Rejoinder to Hoxby

Jesse Rothstein
Princeton University

November 2007

After several rounds of revisions, the *American Economic Review* has accepted a Reply to my Comment (Rothstein, forthcoming) on Caroline Hoxby's "Does Competition Among Public Schools Benefit Students and Taxpayers" (Hoxby 2000). The Reply (Hoxby, forthcoming) is deeply flawed. It badly misstates the import of my comment; only briefly addresses my primary argument; and gets many important facts wrong. In this rejoinder, I attempt to clarify matters.

Although Hoxby purports to refute my Comment, my central conclusions are undisputed:

1) Hoxby has not made available the data that were used for her original paper.
2) Hoxby's basic result holds only when we rely on her specific construction of the "larger streams" variable. She has never presented estimates that do not rely on this variable; all such estimates that I have computed yield small, insignificant effects of choice on test scores.
3) There are several odd aspects of Hoxby's particular larger streams variable that cast doubt on its validity.
4) There are serious errors in both the program code and data that Hoxby has distributed. Hoxby has not released the code that was used in her Reply, but she offers no indication that these errors have been repaired.

The Reply's discussion of these issues is circuitous and masked by extensive, combative digressions about irrelevant points, and even careful readers who lack access to the underlying data may be misled. This rejoinder attempts to clarify these issues. Tables 1-3 present the primary estimates that underlie my claims. Section I details the case for each of the four basic facts. Section II discusses Hoxby's Reply, nearly all of which is irrelevant to the core claims of my Comment. I do not take up the issue of Hoxby's aspersions against my motives and character that appeared in earlier drafts of Hoxby's reply (e.g., Hoxby 2005) and in press accounts (e.g., Hernandez 2005), though I certainly dispute them. Interested readers can evaluate these issues for themselves.

Table 1. Estimates from published paper (Hoxby 2000) and from the data that Hoxby has distributed (Hoxby 2004)

	Published estimates	Distribution data	
		All MSAs	MSAs in IV Sample
	(1)	(2)	(3)
A. Means of stream variables			
Larger streams	8	44	45
Smaller streams	183	84	104
# of MSAs	316	310	184
B. First stage regression coefficients (with standard errors)			
Larger streams (100s)	0.080	0.012	-0.044
	(0.040)	(0.021)	(0.028)
Smaller streams (100s)	0.034	0.096	0.143
	(0.007)	(0.019)	(0.025)
# of MSAs	316	310	184
C. IV estimates of choice effect on 12th grade reading scores			
	5.77		5.30
	(2.21)		(2.94)
# of students	6,119		5,475
# of MSAs	316		184

Notes: Column 1 is from Tables 2 and 4 of Hoxby (2000). Column 2, and Panel C of Column 3, is obtained by running Hoxby's (2004) programs on the data on the distribution CD. Panels A and B of Column 3 derive from slight modifications of Hoxby's code t

Table 2. IV estimates of the choice effect on 12th grade reading scores, alternate samples and instruments

	Instruments: Hoxby larger and smaller streams	Total streams	Inter- and intra-county streams	Longer and shorter streams
	(1)	(2)	(3)	(4)
Sample:				
Published (Hoxby 2000)	5.77			
	(2.21)			
Distribution data (Hoxby 2004)	5.30	-3.36		
	(2.94)	(4.05)		
Replication sample with corrections	4.74	0.87	2.04	1.35
	(2.42)	(2.81)	(2.94)	(2.04)

Notes: Each entry is the IV estimate from a different specification. Clustered standard errors are reported except in the first row, which reports Hoxby's (2000) "Moulton" standard errors. Inter-county and longer streams are not separately reported in

Table 3. First stage estimates, separately for NELS MSAs (which contribute to IV estimates) and non-NELS MSAs (which do not)

Sample	Hoxby distribution data		Replication data	
Streams measure(s)	Hoxby larger and smaller streams	Total streams	Inter- and intra-county streams	Longer and shorter streams
	(1)	(2)	(3)	(4)
Non-NELS MSAs				
Larger streams (100s)	0.101		0.245	0.191
	(0.035)		(0.118)	(0.082)
Smaller or total streams (100s)	0.046	0.072	-0.013	-0.016
	(0.034)	(0.026)	(0.029)	(0.028)
NELS MSAs				
Larger streams (100s)	-0.044		0.243	0.170
	(0.028)		(0.067)	(0.044)
Smaller or total streams (100s)	0.143	0.057	0.026	0.020
	(0.025)	(0.016)	(0.020)	(0.024)
P-value, same streams coefficients	0.001	0.638	0.393	0.576
P-value, same other coefficients	0.300	0.406	0.307	0.280

Notes: All estimates are computed at the MSA level and include Hoxby's vector of MSA-level covariates. A NELS MSA is one that contributes at least one observation to the NELS 12th grade reading sample.

