NBER WORKING PAPER SERIES

COMPETITION AMONG PUBLIC SCHOOLS: A REPLY TO ROTHSTEIN (2004)

Caroline M. Hoxby

Working Paper 11216
http://www.nber.org/papers/w11216

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2005

Competition Among Public Schools: A Reply to Rothstein (2004)

Caroline M. Hoxby
NBER Working Paper No. 11216
March 2005, Revised July 2008
JEL No. H70, I20

ABSTRACT

Rothstein has produced two comments, Rothstein (2003) and Rothstein (2004), on Hoxby "Does Competition Among Public Schools Benefit Students and Taxpayers," American Economic Review, 2000. In this paper, I discuss every claim of any importance in the comments. I show that every claim is wrong. I also discuss a number of Rothstein's innuendos--that is, claims that are made by implication rather than with the support of explicit arguments or evidence. I show that, when held up against the evidence, each innuendo proves to be false. One of the major points of Rothstein (2003) is that lagged school districts are a valid instrumental variable for today's school districts. This is not credible. Another major claim of Rothstein (2003) is that it is better to use highly non-representative achievement data based on students' self-selecting into test-taking than to use nationally representative achievement data. This claim is wrong for multiple reasons. The most important claim of Rothstein (2004) is that the results of Hoxby (2000) are not robust to including private school students in the sample. This is incorrect. While Rothstein appears merely to be adding private school students to the data, he actually substitutes error-prone data for error-free data on all students, generating substantial attenuation bias. He attributes the change in estimates to the addition of the private school students, but I show that the change in estimates is actually due to his using erroneous data for public school students. Another important claim in Rothstein (2004) that the results in Hoxby (2000) are not robust to associating streams with the metropolitan areas through which they flow rather than the metropolitan areas where they have their source. This is false: the results are virtually unchanged when the association is shifted from source to flow. Since 93.5 percent of streams flow only in the metropolitan area where they have their source, it would be surprising if the results did change much. The comments Rothstein (2003) and Rothstein (2004) are without merit. All of the data and code used in Hoxby (2000) are available to other researchers. An easy-to-use CD provides not only extracts and estimation code, but all of the raw data and the code for constructing the dataset.

Caroline M. Hoxby
Department of Economics
Harvard University
Cambridge, MA 02138
and NBER
choxby@harvard.edu

Competition Among Public Schools:

A Reply to Rothstein (2004)

Caroline M. Hoxby*

December 2006

JEL: H70, I20

Hoxby (2000) tests several implications of the Tiebout (1956) model by comparing public schools across metropolitan areas where households can more or less easily choose among school districts ("Tiebout choice"). The article attempts to answer questions such as, when households have more Tiebout choice, do schools have higher productivity, do fewer students attend private school, and do households sort themselves more among school districts? Readers often focus on the achievement results, which suggest that, in a metropolitan area with substantially more Tiebout choice, student achievement is higher all else equal. Some of the most cited results, based on the National Education Longitudinal Study (NELS), suggest that students' reading and mathematics performance would be about 0.3 to 0.5 of a standard deviation higher if they were to attend school in a metropolitan area with the maximum amount of choice observed in the U.S. (more than 100 districts) as opposed to a metropolitan area with the minimum amount of choice observed (no choice; only one district in the metropolitan area).[1] A more realistic change of one standard deviation in Tiebout choice would generate an improvement in reading and mathematics achievement of about 0.1 standard deviations.

A significant share of the paper is devoted to the empirical challenge of identifying a source of exogenous variation in Tiebout choice. Because households are likely to shift from districts with unsuccessful schools to districts with successful schools, the Tiebout choice that we observe in a metropolitan area is endogenous to schools' performance. In particular, in a metropolitan area with a dysfunctional central city district, suburban districts may refuse to consolidate with the central district

and the area may end up–endogenously–with a large number of districts. This challenge is addressed with instrumental variables based on a metropolitan area's number of streams. Essentially, the logic is that many streams generate many natural boundaries, which in history generated many school attendance areas, which left some areas with a lingering, large number of school districts.

All of the raw data, data extracts used for analysis, and the code used to make extracts from the raw data and compute the estimates are available to researchers. Researchers may request the data by contacting their license representative at U.S. Department of Education.[2] The replication dataset was designed and documented for general use by instructors (several of whom had requested it), researchers working on related questions, and replicators. The dataset was created prior to Rothstein's dessemination of his comment, and his claim (p. 4) that the dataset was created in response to his comment is false. Indeed, the internal evidence of his comment shows that he uses the dataset as a resource, even if he sometimes misrepresents its documentation (see below).[3]

Before getting to the substantive elements of Rothstein's comment, it is important to point out that Rothstein's claim (p. 2) that I refused to provide him with the original data is a misrepresentation. Rather, the original data simply do not exist and for very good reason. I was trained to (and still believe in) writing code that takes a researcher all the way from the raw data to estimation. Such code may create intermediate datasets along the way, but they are replaced every time the code is run. Using this procedure has the important advantage that, once a correction or update is made, it feeds through completely. Obsolete datasets are not left sitting around to be used later, accidentally. The procedure prevents the unwitting propagation of erroneous or superseded data and code.

One of the raw datasets used in Hoxby (2000)--the *Common Core of Data*, a sort of directory of public schools--is a dataset that I use routinely.[4] (I have used it in more than a dozen papers subsequent to Hoxby (2000).) After the work in Hoxby (2000) was completed, the National Center for Education Statistics released a corrected version of some of the raw data. I had been aware that there were errors in

the original geographic codes of these raw data, and I had, in fact, fixed the errors that I could find. When the corrected raw data were released, I naturally substituted them for the original (erroneous) raw data. I checked that the corrected geographic codes--which generated small changes in the metropolitan areas, students, and streams included in the sample--yielded approximately the same results. The fact that the correct geographics codes (which are in the replication data set available from NCES) and uncorrected codes yield approximately the same estimates is a good sign. It indicates that the the results are robust to small changes in the sample. The important points to take away are (a) that Rothstein was not refused anything that existed: he asked for an intermediate dataset that no longer existed; and (b) that Rothstein was aware of why the intermediate dataset no longer existed. Thus, if Rothstein's comment makes it appear that he was, without explanation, refused access to data that exists, his comment is misleading.

Rothstein's comment contains a variety of criticisms. In rough order of most to least important, they are as follows. (1) He argues that the results change when students who attend private schools are included in the sample. (2) He argues that the streams variable should be measured in different ways and that the results are sensitive to the changes. (3) He argues that there are errors in the data and that they substantially affect the results. (4) He argues for a different manner of estimating the first-stage equation,

It is easy to summarize my response to the main criticisms. In each case where he argues that a correction or reasonable change would alter the results substantially (in claims 1, 2 and 3), he is incorrect. Such changes alter the results hardly or not at all. Moreover, several of the "corrections" he identifies are wrong or are presented in a very misleading way. When Rothstein *is* able to alter the results, it is only by introducing substantial errors. For example, in the case of the private school students, he obtains different results not because he includes private school students--in fact, they make no difference--but because he uses the private school issue to motivate the introduction of an incorrect

method of assigning students to their home localities. In doing so, he introduces errors and discards a substantial amount of data. It is not surprising that his wrong method generates wrong results. Since Tiebout's model suggests that the jurisdictional makeup of a student's locality affects his school, one cannot test the theory using a method in which students are assigned to the wrong localities and localities are systemically dropped. Moreover, there is no reason to assign students incorrectly: the raw data contain codes that allow researchers to assign students correctly. Rothstein ignored or set aside the documentation that describes the codes. Another example is the pair of stream variables. The reasonable alternative variables he suggests make no difference to the results; he is able to generate different results only when he constructs variables that measure something other than what is intended. With regard to claim 4, Rothstein does not estimate the appropriate first-stage equation and misinterprets key coefficients from the first-stage equation he does estimate.

