Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)

Jesse Rothstein[*]
Princeton University

December 15, 2004

[*] Industrial Relations Section, Firestone Library, Princeton, NJ 08544; jrothst@princeton.edu. Phone (609) 258-4045; fax (609) 258-2907. I thank Orley Ashenfelter, David Card, Tom Davidoff, Olivier Deschenes, Hank Farber, Alan Krueger, Steve Levitt, Robert McMillan, Cecilia Rouse, Lori Taylor, Miguel Urquiola, Jacob Vigdor, Diane Whitmore, three anonymous referees, and seminar participants at Berkeley, Princeton, and the Southern Economics Association meetings for helpful discussions. Anna Huang provided excellent research assistance. I am grateful to the National Center for Education Statistics for access to confidential data and to Caroline Hoxby for generously making available her data and program code. I also gratefully acknowledge financial support from the Princeton Industrial Relations Section and a Graduate Research Fellowship from the National Science Foundation. Any errors, opinions, findings, conclusions, or recommendations are mine alone and do not necessarily reflect the views of the NSF or NCES.

Does Competition Among Public Schools Benefit Students and Taxpayers?
A Comment on Hoxby (2000)

I. Introduction

School choice policies promise to align the incentives of school administrators with the demands of parents, and may therefore lead to more efficient educational production (Friedman, 1962; Brennan and Buchanan, 1980; Chubb and Moe, 1990). Absent a large-scale school voucher program in the United States, however, this prediction has been difficult to test. Several authors (e.g. Borland and Howsen, 1992; Belfield and Levin, 2002) have suggested studying the effects of "Tiebout choice," the use of the residential location decision to select among local monopoly education providers. The idea here is that fragmented governance induces competition among school districts analogous to that which would occur among schools with non-residential choice.

In an influential paper, Hoxby (2000) points out that current governance structures are potentially endogenous to school productivity, and proposes that variation in topography, which may have influenced optimal jurisdiction size before modern transportation technologies, provides a source of exogenous variation. She estimates instrumental variables regressions of individual test scores and school spending on a metropolitan-level Tiebout choice index, defined as one minus a Herfindahl concentration index with districts' enrollments as their "market shares," using as excluded instruments the number of larger and smaller streams in the area. She reports substantial positive effects of district fragmentation on student test scores and negative effects on spending.

This comment presents a reanalysis of Hoxby's test score results, which form the core of her empirical analysis. These results turn out to be quite sensitive to plausible alterations to Hoxby's specification. In particular, the large, significant effect of choice on achievement obtains only with Hoxby's particular streams variables. When I substitute alternative and arguably better constructions of the same variables, I obtain smaller estimates that are never significant. There is also some evidence of sample selection bias, deriving from Hoxby's decision to exclude private school students from the analysis. I conclude that Hoxby's positive estimated effect of interdistrict competition on student achievement is not robust, and that a fair reading of the evidence does not support claims of a large or significant effect. Similarly, I find little compelling evidence of endogeneity of the choice index to school quality, suggesting that the more precise OLS estimate of zero choice effect on test scores should be preferred to less precise IV estimates. The evidence that competition among schools will improve academic outcomes is thus substantially weaker than it might have appeared.

Professor Hoxby's response to this Comment follows. I dispute many of the claims made there. A discussion (Rothstein, 2007) of her Reply is available at my web site (http://www.princeton.edu/~jrothst/hoxby/index.html).

Section II focuses on replication. Despite several requests, Hoxby has not provided the precise data set from which her published results were derived. She has, however, made available a corrected data set (Hoxby, 2004a). The new data generate results that deviate in important ways from those that were published. In particular, the first stage coefficients, and even basic summary statistics for the streams variables, are substantially different. Moreover, there appear to be errors remaining in Hoxby's data and computer programs, causing some students to be assigned to the wrong metropolitan statistical areas (MSAs) and some others to be randomly assigned to districts and MSAs. When I correct these errors, I obtain somewhat weaker results. In what I consider the best replication sample, Hoxby's specification and instruments indicate an insignificant or marginally significant effect of choice (i.e., district fragmentation) on student achievement.

In Section III, I consider the sensitivity of the results to the particular instrumental variables used. Hoxby's discussion does not make clear precisely how her larger and smaller streams counts are defined. In particular, though Hoxby writes that the source of her smaller streams variable provides "the longitude and latitude of [each stream's] origin and destination" (2000, p. 1222), she actually uses only streams' destinations to assign them to MSAs. A stream that flows through an MSA but ends elsewhere is not included in the MSA's count. I present results using an alternate variable that counts all streams flowing through each MSA, regardless of where they end. I also demonstrate that Hoxby's larger streams variable is key to the results, and that it plays a substantially different role in the first stage to the individual-level IV model than in the MSA-level model that Hoxby presents as "the implied first-stage regression" (2000, p. 1224-5).[1] The choice coefficient shrinks by 45 to 85% and ceases to be significant when the larger streams variable is excluded. I obtain similarly small and insignificant coefficients when I substitute alternative larger streams counts that, unlike Hoxby's subjectively coded variable, are readily replicable using public-use data.

Finally, Section IV explores the implications of Hoxby's exclusion of private school students from her sample. Hoxby documents a negative relationship between the Tiebout choice index and the metropolitan private enrollment rate. This may produce selection bias in specifications, like Hoxby's, that are estimated only on public sector students (Hsieh and Urquiola, 2006). Estimates from samples that include both public and private school students are free of this potential sample selection bias, and are notably smaller than those from public-sector samples. None are significantly different from zero, even with Hoxby's instruments.