I. Four Important Facts

1) Hoxby has not made available the data that were used for her original paper.

Even after several years of requests, Hoxby has not made available data that produce the results in her published paper. While she has released a data set (Hoxby 2004), it is different in several important ways from that which generated the published results. These differences cannot be attributed to updates in the source data sets; Hoxby has clearly made intentional changes.

Table 1 reports two versions of the key statistics from Hoxby's study. Those in the first column are taken from the published paper (Hoxby 2000), while those in the second are reported by Hoxby's (2004) programs when executed on the data distributed with her CD.[1] Panel A reports simple MSA-level means of the "streams" instruments; Panel B reports first stage coefficients on these instruments, computed at the MSA level and including all available MSAs, as in Hoxby (2000); and Panel C reports IV estimates of the key coefficients.

The streams variables are clearly different from those used in the published paper: The mean of the "larger streams" variable is five times larger in Hoxby's data set than was reported in her paper, while the mean number of total streams (larger plus smaller) is a third smaller. The first stage coefficients are also notably changed, with the larger streams coefficient shrunken by about six sevenths (and no longer significant) and the smaller streams coefficient nearly trebled. Importantly, the first stage effect of smaller streams is now far bigger than the effect of larger streams. Finally, although the IV estimates of the choice effect on student test scores are comparable in magnitude, the standard errors are notably increased, enough so that neither Hoxby's original point estimates nor the new ones are significant at the 0.05 level.[2]

---

[1] As noted in my Comment, Hoxby's (2004) programs produce a slightly different data set each time they are run. All statistics reported here use the data sets distributed on the CD, which I take to be one draw from the underlying stochastic distribution. In any case, this problem—which arises via a many-to-many merge—cannot account for the dramatic differences in the summary statistics that I document here.

[2] Hoxby (Reply, p. 24) attributes the change in the standard errors to her shift from "Moulton" standard errors in Hoxby (2000) to "clustered" standard errors in her distributed programs, but offers no evidence for this. I cannot confirm her

Hoxby (Reply, pp. 2-3) attributes all differences in her data from what was used in her original paper to small changes in the geographic codes in the Common Core of Data (CCD).[3] This cannot account for the dramatic changes in the basic summary statistics for the streams variables, for which Hoxby has offered no explanation. Readers may or may not judge the differences between the published estimates and those that Hoxby's data produce to be substantively important. But they certainly invalidate the Reply's claim (p. 2) that "All of the raw data and code used to make extracts from the raw data and compute the estimates [in Hoxby (2000)] are available to researchers."

2) *Hoxby's basic result holds only when we rely on her specific construction of the "larger streams" variable. She has never presented estimates that do not rely on this variable; all such estimates that I have computed yield small, insignificant effects of choice on test scores.*

The central result of my Comment is that Hoxby's results are extraordinarily sensitive to the way that streams are defined, with the positive, significant effect of choice on test scores appearing only with Hoxby's particular "larger streams" measure. When I substitute alternative constructions of this variable, I obtain estimates of the choice effect on test scores that are uniformly smaller than those obtained using Hoxby's variable and are never significantly different from zero. Table 2 reports estimates of the choice effect on test scores using four measures of streams: Hoxby's larger and smaller streams; total streams, the simple sum of Hoxby's two variables, entered as a single instrument; inter- and intra-county streams, used by Hoxby in an early pre-publication working paper (Hoxby 1994); and longer—greater than 3.5 miles—and shorter streams. Each of these is constructed from the same "total streams" measure; the only difference is the way that this measure

---

claim: When I estimate "Moulton" standard errors – Hoxby has never provided her code for this, so I use my own – they are larger than those that Hoxby (2000) reports, and the difference between Moulton and clustered standard errors is always quite small.

[3] In fact, there have been no changes in the relevant parts of the CCD since its original release in the mid-1990s. I discuss this in Section II.

is divided into subcomponents. Using each alternative instrument set, the choice coefficient is dramatically reduced from that produced by Hoxby's variables and is far from significant.