In addition to his criticisms, Rothstein introduces a "preferred" sample and specification, making numerous choices almost entirely without justification. I show below that these choices do not, in fact, bear scrutiny. Some introduce errors; others make the sample less representative. Nevertheless, even with a free hand to make changes without much or any justification--a clear opportunity for specification searching--Rothstein alters the results only marginally.

II. Private School Students and Assigning Students to Localities

The number and composition of students who attend public schools in a metropolitan area may be endogenous to the absence or presence of Tiebout competition. One section of Hoxby (2000) shows that, where Tiebout choice is greater, a larger share of students attend public, as opposed to private, schools. A reasonable query is whether the positive effect of Tiebout competition on public school students' achievement arises partly because it induces "good" students to stay in the public schools.

Scholars may be interested in a parameter that cannot be identified econometrically: the "pure"

effect of competition on public school students. By this I mean the following. Suppose that competition increases because of an exogenous increase in choice and that public schools improve or deterioratein response. The pure effect of competition is the effect of increased choice on the outcomes of students who would have attended the public schools in the counterfactual where choice had not increased. We cannot identify this parameter because we never simultaneously observe the counterfactual and the increase in choice. Similarly, we cannot identify the pure effect of competition on *private* school students, which is a mirror image of the pure effect of competition on public school students.

Fortunately, we can identify the general equilibrium effect *on public schools*, which is the difference between the outcomes of students who attend the public schools after an exogenous increase in choice and the outcomes of students who attend the public schools before. This effect aggregates the pure effect of competition on public school students, the change in the students who attend the public schools, the changes in peer effects in public schools associated with the change in students, the changes in public school parents, the change in voters' support for public schools, the change in who teaches in public schools, and so on. In other words, the general equilibrium effect provides a reply to the typical policy maker when he asks, "When choice increases, won't able students desert my public schools, depriving the remaining students of good peers? Won't motivated parents desert too, depriving the schools of strong advocates and good governance? Won't voters who previously supported tax levies or took an interest in school board elections fall away? Won't talented people who taught in my public schools follow their children elsewhere?" And so on. Most debates on school choice do not focus exclusively on the pure effect of competition: they include the fully panoply of general equilibrium effects.

We can also identify the general equilibrium effect of choice on *all* students, public and private. This is the difference between the outcomes of all students before and after an exogenous increase in choice. This effect aggregates the pure effect of competition on public school students; the pure effect of

competition on *private* school students; the change in peer effects in public schools; the changes in parents, voters, teachers, and so on in the *public* schools; the change in peer effects in *private* schools; the change in *private* school parents' advocacy, monitoring, and financial resources; the change in donors to the *private* schools; the change in who teaches in the *private* schools, and so on.

Rothstein argues (pp. 12-14) that, in Hoxby (2000), I produce a "biased" estimate of the pure effect of competition on public school students, but this is a misrepresentation. I do not attempt to estimate the pure effect of competition (which is not identifiable), I estimate the general equilibrium effect on public school students. That the estimation choice is knowing, and not an unwitting error, is obvious: the paper contains an entire section that describes why more Tiebout choice might cause students to switch from private to public school. Moreover, the section in question actually estimates the share of students who shift from private to public schools as Tiebout choice rises.

Because, in Hoxby (2000), I actually estimate the shift, it is easy to see that the general equilibrium effect on public schools will be very similar in practice to the pure effect of competition on public schools. A two standard deviation increase in choice among public schools induces two percent of students to shift from private to public schools. In a typical metropolitan area, this means that 91 percent, rather than 89 percent, of students will attend public schools. The two percent of students who shift would have to have extraordinary scores, be extraordinary peers, have extraordinary parents, and attract extraordinary voters and teachers to change achievement significantly among public school students. Thus, even though one cannot estimate the pure effect of competition of public schools, it constitutes most of the general equilibrium effect on public schools.

Rothstein, instead of realizing that the pure competition and general equilibrium effects on public schools are different parameters, argues that general equilibrium effect on public schools is a "biased" estimate of the pure competition effect. In a bizarre twist, he then asserts that"bias can be easily avoided," implying that he has a method of estimating the pure effect of competition on public schools.

In fact, he does not do this at all. Instead, he estimates the general equilibrium effect on *all* students, a parameter that contains many more elements than the pure effect of competition on public schools. In particular, it contains many changes in private schools that could be important. This is because an increase in choice induces a movement of students from private to public schools that is tiny by public school standards but large by private school standards, where a fifth students might depart. Thus, private schools' peer effects would change, their parents would change, their finances would change, and so on. The general equilibrium effect on *all* students is interesting, but it is certainly not an unbiased estimator of the pure effect of competition on public schools. Rothstein's statements about bias are confused and may lead future researchers, who need to keep track of the parameter they are estimating, astray

In any case, Table 1 shows that, in the NELS sample, estimates of the general equilibrium effect on public schools (top row) are very similar to estimates of the general equilibrium effect on all students (bottom row). The estimated coefficient for the eighth grade reading score, for instance, changes from 6.64 to 5.36; and the estimated coefficient for the tenth grade reading score, for instance, changes from 8.50 to 8.16. In no case do the estimates suggest that the general equilibrium effect on public schools students differs from the effect on all students. (That is, all of the changes are far from being statistically significant. The changes mentioned are about 0.30, but the standard errors indicate that a coefficient would have to change by about 5.50 for the change to be statistically significant.)

Given these results, why does Rothstein apparently generate such different results in his Table 5 when he also estimates the general equilibrium effect for all students? He generates different results because he reassigns the locality of every student in the NELS using a "zipcode-backing-out method" that is error-prone and that does not properly assign students to many districts. This zipcode-backing-out method is unnecessary because, as described in its documentation, the NELS contains codes that identify all schools in the study, public and private.[5] Rothstein ignored or set aside the documentation and thereby--ostensibly--justified his use of a method that makes error-prone assignments of students to

locations, drops covariates, and uses only sub-samples of the data. It is important to note from his Table 5 that it is these actions and *not* the inclusion of private school students that cause him to obtain results substantially different from Hoxby (2000). The top row of his Table 5 excludes private school students and the bottom row includes them, but the difference between the two rows is inconsequential. It is his introduction of error, discarding of data, and dropping of covariates that cause him to report that the "[e]stimates are substantially smaller than those presented earlier." Misleadingly, he presents this conclusion as though it were related to the inclusion of private school students when, in fact, it has nothing to do with them.

The zipcode-backing-out method works as follows. There are several Census variables in the restricted-access NELS that are associated with each school's zipcode, but students' zipcodes are *not* provided in the NELS. Rather, NELS provides a few variables derived from the Census that describe the zipcode of a student's school, not his residence.[6] By cross-referencing these variables, one can back out a unique school district location for some (only some) students. Unfortunately, this method associates other students with the incorrect district or multiple districts (as many as nine). Two errors are generated by this procedure. First, the method does not always generate a unique zipcode: multiple zipcodes may have the same values for the descriptive variables provided by NELS. Second, zipcodes are not aligned with school districts in the U.S., and many zipcodes cross school district boundaries.