II. Replication

Table 1 presents IV estimates of the district fragmentation effect on each of two test scores, using Hoxby's streams variables as instruments.[2] The first column reproduces the estimates from Hoxby's Tables 3 and 4. Hoxby's preferred specification is that for 12th grade reading scores in Panel A, although I analyze 8th grade scores as well (in Panel B) because the sample sizes are so

[1] The IV model could be estimated at the MSA level as well, as both the endogenous variable (choice) and instruments (streams) vary only across MSAs. Hoxby (2000, p. 1219) claims that her specification "is most efficiently estimated at the individual level." I follow this decision throughout, though I present MSA-level estimates in the appendix.

[2] The student test score data are drawn from the National Educational Longitudinal Study (NELS). Details of the data set construction, along with summary statistics, control variable coefficients, and alternative specifications, are in an appendix available from the author.

much larger.[3] Hoxby assumes that the student-level error term is composed of three homoskedastic components, one common to all students in the same metropolitan area, another common within the district, and the last specific to the student. She computes standard errors using an FGLS estimator, due to Moulton (1986), that accounts for the implied student-level serial correlation. The estimated choice effect is positive and significant in each panel.

An earlier version of this comment discussed several alternative algorithms for assigning students in the NELS data to school districts and metropolitan areas (MSAs), as Hoxby's (2000) discussion did not specify her approach. In response to that draft, Hoxby re-evaluated her assignment algorithm and discovered some errors (Hoxby, 2004c). She has made available, via the National Center for Education Statistics (NCES), a corrected data set that uses a new crosswalk.[4] Column 2 reports estimates from the Hoxby/NCES data, which provide substantially smaller samples than were used in the published results. Hoxby's computer program, also provided (Hoxby, 2004b), does not compute the "Moulton" standard errors that were used in the published paper, but instead uses Stata's "cluster" option to generate standard errors which are consistent in the presence of arbitrary heteroskedasticity and within-MSA serial correlation. I have implemented the Moulton estimator, and I report both Moulton and clustered standard errors for each specification in Table 1.[5] Estimates from Hoxby's corrected data (hereafter, the "Hoxby/NCES" data) have somewhat

[3] I prefer the 8th grade sample, as its design is much more straightforward than in later waves. Students were randomly sampled from within their schools in the 8th grade, then followed across schools in successive waves. As a result, the follow-up samples are not representative of the schools their students attend, nor of their districts or metropolitan areas, though they remain nationally representative. Also, as with any panel data, sample attrition is a potential problem in later survey waves.

[4] The corrected data set and the programs used to construct it are available from NCES to researchers who are licensed for access to the restricted-use NELS data.

[5] Hoxby writes that "Robust [clustered] standard errors are larger than standard errors calculated using the Moulton method" (Hoxby 2004b). Both estimators are consistent (with asymptotics in the number of MSAs) under the error components model, and there is no model in which the Moulton estimator is consistent but the cluster estimator is not. A difference between the two estimators may indicate that the error components assumption is incorrect; in that case, cluster is consistent but the Moulton estimator is not. Further discussion of the two estimators, and of my implementation of the Moulton estimator, is in the appendix.

larger standard errors than did those in the published paper, and the 12th grade coefficient ceases to be significant (at the 5% level) when clustered standard errors are used.

In examining the Hoxby/NCES data and code, I have found several remaining glitches. First, some errors remain in the new district-MSA crosswalk: Several Ohio school districts are assigned to the Raleigh-Durham MSA; several additional districts have incorrect, invalid or obsolete MSA codes; and over one quarter of metropolitan districts are missing MSA codes. Second, though the clear intent is to use all three waves of the NELS survey to assign students to districts, due to an apparent coding error information about students' second- and third-wave schools is ignored.[6]

Finally, students with missing school IDs from the first wave of the NELS survey—the sample was freshened in later waves—are randomly assigned to schools that entered the survey in later waves. This occurs because Hoxby's program fails to exclude observations with missing IDs when merging the student and school files. Stata's sort algorithm breaks ties randomly when, as here, a unique sort order is not specified. Stata's merge procedure then assigns the first observation with a missing ID from the "master" data set to the first similar observation from the "using" data set, the second to the second, and so on. Because ties among students and schools with missing IDs are broken differently every time the sort command is run, each execution of Hoxby's program produces a different data set, and different estimated choice effects.[7] To gauge the severity of this unintended stochasticity, I executed Hoxby's data construction program 10,000 times, tabulating the estimated choice effect from each resulting data set. The histogram is available as Appendix Figure A1. The mean choice effect for 12th grade scores is 5.39, quite close to the 5.30 computed from the

[6] Hoxby merges the NELS student file to the NELS school file three times in succession, using school ID variables from each of the three survey waves. By the second merge, all variables from the school file exist on the student file. Without specific instruction (which is not provided), the merge command in Stata does not overwrite variables that already exist on the "master" file, so nothing on the student file is altered by the second and third merges.

[7] Hoxby's program also fails to account for Stata's tie-breaking procedure when creating the MSA-level data set used for her first stage model, and her program thus assigns the Raleigh MSA to the East North Central division (which contains Ohio; see above) 36% of the times it is executed; the Hoxby/NCES data set is one such draw from the distribution.

Hoxby/NCES data. The standard deviation across iterations (0.47) is not particularly large, but the range is quite wide: I obtained estimates as small as 2.17 and as large as 8.15.