A crucial point here is that the streams variables serve only as instruments in Hoxby's model. The Reply (pp. 11-12) claims that her larger streams construct better corresponds to the relevant variable, navigable streams, than do any of my alternatives. I grant her claim that my constructs are only proxies for navigable streams, not perfect measures. But there is absolutely no reason to think either that the measurement error in my variables is systematic or that her variable is free of error. Evidently, there is some systematic, non-random difference between Hoxby's measure and my alternatives. Below, I offer suggestive evidence that the problem lies with Hoxby's variable.

Table 3 of the Reply reports estimates that use Hoxby's smaller streams variable as the only instrument. It is important to realize that this variable depends directly on Hoxby's hand-coded larger streams measure: Hoxby measures smaller streams as the difference between total streams (computed from the Geographic Names Information System data) and larger streams, so any errors in the larger streams measure infect the smaller streams measure as well. Thus, this specification does nothing to counter my claim that Hoxby's results depend on her measurement of larger streams.[4]

3) There are several odd aspects of Hoxby's particular larger streams variable that cast doubt on its validity.

Hoxby's defense of her larger streams variable is that it is a better measure of navigable rivers than any conceivable proxy, and that we should therefore rely on her judgment of what constitutes a navigable stream rather than on reproducible proxies. There are several reasons to be wary of this claim.

[4] When I use any of my alternative "smaller streams" variables as the sole instrument, I obtain small, insignificant choice effects.

First, Hoxby's distribution data (Hoxby 2005a) reports 44 larger streams in the average MSA. It is simply not credible that the average MSA has 44 navigable rivers. The published paper reported a more plausible mean of 8. Hoxby has offered no account for this dramatic change.

Second, a close examination of the first stage model raises questions about Hoxby's variable. Even the first stage estimates presented in Column 2 of Table 1 indicate that the large streams effect is oddly small: The smaller streams effect is much bigger, so the estimates indicate that changing a small stream to a large, navigable river *reduces* the amount of choice in the metropolitan area. This counterintuitive result does not arise with either of my alternative categorizations of streams (nor, for that matter, in the first stage coefficients published in Hoxby's 2000 paper), each of which indicates that large streams have much bigger effects on choice than do small streams.

Even more troubling is the substantial difference between the first stage coefficients obtained when the model is estimated on the full set of MSAs (Column 2 of Table 1) and those obtained using just the NELS MSAs that are relevant for Hoxby's IV estimates (Column 3).[5] While larger streams have a positive (but insignificant) relationship with choice in the full sample, in the subsample that is relevant for Hoxby's IV analysis the larger streams coefficient has the *wrong sign*, indicating that metropolitan areas with more "navigable" streams have *less* choice than do those with fewer.

The NELS MSAs are a random subset of MSAs, as the NELS sampled schools randomly from across the country.[6] Thus, the streams effect on inter-district choice should be the same, up to sampling error, in NELS as in non-NELS MSAs. Table 3 reports estimates of the first stage coefficients for Hoxby's instruments and for the alternative instruments, separately for NELS and non-NELS MSAs. Column 1 uses Hoxby's measures. Note that there is a highly significant *positive*

[5] These estimates are computed at the MSA level.
[6] Large MSAs, with many schools, are more likely to appear in the NELS sample. Exploration of specifications that include controls for and interactions of a measure of MSA size indicates that this cannot account for the differences between the first-stage specifications estimated on NELS and non-NELS MSAs.

effect of larger streams on choice among non-NELS MSAs, in contrast with the negative effect among NELS MSAs. A test for equality of the streams coefficients between the two subsets of MSAs rejects decisively, at the 0.001 level. All other coefficients, however, are similar in the two subsamples.

The remaining columns of the table repeat the exercise for the alternative streams variables. In no case is there any indication that parameters vary between NELS and non-NELS MSAs, nor, in any column, are there differences in the control variable coefficients between the two subsamples. It thus appears that the relationship between Hoxby's larger streams measure (and only her measure) and choice is systematically different in the random subsample of MSAs that appear in the NELS data than in the complementary subsample.

Hoxby's response to this extremely troubling result is not convincing. The Reply states (p. 20) that the observed differences "were only to be expected since he [Rothstein] is dropping large parts of the sample" and attributes this to "a non-monotonic relationship" between larger streams and choice. This seriously misunderstands random sampling: Dropping a *random* subset of the sample should not produce large changes in regression coefficients. The only plausible explanation is that the larger streams variable was constructed in systematically different ways for NELS and non-NELS MSAs.

