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

The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Competition Among Public Schools: A Reply to Rothstein (2004)
Caroline M. Hoxby
NBER Working Paper No. 11216
March 2005
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

In order to be productive, replication should be carried out in an objective way, as it is in some scientific fields. A replicator should go into the enterprise equally ready to find evidence confirming or contradicting a previous study. He should put himself in the shoes of the researcher and try to make good decisions. He should not countenance poor decisions or introduce erroneous data in a search for results that conflict with established results. Unfortunately, economics lacks a strong tradition of scientific replication and, perhaps as result, some work that purports to be replication is not. Some economists even seem to believe, confusedly, that replication is not worthwhile unless it upsets a previous result.[1]

Rothstein may suffer from such confusion. In 2003, he produced a comment (hereafter Rothstein 2003) that purported to be a replication of Hoxby (2000). The comment is not a replication, and each of its points is incorrect. One of its major points is that lagged school districts are a valid instrumental variable for today's school districts. This is not credible, as discussed below. Also, Rothstein argues 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 argument is wrong for multiple reasons. I examine these and other claims from Rothstein (2003) below. I responded comprehensively to that comment in correspondence through the *American Economic Review (AER)* .

Perhaps unable to find counterarguments to my responses, Rothstein produced another comment (hereafter, Rothstein 2004) that has almost no overlap with the first. Rothstein (2004) makes several claims, each of which is wrong and/or misleading. For instance, one of his important claims 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 below that the

[1] Journals that only publish replications that upset previous results may unintentionally nurture such beliefs.

change in estimates is actually due to his using erroneous data for public school students. Another of Rothstein's claims is 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. Indeed, 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. In the sections that follow, I discuss every claim of any importance made by Rothstein. 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. 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. As described in Section I, I went to some lengths to make them available when serious researchers asked me whether they could use the data for econometric and other exercises. Although many authors make extracts available to fellow researchers, I believe that what I have made available is almost unique in its thoroughness. I have provided complete data and code, not merely extracts and estimation code, but all of the raw data and the code for constructing the dataset.

I. A Data CD Related to Hoxby (2000)

Hoxby (2000) uses restricted-access data from the National Education Longitudinal Study (NELS) that may not be distributed by me, even to other researchers who hold a restricted-access license for the same raw data. The construction of the dataset in Hoxby (2000) is, moreover, complicated because the data are drawn from several sources and matched on geography. The complicated construction would suggest that freely distributing a fully-formed extract would be best. The restricted-access arrangements precluded such distribution, however.

For these reasons, I asked the National Center for Education Statistics (NCES) whether I might create a CD that they could redistribute to holders of restricted-access licenses (hereafter, the "redistribution CD"). Because I think that good empirical work involves thinking through the decisions that produce a dataset, I also wanted to provide the code for dataset construction. Finally, I wanted to produce a CD that was accessible to users.

Having obtained a preliminary go-ahead from NCES, I created a redistribution CD for the Hoxby (2000) dataset. It contains extracts and code for generating the results in the paper. Moreover, it contains all of the raw data from which the extracts were made, except that the researcher is expected to have his own NELS data CD (since he would have had already to obtain it in the process of getting a restricted-access license). The redistribution CD contains code to read the raw data into Stata, generate the variables, match on geography, and run the regressions. With more 80 pages of code, almost a gigabyte of data when uncompressed, several pages of documentation, and citations for all the original data, the CD is a self-contained and thorough resource for people interested in the dataset. In order to ensure that the code was not mysterious, I reorganized my original code so that the steps were easier to follow. I also rewrote code that seemed at all cryptic and annotated it heavily, explaining decisions whenever they were not obvious. I created a master do-file that runs all of the subordinate do-files so that a researcher can automate the entire process of getting from raw data to results.

Jeff Owings at NCES was kind enough to guide the CD through the process of being made available. Bruce Daniel at Pinkerton Computer Consultants Incorporated (a contractor for NCES) examined the code, documentation, and data. I believe that I was the first person ever to ask NCES to redistribute data in this way, and I hope that I will not be the last (although, obviously, their willingness to repeat the experiment may depend on whether people make legitimate use of the resource). Owings' and Daniel's help was especially generous because the way that economists combine data from an array of sources is unusual in education research and is not entirely comfortable for them. We corresponded on code, variables, and data. I rewrote a some code that they thought could be misinterpreted by social scientists who are unfamiliar with economists' methods of doing things. With such a vast amount of code and with reorganization and rewriting to make the code more accessible to others, it is perhaps inevitable that there are a few typos, even after numerous, scrupulous proof-readings. Of course, I am glad to have any mistakes revealed to me and will correct them when I learn of them.

