces the usable variation in the data. .[11]

2.2 A Serious Problem with the NY Comparison Group for ENTCs

In terms of ENTCs, NY is the only available study estimating their impacts besides HSIS. However, as noted above, HSIS' and NY's comparison groups for ENTCs are not comparable. HSIS' main comparison group consists of untreated tracts that are relatively near the ENTC

[11] NY correctly note that we did not justify our choice of instrument. We assumed that the reason for our choice of instrument was fairly obvious but in hindsight, we should have discussed it. We were concerned that treatment occurred in response to a large idiosyncratic shock in the 1990 outcome variables, and hence it would be inappropriate to use 1990 values of the other outcomes as instrument. Of course, we would not want to use 2000 values of the other variables because they could be driven by the 1990 shocks. Hence, in responding to a referee's request for IV estimates, we were left with using the 1980 values of the other variables as IVs for the outcome in question. (1970 values were not available.) For these 1980 values of the other outcome variables to be valid IV, we needed to assume that conditional on triple differencing, these variables are independent of the unobserved shocks to the 1990 value of the outcome in question.

tracts.[12] NY use unsuccessful applicants for the *EMPZ* program who were not covered by any other program as their comparison group to judge the effectiveness of *ENTC* designation. However, we cannot think of a good justification of the use of unsuccessful candidates for one program as a comparison group for a *different* program.

Further, ignoring the problems with their comparison group, they obtain an insignificant big zero estimated impact for the ENTC program, which they argue negates the precise and significant estimate in HSIS. But of course this is just another example of the error they made in comparing the ENTZ estimated impacts across studies. Moreover, there is substantial overlap in the CIs for the NY ENTC impact of and the HSIS ENTC impact.

2.3 A Different, and an Even More Serious, Problem with the NY Comparison Group for EMPZs

In assessing the effectiveness of EMPZs, NY use tracts in unsuccessful EMPZ applicants as a comparison group. However, 34 percent of these comparison tracts were covered by ENTCs and 16 percent were covered by ENTZs. Of course including tracts covered by ENTC and ENTZ programs in the comparison group is only valid if these programs do not have an effect, but NY justify using tracts treated by ENTCs and ENTZs by arguing either that the benefits of these programs are i) small compared to the benefits of being covered by an EMPZ or ii) are actually zero. However, even if the effects of these programs are small, one still cannot ignore the fact the exist in estimating the EMPZ treatment effect with this comparison group, so the question becomes whether it is sensible to set the effects of ENTCs and ENTZs to zero *a priori*.

[12] HSIS also used untreated tracts from the rest of the state as a comparison group when they found that estimated impacts from this larger comparison group were similar to those from a comparison group based on near-by tracts (using a Hausman test).

To assess the case for setting the ENTZ and ENTC effects to zero *a priori*, we consider the summary of the provisions of ENTCs and EMPZs compiled in Ham and Song (2018; hereafter HS). The benefits of EMPZs are:

A1. EMPZs receive large grants of around $100 million to help jump-start the local economy;

A2. Employers may receive up to $3,000 for each employee who is a resident of the zone;

A3. Employers may receive up to $2,400 for each new worker hired between the ages of 18 and 24 who is a zone resident under the Work Opportunity Tax Credit (WOTC), but may not combine this with the subsidy in A2 for the same individual;

A4. EMPZs may receive subsidized loans, and they may provide, in turn, subsidized loans to Community non-profit organizations for buildings that are of benefit to the community. They may also use these loans to help private builders finance construction and renovation in the zone.

A5. Firms in EMPZs received increased tax deductions for depreciable, tangible property owned by businesses.

On the other hand, the opportunities open to ENTCs consist of:

B1. ENTCs receive small grants of around $3 million to help jump-start the local economy, except for five Enhanced ENTCs that received about approximately $15 million each.

B2. Employers may receive up to $2,400 WOTC subsidy for each new worker between the ages of 18 and 24 who is a zone resident.

B3. ENTCs may receive subsidized loans, and they may provide subsidized loans to Community non-profit organizations for buildings that are of benefit to the community. They may also use these loans to help private builders finance construction and renovation in the zone.