After discovering these anomalies, I re-wrote Hoxby's data assembly program, fixing errors in the district-MSA crosswalk and taking care to correctly match students, schools, districts, and metropolitan areas. I attempted to follow Hoxby's algorithm as closely as possible.[8] I did not at this point attempt to reproduce the "larger streams" variable, but simply relied on the MSA-level count that Hoxby provided and discarded MSAs that were excluded from her tabulation.[9] Results are presented in column 3 of Table 1. Sample sizes are somewhat larger—correctly assigning districts that were previously classified as non-metropolitan more than offsets the loss of students who are reclassified to an MSA with a missing larger streams value—and approach those seen in Hoxby's Table 4. Coefficients resemble those found in the Hoxby/NCES data, somewhat smaller for 12th grade scores and somewhat larger for 8th grade scores, with similar patterns of significance.

Column 4 represents a somewhat more expansive interpretation of replication. I retain Hoxby's specification, but I follow my own judgment in sample and covariate construction rather than directly following her algorithm. Where Hoxby assigns each student to a single district for all three waves even if the student moved between waves, for this sample I use only contemporaneous information to construct distinct assignments for each wave. There are also minor differences in

[8] There were some ambiguities. In particular, each student has nine potential district codes, as each student may have a school code in each of three waves and each school may have different district codes in each wave. Hoxby attempts to assign a single district code for each student, to be used with data from all three waves, but the aforementioned coding errors mean that only the three district codes from the first-wave school are considered. It is not clear how she would resolve discrepancies among the larger set. I assign each student to a separate district for each wave, using only contemporaneous information from the student and school files, then use Hoxby's majority rule algorithm to select among the three resulting assignments.

[9] Hoxby uses 1990 MSA definitions. Puzzlingly, she does not provide counts of larger streams for all of the MSAs included in these definitions, but does provide counts for some obsolete MSA codes—from the 1983 or 1981 MSA definitions—that appear in her faulty crosswalk. For example, 19 larger streams are reported for MSA number 3755, which corresponded to the Kansas City, KS PMSA in 1983 but was included in the Kansas City MO-KS MSA (number 3760) in 1990; there is also an entry of 37 larger streams in MSA 3760. It is not clear what algorithm might have produced this redundancy, nor whether the latter count includes the streams attributed to the former.

variable definitions.[10] Choice effect estimates are smaller with this sample. For 12th grade scores, the choice effect is insignificant regardless of the standard error computation; for 8th grade scores, it is insignificant with the random effects standard errors but significant when the errors are clustered.

Panel A of Table 2 reports mean values of the streams variables. Column 1 is from Hoxby's Table 2, while columns 2 and 3 are computed from the Hoxby/NCES data set and from my replication sample, respectively. There are substantial differences between columns 1 and 2. For some reason, the mean of the larger streams variable is more than five times larger than that reported in the published paper, while the average MSA has only two thirds as many total—larger plus smaller—streams as is indicated by Hoxby's (2000) Table 2.

Both the streams variables and the potentially endogenous choice measure vary only at the MSA level. Though Hoxby's IV estimates are computed at the student level, Hoxby reports only an MSA-level "implied first-stage regression." I reproduce this specification in Panel B, with the published estimates in column 1, those from the Hoxby/NCES data in Column 2, and those from the replication samples in 3 and 4.[11] All of the replication estimates are substantially different from those in the published paper. Comparing the Hoxby/NCES estimates to the published results, the larger streams coefficient has fallen by more than 80% and is no longer remotely significant, while the smaller streams coefficient has tripled. Though both of these findings are somewhat attenuated in the replication data sets, they remain worrisome: The logic of the argument for Hoxby's instruments is that streams once represented impediments to travel, and one would expect this to be far more true for larger than for smaller streams, particularly when the threshold for being a "larger"

[10] The largest difference is in what Hoxby calls the "mean of log(income) of metropolitan area" variable. She uses an arithmetic weighted average of the log of each district's mean income; I use instead the log of the MSA mean income. There are also minor differences in the Gini coefficient and the racial composition variables. Finally, I compute the choice index over 8th grade enrollment, where Hoxby uses total enrollment, reasoning that parents cannot be said to choose between overlapping elementary and secondary districts (Urquiola, 2005). Further details are in the appendix.

[11] The replication data sample sizes are somewhat smaller, as several invalid MSA codes that were on the Common Core of Data file from which Hoxby took her district-MSA assignments are no longer present and some newly added MSA codes must be excluded for lack of the larger streams variable.

stream is set low enough to include over 40 streams from the average MSA (rather than the 8 indicated in the published paper).

As noted above, the MSA-level estimates are not the actual first stages for the individual-level models in Table 1. The actual first stages are reported in Panel C (for the 12th grade samples) and D (for the 8th grade samples). The streams coefficients are dramatically different: Larger streams are now *negatively* related to choice in five of the six samples, once significantly and once nearly so.[12] Again, this is difficult to reconcile with the story behind the identification strategy.

III. Counting Streams

There are several reasons to worry about the validity of Hoxby's larger streams variable: It derives from Hoxby's subjective count from printed maps—she describes counting streams "of a certain width on the map," (2000, p. 1222), but does not elaborate; it is missing for several MSAs that were inadvertently excluded from Hoxby's sample;[13] and, as Hoxby writes, "one has more a priori confidence in the exogeneity of the smaller streams variable because smaller streams are too small to affect modern life," (2000, p. 1230). Given the evident differences between the larger streams variable described in the published paper and the one included in the Hoxby/NCES data, it is unclear whether the discussion in Hoxby's text even applies to the latter variable.