Finally, there are several cases where I can document clear errors in Hoxby's larger streams measure. I have not been able to replicate Hoxby's construction, and the nature of her methods combined with the vagueness of her description of them suggests that this is an impossible task. But there are two categories of errors in Hoxby's measures that are immediately apparent without the need to replicate her methods. Recall that Hoxby (forthcoming, p. 11) characterizes the object of interest as those streams "that are potentially navigable for the purposes of commerce."

In the first category, there are a few MSAs for which Hoxby counts more large streams than there are total streams. An example is Topeka, Kansas, where Hoxby counts 82 larger streams but there are only a total of 41 streams—including small streams—in the metropolitan area. The second category of demonstrable errors consists of metropolitan areas for which Hoxby counts zero large streams when there are well-known rivers that have been used for active commerce. These include Miami, Florida (which grew from a tribal settlement on the banks of the Miami River)[7] and Cedar Rapids, Iowa (whose city hall is on an island in the middle of the Cedar River).[8] Hoxby would have us treat her measure as clearly superior to alternative constructions. Given evidence of unexplained errors in it, her argument must rest on a claim about the relative reliability of her approach and the alternatives. She offers no evidence whatsoever about this.

4) There are serious errors in both the program code and data that Hoxby has distributed. Hoxby has not released the code that was used in her Reply, but she offers no indication that these errors have been repaired.

Hoxby acknowledges just one error in her code: "all that he has found is that the word 'update' is missing from two lines" (Reply, p. 18). In fact, this is the most minor of several errors that I found when attempting to reproduce her results. More important—and easier to locate, because it results in coefficient estimates that are different every time the program is run—is Hoxby's failure to correctly handle "missing" ID variables when performing merges. Inserting "update" into Hoxby's code in the appropriate places will not solve this; her programs will continue to produce a different, erroneous data set every time that they are run.

I have not been able to obtain the programs that Hoxby uses in her Reply. Given her continued failure to acknowledge the serious problems in her earlier code, it is difficult to be

[7] Indeed, Hoxby counts zero *total* streams in Miami. This error derives from Hoxby's failure to standardize county codes across data sets, so that data for the old Dade county is not matched to that from the renamed Miami-Dade County.
[8] A history of steamboating in Iowa states that "[i]n 1859, the Blackhawk made 29 round trips between Cedar Rapids and Waterloo on the Cedar River." See http://www.iptv.org/IowaPathways/mypath.cfm?ounid=ob_000218.

confident that these have been repaired. Until the results reported in the Reply can be replicated, they should be treated as unreliable.

The source data that Hoxby relies on also continue to be plagued by errors. My comment identified one source of many of these errors: Hoxby uses the Common Core of Data (CCD) to assign school districts to metropolitan areas, despite numerous errors in the MSA codes in these data.[9] Hoxby's Reply (pp. 3 & 16) attributes changes in her data since publication to corrections in the CCD MSA codes. This is simply incorrect: The data that she uses have not been corrected since their original release.[10]

Hoxby goes on to dispute my claim that the CCD has many obsolete MSA codes, and writes that "Rothstein's declaring codes to be obsolete amounts to nothing more than his having arbitrarily picked a later year's metropolitan area definitions and saying that codes are obsolete if they do not match that later year's" (p. 17). She evidently misunderstands my Comment: I do not argue that the CCD codes are obsolete because they do not conform to *current* MSA definitions, but because they fail to conform to the *contemporaneous* definitions that were intended to be used. An example is the Bangor, Maine school district. According to the Census, Bangor, Maine is a metropolitan area, and has been one since 1981. According to Hoxby's own description of her variable definitions ("metropolitan area codes that were current at the time the NELS was conducted," Reply, p. 17), schools in Bangor should be assigned to the Bangor MSA. The CCD data sets that Hoxby uses, however, coded the Bangor schools as non-metropolitan, so Hoxby excludes these schools from her

[9] Even this is changed from the original paper. Hoxby's own documentation (Hoxby, 2004, "construct.do" program) states that "The School District Data Book (SDDB) had some erroneous geographic codes. In the original version, the SDDB codes were used….In this version, the Common Core of Data (CCD) has been used for geographic codes because it appears to be more consistently correct." Hoxby's Reply nevertheless excoriates me for writing that she had changed the mechanism from that used in the published paper, stating that instead "the program just assigns each district to the metropolitan area in which it is located" (p. 16).

[10] The CCD is an annual census of public schools; Hoxby's programs rely on the 1987-88, 1989-90, and 1991-92 editions of the data. I compared the initial release of these data (on a CD in the Princeton library's collection) with a version that I downloaded from the CCD web page in Spring 2007. In the entire nation, there were only two school districts whose MSA codes differed, both special education districts in New Mexico.

samples. I provide several additional examples in my Comment. I do not fault Hoxby for the errors in the CCD data, but neither does it seem appropriate to continue relying on the faulty data when a more accurate procedure is available.