Thus, the relevant question for NY's comparison group is whether, based on B1-B3, we can reasonably assume ENTCs will have no impact *a priori*, i.e., without even measuring their impact. NY argue that this assumption is reasonable since "ENTCs did not have hiring tax credits; they only received $3 million in Social Services Block Grant (SSBG) funds and were eligible for tax-exempt bond financing" (p.13). From B1-B3, it is clear that NY miss the fact that firms hiring in an ENTC could also receive a Work Opportunity Tax Credit for residents aged 18-24.[13] Further, given B1-B3, economic theory would certainly predict ENTCs will have a non-zero impact. (Of course estimating positive impacts of ENTCs is very different from saying that the gross benefits of ENTCs justify their cost.) Moreover, since the ENTC effect can be estimated, arbitrarily setting it to zero is antithetical to the spirit of program evaluation and evidence-based policy evaluation. Note that NY could have avoided this problem by using tracts that were denied EMPZ status, but were not covered by other programs, as an alternative to HSIS' comparison group for EMPZs. However, an important argument against even this alternative comparison group is that because they are farther away from the treatments than HSIS' comparison group, it will be harder to control for unobserved differences between the treatments and comparisons. In fact, Smith and Todd (2005, p. 306) argue that when evaluating training application using propensity score matching, both the treatments and comparisons should be drawn from the same labor market, which favors HSIS' choice of comparison group.

Surprisingly, NY argue that their inclusion of ENTCs and ENTZs in the comparison group for EMPZs is valid because BGK used the same comparison group and must have done so because

[13] Hanson (2009) also makes this error. For an accessible treatment of the WOTC grants, see Hamersma (2005); she also notes that WOTCs may be used to provide those aged 16 and 17 with summer employment in ENTCs.

ENTCs have weak provisions. This argument is clearly an example of circular reasoning. The fact that BGK did something is a compelling argument only if BGK present a strong case for using this comparison group. However, BGK do not even mention any problems with their comparison group *for the standard treatment effect where the alternative is not being covered by any program*. Neither do they mention the estimation of significant positive impacts for ENTCs and ENTZs in the HSIS paper. Instead BGK only state that they think that A1. and A2. are the most important benefits of being an EMPZ (p. 900-901), and that ENTCs received a small level of development funds and loans with reduced interest rates (p. 900, fn. 5). BGK also did not mention the fact that employers in ENTCs were eligible for WOTC grants for residents aged 18 to 24 years. Thus, BGK offer no additional justification for the choice of comparison group used by NY.

Given that HSIS find significant positive effects for ENTCs and ENTZs on employment, employment in the NY comparison group is biased upward, and the NY estimated impacts of EMPZs on employment are biased downward. Since BGK have the same problem, we can infer the expected bias in the NY estimates from the calculations in Ham and Song's (2018, HS) estimate of bias in the BGK's estimates. HSIS' estimates imply that ENTCs and ENTZs raised employment by 109.0 (standard error of 51.3) individuals, and by 68.9 (standard error of 32.6) individuals, respectively. Further, 34.2 percent and 16.0 percent of the NY comparisons are in ENTCs and ENTZs, respectively.[14] Thus, the NY employment impacts are biased downward by

$\boldsymbol{BIAS} = (0.342 * 109.0) + (0.16 * 68.9) = 48.32$ Individuals.

[14] HS obtain the 0.342 figure using Table A1 in BGK, which reports the number of Census tracts in their comparison group that belong to the round 1 ENTC. To obtain the 0.16 figure, HS matched BGK's census tracts in their comparison group with the 1990–1997 nationwide designated ENTZ census tracts.

This is a non-trivial bias correction term given that NY's estimated EMPZ effect on employment is 89.49 (standard error of 33.84) individuals. Applying the bias correction implies that the EMPZ corrected effect rises to $89.49 + 48.32 = 137.81$ individuals. In other words, the estimated impact is raised by over 50 percent. Unfortunately, to construct a standard error for the corrected effect, we would need the covariance between: (i) the NY EMPZ effect on employment, and (ii) the HSIS ENTC and ENTZ effects on employment. However, none of the (estimated) covariance terms are available from the papers. Note that we could make a similar adjustment to NY's estimates of the EMPZ impacts on other outcome variables.

3. Conclusion

We appreciate that NY spotted an error in the HSIS 1990 poverty rate data. While the new results based on the corrected data are smaller, they are still quite sizable and statistically significant. We also appreciate knowing that NY obtained similar results to ours with the NCBD data, which are based on a different approach than that used by HSIS to deal with the changing borders over time of some Census tracts.