These concerns cannot be addressed by using the smaller streams variable as the sole instrument, however. Hoxby uses the U.S. Geologic Survey's Geographic Names Information System (GNIS) to count total streams, and defines smaller streams as the number of total streams

[12] The divergence between the MSA-level results in Panel B and the individual-level results in Panels C and D appears to derive from differences in the set of MSAs included. Hoxby's first stage estimates and those that I report in Panel B include all MSAs, regardless of whether they contain NELS sample students. When I restrict the sample to those in the NELS data (Appendix Table D5), coefficients are similar to those in Panels C and D. Efficiency can be improved with two-sample IV, using the full sample of MSAs to estimate the first stage. In the Hoxby/NCES data, this yields choice coefficients of 3.68 for 8th grade scores and 2.14 for 12th grade scores, both substantially shrunken from the estimates in Table 1 and neither significant (Appendix Table D6).

[13] One indication that there may be problems with Hoxby's larger streams count is that when I correct Hoxby's code to correctly assign total streams to MSAs—her incorrect district-MSA crosswalk is used here as well—there are several MSAs with fewer total streams than larger streams. Hoxby writes that the hand counts were "checked against" the GNIS data (2000, p. 1222), but appears not to have caught all discrepancies. Though I argue below that Hoxby systematically undercounts total streams, my correction of this problem reduces but does not eliminate the discrepancies.

less the count of larger streams. As a result, any errors in the larger streams variable appear as errors of the opposite sign in the smaller streams count. To avoid reliance on Hoxby's larger streams count, I present estimates that use the total streams count—which can be produced using Hoxby's code from the public-use GNIS data set—as the single instrument.

I also explore an alternative specification for the "total streams" variable. Despite her reference to GNIS variables describing the longitude and latitude of streams' origins and destinations, Hoxby's code uses only a variable indicating the county where a stream's destination (mouth) is located to assign streams to MSAs. To illustrate the consequences of this, the Mississippi River is attributed only to the non-metropolitan Plaquemines Parish, Louisiana, and not to any of the eight metropolitan areas along its banks.[14] There is little reason to think that a stream's destination is the key to either its past effects on travel costs or to its current effects on district structure. The USGS distributes an alternative version of the GNIS data that codes each county through which each stream flows, from origin to destination. Using this data file, I construct a "total streams" measure that counts toward an MSA's total any stream flowing through it.[15]

Finally, I explore alternative classifications of streams into "larger" and "smaller" groups. First, following Hoxby (1994), I compute separate counts of inter-county and intra-county streams and enter them as separate instruments. I also categorize streams based on their lengths, computed as the distance between their sources and mouths, following Hoxby (2000) in requiring a larger stream to exceed 3.5 miles. Each is a crude measure for the variation of interest, but it is difficult to see how either might be endogenous; as a result, either should provide consistent IV estimates of the

[14] This is not documented in the published paper. It does not automatically mean that inland cities lack streams, as a smaller stream's mouth might be located where it feeds into a larger river. Note that the Mississippi may be included in the *larger* streams counts for the relevant MSA's, though it is not counted toward the *total* streams. This appears to account for some but not all of the negative smaller streams counts discussed in footnote 13.

[15] In most of the country, MSAs are composed of whole counties. In New England, however, towns are the basic unit, and some counties are split among several MSAs. Hoxby assigns all of each county's streams to the MSA containing the plurality of its population. When I reproduce her stream mouths variable, I follow her all-or-nothing rule; my total streams count instead assigns streams fractionally to MSAs in proportion to the MSAs' shares of the county population.

choice effect.[16] These estimates provide a check on the robustness of the earlier estimates, and have the virtue of being easily replicable using the public-use GNIS data.

Table 3 presents instrument means (Panel A) and first-stage estimates (Panels B-D, using the close replication sample) for several instrument sets. As before, the first stage is computed at both the MSA and individual levels; corresponding estimates using my alternative sample and covariate definitions are similar and are reported in the appendix. For a benchmark, Column 1 reproduces the estimate from Column 3 of Table 2, using Hoxby's streams variables. Column 2 uses only total streams (by Hoxby's definition, counting only stream mouths), which have positive coefficients at both the MSA and individual levels. Columns 3 and 4 repeat these specifications, using the count of all streams flowing through each MSA in place of the count of stream mouths. This change has little effect on the estimates, with the negative larger streams coefficient still evident in the individual-level model. Columns 5 and 6 use alternative definitions for "larger" streams, first as inter-county streams and second as streams exceeding 3.5 miles in length. Using either definition and in both the MSA and individual samples, the larger streams variable accounts for the full effect of streams on choice, a result that is consistent with the idea that the role of streams derives from their importance as natural barriers to travel.

For each set of instruments, Table 4 reports IV estimates of the choice effect on 12th and 8th grade reading scores, Moulton and clustered standard errors, and p-values for tests of the exogeneity of the choice variable (using the cluster estimator).[17] I also report OLS estimates, each of which indicates a negligible choice effect.

The choice effects are consistently positive and exogeneity of the choice variable is consistently rejected when Hoxby's larger streams count is included as an instrument. Neither of

[16] Measurement error in instruments, so long as it is uncorrelated with the endogenous variable, reduces the precision of IV estimates but does not affect consistency as long as the measures are sufficiently reliable to avoid so-called "weak instruments" problems. As I show below, the first stages are quite strong.

[17] I obtain similar results with Moulton standard errors or when I use the preferred replication sample and covariates.

these results holds in any of the specifications that exclude Hoxby's larger streams variable, however. This is partly because the latter estimates are less precise, but this is not the whole story: The coefficient estimates are also uniformly smaller, generally less than half as large, when Hoxby's larger streams variable is excluded.