How valid are the localities assigned by the zipcode-backing-out method? Because the NELS contains an identifying code for all of its schools, they can be associated with the school district in which they are actually located.[7] Thus, by following NELS documentation, students can be assigned to the actual school district in which they live or, in the case of private schools, actually attend school.[8] Table 2 compares the actual school districts in which NELS base year schools are located to the districts that the zipcode-backing-out method associates with them. Table 2 shows that 39 percent of the public schools have problematic (missing, non-unique, or apparently unique but actually incorrect) "backed-out" codes.

The table also shows that 41 percent of the private schools have problematic "backed-out" codes. Detailed break-downs are in the table.

In short, Rothstein obtains results that differ from those in Hoxby (2000) not because he adds private school students to the sample but because his zipcode-backing-out method mis-assigns about 40 percent of all students.

III. Measuring Streams by Flow Rather than by Primary Location and Related Issues

In Hoxby (2000), two streams variables, smaller streams and larger streams are used as instruments. For reasons discussed in my original paper (p. 1222) and repeated below, it is important to divide streams into those that are more and less suitable for commercial navigation. However, the essential logic behind using streams as instruments is simple: streams cause variation in the number of school districts because, during the settlement of America, district boundaries often *were* streams. In fact, early American laws often stated that students should not have to cross streams to get to school. In other words, real walking distance, not distance as the crow flies, was what mattered for students' travel. The streams variables work as instruments because they affected initial district boundaries and there is substantial inertia in boundaries. (The idea evidently has very old origins: Maimonides' Rule also stated that students should not have to cross streams to attend school.)

The United States Geological Survey (USGS) gathers accurate information on streams and conveys it in two forms that are relevant: topographic quadrangle maps and the Geographic Names Information System (GNIS).[9] The maps show every feature known to the USGS that can be displayed at 1:24000 resolution. The GNIS is a list of certain features shown on the maps. Every USGS feature is classified (as a "stream", "reservoir," "lake," "summit," *et cetera*) and its location is described by its latitude and longitude. A stream's location is described by the latitude and longitude of both its "primary location" and its source. (Rothstein uses the word "mouth" for primary locaton, but I use the USGS

terminology.) The vast majority, 93.5 percent, of streams have their source in the same metropolitan area as their primary location.

A. Streams by Primary Location and Streams by Flow

Hoxby (2000) associates streams with the metropolitan area in which they have their primary location, as defined by the United States Geological Service. At the time the paper was written, this was the form in which the GNIS data were available. Rothstein argues that streams ought to be associated with all the metropolitan areas in which they flow, and he proposes use of a GNIS dataset made available since Hoxby (2000) was written.[10] The shift in classification (from streams-by-primary-location to streams-by-flow) is reasonable, but is unlikely to affect the results much because 93.5 of streams are classified identically under the two methods. The remaining 6.5 percent of streams are not removed from the count for the metropolitan areas that are their primary locations; they are simply added to the counts of other metropolitan areas through which they flow. Thus, the change in classification scheme produces only a small change in the instrumental variables.

The top two rows of Table 3 show that the estimates based on streams-by-primary-location and streams-by-flow are extremely similar. (As described below, these estimates use as instruments smaller and larger streams, where the latter variable is measured as described below. The instruments can be computed either by primary location or flow.) The estimated coefficient for the eighth grade reading score, for instance, changes from 6.64 to 6.61; and the estimated coefficient for the tenth grade math score, for instance, changes from 7.98 to 7.44. (The changes mentioned average 0.29, but the standard errors indicate that a coefficient would have to change by about 5.50 for the change to be statistically significant.) Indeed, Rothstein's own estimates (see his Table 4) show that using streams-by-flow does not affect the results. It is difficult to see, therefore, why he raises the matter.

B. Measuring Navigable Larger Streams

As described in Hoxby (2000) and at some length below, it is important to create separate counts

of smaller streams and larger streams, where larger streams are defined as those that are potentially navigable for the purposes of commerce. Because width as well as length is important for assessing commercial navigability, I used the USGS topographic quadrangle maps to measure larger streams. There is no single rule for the width that makes a stream navigable for the purposes of commerce, but states where width is explicitly considered in the determination of navigability include Texas (30 feet wide), Washington (40 feet wide), Georgia (40 to 45 feet wide), and Arizona (similar to Georgia).[11] Thus, I chose 40 feet wide as a typical standard and looked for streams that met this criterion and that were at least 3.5 miles in length (since shorter streams could not plausibly connect two trading centers even if they were wide). Curvilinear bodies of water that met these criteria were noted, checked to ensure that they were USGS designated streams (to exclude non-streams such as manmade bodies of water), and counted. Such measurement is painstaking, essentially because one has to carry information correctly over the edges of maps, but it is not fundamentally difficult. The count of larger streams is subtracted from the total number of streams to obtain a measure of smaller streams.

Rothstein argues that my method of processing USGS maps is subjective, and much of his discussion of subjectivity is devoted to a peculiar example meant to illustrate it. Focusing on the Fort Lauderdale metropolitan area, he counts manmade canals (which, straight or not, were created by dredging) and the Atlantic Intercoastal Waterway, a engineered channel. Considering that Fort Lauderdale's municipal and tourist offices heavily advertise the city's being known as the "Venice of America" and emphasize how its waterways were artificially created from the marsh, it is obviously a location where manmade water features--which are not streams and which blatantly violate the spirit of the instrumental variable--abound.[12] It can be no surprise that, in a location like this, he comes up with a count that differs from mine: he is knowingly counting bodies of water that are not defined as streams by the USGS.[13] He need not guess: he need only acknowledge the USGS designations. Rothstein is not identifying a problem of subjectivity. He is merely demonstrating that, by ignoring information that is

pertinent and then focusing on the very location where the pertinent information is most useful, he can generate mismeasurement.

Moreover, Rothstein's argument that only GNIS data, and not map-based data, should be used is based on a fundamental misunderstanding. The GNIS data are *derived* from USGS maps. The difference between the maps and the GNIS is that the GNIS contains only a tiny fraction of the information on the maps. Thus, if a researcher devotes time to the maps, it is repaid with accurately measured variables. In contrast, there is only a certain amount a researcher can derive from the sparse set of variables available in the GNIS. To put it another way, suppose that the GNIS contained not only the latitude and longitude of a stream's origin and destination, but also the latitude, longitude, and shore-to-shore width at every turning point in a stream's course. A researcher could then measure a stream's size looking for a combination of length and width that is continuous (not interrupted by narrow stretches). This is what the maps allow one to do except that the maps are superior because they show width at all points, not merely occasional points. Another advantage of using maps is that it is so time-consuming to measure streams using them that the equivalent of "specification searching" is impracticable. One must adopt a metric for counting streams and stick with it.

C. Larger (Potentially Navigable) Streams

As argued in Hoxby (2000, p. 1222), the one shortcoming of streams as instruments is that large streams that are navigable for shipping purposes may be (or have been) important channels for trade. If large rivers attract commerce and cities are built where commerce thrives, then we might expect to find big city districts around important rivers. To take a particularly obvious example, consider Pittsburgh, a city that would not exist if it were not for the confluence of the Allegheny, Monongahela, and Ohio rivers. In other words, a large navigable river might attract commerce to a particular location. The dense population that gathers in the location may cause a central city jurisdiction to arise, and it may have the political power to swallow up others. In short, large streams may, like small streams, generate more

initial jurisdictions, but large streams may also create the conditions in which jurisdictions are more likely to consolidate into "central city" ones. Therefore, we would not expect that large and small streams should necessarily have the same effect on the choice index.