However, NY's criticism of the HSIS results for ENTZs and ENTCs arises from their making a fundamental error in hypothesis testing. Further, careful examination of the estimated individual state ENTZ effects and the average state effects, as well as the analysis in Conley and Taber (2011), indicate that attention should be focused on the average state effects. Finally, NY uses a comparison group for EMPZs in which half of the comparison tracts are impacted by other labor market programs, leading to downward biased estimates of the EMPZ impacts. Using estimates from HSIS, we find that this bias is about 50 percent of their estimated treatment effect for EMPZs.

Table 1: Triple-Difference Random Effects Estimates for State Enterprise Zone Analysis: Comparing Estimates using the NCDB and HSIS Data

Dependent Variable: E [{ENTZ(Δ00)-NENTZ(Δ00)} - {ENTZ(Δ90)-NENTZ(Δ90)}]

Panel A. Unemployment rate (%)

Data	NCDB	HSIS
Comparison	Contiguous	Contiguous
	-1.88***	-1.64***
	(0.25)	(0.23)
Observations	1,158	1,227
Number of ENTZs	1,158	1,227
Number of counties	90	112

Panel B. Poverty rate (%)

Data	NCDB	HSIS
Comparison	All	Contiguous
	-1.25***	-1.665***
	(0.32)	(0.368)
Observations	23,151	1,265
Number of ENTZs	1,290	1,265
Number of counties	317	112

Panel C. Fraction of households with wage and salary income (%)

Data	NCDB	HSIS
Comparison	Nearest	Contiguous
	0.38	0.45
	(0.52)	(0.30)
Observations	1,153	1,241
Number of ENTZs	1,153	1,241
Number of counties	90	112

Table 1 (continued): Triple-Difference Random Effects Estimates for State Enterprise Zone Analysis: Comparing Estimates using the NCDB and HSIS Data

Panel D. Average wage and salary income ($2000)

Data	NCDB	HSIS
Comparison	Nearest	Nearest
	614.6	703.0*
	(429)	(387)
Observations	1,124	1,212
Number of ENTZs	1,124	1,212
Number of counties	90	112

Panel E. Employment

Data	NCDB	HSIS
Comparison	All	Contiguous
	29.53	68.91**
	(21.14)	(32.57)
Observations	23,230	1,264
Number of ENTZs	1,296	1,264
Number of counties	317	112

Table 2: Triple-Difference Estimates for Federal Enterprise Zone Analysis: Comparing Estimated Effects on Poverty Rates Using the NCDB Data and Corrected HSIS Data
Dependent Variable: E [{ENTZ(Δ00)-NENTZ(Δ00)}- {ENTZ(Δ90)-NENTZ(Δ90)}]

	Empowerment Zones (EMPZ)		Enterprise Communities (ENTC)	
Data	NCDB	HSIS	NCDB	HSIS
Comparison	Contiguous	Contiguous	All	All
	-9.60**	-8.213***	-11.54***	-11.67***
	(1.84)	(1.511)	(0.53)	(0.515)
Observations	264	271	27,520	29,615
Number of EMPZs/ENTCs	264	271	374	412
Number of counties	9	14	533	960

Note: Asymptotic standard errors in parentheses in both tables.

Figure 1: 95% Confidence Intervals for Different Estimates of the Impact Of Enterprise Zones on Log Employment