A large amount of work went into the redistibution CD. It is tangible evidence of my commitment to furthering research on the topic. Although I have worked with extracts produced by other researchers, I have never myself seen such a thorough effort to make data accessible and comprehensible.

It is noteworthy that nowhere in his comment does Rothstein describe the redistribution CD in detail or how much work and care it evinces. Indeed, he implies that whatever data I provided was inadequate, and that *he* did all of the work. Nothing could be further from the truth. Indeed, it is dismaying to see a great deal of my work (with strategic changes made by him) appearing on Rothstein's website, presented as though it were mainly if not wholly his work. Rothstein's use of the CD suggests that there may be a danger inherent in making complete data and code available to researchers who have not internalized the ideal of scientific replication. Having so many resources made available to him appears to have enabled Rothstein to focus his time on hunting for typos and rewriting code to conduct specification searches.

II. Some Claims About Data and Econometrics

Rothstein makes several claims that rely largely on innuendo for their force. We shall see that when we consider the evidence, the innuendoes are false.

A. Metropolitan Area Codes

Rothstein claims (page 4) that I changed my "assignment algorithm" for matching NELS data to geographic codes for metropolitan areas. This is untrue: there is no algorithm. School districts are simply matched to the metropolitan area in which they exist.

His claim about the "algorithm" is part of a larger innuendo (pages 2 and 4) that the metropolitan area codes on the CD are dubious because they are corrected. (Strikingly, he simultaneously implies that the metropolitan area codes in Hoxby (2000) are dubious because they are uncorrected.) What Rothstein does not explain is that the only corrections are occasioned by the fact that the Bureau of the Census and NCES improved their geographic coding of school districts between the 1990 and 2000 censuses. Codes that were missing or incorrect in the data I originally used, which were the best available at the time, are now available in more correct form. I naturally include the more correct data on the CD. I could hardly ask NCES to distribute data that is not as correct as what NCES makes available.

When I wrote the paper, the *School District Data Book (SDDB)*, which is based on the 1990 Census, was the just-released state of the art for combining school district data with Census data. (Note that, owing to delays in the editorial process, I wrote the paper almost entirely before 1996 even though it was published only in 2000.) The *SDDB* was in some ways a vast improvement on the parallel dataset for the 1980 Census (*Summary Tape File 3F*). However, the *SDDB* was poorly documented and difficult to use: extracting a single variable could require reading 44 CDs.[2] Although the metropolitan area codes in the *SDDB* were correct for the vast majority of districts, some districts had missing or incorrect codes. I gave a research assistant the task of filling in missing codes and correcting any erroneous codes he discovered. While he did his best, it is unrealistic to suppose that he caught every error. After all, the contractor for Census and NCES --with many more resources at its disposal-- had evidently had a difficult time producing correct metropolitan area codes for the 15,274 districts.[3]

[2] See http://www.nber.org/sddb/ for a description of the issues that arose in the National Bureau of Economic Research's attempt to make the whole of the *SDDB* available in a convenient format for researchers. Their version is based on the dataset held by the National Archives, but it is apparently missing some files that were on the CDs, which I used.

[3] The research assistant used a combination of county codes, mapped to metropolitan area codes, and district names and addresses. The contractor was Proximity One. Producing correct geographic codes for school districts is not a trivial task, especially because school districts and other jurisdictions change over time. The task is especially complicated in states where there are multiple districts with the

Between the 1990 and 2000 Censuses, the Bureau of the Census and the NCES began providing more accurate and detailed geographic codes for school districts. The improved codes were made available in new revisions of the *Common Core of Data*. I used the improved codes as they became available and stopped using the *SDDB* for geographic information. When I created the redistribution CD, I used the improved codes from *Common Core of Data* as I have been doing for the last few years. Compared to results based on the *SDDB,* the estimates based on the improved codes are slightly *more* consistent across grades and subjects of achievement, probably because some measurement error of unknown form was eliminated. The key results are shown in Table 1 below.