Taking the estimates in Table 4 together, it is clear that Hoxby's conclusions depend critically on her count of larger streams. I attempted my own count for several MSAs that contribute most to the large choice effect estimates, using the same 1/24,000 quadrangle maps that Hoxby reported using. It quickly became apparent that counting streams involves many subjective judgments.[18] Hoxby describes larger streams as those that "were at least 3.5 miles long and of a certain width on the map" (2000, p. 1222), but does not specify what constitutes "a certain width" nor where in a stream's course the width is to be measured. I began with Fort Lauderdale, which may be a particularly difficult case as much of the MSA is swampland and much of the remainder was recovered from swampland by a system of man-made canals. (Even today, airboat trails are more common through much of the MSA than is dry land; it seems unlikely to have been settled by people who viewed water as an obstacle to travel.) I decided not to count canals which ran perfectly straight, generally exactly West to East, but I did count canals which took irregular paths, reasoning that the latter were more likely to correspond to pre-existing rivers. I also counted branches of streams as separate from their parents when they had distinct names (such as the North and South Forks of the Middle River), and counted the intracoastal waterway, which separates the easternmost portion of the Florida coast from the mainland, as a stream for its similar effect on the ease of travel. Where Hoxby reports 5 larger streams in Fort Lauderdale, I counted 12, and a research assistant—working independently—counted 15.

[18] I worked without reference to Hoxby's counts, to prevent being influenced by these. Hoxby's text is confusing about whether linear bodies of water other than streams are included in her count. Her footnote 24 seems to suggest that they are not, but her footnote 16 indicates that she counts "inlets, lakes, ponds, marshes, and swamps" "*if they are roughly curvilinear in form*" (emphasis in original). I followed the latter rule.

I had a similarly difficult experience with other MSAs, finding that many rivers divide and recombine multiple times, become wider and narrower, and are interrupted by man-made structures throughout their courses. My counts were correlated with Hoxby's, but generally not identical. The exercise makes clear that Hoxby's larger streams variable is subjective and unverifiable without a list of the particular rivers coded as large. In the absence of such a list, which Hoxby has not provided, no two researchers would come up with identical counts. As I have only counted streams for a few MSAs, however, I cannot be certain of the sensitivity of Hoxby's results to the differences that would inevitably arise.

IV. Private Enrollment and Selection Bias

I have concerned myself thus far with replication of Hoxby's primary specification, and with its robustness to plausible alternative decisions about sample and variable construction. In this section, I turn to another issue: Hoxby's specification may not provide consistent estimates of the effect of interest, that of choice on public school productivity, because her sample excludes private school students. In her Table 6, she documents that choice has a significant negative effect on the metropolitan private enrollment share.[19] As a result, Hoxby's specification may be subject to selection bias even with valid instruments (Hsieh and Urquiola, 2006). The reasoning is simple: Suppose that the distribution of student test scores is identical across MSAs when both public and private school students are included, but that MSAs vary in private enrollment patterns. In particular, suppose that some relatively high-scoring students would choose private schools in a low-choice market but would remain in the public sector when Tiebout choice is sufficient to provide public schools with desired characteristics (Rothstein, 2006). Then the average test score among

[19] Using both of her streams instruments in a district-level regression, Hoxby (2000, Table 6) estimates that a one-unit increase in choice leads to a 4.2% (s.e. 1.2%) reduction in private enrollment. Hoxby's SDDB data set double-counts students in areas served by separate elementary and secondary districts. When I instead estimate the relationship at the MSA level, I estimate a choice effect of -4.8% (s.e. 2.4%), though this result is somewhat sensitive to the sample and covariate construction.

public school students will tend to be higher in high-choice markets purely as a result of differential sample selection.

Any resulting bias is present in both OLS and IV estimates, though its sign and magnitude depend on whether the marginal private school student is positively or negatively selected. If the average score is higher among students drawn into the public sector by expansions of choice than among inframarginal public school students, estimates from public school students are (asymptotically) upward-biased; if the average score is lower among marginal students than among the inframarginal, these estimates are downward-biased.[20] Hoxby seems to make the former claim when she discusses the consequences of "families with a strong taste for education leav[ing] the public sector by shifting their children into private schools" (2000, p. 1233).

As the NELS survey includes both public and private school students, this potential bias can be easily avoided by simply including both groups in the sample. [21] The only hurdle is that the CCD cannot be used to assign private schools to school districts and MSAs. As an alternative, I use NELS variables characterizing the demographic composition of the school's zip code to uniquely assign the vast majority of schools to zip codes, and via these to MSAs. [22] As many zip codes span school districts, I cannot use this strategy to assign school districts, and I therefore must exclude district-level covariates from the specification.[23]

[20] NELS private school students score nearly half a standard deviation higher on the 8th grade reading test than do public school students. This is not particularly informative, however, as the students whose sectoral decision is sensitive to Tiebout choice are likely atypical of the inframarginal private school population.

[21] Under fairly strong assumptions—including that private schools are not systematically better or worse than public schools; that competition has similar effects on the productivity of public and private schools; and that any peer effects are linear and additive, so that stratification does not have an independent effect on average scores— an unbiased estimate of the choice effect on average school productivity can be obtained by estimating Hoxby's specification on a pooled sample of public and private school students (Hsieh and Urquiola, 2006). Hoxby (1994) uses exactly this strategy to test for selection bias from private school enrollment.