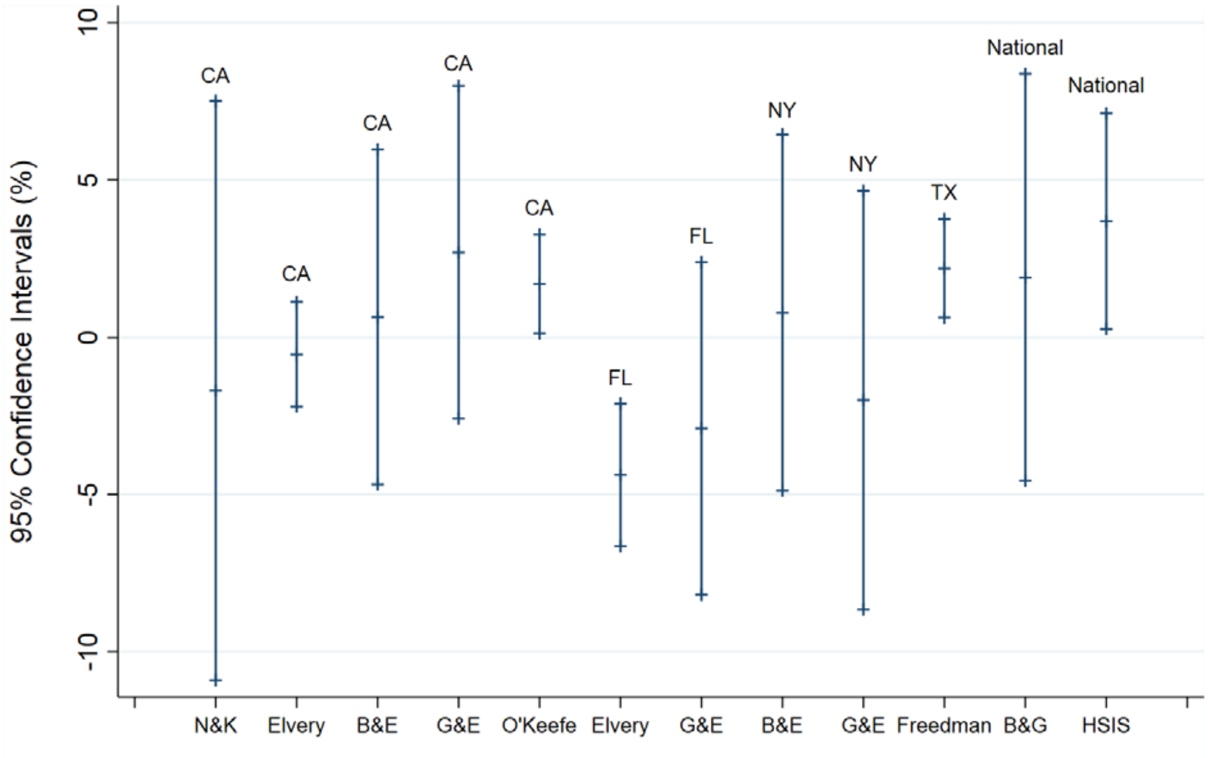
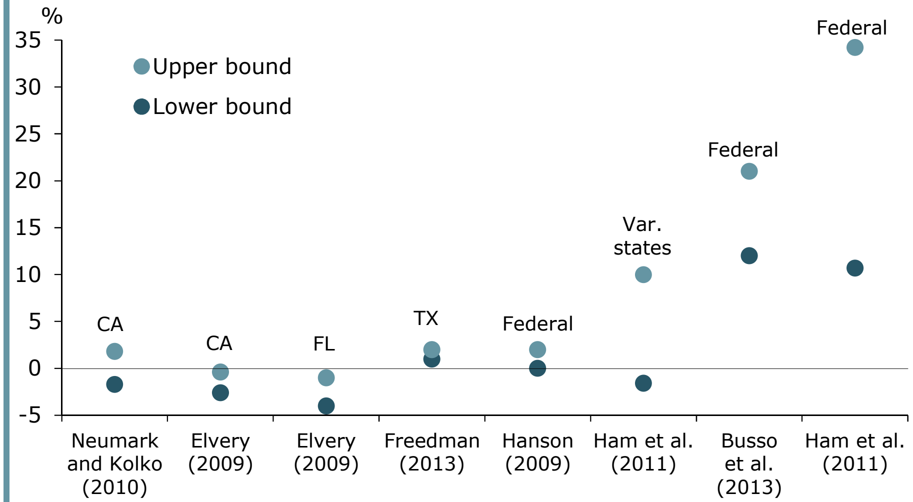


Figure 2 - Neumark and Simpson Summary of ENTZ Effects

programs are even larger, yet this study finds some effects on other outcomes, particularly in reducing poverty, that are larger for federal Enterprise Communities, which had more restricted hiring credits and did not receive major block grants. In our view this casts doubt on the study's findings, and we omit it from some of the discussion below. Finally, the Hanson (2009) study also examines federal Empowerment Zones and finds little evidence of an employment effect.