A separate concern about large streams is that they may indirectly affect achievement by affecting commerce which, in turn, may affect the type of person who decides to live in an area. The direction of such effects is unclear and need not be uniform across metropolitan areas. For instance, large streams may make one metropolitan area a center for agricultural goods, another a center for industrial goods, and a third a center for finance. The types of people who work in agriculture, industry, and finance are likely not the same.

In short, although all streams are created by nature, there is reason to expect that large and small streams will have different effects on the number of jurisdictions in an area. Moreover, large streams may have non-monotonic effects: positive in some areas, negative in others. Thus, in Hoxby (2000), I argue that small streams (streams too small to be commercially navigable) are more credible instruments than are commercially navigable rivers. It is very credible that the number of smaller streams fulfils the requirement that an instrumental variable be uncorrelated with the unobserved determinants of achievement.

Fortunately, the numbers of smaller and larger streams are not highly collinear: the correlation is only 0.41. Thus, in Hoxby (2000), I let the two streams variables enter the first-stage equation separately to see whether their coefficients are the same (they are not). Moreover, in that paper, I compute estimates that rely solely on smaller streams in order to determine whether they differ statistically significantly from the estimates that rely on both streams variables. They do not. (See the bottom rows of Tables 4 and 6 of Hoxby (2000).) Similarly, the bottom panel of Table 3 below shows that the estimates that rely on only the smaller streams variable are very similar to the estimates that rely on both stream variables. For instance, the estimated coefficient for the eighth grade reading score, for instance,

changes from 6.64 to 5.86; and the estimated coefficient for the tenth grade math score, for instance, changes from 8.50 to 7.82. (The changes mentioned average 0.7, but the standard errors indicate that a coefficient would have to change by about 6 for the change to be statistically significant.)

As noted in Hoxby (2000), the two streams instruments have different coefficients in the first stage regression. This may be an indication that larger streams have the offsetting effects mentioned earlier: more large streams mean more initial jurisdictions, but more large streams also create conditions favorable to consolidation. Also, while the first stage regression that includes all metropolitan areas produces positive coefficients on both streams variables, a regression run on certain subsets of the metropolitan areas produces a coefficient on the larger streams variable that is not statistically significantly different from zero and occasionally has a point estimate with a negative sign. This is a hint–though not firm statistical evidence–that the larger streams variable is associated with commerce and central city consolidation in some metropolitan areas. Put another way, this is a hint that the larger streams variable may not fulfil the monotonicity condition for an instrumental variable. This would be another reason to think that it is the smaller streams variable that should be regarded as the more reliable instrument, as suggested in Hoxby (2000).

Rothstein argues that one should be able to add the large and small streams variables and use total streams as an instrument. But, as noted above, the two streams variables do not have similar coefficients. If there are potential monotonicity issues with the larger streams variable, adding it to the smaller streams variable produces a single contaminated variable–aggravating rather than remedying the problem. Rothstein also argues that one should be able to measure larger streams based solely on criteria such as length or whether a stream crosses a county boundary. But, as noted above, the goal is to separate streams that are useful for commercial purposes from those that are not. Merely being 3.5 miles long or a stream's happening to cross a boundary does not make it useful for commerce. It is *width* that is usually the binding criterion, not the 3.5 mile length, which is a simple minimum. There are literally

thousands of creeks that run for 10 miles but that are only a few feet wide, and it is obvious that they can serve no real commercial purpose (regardless of whether they cross a boundary). It is no wonder that, by introducing variables (streams that are merely longer than 3.5 miles, streams that merely cross boundaries) that do not even attempt to measure what one needs to measure, Rothstein is able to obtain results in his Table 4 that differ from mine.

IV. A Variety of Claims about Errors or Corrections

Rothstein makes several claims that errors and/or changes in the program and variables contained in the replication data set have a substantial effect on the results. These claims are incorrect. None of the issues he describes has much effect on the results. Moreover, his discussion of the errors and/or changes is consistently misleading, incorrect, or both.

A. Raw Data Corrected by the Census

Rothstein states that I "refused" to provide him with the data from Hoxby (2000). As mentioned in the introduction, I refused him nothing that exists but provided him with all of the code for estimating results from the raw data. One particular source of raw data has been corrected since my work on the original paper: the coding of the metropolitan areas associated with school districts. This complicated geographic coding is performed by the Bureau of the Census in coordination with NCES. The codes become part of the *Common Core of Data*, which is a directory-like source routinely used by education researchers. When corrected *Common Core* data becomes available, it is released by NCES. Because I use the *Common Core* many times each year, I download the corrected data as a matter of routine. After writing Hoxby (2000), I was aware of the need to do this because, in the course of working on that paper and others, I had discovered some errors in the geographic codes of then-available *Common Core* data. One of my research assistants had attempted to identify errors and correct them, but we can not claim to have had the resources or expertise that the Census regularly devotes to the process of correction.[14] In

short, when the code in the replication dataset calls the raw data from the *Common Core*, it calls the more correct data available now.

On this point, Rothstein's comment may mislead readers. First, he implies that he was refused access to some intermediate data set that exists. In fact, such an intermediate dataset never existed independently of the raw data: it was replaced every time the code was re-run with the raw data. Second, he claims that I changed an "assignment algorithm" for matching school districts to geographic codes for metropolitan areas (p. 4). In fact, the only thing that changed was the raw data. The word "algorithm" is itself misleading: the program just assigns each district to the metropolitan area in which it is located. Third, his language (p. 4) suggests that I generated the corrections to the data, and that I did it in response to his comment. In fact, the raw data was corrected by the Census, and my downloading it predated his comment by years.[15]

In any case, despite Rothstein's repeated insistence on the issue, there is not much difference between using the codes available for the original paper and using the codes corrected by the Census. A comparison of the top row of Table 1 (above) and the top row of Table 4 in Hoxby (2000) reveals very similar results. The similarity is also shown in Rothstein's own Table 1 (compare columns 1 and 2), which makes it hard to see why he raises the issue.

B. Four Ohio School Districts

Rothstein argues that the program incorrectly assigns four school districts in Ohio to a North Carolina metropolitan area. In fact, the code is correct. It is the raw data–the *Common Core of Data*–that contain an error and associates the Ohio districts with North Carolina. More importantly, the four districts' being misassigned has no effect on the results because there are no NELS students in them.

Just for completeness, Table 4 shows the results with and without the correction for the four Ohio districts. Rhe results are identical; the error in the raw data is harmless. Rothstein's discussion of the Ohio districts (pp. 4-5) is misleading: he implies that they are included in the NELS estimation

sample and affect the results.

C. Use of Contemporaneous Metropolitan Area Codes

Rothstein argues that several school districts have "incorrect, invalid, or obsolete" metropolitan area codes (p. 5). The claim about "obsolete" codes is groundless. It is apparently based on his mistaken belief that districts should match current Census codes rather the codes that existed at the time of the NELS survey. This is wrong: the Bureau of the Census routinely updates the definitions of metropolitan areas.[16] Thus, school districts in the NELS *rightly* have metropolitan area codes that were current at the time the NELS was conducted.

The base year of the NELS data is 1987-88, so I first matched NELS districts to the 1987-88 *Common Core of Data*. The first follow-up year of the NELS is 1989-90, and the Census updated a few metropolitan areas to include new districts between the base and first follow-up years. Thus, NELS districts that remain unmatched after the 1987-88 sweep are matched to the 1989-90 *Common Core*. The second follow-up year of the NELS is 1991-92, and the Census updated a few metropolitan areas between 1989-90 and 1991-92. Thus, a NELS district that remains unmatched after the first two sweeps is matched to the 1991-92 *Common Core.*

In fact, the number of districts that change codes between 1987-88 and 1991-92 is so small that it does not matter which of the three survey years' codes are used. This is shown in Table 5, which reveals that the coefficients hardly vary with the year of the codes.