It is also worth noting that problems in the CCD MSA codes cascade into Hoxby's streams variables themselves. Hoxby uses the CCD to assign county-level stream counts (including both larger and smaller streams) to MSAs. Errors in the CCD data lead her to mis-assign streams to MSAs. The exact misassignment depends on Stata's sort algorithm, so even the number of streams in many MSAs changes each time Hoxby's programs are run.

II. The Remainder of Hoxby's Reply

Hoxby's Reply takes up the central claims of my Comment, as detailed above, only briefly. The vast majority of the Reply is instead devoted to ancillary arguments that that have little bearing on my central claims and, often, little relationship to anything that appears in my Comment. I do not take up here Hoxby's false characterizations of my Comment, as readers can verify these for themselves simply by comparing what is actually written in my Comment with Hoxby's description of it. Neither is it worthwhile to hash out each of the many factual points that Hoxby gets wrong; it would take pages to document all of the Reply's inaccuracies.

To illustrate the problems, I describe here the single most egregious diversion, concerning my analysis of public and private school students. The NELS allows private schools to be assigned to the zip codes in which they are located, and I use this information to assign the private schools to MSAs. For consistency, in analyses that include the private schools I use the zip code to assign public school students to MSAs as well. Hoxby (Reply, p. 8) claims that I introduce substantial error by doing this, because zip codes cannot always be uniquely assigned to school districts. But this mis-states what I do with the zip code data. I do not attempt to assign zip codes to districts, but

only to MSAs. There are no zip codes in the sample that span multiple MSAs. That Hoxby is unable to successfully do something that I never attempted to do – assign zip codes to school districts – says nothing about the results of my Comment.

Indeed, there is not a single public school in the sample that is assigned to a different MSA via the zip code assignment rule than via the CCD-based rule, once the errors in the CCD are corrected.[11] It is not the use of zip codes that leads to changes in the results in Table 5 of my Comment, the only place that the zip codes are used. Rather, the differences in results are attributable to two other differences in the specifications. First, the zip code analysis uses only the *current* school's zip code to assign students to MSAs, where Hoxby's code and my close replication sample assign students to the same MSA for all three waves of the NELS. Thus, a student who attends school in Chicago in grade 8 but attends a non-metropolitan school in grades 10 and 12 is included as a Chicago student in Hoxby's analyses of test scores in all three grades, but is excluded from the grade-12 sample used for Table 5. Second, the zip code analysis excludes the district-level covariates. Hoxby (2000, pp. 1217-1220) argues correctly that these covariates are not necessary for consistent identification of the choice effect.

In neither case is one specification decision clearly superior to the other. There is no reason to think that the specification in Panel A of Table 5 of my Comment is in any way inferior to those that Hoxby presents.

Hoxby's discussion of zip code matching is but one of many misleading aspects of her Reply. Readers are encouraged to read her Reply carefully and in particular to verify for themselves the Reply's characterization of the contents of my Comment.

[11] There are a few schools – accounting for less than 3% of observations – that are coded as non-metropolitan by one rule but not by the other. This seems to occur when the school is in a different county than the district headquarters. The zip code used is that for the school, while the county code used for CCD-based matching is at the district level.

References:

Hernandez, Javier C. (2005). "Star Ec Prof Caught in Academic Feud." *The Harvard Crimson*, July 8, http://www.thecrimson.com/article.aspx?ref=508253.

Hoxby, Caroline M. (1994). "Does Competition Among Public Schools Benefit Students and Taxpayers?" NBER Working Paper 4978, December.

Hoxby, Caroline M. (2000). "Does Competition Among Public Schools Benefit Students and Taxpayers?" *American Economic Review* 90 (5), December, 1209-38.

Hoxby, Caroline M. (2004). "District-Level and Metropolitan-Area Variables Merged with NELS Data." Electronic Media (CD), National Center for Education Statistics, September 2.

Hoxby, Caroline M. (2005). "Competition Among Public Schools: A Reply to Rothstein (2004)." NBER Working Paper 11216, March.

Hoxby, Caroline M. (forthcoming). "Competition Among Public Schools: A Reply to Rothstein." *American Economic Review*.

Rothstein, Jesse (forthcoming). "Does Competition Among Public Schools Benefit Students and Taxpayers? Comment." *American Economic Review*.