Figure 1
Range of enterprize zone effects on employment



Note: standard errors play no role in this Figure

Table 1
Effects of enterprize zones on poverty and house prices

Program	Study	Findings
		Poverty
Federal Empowerment Zones	Hanson (2009)	Insignificant positive effect (2 percentage points)
	Reynolds and Rohlin (2013)	No significant effect (−1 percentage point)
		Significant increase in proportion of households below one-half the poverty line (1.1 percentage points)
		Significant increase in proportion of households more than twice the poverty line (1.9 percentage points)
		House prices
Texas enterprise zones	Freedman (2013)	Significant positive effect on median home value (10.7%)
Federal Empowerment Zones	Busso et al. (2013)	Large significant positive effects on house values (28–37%)
	Reynolds and Rohlin (2013)	Increases in value for houses valued $100,000 or more, extending above $300,000

Even if there is some evidence that federal enterprise zones create jobs, assessments of their effectiveness must be tempered by other research findings summarized in Table 1. First, even though some research on federal Empowerment Zones finds some evidence of positive employment effects, other research fails to find evidence of reduced poverty, and points to some increases in the share of households falling below other low income thresholds. Second, there is consistent evidence of housing price increases, implying that benefits are received by unintended recipients. Other results not included in the table sometimes point to negative spillover effects on nearby areas, suggesting that enterprise zones largely rearrange the location of jobs rather than creating more of them.

Our overall view of the evidence is that state enterprise zone programs have generally not been effective at creating jobs. The jury is still out on federal programs—Empowerment Zones in particular—and we need more research to understand what features of enterprise zones help spur job creation. Moreover, even if there is job creation, it is hard to make the case that enterprise zones have furthered distributional goals of reducing poverty in the zones, and it is likely that they have generated benefits for real estate owners, who are not the intended beneficiaries.

Table A: Input for Figure 1

	Point Estimate	Standard Error	Source	Analyzed Enterprise Zone periods
N&K (CA)	-1.7	4.7	Neumark and Kolko (2010): We use the estimate in column 1 and row A of Table 6 on p. 12.	1992~2004
Elvery (CA)	-0.5	0.9	Elvery (2009): We use the estimate in column 2 and row 2 for all men and women of Table 5 on p. 55.	1987~1990
B&E (CA)	0.6	2.7	Bondonio and Engberg (2000): We use the estimate in column 1 and rows 2–3 of Table 7 on p. 537 and column 1 and row 1 of Table 1 on p. 523.	1981~1994
G&E (CA)	2.7	2.7	Greenbaum and Engberg (2000): We use the estimate in column 7 and row 1 of Table 6 on p. 43.	1986~1990
O'Keefe (CA)	1.7	0.8	O'Keefe (2004): We use the estimate in column 2 and row 1 of Table 5 on p. 145.	1986~1992
Elvery (FL)	-4.4	1.2	Elvery (2009): We use the estimate in column 5 and row 2 for all men and women of Table 5 on p. 55.	1987~1990
G&E (FL)	-2.9	2.7	Greenberg and Engberg (2000): We use the estimate in column 7 and row 3 of Table 6 on p. 43.	1986~1990
B&E (NY)	0.8	2.9	Bondonio and Engberg (2000): We use the estimate in column 1 and rows 2–3 of Table 7 on p. 537 and column 3 and row 1 of Table 1 on p. 523.	1981~1994
G&E (NY)	-2.0	3.4	Greenbaum and Engberg (2000): We use the estimate in column 7 and row 7 of Table 6 on p. 43.	1987~1990
Freedman (TX)	2.2	0.8	Freedman (2013): We use the estimate in column 10 and row 1 of Table 2 on p. 328.	2002~2009
B&G (National)	1.9	3.3	Bondonio and Greenbaum (2007): We use the estimate in column 1 and row 1 of Table 2 on p. 130.	1982~1992

Table A (continued): Input for Figure 1

	3.7	1.7	Ham, Swenson, İmrohoroğlu, and Song (2011): Since HSIS estimate is in levels, we get a percentage impact by using the estimate for average 1990 employment in ENTZs of Table 1 on p. 784, and the coefficient and standard error in column 5 and row 1 of Table 2 on p. 790.	1990~1997

Note: Bondonio and Engberg (2000) estimate $\ln E_{it} = \delta ENTZ_{it} + \lambda ENTZ_{it} \times \text{policy}_{it} + \cdots$.

We measured the effect on the employment change by $\hat{\delta} + \hat{\lambda} \times \overline{\text{policy}}_{it}$. Here, we focused on the monetary incentive policy effect of ENTZ and its mean value 0.115 as in column 1 and row 1 of Table 1 on p. 523. Hence, the estimated effect on employment is $0.4 + 2.1 \times 0.115 = 0.6415$ in the case of CA. We calculate the standard error for this estimate using $\left[\text{var}(\hat{\delta}) + 0.115^2 \times \text{var}(\hat{\lambda})\right]^{1/2}$; here we made the simplifying assumption $cov(\hat{\delta}, \hat{\lambda}) = 0$ since we cannot retrieve $cov(\hat{\delta}, \hat{\lambda})$ from Bondonio and Engberg's California results. We proceeded in the same manner for Bondonio and Engberg's New York results.