Table 1

Instrumental Variables Estimates of the Coefficient on the Choice Index

8th Grade		10th Grade		12th Grade	
Reading	Math	Reading	Math	Reading	Math
4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

It is in no way dubious to put accurate geographic codes on the redistribution CD. Indeed, I wonder if Rothstein would not have more critical if I had knowingly put slightly inaccurate codes on the CD.[4]

same name, non-standard districts, or where district boundaries are not contiguous with the boundaries of municipalities. We will see later that, even after all of the improvements in geographic coding in the 1990s, NCES datasets still contain a small number of erroneous geographic codes.

[4] Indeed, we will see that Rothstein later faults me for *not* using a improved dataset that is newly available. He should not be able to have it both ways.

B. The First Stage Regression

Rothstein (page 7) suggests that it is deceptive for me to show a first-stage regression that is run at the metropolitan area level and that includes all metropolitan areas. He seems to feel that I should show a first-stage regression where the observations are at the student level. He also seems to feel that the regression should include only those metropolitan areas for which there are NELS achievement data. However, showing the first-stage regression that I show was a thoughtful decision made with the advice of the *AER* editor with whom I mainly dealt, Dennis Epple. Epple, for good reason, wanted readers to understand 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. He and I agreed that readers would be confused if they saw thousands of observations that corresponded to students in a regression that really had only as much variation as there are metropolitan areas.[5]

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.[6] But, in second-stage regressions where variables like school spending or private school attendance are the dependent variables, 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 that was most general because it included all metropolitan areas. (Of course, the code on the redistribution CD runs all the variants of the first-stage regression --the specification stays the same but the metropolitan areas covered differ.)

Showing the most general first-stage regression in the form that people would most readily

[5] Indeed, Epple encouraged me to provide results of regressions that are run wholly at the metropolitan area level, where achievement is aggregated up. Such regressions are shown in Hoxby (2000) and are run in the code on the redistribution CD.

[6] The subset of metropolitan areas differs by the grade at which achievement in measured: eighth, tenth, or twelfth.

understand was the right thing to do. Rothstein's objection to it has no merit.

C. Stata's Robust Clustered Standard Errors

Rothstein (page 4) portrays as obfuscatory the fact that, on the redistibution CD, I provide code that calculates Stata's robust clustered standard errors, rather than the standard errors due to Moulton (1986).[7] The portrayal is the opposite of what is true. I deliberately employed robust standard errors, despite the apparent loss of statistical significance, in order make the CD more accessible and transparent to users.

There is no "canned" Stata command that calculates Moulton standard errors. In order to calculate them, researchers must export their data to a matrix programming language. In the case of the standard errors for Hoxby (2000), the matrix programming is complex partly because instrumental variables are employed but mainly because the observations must be weighted properly to produce estimates that are representative of U.S. metropolitan areas. In an effort to make the redistribution CD accessible to the average user, I provided code for a user to employ Stata's "robust cluster" option to address the problem of clustering. The "robust cluster" standard errors are systematically *larger*, not smaller, than Moulton standard errors (see below). Thus, I knowingly reduced the apparent statistical significance of the estimates in order to make the code convenient for others. Rothstein knew that I employed "robust cluster" to make the CD accessible to others because the documentation on the CD says so.

The "robust cluster" option is commonly employed by applied economists. Its main advantages are that it is readily available and that it produces standard errors robust to arbitrary forms of intra-cluster correlation. For instance, if the clusters are states and the observations within each cluster are years, then there is likely to be serial correlation among the yearly observations within each state. For another

[7] Stata's robust clustered errors are Huber-White sandwich estimators. See Huber (1967), White (1980), Rogers (1993), and Wooldridge (2002).

example, suppose that the clusters are establishments and the observations within each cluster are employees. If a subgroup of employees in an establishment experiences a common shock, then there will be correlation in their disturbances. Stata's "robust cluster" option is designed to deal with intra-cluster serial correlation, correlation generated by common shocks, and all other forms of intra-cluster correlation.