In short, I match NELS survey data to contemporaneous administrative data. In this way, I accept the Census's determination of which districts were in metropolitan areas in the survey years. This is the accurate way to assign districts to metropolitan areas. 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.

D. A Typographical Error in the Program

Rothstein makes much of a typographical error in a program in the replication dataset--his discussion occupies all of his page 5, half of page 6, and footnotes 6, 7, 8. The length and intensity of this discussion suggest that Rothstein has found a serious error related to missing school district codes, but actually all that he has found is that the word "update" is missing from two lines. Moreover, this typographical error has nothing to do with Hoxby (2000) because it was introduced when I was finalizing my creation of the replication dataset. (I made superficial changes in an effort to make the code transparent to users while complying with NCES's wishes regarding methods of indicating missing observations. The typographical error was introduced at that point).[17] In his lengthy discussion, Rothstein suggests that the typographical error has a substantial effect on the results. This is incorrect: the typographical error has no statistically significant effect on the results and, if anything, weakens them. This is shown by comparing the top and bottom rows of Table 6. For instance, the estimated coefficient for the eighth grade reading score is 6.64 without the typographical error and 4.41 with it. The estimated coefficient for the tenth grade reading score is 8.50 without the error and 6.57 with it. (The changes in the table average about 1.1, but the standard errors indicate that a coefficient would have to change by about 5 for the change to be statistically significant.) Rothstein himself says that, with the typographical error, "The mean choice effect for 12th grade scores is 5.39, quite close to the 5.30 computed from the Hoxby/NCES data." His related Appendix Figure also shows that vast majority of estimates generated with the typographical error in place are so close to 5.30 that the difference is unimportant.

E. Summing Up Issues Related to Errors and Corrections

Tables 4 through 6 demonstrate that each of Rothstein's issues regarding an error or correction has no effect or little effect on the results. Indeed, although one would not know it from his discussion, Rothstein's own results reveal the same thing (compare columns (1) to (3) of his Table 1). Rothstein's long discussion of data issues that make no difference to the results creates an impression of numerous

errors. The barrage of assertions with regard to errors simply sows confusion in readers' minds and thereby provides a justification for Rothstein's presenting a "preferred" specification (discussed below).

V. The First Stage Regression

Rothstein argues that Hoxby (2000) was wrong to show a first-stage regression that is run at the metropolitan area level and that includes all metropolitan areas. He argues that the article should have shown a first-stage regression that uses student level observations and that includes only metropolitan areas in which NELS students live. However, the first-stage regression shown was a deliberate decision made as part of the editorial and refereeing process at this journal. It was intended to make it clear to readers that the first-stage regression only made use of metropolitan area variation because the dependent variable (the index of choice) varies only at the metropolitan area level. Readers could have been confused if they saw thousands of observations in a regression that really had only as much variation as there are metropolitan areas.[18]

Moreover, the same first-stage specification is used for several second-stage regressions in the paper. In the second-stage regressions where NELS achievement variables are the dependent variables, only some metropolitan areas (about 60 percent) are included: the number depends on the locations of the students whose achievement is being considered. But, in numerous second-stage regressions where variables like district-level variables were the dependent variables (for instance, school spending and the private school share), all metropolitan areas are included. Given space constraints in the journal, it was clear that only one first-stage regression could be shown. It was logical to show the one used for district-level dependent variables because it was most general since it included all metropolitan areas. The coefficients in this regression are as expected: there are more jurisdictions in areas with more streams, and smaller streams exert a more powerful influence on school district boundaries than larger streams–probably for the reasons described above. For those who are interested, the program in the

replication dataset runs all the variants of the first-stage regression: the specification stays the same but the sample of the metropolitan areas that are included varies with the students in the regression.

Rothstein (Tables 2 and 3) shows first-stage results that drop many metropolitan areas and it is this drop that produces results that he can claim are "dramatically different." For instance, in panel C of his Table 2, he drops 37 percent of metropolitan areas and, in panel D, he drops 39 percent. His discussion of the first-stage results is therefore misleading: he does not mention how much of the sample is dropped and implies that the differences are due to changes in the data, not the sample drops. Also misleading is his commentary on how the coefficient on the larger streams variable changes (pages 6-7 of his manuscript). He suggests that the changes in the coefficient are very worrisome, but--in fact--they were only to be expected since he is dropping large parts of the sample. As described above and in Hoxby (2000), the larger streams variable may exhibit a non-monotonic relationship with the choice variable. Therefore, as one changes the sample, the coefficient may change. This is the reason for constructing measures of both larger and smaller streams.

VI. Rothstein's "Preferred" Specification

Rothstein introduces a "preferred" specification, making numerous changes to the variables and the sample. Four of these changes are mentioned in footnote 10 but the vast majority are mentioned only in his appendices. A count of the changes cannot be constructed from his cursory descriptions, but it would be scores--hundreds if multiple changes to the same variable are counted separately. Moreover, the descriptions of the changes--if the reader gets to them--provide little justification or discussion of the implications. This makes it hard for a reader to assess the validity of his "preferred" estimates. In this section, I consider just a few of the many changes to illustrate the problem created.

A. The Construction of the Metropolitan Income Variable

Rothstein (footnote 10) substitutes, without explanation, an incorrect measure of metropolitan

income: the natural log of the mean income of districts. The correct measure is the mean of the districts' log mean incomes. Essentially, one must include a metropolitan-level mean of any district-level covariate that is included in the regression. If this is not done, the district level variables can be correlated with metropolitan area variables, and such correlation produces a phenomena sometimes described as "Tiebout bias." The bias is eliminated by including proper metropolitan level aggregates. Substituting the incorrect variable for the correct one generates a misspecification error. It is hard to understand why Rothstein introduces the error because he does not offer an argument for the substitution. This is despite the fact that there is a description of the correct measure in Hoxby (2000, pp. 1217-1218) and an even more detailed description in the code contained in the replication dataset.

The substitution of the incorrect metropolitan income variable does not, in itself, make a difference to the results. This is shown in Table 7, the top row of which uses the proper variable and second row of which uses the misspecified variable. The differences in the coefficients on the choice index are extremely small: 6.64 becomes 6.74, 5.36 becomes 5.39, and so on. (These changes average 0.7, but the changes would need to be about 5 to be statistically significant.)

B. Family Background Variables

In constructing his "preferred" estimates, Rothstein changes numerous family background variables (described only in his appendices).

Before considering these changes, let us remember why the estimation included family background variables at all. As emphasized in Hoxby (2000, pp. 1226-7), they were *not* included because their coefficients bore causal interpretations (they do not). Rather, such variables were included because the NELS sample was only designed to be nationally representative. It is not representative of individual metropolitan areas. Thus, family background variables are used to adjust the outcomes for sampling differences among metropolitan areas. For example, suppose that poor and middle-income districts are sampled in metropolitan area A and that middle-income and affluent districts are sampled in

metropolitan area B. Metropolitan area B should not get "credit" for the better outcomes associated with affluent, as opposed to poor, students. Adjusting outcomes to account for sampling differences is important if we are to compare metropolitan areas A and B fairly.

Because my primary concern, in including family background variables, was making the sample maximally representative, it was important *not* to use measures of background variables that had been constructed in such a way that they would--by their inclusion--make the sample less representative.