[22] In the rare cases where a zip code spans multiple MSAs, I assign each student attending school in that zip code to each MSA, with weights proportional to the fraction of the zip code population in each MSA.

[23] Hoxby (2000, Section 7) argues at great length that the inclusion of district-level variables improves the precision but does not affect the coefficients on MSA-level variables as long as MSA-level means are included in the specification. Strictly, this is only true in the limit, as it relies on the assumption that the district-level variables aggregate exactly within

Panel A of Table 5 reports estimates from public school students who have been matched to MSAs via their schools' zip codes, using both the "close" and "preferred" covariate definitions. Estimates are substantially smaller than those presented earlier, with the divergence due more to the different methods of assigning MSAs than to the exclusion of district-level covariates.[24] Panel B adds the private school students to the sample. The choice effect estimates fall notably farther here, and *t*-statistics are uniformly less than one.

I read the estimates in Table 5 as suggesting, but not conclusively demonstrating, that the students drawn into the public sector by expansions of choice are somewhat positively selected.[25] While much of the difference from earlier estimates appears to derive from sensitivity of the results to the exclusion of district-level covariates and to the method by which schools are assigned to MSAs, point estimates do fall even farther when private school students are added to the sample.

V. Discussion

Hoxby's analysis has been very influential, providing what many (e.g. Howell and Peterson, 2002; Maranto, 2001; Bast and Walberg, 2004) have seen as some of the most compelling extant evidence in favor of the proposition that school choice will lead to improvements in the efficiency of educational production. Unfortunately, Hoxby's key results do not seem to be robust to small, reasonable alterations to the sample or to the instrumental variables used. Interested readers are invited to explore alternative specifications beyond those considered here; code to construct both of

the sample to the MSA-level means. In small samples this is not likely to hold, and the choice coefficient is somewhat smaller (more negative) when district-level covariates are excluded from Hoxby's specification (Appendix Table D3).

[24] The declines are largest in the close replication sample, as my zip code matching algorithm, which uses only the contemporaneous school, is more similar to that used in the preferred sample. Students in the close replication sample who were assigned to MSAs based on their 8th or 10th grade school's district code in Panel B are assigned using the 12th grade school's zip code in Panel C.

[25] As an alternative test for selection bias, I have estimated a version of Hoxby's specification (using only public school students) that includes a control for an inverse Mill's ratio computed from the MSA private enrollment rate, in the spirit of normal-distribution selection corrections (Gronau, 1974; Heckman, 1979; Card and Payne, 2002). Estimates of the selectivity parameter were extremely imprecisely estimated, and the selection correction had little effect on the estimated choice coefficients.

my replication samples and to perform all analyses is available from my web page, as are all data components that I am at liberty to distribute.

As I document above, there are several problems with the Hoxby/NCES data set. When these are remedied, I estimate somewhat weaker effects of choice on student performance than those that Hoxby reports.[26] When I consider slight adjustments to her specification of the streams variables—such as replacing them with plausible, replicable alternative measures—or when I alter the sample to avoid potential selection bias from private enrollment, the significant effect of Tiebout competition on student scores is greatly attenuated and not statistically distinguishable from zero. In my specification including private school students, using my preferred sample, and instrumenting with inter- and intra-county streams (Table 5, Panel B, Column 6), I estimate that a one standard deviation increase in choice raises test scores by just under 0.05 standard deviations, with a standard error somewhat larger than that. This compares unfavorably to, for example, the 0.22 standard deviations that Krueger (1999) estimates as the effect of reducing elementary school class sizes from 22 to 15 students in the Tennessee STAR experiment.

I do not find support, in any of the alternative specifications that I consider, for Hoxby's claim that "naïve estimates (like OLS) that do not account for the endogeneity of school districts are biased toward finding no effects" (2000, p. 1236), nor for her conclusion that "Tiebout choice raises productivity by simultaneously raising achievement and lowering spending" (p. 1236-7). Any relationship between choice and student test scores is too imprecisely estimated to be robustly distinguishable from zero. Hoxby's results for the effect of district fragmentation on school spending, which I examine in the appendix, are only slightly more robust.[27]

[26] The current analysis has not considered Hoxby's analysis of the NLSY, which echoes her NELS analysis in indicating a salutary effect of interdistrict competition on attainment. Hoxby seems to find her NELS estimates the most compelling, however, and focuses her discussion on these.

[27] Hoxby (2000, Table 5) reports a choice effect on the log of per pupil spending of -0.076 (Moulton standard error 0.034). The Hoxby/NCES data yield an estimate of -0.074 (0.141); IV estimates in the replication samples similarly fail to reject zero, although OLS estimates are significantly negative.

There are only a few hundred metropolitan areas in the United States, and this is evidently too few to precisely estimate any relationship that may exist between jurisdictional fragmentation and either student performance or school spending. One cannot reject large effects of competition, but neither is there strong evidence against a hypothesis of zero effect. It would be premature to conclude that schools respond to Tiebout competition by raising productivity, nor to use such a conclusion as justification for policies that expand non-residential forms of school choice.

References

Bast, Joseph L. and Walberg, Herbert J. "Can Parents Choose the Best Schools for Their Children?" *Economics of Education Review*, August 2004, *23*(4), pp. 431-40.

Belfield, Clive R. and Levin, Henry M. "The Effects of Competition between Schools on Educational Outcomes: A Review for the United States." *Review of Educational Research*, Summer 2002, *72*(2), pp. 279-341.

Borland, Melvin V. and Howsen, Roy M. "Student Academic Achievement and the Degree of Market Concentration in Education." *Economics of Education Review*, March 1992, *11*(1), pp. 31-39.