In contrast, Moulton standard errors are designed to deal with a specific problem: the intra-cluster correlation generated when otherwise independent observations share a common, cluster-level covariate. This is the case described in Hoxby (2000) where students in the same cluster are independent but share a metropolitan-area level choice index and other aggregate covariates. As demonstrated by Moulton (1986), shared cluster-level covariates generate a cluster-specific random effect. Because any two students in the same cluster share the choice index to the same degree, any two students drawn from a cluster can be expected to have the same positive correlation. This is a highly specific form of intra-cluster correlation. Many other forms of intra-cluster correlation, such as serial correlation, are probably not present in the NELS data. There is little reason to think that just because two students have adjacent identification numbers, they are more similar than two students with non-adjacent identification numbers.

The Moulton method is more efficient but less robust than the "robust cluster" method. That is, the Moulton standard errors are smaller when they are applied in a setting where observations in the same cluster share a common covariate. See Wooldridge (2002) and Baum, Schaeffer, and Stillman (2003). Consider the logic. Moulton standard errors put structure on the disturbance covariance matrix that the "robust cluster" option does not. Thus, with Moulton standard errors, fewer forms of intra-cluster covariance may appear in the estimated disturbance covariance matrix. Having fewer forms of inter-cluster covariance means smaller standard errors in applications like this one.

Formally, suppose the equation to be estimated is

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u} \tag{1}$$

with instrument vector Z. Observations are indexed by i and grouped in clusters indexed by j. There are N_j observations in cluster j. Denote by $\boldsymbol{u}_j$ the $N_j x 1$ vector of disturbances for cluster j Define $\boldsymbol{\Sigma}_j \equiv \boldsymbol{u}_j \boldsymbol{u}_j'$, the $N_j x N_j$ intra-cluster correlation matrix for cluster j. Note that the off-diagonal elements in $\boldsymbol{\Sigma}_j$ represent intra-cluster serial correlation and any other type of intra-cluster correlation. Note that the $\boldsymbol{\Sigma}_j$ can vary from cluster to cluster. This is a cluster analog of heteroskedasticity. The covariance matrix in the presence of clustering is:

$$\Omega = \begin{pmatrix} \Sigma_1 & & & & 0 \\ & \ddots & & & \\ & & \Sigma_j & & \\ & & & \ddots & \\ 0 & & & & \Sigma_J \end{pmatrix} \tag{2}$$

For Stata's "robust cluster" errors, obtain a consistent estimate of $\boldsymbol{u}_j$ --for instance, from the instrumentals variables residuals. Call it $\hat{\boldsymbol{u}}_j$. Compute $\hat{\boldsymbol{\Sigma}}_j \equiv \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j'$. Then, compute

$$\hat{\Omega} = \begin{pmatrix} \hat{\Sigma}_1 & & & & 0 \\ & \ddots & & & \\ & & \hat{\Sigma}_j & & \\ & & & \ddots & \\ 0 & & & & \hat{\Sigma}_J \end{pmatrix} \tag{3}$$

Finally, compute the estimated variance-covariance matrix for the instrumental variables estimator that is robust to arbitrary forms of intra-cluster correction and cluster-based heteroskedasticity:

$$V(\hat{\beta}^{IV,robust}) = (X'P_Z X)^{-1} (X'Z(Z'Z)^{-1} (Z'\hat{\Omega}Z) (Z'Z)^{-1} Z'X)(X'P_Z X)^{-1}. \tag{4}$$

For Moulton standard errors, assume an error components model of u such that $u_{ij} = \nu_j + \epsilon_i$, where ν_j is the cluster-specific random effect and ϵ_i is the individual's random effect. Then, the covariance matrix in the presence of clustering is given by equation (2) except that the intra-cluster correlation matrix for each cluster j is:

$$\Sigma_j^{Moulton} = \begin{pmatrix} \sigma_\nu^2 + \sigma_\varepsilon^2 & & & & \sigma_\nu^2 \\ & \ddots & & & \\ & & \sigma_\nu^2 + \sigma_\varepsilon^2 & & \\ & & & \ddots & \\ \sigma_\nu^2 & & & & \sigma_\nu^2 + \sigma_\varepsilon^2 \end{pmatrix} \tag{5}$$

Estimate equation (1) by instrumental variables with cluster-specific random effects. Using the estimated random effects and estimated individual-level residuals, compute $\hat{\sigma}_\nu^2$ and $\hat{\sigma}_\epsilon^2$ and construct $\hat{\Sigma}_j^{Moulton}$. Finally, construct $\hat{\Omega}^{Moulton}$, the estimated analog of the matrix in equation (3). Then the Moulton estimated variance-covariance matrix is:

$$V(\hat{\beta}^{IV,Moulton}) = (X'P_ZX)^{-1}(X'Z(Z'Z)^{-1}(Z'\hat{\Omega}^{Moulton}Z)(Z'Z)^{-1}Z'X)(X'P_ZX)^{-1}. \tag{6}$$