Rothstein substitutes parental education information from the parent survey for parallel information from the student survey. This choice makes the sample less representative, largely because only a subset of parents responded to the survey. The NELS documentation says, "...only the student and school data sets constitute fully representative national samples. While in various respects the parent data set resembles a representative or probability sample...several features of the NELS:88 parent component depart from the strict requirements for a probability sample."[19] The documentation also points out that answers in the parent survey could be biased depending on the identity of the parent who responded to the survey (mother, father, step-parent, and so on) and on family composition (two parents at home, divorced parents, and so on). Given that a parental education variable was available in the student survey, that this variable generated no sampling problems or bias, and that parental education did not have a coefficient of interest but was merely added to account for sampling differences, it makes sense to use parental education based on the student survey.

Rothstein changes the construction of several family background variables--for instance, using family income from the second follow-up--in order to include students who were not part of the NELS base year sample. However, as explained by Grogger and Neal (2000), only the base year sampling used by NELS is representative. Its follow-up sampling is very problematic for applications related to school choice because the follow-ups do *not* sample randomly but systemically drop students who attend a tenth grade school that relatively few of their former eighth grade (base year) classmates attend.[20] For

instance, consider a public middle school that performs poorly and thereby alienates the parents of its eighth graders. In response, some parents might send their children to private schools for grades nine through twelve, other parents might enroll their children in magnet public schools for grades nine through twelve, and the remaining parents might send their children to the public high school that is the default. The students who attend the private and magnets schools have a high probability of being dropped from the sample, and the default public high school has a fair probability of being dropped too. Thus, the follow-up sampling is not only non-representative, but is non-representative in a way that is endogenous to the performance of schools in the base year.[21] The "freshening" of the sample in the follow-ups does not relieve, but instead aggravates this situation by adding new students from schools that are already "over-sampled" relative to, say, the schools in the example just provided.[22]

Changing a single family background variables affects the results only trivially. For instance, Table 7 shows (third row) that using parental education data from the parent survey produces only slight differences relative to the top row.[23] The estimated coefficient for the eighth grade reading score is 5.61 instead of 6.64, for instance. Similarly, using family income from the second follow-up (twelfth grade) survey, as Rothstein does, produces only slight differences.[24] See the fourth row of Table 7 which shows, for example, that the estimated coefficient for the eighth grade reading score is 6.52 instead of 6.64.

C. Summing Up the "Preferred" Specification

To get his "preferred" specification, Rothstein makes numerous changes to covariates that are hard to justify if scrutinized. If we take the changes one at a time, each has little or no effect on the results. However, Rothstein makes *many* changes to the sample and to covariates, and it is unknown how many other changes were tested and rejected. Testing a large number of changes without much justification runs the danger of either intentional or unintentional pretest bias by leading an investigator to find the specification that best fits his priors. In any case, even with the large number of changes he made, Rothstein merely reduces the coefficient to being marginally statistically significant (column 4 of

Table 1).

VII. A Note on the Standard Errors

Rothstein gives prominence to the fact that, in the replication data set, I provide code for Stata's robust cluster standards but not for Moulton (1986) standard errors as in Hoxby (2000). He then immediately goes on to say that the data from the replication data set produces larger standard errors. The implication is that, somehow, the replication dataset has problematic standard errors. Yet, Rothstein is perfectly well aware of the fact that it is *because* the replication dataset uses Stata's robust clustered standard errors that it produces larger standard errors. He knows that the Moulton standard errors are consistently *smaller* than Stata's robust clustered standard errors--the fact is demonstrated throughout his tables. Rothstein is also perfectly well aware of why I used Stata's robust clustered errors in the replication dataset, despite the apparent loss of precision: I was concerned that the typical user would be unable to compute the Moulton standard errors with comfort since the computations involve some complicated decisions. My reasoning is clearly documented in the replication dataset. Rothstein even *agrees* that Moulton standard errors are beyond many users--this agreement is buried in his appendix. Rothstein lays out none of this relevant information for readers. Instead, they are likely to be left with the impression that there is something wrong with the replication dataset.

VII. Conclusions

In each case where Rothstein argues that a reasonable change to the specifications estimated in Hoxby (2000) would substantially affect the results, I have shown that the argument is either incorrect or has negligible effects on the results (the latter is usually shown in his Comment as well). He is able to obtain substantially different results only by introducing important errors such as the error-prone zipcode-backing-out method. Similarly, reasonable alterations to the instrumental variables for streams

do not change the results. The results change only when Rothstein introduces variables that do not measure what the instruments are intended to measure. Rothstein discusses various "errors" and "corrections" in the data, but all of these have been shown here to have no effect or only a trivial effect on the results. Rothstein's own tables show the same lack of effect. So, why raise these points at all? Rothstein's style of discussion, which raises a long series of issue which are claimed to be important but turn out not to be, is consistently misleading. Again and again, his comment gives the *impression* that an important error has occurred or an important change needs to be made when, in fact, his own estimates show nothing of the kind. There is not a single case, however, in which he alerts the reader to the fact that an issue he has raised--sometimes at length--has negligible consequences or no consequences at all. Instead, his discussion suggests that he has found a slew of errors and misjudgements in Hoxby (2000)--all of which affect its results to some degree. What his discussion disguises is the sharp discontinuity that really characterizes his results. The reasonable changes that he suggests produce results that differ trivially or not at all from Hoxby (2000). Only when he introduces major errors does he generate results that differ meaningfully. As readers may confirm for themselves (since the replication dataset is available), the results in Hoxby (2000) can be replicated well and are robust to a wide array of reasonable changes in the variables and sample.

References

Dataware. *GNIS Digital Gazetteer*. Software and data. Reston, VA: Issued by the United States Geological Survey for Dataware, 1999.

Gillis, Susan. *Fort Lauderdale: The Venice of America*. Mount Pleasant, SC: Arcadia Publishing, 2004.

Grogger, Jeffrey, and Derek Neal, "Further Evidence on the Effects of Catholic Secondary Schooling," in William G. Gale and Janet Rothenberg Pack, eds. *Brookings-Wharton: Papers on Urban Affairs 2000*. Washington, DC: The Brookings Institution Press, 2000, pp. 151-202.

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

Hoxby, Caroline M. "Competition Among Public Schools: A Reply to Rothstein (2004)," National Bureau of Economic Research Working Paper No. 11216 (2004), reissued 2007.

Moulton, Brent R., "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, August 1986, *32*(3), pp. 385-397.

Parkman, Aubrey. *History of the Waterways of the Atlantic Coast of the United States*. A National Waterways Study commissioned by the United States Army Engineer. Washington, DC: United States

Army Engineer Support Center, Institute for Water Resources, 1983.

Quality Education Data. *National Education Database™*. Electronic data. Denver: Quality Education Data, 1988.

Rothstein, Jesse. 2003. "Does Competition among Public Schools Benefit Students and Taxpayers? Comment." Princeton University unpublished manuscript.

Rothstein, Jesse. 200?. "Does Competition among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)." American Economic Review [editor: please fill in information about volume, date, and pages appropriately].

United States Department of Commerce, Bureau of the Census, Population Division. *Historical Metropolitan Area Definitions*. Text file released on the internet. Revised March 2005.

United States Department of Commerce, Bureau of the Census. *Metropolitan Areas and Components, 1990 with FIPS Codes*. Text file released on the internet. Revised April 1999.

United States Department of Commerce, Bureau of the Census. *Race and Hispanic Origin of Householder--Families by Median and Mean Income: 1947 to 2004*, Detailed Table F-5. Text file released on the internet, 2006 version.