Brennan, Geoffrey and Buchanan, James. *The Power to Tax: Analytical Foundation of a Fiscal Constitution.* Cambridge: Cambridge University Press, 1980.

Card, David and Payne, A. Abigail. "School Finance Reform, the Distribution of School Spending, and the Distribution of SAT Scores." *Journal of Public Economics*, 2002, *83*(1), pp. 49-82.

Chubb, John and Moe, Terry M. *Politics, Markets, and America's Schools.* Washington, D.C.: The Brookings Institution, 1990.

Friedman, Milton. *Capitalism and Freedom.* Chicago: University of Chicago Press, 1962.

Gronau, Reuben. "Wage Comparisons--a Selectivity Bias." *The Journal of Political Economy*, Nov. - Dec. 1974, *82*(6), pp. 1119-43.

Heckman, James J. "Sample Selection Bias as a Specification Error." *Econometrica*, 1979, *47*, pp. 153-61.

Howell, William G. and Peterson, Peter E. "Impact of School Voucher Research." *PS-Political Science & Politics*, December 2002, *35*(4), pp. 659-60.

Hoxby, Caroline M. "Do Private Schools Provide Competition for Public Schools?" National Bureau of Economic Research Working Paper #4978, December 1994a.

____. "Does Competition among Public Schools Benefit Students and Taxpayers?" National Bureau of Economic Research Working Paper 4979, December 1994b.

____. "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review*, December 2000, *90*(5), pp. 1209-38.

____. "District-Level and Metropolitan-Area Variables Merged with NELS Data." CD, National Center for Education Statistics, September 2, 2004a.

____. "Documentation to 'District-Level and Metropolitan-Area Variables Merged with NELS Data': Construct.Do." National Center for Education Statistics, September 2, 2004b.

____. "Documentation to 'District-Level and Metropolitan-Area Variables Merged with NELS Data': Runregressions.Do." National Center for Education Statistics, September 2, 2004c.

Hsieh, Chang-Tai and Urquiola, Miguel. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's School Voucher Program." *Journal of Public Economics*, September 2006, *90*(8-9), pp. 1477-1503.

Krueger, Alan B. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics*, May 1999, *114*(2), pp. 497-532.

Maranto, Robert. "Finishing Touches." *Education Next*, Winter 2001, *2001*(4), pp. 20-25.

Moulton, Brent R. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, August 1986, *32*(3), pp. 385-97.

Rothstein, Jesse M. "Good Principals or Good Peers? Parental Valuation of School Characteristics, Tiebout Equilibrium, and the Incentive Effects of Competition among Jurisdictions." *American Economic Review*, September 2006, *96*(4), pp. 1333-1350.

____. "Rejoinder to Hoxby." Unpublished manuscript, posted at http://www.princeton.edu/~jrothst/hoxby/rejoinder.pdf, 2007.

Urquiola, Miguel. "Does School Choice Lead to Sorting? Evidence from Tiebout Variation." *American Economic Review*, September 2005, *95*(4), pp. 1310-1326.

Table 1: IV estimates of choice effect on NELS 8th and 12th grade reading scores in several samples, Hoxby specification

	Published	Hoxby/ NCES data	Close replication sample	Preferred sample and covariates
	(1)	**(2)**	**(3)**	**(4)**
Panel A: 12th grade reading scores				
# of students	6,119	5,475	5,934	6,688
# of MSAs	209	184	194	199
Choice index coefficient	5.77	5.30	4.74	3.29
S.E. (Moulton)	**(2.21)**	**(2.36)**	**(1.98)**	(1.83)
S.E. (Cluster)		(2.94)	**(2.42)**	(2.56)
P-values, exogeneity test (clustered)		0.02	0.02	0.20
Panel B: 8th grade reading scores				
# of students	10,790	10,175	10,429	11,719
# of MSAs	211	185	186	184
Choice index coefficient	3.82	4.45	5.93	2.93
S.E. (Moulton)	**(1.59)**	**(1.87)**	**(2.10)**	(1.58)
S.E. (Cluster)		**(1.99)**	**(2.32)**	**(1.40)**
P-values, exogeneity test (clustered)		0.00	0.00	0.00

Notes: See Hoxby (2000) and data appendix for description of data, samples, and covariates. Column 1 is from Hoxby (2000), Table 4. Standard error estimators and exogeneity tests are described in the appendix. Following Hoxby, all analyses use NELS sa

Table 2: Overview of first stage estimates, different samples
Dependent variable is MSA-level choice index (1- index of concentration across districts)

	Published	Hoxby/ NCES data	Close replication sample	Preferred sample and covariates
	(1)	(2)	(3)	(4)
Panel A: MSA-level sample means				
Larger streams	8	44	45	
Smaller streams	183	84	80	
Panel B: MSA-level first stage estimates				
Larger streams (100s)	0.080	0.012	0.040	0.043
	(0.040)	(0.021)	(0.021)	(0.021)
Smaller streams (100s)	0.034	0.096	0.093	0.091
	(0.007)	(0.019)	(0.018)	(0.018)
N	316	310	304	304
F statistic (instruments)	24.4	14.8	16.2	16.3
Panel C: Individual-level first stage estimates (12th grade reading sample)				
Larger streams (100s)	nr	-0.043	-0.024	0.015
		(0.023)	(0.020)	(0.020)
Smaller streams (100s)	nr	0.133	0.133	0.114
		(0.021)	(0.017)	(0.018)
N	nr	5,475	5,934	6,688
F statistic (instruments)	nr	20.5	31.3	28.4
Panel D: Individual-level first stage estimates (8th grade reading sample)				
Larger streams (100s)	nr	-0.045	-0.033	-0.012
		(0.021)	(0.018)	(0.018)
Smaller streams (100s)	nr	0.131	0.130	0.132
		(0.022)	(0.017)	(0.017)
N	nr	10,175	10,429	11,719
F statistic (instruments)	nr	17.6	30.7	32.1