References

Bondonio, D. and J. Engberg. 2000. "Enterprise Zones and Local Employment: Evidence from the States' Programs." *Regional Science and Urban Economics* 30(5), 519-549.

Bondonio, D. and R. Greenbaum. 2007. "Do Local Tax Incentives Affect Economic Growth? What Mean Impacts Miss in the Analysis of Enterprise Zone Policies." *Regional Science and Urban Economics* 37(1), 121-136.

Busso, M., J. Gregory, and P. Kline. 2013. "Assessing the Incidence and Efficiency of a Prominent Place Based Policy." *American Economic Review* 103(2), 897-947.

Conley, T. and C. Taber. 2011. "Inference with 'Differences in Differences' with a Small Number of Policy Changes." *The Review of Economics and Statistics* 93(1), 113-125.

Dehejia, R. and Wahba, S. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448), 1053-1062.

________________. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics* 84(1), 151-161.

Elvery, J. 2009. "The Impact of Enterprise Zones on Residential Employment: An Evaluation of the Enterprise Zone Programs of California and Florida." *Economic Development Quarterly* 23(1), 44-59.

Engle, R. 1984. "Chapter 13 Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics." In *The Handbook of Econometrics*, Volume 2, Amsterdam: Elsevier Science, pp. 775-826.

Freedman, M. 2013. "Targeted Business Incentives and Local Labor Markets." *Journal of Human Resources* 48(2), 311-344.

Greenbaum, R. and J. Engberg. 2000. "An Evaluation of State Enterprise Zone Policies." *Review of Policy Research* 17(2-3), 29-45.

Hahn, J., J. Ham, and H. R. Moon. 2011. "Test of Random Versus Fixed Effects with Small Within Variation." *Economics Letters* 112(3), 293-297.

Ham, J., C. Swenson, A. İmrohoroğlu, and H. Song. 2011. "Government Programs Can Improve Local Labor Markets: Evidence from State Enterprise Zones, Federal Empowerment Zones and Federal Enterprise Communities." *Journal of Public Economics* 95(7), 779-797.

Ham, J. C. and H. Song. 2018. "Consistent Estimation of the Employment Impact of the U.S. Federal Empowerment Zone Program." Mimeo, NYU Abu Dhabi.

Hamersma, S. 2005. "The Work Opportunity and Welfare-to-Work Tax Credits." Urban-Brookings Tax Policy Center, The Urban Institute, 15, 1-7.

Hanson, A. 2009. "Local Employment, Poverty, and Property Value Effects of Geographically-Targeted Tax Incentives: An Instrumental Variables Approach." *Regional Science and Urban Economics* 39(6), 721-731.

Neumark, D. and J. Kolko. 2010. "Do Enterprise Zones Create Jobs? Evidence from California's Enterprise Zone Program." *Journal of Urban Economics* 68(1), 1-19.

Neumark, D. and H. Simpson. 2015. "Do Place-Based Policies Matter?" *Federal Reserve Bank of San Francisco Economic Letter*, March 2, 2015.

Neumark, D. and T. Young. 2017. "Government Programs Can Improve Local Labor Markets, But Do They? A Re-Analysis of Ham, Swenson, İmrohoroğlu, and Song (2011)." IZA Working Paper 11168.

O'Keefe, S. 2004. "Job Creation in California's Enterprise Zones: A Comparison Using a Propensity Score Matching Model." *Journal of Urban Economics* 55(1), 131-150.

Oakley, D. and H. Tsao. 2006. "A New Way of Revitalizing Distressed Urban Communities? Assessing the Impact of the Federal Empowerment Zone Program." *Journal of Urban Affairs* 28(5), 443-471.

Romano, J., A. Shaikh and M. Wolf (2010). "Hypothesis Testing in Econometrics." *Annual Review of Economics* 2:75–104.

Smith, J. and P. Todd. 2005. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125(1-2), 305-353.

Staiger, D. and J. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3), 557-586.

Stock, J. H. and M. Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In: Andrews, D.W.K. and J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* New York: Cambridge University Press, pp. 80-108.

Wooldridge, J. 2016. Introductory Econometrics: A Modern Approach. Boston: Cengage Learning.