United States Department of Education, National Center for Education Statistics. *National Education Longitudinal Study of 1988, Base Year: Parent Component Data File User's Manual*. NCES publication

90-466. Washington, DC: 1990.

United States Department of Education, National Center for Education Statistics. *National Education Longitudinal Study of 1988, Base Year Sample Design Report*. NCES publication 90-463. Washington, DC: 1990.

United States Department of Education, National Center for Education Statistics. *National Education Longitudinal Study of 1988, First Follow-Up: Student Component Data File User's Manual, Volume I*. NCES publication 92-030. Washington, DC: 1992.

United States Department of Education, National Center for Education Statistics. *National Education Longitudinal Study of 1988, NELS 88/94 Restricted-Access cd-rom*. Washington, DC: National Center for Education Statistics, 1996.

United States Department of Education, National Center for Education Statistics. *Private School Universe Survey*. 1989-90, 1991-92, 1993-94, 1995-96 editions. Electronic data. Washington DC: United States Department of Education, 2003.

United States Department of Education, National Center for Education Statistics. *School Locale Codes 1987 – 2000*. NCES publication 2002-02, by Nancy Speicher. Arnold A. Goldstein, project officer. Washington, DC: 2002.

United States Department of Education, National Center for Education Statistics. *The Common Core of Data*. 1987-88, 1989-90, and 1991-92 editions. Electronic data. Washington DC: United States

Department of Education, 2004.

United States Geological Survey. 1:24000 Series. ...Quadrangle, [State Name]. Maps (serial). As available in 1993.

United States Geological Survey. Frequently Asked Questions about GNIS [Geographic Names Information System]. Electronic file posted on the internet at www.usgs.gov, 2006.

United States Geological Survey. Geographic Names Information System, State Files. Electronic files. 2004.

*. Department of Economics, Harvard University.

1. All references to the National Education Longitudinal Study are to United States Department of Education (1996).

2. Hoxby (2000) uses data from NELS matched to data from the United States Census, school districts' administrative files, and the United States Geological Survey. Researchers require a license to obtain NELS data that includes the geographic codes on which matching is based. Many researchers who study education already hold such licenses but others may apply to the National Center for Education Statistics (NCES).

I am greatly indebted to Bruce Daniel at Pinkerton Computer Consultants Incorporated (a contractor for NCES) for examining the code and data in the replication dataset. I am also greatly indebted to Jeffrey A. Owings and Cynthia Barton at NCES for working out a procedure for distributing the replication dataset.

3. Although the original replication dataset contained only data and code associated with the original Hoxby (2000) paper, I have subsequently added to it all of the raw data and code used to compute the specification tests described in this reply.

4. For the *Common Core of Data*, see United States Department of Education (2004).

5. See p. 12 of United States Department of Education, *National Education Longitudinal Study: Base Year Sample Design Report* (1990).

6. The Census variables are drawn from 1990 Summary Tape File 3B.

7. Quality Education Data codes are used for both public and private schools in the NELS data. Public schools are also associated with their National Center for Education Statistics code. The codes can be used to match the NELS to relevant directory-type data--either the Quality Education Data *National Database* (1988) or the combination of the public school directory (the *Common Core of Data*) and private school directory (*Private School Universe Survey*). Both of the latter directories have been

published regularly by the United States Department of Education since the late 1980s. See United States Department of Education (2004 and 2003, respectively). Private schools' addresses are used to determine the school district in which they are located. This is standard geocoding.

For the purposes of this paper, a tiny number of students (18) who may be attending private school have problematic Quality Education Data codes and an even small number (9) have a missing Quality Education Data code. To put these numbers in perspective, consider that, together, these students represent 0.1 percent of those who completed the base year NELS survey.

8. It would be preferable to associate private school students with the district where they live, as opposed to the district where they attend school. Although the two districts are the same for many students, they are likely to be different for cities in which prestigious private schools and Catholic schools are located, for historical reasons, in neighborhoods that are more urban and poorer than the areas in which their students reside. By associating some private school students with the "wrong" localities, one introduces some degree of measurement error that is probably systematic rather than random. Given the small number of students affected, the resulting bias is likely to be small. It affects only the estimate of the second general equilibrium effect. The estimate of the first general equilibrium effect is unbiased.

9. See United States Geological Survey (1993) and United States Geological Survey (2004).

10. The new dataset is Dataware (1999). The dataset became available after the paper was written though before the year of publication (2000) owing to standard publication delays.

11. The federal standard for navigability is based on a Supreme Court case known as The Daniel Ball, 77 U.S. (1870). In it, streams are defined as navigable-in-fact "when they are used or susceptible of being used, in their ordinary condition, as highways for commerce, over which trade and travel are or may be conducted in the customary modes of trade and travel on water." The federal test for navigability is comprised of four criteria: (1) the stream must be susceptible to navigation; (2) the navigation should be for commercial purposes, not merely navigation for any purpose; (3) the stream should be susceptible to

navigation in its ordinary condition; and (4) the stream should be navigable by the customary mode of commercial transportation in the area. Most states, unless they have recently adopted a new definition of navigability based on recreation (irrelevant for the purposes of this paper), use the federal definition. The Texas Natural Resources Code § 21.001(3) defines as navigable a stream that "retains an average width of 30 feet from the mouth up." In the state of Washington, the precedent setting case is Griffith versus Holman (1900, Wash. 347,63 P. 239, 83 AmSt.Rep. 821 s), which says that a stream "averaging in width about 40 feet" or less is non-navigable. Georgia and Arizona require that a stream be navigable by barges that were commonly used for shipping commercial goods in, respectively, 1863 and 1912. Givens versus Ichauway, Inc. (S97A1074., 268 Ga. 710, 493 SE2d 148, 1997) established that the width of the smallest such barge was 35 feet. Thus, in practice, Georgia and Arizona look for streams with a width of 40 to 45 feet.

12. The canals of Fort Lauderdale were deliberately dug to create housing developments and facilitate commercial traffic. Thus, they are a conspicuous example of what ought *not* to count as a stream if the instrumental variable is to work as intended. Even Fort Lauderdale's port, which is linked to its manmade canals and channels, is artificially constructed. For information and history on the canals of the Atlantic coast, including the Altantic Intercoastal Waterway, see Parkman (1983), especially pages 83-87. For a history, including descriptions of the canal and port creation, see Gillis (2004), especially pages 30 and 39.

13. The USGS standard for what constitutes a stream is "a linear flowing body of water." See USGS (2006).

14. There are still a few errors remaining in the *Common Core*. These are corrected by the code in the replication dataset wherever they are found. Compared to the data available in 1993, however, the currently available data are substantially error-free. Also, the data available in 1993 were a vast improvement on the parallel dataset for the 1980 Census: *Summary Tape File 3F*.

For more information on changes to the geographic codes, see United States Department of Commerce (1999) and United States Department of Education (2002).

15. For clarity, it is useful to distinguish between codes that were *corrected* and codes that were *updated*. When the Census and NCES correct a code in, say, the 1987-88 administrative data, they give it the code it should have had at the time. When the Census and NCES update a code between, say, the 1987-88 and the 1991-92 administrative years, they are switching from one correct code to another correct code in order to reflect changed circumstances. Thus, when I say that the program calls the correct geographic codes, it calls the correct codes for the year–not some later, updated year.

16. For information, see *Historic Metropolitan Area Definitions*, United States Department of Commerce, 2005.

17. A variety of missing value indicators are used in the NELS data. In an attempt to make the data transparent for the general user while still respecting NCES distinctions about missing data, I changed a few lines. When I did so, the word "update" should have been added to lines 59 and 69 of the main program.