Notes: "nr"=not reported. Column 1 is from Hoxby (2000), Table 2. Sample sizes in Panels C and D are identical to those in the corresponding columns of Table 1, Panels A and B respectively. Standard errors are clustered in Panels C and D, but are conv

Table 3: First-stage estimates for alternative instruments, using "close replication" sample and covariates

	(1)	(2)	(3)	(4)	(5)	(6)
Total stream definition:	**Stream mouths**		**All streams**			
Larger stream definition:	**Hoxby**	**n/a**	**Hoxby**	**n/a**	**Inter-county**	**>3.5 miles**
Panel A: MSA-level sample means						
Larger streams	45		45		41	70
Smaller streams	80		108		107	75
Total streams		124		148		
Panel B: MSA-level first stage estimates						
Larger streams (100s)	0.040		0.037		0.260	0.177
	(0.021)		(0.021)		(0.055)	(0.036)
Smaller streams (100s)	0.093		0.069		0.014	0.013
	(0.018)		(0.013)		(0.016)	(0.017)
Total streams (100s)		0.071		0.061		
		(0.013)		(0.010)		
F statistic (instruments)	16.2	30.9	17.5	36.5	25.8	23.9
Panel C: Individual-level first stage estimates (12th grade reading sample)						
Larger streams (100s)	-0.024		-0.030		0.240	0.190
	(0.020)		(0.019)		(0.047)	(0.029)
Smaller streams (100s)	0.133		0.104		0.015	0.001
	(0.017)		(0.013)		(0.013)	(0.013)
Total streams (100s)		0.064		0.058		
		(0.011)		(0.009)		
F statistic (instruments)	31.3	32.0	35.0	37.0	27.5	33.7
Panel D: Individual-level first stage estimates (8th grade reading sample)						
Larger streams (100s)	-0.033		-0.036		0.243	0.177
	(0.018)		(0.017)		(0.046)	(0.029)
Smaller streams (100s)	0.130		0.101		0.011	0.001
	(0.017)		(0.012)		(0.012)	(0.014)
Total streams (100s)		0.059		0.054		
		(0.011)		(0.009)		
F statistic (instruments)	30.7	28.9	34.8	34.7	26.5	30.1

Notes: Base samples are those from Column 3 of Tables 1 (individual level) and 2 (Panel B; MSA level), though some observations that were excluded from those samples for missing data on larger streams are included here in Columns 2, 4, 5, and 6. Alterna

Table 4: IV estimates of choice effect, using alternative instruments and "close replication" sample

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
	OLS	**IV**					
Total stream definition	**n/a**	**Stream mouths**		**All streams**			
Larger stream definition	**n/a**	**Hoxby**	**none**	**Hoxby**	**none**	**Inter-county**	**>3.5 miles**
Panel A: 12th grade reading scores							
Choice index coefficient	-0.25	4.74	0.68	4.38	0.87	2.04	1.35
S.E. (Moulton)	(0.79)	**(1.98)**	(2.79)	**(1.98)**	(2.59)	(2.36)	(2.30)
S.E. (Cluster)	(0.94)	**(2.42)**	(3.12)	**(2.15)**	(2.81)	(2.94)	(2.04)
p-value, exog. test	--	0.02	0.70	0.02	0.66	0.37	0.38
Panel B: 8th grade reading scores							
Choice index coefficient	-0.06	5.93	2.76	5.17	2.78	1.67	0.91
S.E. (Moulton)	(0.70)	**(2.10)**	(2.54)	**(2.01)**	(2.33)	(2.09)	(1.93)
S.E. (Cluster)	(0.82)	**(2.32)**	(3.19)	**(2.02)**	(2.84)	(1.77)	(1.81)
p-value, exog. test	--	0.00	0.30	0.00	0.24	0.21	0.51

Notes: Base samples are those from Column 3 of Table 1, though some observations that were excluded from that sample for missing data on larger streams are included here in Columns 3 and 5-7. Alternative specifications that use the preferred covariates

Table 5. Exploration of potential bias from exclusion of private school students, 12th grade reading scores

Covariate specification	Close replication			Preferred replication		
Streams instruments	OLS	Hoxby	Inter- and intra-cnty	OLS	Hoxby	Inter- and intra-cnty
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Public school students in zip-code matched sample (no district covariates)						
Choice index coefficient	-0.93	1.40	1.10	-0.76	1.97	2.25
S.E. (Cluster)	(1.05)	(2.44)	(2.66)	(0.97)	(2.20)	(2.30)
N	5,631	5,445	5,631	6,976	6,729	6,976
p-value, exog. test		0.35	0.36		0.22	0.12
Panel B: Public and private school students in zip code-matched sample						
Choice index coefficient	-0.71	0.68	0.84	-0.41	1.35	1.81
S.E. (Cluster)	(0.98)	(2.59)	(2.35)	(0.92)	(2.32)	(2.14)
N	6,900	6,670	6,900	8,553	8,259	8,553
p-value, exog. test		0.63	0.43		0.45	0.22

Notes: Clustered standard errors and test statistics are reported.