18. Hoxby (2000) also shows results of regressions that are run wholly at the metropolitan area level. Code that estimates such results is in the replication dataset.

19. See United States Department of Education (1990), pp. 1-2

20. See United States Department of Education (1992), p. 39.

21. For the reasons discussed in this and the above paragraph, the results shown for the base year (eighth grade) tests are the most reliable. A reader may wish to focus on them rather than the results for the tenth grade and twelfth grade tests, which are less representative because of the manner in which NELS drops students. (The reweighting of students from the base year to the follow-up years may help somewhat to counteract the loss of representativeness, but the reweighting cannot do the job well. This is because the sort of student who remains in a school that students tend to leave is not the sort of student who leaves a

school that students tend to leave. Even if they are alike on observable characteristics, "leavers" and "stayers" must differ in their unobservable characteristics.) Despite the fact that the eighth grade results are the most reliable, I show results for all of the available grades in the spirit of transparency.

22. See United States Department of Education (1992), p. 40. Essentially, if a student is already being sampled in the follow-up, a hot-deck procedure based on his school's roster is used to add classmates who were not attending eighth grade in the U.S. in 1988, either because they were abroad or were attending a grade other than eighth. Thus, a school that is sampled may have its sample augmented by the freshening. If a student has already been dropped from the follow-up because he attends, say, a private school that is not his eighth grade school, not only does he have no chance to be represented but his school is also denied the opportunity to add students in the freshening procedure. Thus, the freshened sample aggravates the over-sampling of certain schools and under-sampling of others.

23. Following Rothstein, data from the parent survey is used if available and data from the student survey is used if parent survey data are missing.

24. The twelfth grade family income variable has some peculiar problems. By twelfth grade, students can be contributing to their family's income and such income is simultaneously determined with their achievement. Also, there is no accurate way to adjust the family income of a twelfth grader to make it comparable with the income his family had when he was an eighth grader. Yet, the two sources of income must somehow be made comparable because numerous NELS students have been dropped by the second follow-up, in accordance with the problematic sampling procedure described. I adjust incomes for the change in time between the base year and the second follow-up using Bureau of Labor Statistics computations of mean family income.

Table 1

Instrumental Variables Estimates of the Coefficient on the Choice Index,
Public School Students and All Students

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Public school students only (the "first general equilibrium effect")	6.64** (2.69)	5.36** (2.11)	8.50** (2.91)	7.98** (2.50)	5.92* (3.42)	4.12 (2.59)
All, public and private, school students (the "second general equilibrium effect")	5.76** (2.73)	3.51 (2.16)	8.16** (2.89)	7.77** (2.55)	5.55* `(3.39)	4.41* (2.56)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations. The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).

Table 2

Percentage of NELS Base Year Schools for Which the Zipcode-Backing-Out Method Produces a Missing, Non-Unique, or Apparently Unique but Incorrect District Code

	Missing District Location	Non-Unique District Location	Apparently Unique but Actually Incorrect District Location	Total
Public schools	10 %	28 %	1 %	39 %
Private schools	18 %	22 %	1 %	41 %

Notes: The table shows the percentage of base year schools in the NELS for which the zipcode-backing-out method, described in the text, produces a school district code that is missing, non-unique, or apparently unique but incorrect. Comparison is made to the actual school district locations based on school codes provided by the NELS, as described in text. Source: author's calculations.

Table 3

Instrumental Variables Estimates of the Coefficient on the Choice Index
Using Alternative Schemes for Classifying Streams

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Larger and smaller streams, streams classified by primary location (base case estimates)	6.64** (2.69)	5.36** (2.11)	8.50** (2.91)	7.98** (2.50)	5.92* (3.42)	4.12 (2.59)
Larger and smaller streams, streams classified by flow	6.11** (2.50)	4.98** (2.05)	7.98** (2.78)	7.44** (2.38)	5.80* (3.32)	4.14 (2.60)
Smaller streams only	5.86* (2.81)	4.21** (2.03)	7.82** (3.12)	6.28** (2.49)	3.88 (3.19)	2.35 (2.32)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations. The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).

Table 4

Instrumental Variables Estimates of the Coefficient on the Choice Index
With Ohio Districts as in Raw Data and as Fixed

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
4 Ohio districts left as in raw Common Core data	6.64**	5.36**	8.50**	7.98**	5.92*	4.12
	(2.69)	(2.11)	(2.91)	(2.50)	(3.42)	(2.59)
Common Core data fixed for 4 Ohio districts	6.64**	5.36**	8.50**	7.98**	5.92*	4.12
	(2.69)	(2.11)	(2.91)	(2.50)	(3.42)	(2.59)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations. The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).

Table 5

Instrumental Variables Estimates of the Coefficient on the Choice Index
Using Metropolitan Area Codes from Various Contemporaneous Years

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Using the 1987-88 metropolitan area codes	6.64** (2.69)	5.36** (2.11)	8.50** (2.91)	7.98** (2.50)	5.92* (3.42)	4.12 (2.59)
Using the 1989-90 metropolitan area codes	6.56** (2.66)	5.40** (2.11)	8.32** (2.86)	7.81** (2.47)	5.68* (3.37)	4.00 (2.57)
Using the 1991-92 metropolitan area codes	6.58** (2.65)	5.50** (2.10)	8.39** (2.87)	7.88** (2.47)	5.62* (3.36)	4.02 (2.57)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations. The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).

Table 6

Instrumental Variables Estimates of the Coefficient on the Choice Index
With and Without Typographical Error

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
without typographical error	6.64**	5.36**	8.50**	7.98**	5.92*	4.12
	(2.69)	(2.11)	(2.91)	(2.50)	(3.42)	(2.59)
with typographical error that was not in Hoxby (2000) anyway	4.41**	4.18**	6.57**	7.84**	5.25*	3.59
	(1.99)	(1.79)	(2.44)	(2.18)	(2.88)	(2.29)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations. The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).

Table 7

Instrumental Variables Estimates of the Coefficient on the Choice Index,
Proper Specification and Most Representative Sample versus
Various Specification Changes and Errors Introduced by Rothstein

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Properly specified metropolitan income variable and most representative sample available	6.64** (2.69)	5.36** (2.11)	8.50** (2.91)	7.98** (2.50)	5.92* (3.42)	4.12 (2.59)
Misspecified metropolitan income variable recommended by Rothstein (the log of the metropolitan mean household income)	6.74** (2.68)	5.39** (2.11)	8.45** (2.90)	7.98** (2.46)	5.91* `(3.38)	4.14 (2.56)
Parental education from the parent survey (when available, makes sample less representative, see text)	5.61** (2.37)	4.87** (2.08)	7.50** (2.68)	6.95** (2.39)	5.07 (3.20)	3.38 (2.57)
Family income from surveys other than the base year survey (when no base year income available, makes sample less representative, see text)	6.52** (2.60)	5.29** (2.11)	8.43** (2.76)	8.13** (2.40)	5.67* (3.26)	4.32* (2.48)

Notes: The table shows the instrumental variables estimate of the coefficient on the choice index for regressions in which observations are students in the NELS data. Stata's robust clustered standard errors are in parentheses. The clustering unit is the metropolitan area. ** (*) indicates statistical significance at 5 percent (10 percent). Source: author's calculations (see text for a description of each row). The data and code used to produce these estimates are available in the replication dataset. The specification is analogous to the "Base IV" regressions in Table 4 of Hoxby (2000).