The essential difference between $V(\hat{\beta}^{IV,robust})$ and $V(\hat{\beta}^{IV,Moulton})$ is, of course, that the off-diagonal elements in $\hat{\Sigma}$ are larger than those in $\hat{\Sigma}_j^{Moulton}$ under all plausible conditions. This is because the off-diagonal elements contain $\hat{\sigma}_\nu^2$ in both cases, but in the "robust cluster" case they may also contain other arbitrary forms of correlation among students' observations --for instance, common shocks.[8]

[8] The only case in which the off-diagonal elements would be smaller is the case of systematic negative correlation in students' residuals. Given the NELS sampling procedures (first sampling a school, then sampling students within a school), the opposite problem of common shocks is the real concern.

Systematic negative correlation would only occur if, for instance, students' performance was measured in relative terms around a fixed mean: student A could not do better without student B doing worse. However, this situation does not apply to NELS data. The exams measure absolute achivement, and every student in a cluster could receive the top score on the exam.

The bottom line is that the "robust cluster" standard errors are presented for the convenience of CD users. I made a sacrifice of statistical significance that seems reasonable in light of the ready availability of "robust cluster" code and the fact that the "robust cluster" estimates are more robust to arbitrary intra-cluster correlation. It is extraordinary to portray as obfuscatory a sacrifice whose only benefits were accessibility and transparency.

D. Four Ohio School Districts

Rothstein (pages 4-5) makes much of the fact that four school districts in Ohio are assigned to a North Carolina metropolitan area. He alleges that this is due to incorrect coding on my part. The truth is that the erroneous metropolitan area codes are in the raw data from the *Common Core of Data*. Moreover, Rothstein avoids revealing the fact that the four districts in question cannot affect the achievement results anyway because there are no NELS students in them.

Just for completeness, Table 2 below shows the results with and without the correction for the four Ohio districts. Obviously, the results are identical.

Table 2

Instrumental Variables Estimates of the Coefficient on the Choice Index

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
4 Ohio districts left as in raw Common Core data	4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (2.29)
Common Core data fixed for 4 Ohio districts	4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

While I certainly attempted to check and correct geographic codes in the *Common Core of Data* (see lines 580-584 and 627-631 of the main do-file on the CD for problem codes that I caught and corrected), surely the blame for problem codes ought to be directed toward the creators of the raw data. This is especially true given that the districts in question cannot affect the achievement results about which Rothstein is concerned. Of course, now that the problem in the raw data has been pointed out to me, I am glad to add the corrections to the CD. A normal reaction to spotting a harmless error in the raw data would be to send a message to the data creators and to me through the *AER*. Rothstein, instead, builds an innuendo around it.

E. Use of Contemporaneous Metropolitan Area Codes

Rothstein (page 5) claims that several additional school districts have "incorrect, invalid, or obsolete" metropolitan area codes. In fact, I match districts in the NELS data, which has a base year of 1987-88, to the metropolitan area codes in the corresponding year (1987-88) of the *Common Core of Data*. If a district does not have a metropolitan area code after this match, I then attempt to match it with the metropolitan area codes in the 1989-90 *Common Core*, which corresponds with the first follow-up year of NELS. Thus, if the Census updated or created a metropolitan area so that it included new districts between the NELS base year and first follow-up year, such updates were included. Finally, if a district does not have a metropolitan area code after the first two matches, I attempt to match it with the metropolitan area codes in the 1991-92 *Common Core*, which corresponds to the second follow-up year of NELS. Thus, if the Census updated or created a metropolitan area so that it included new districts between the NELS first follow-up and second follow-up, such updates were included. The Bureau of the Census routinely updates the definitions of metropolitan areas. For information, see the page entitled "Historic Metropolitan Area Definitions" on the website of the Bureau of the Census.

In short, I follow the data as closely as possible, matching NELS data to the administrative data from the years covered by the survey. This is the most natural method: taking the years as given and letting the Census and NCES determine which districts were in metropolitan areas at the time.

Rothstein may not like the ways that metropolitan areas are updated and included in the *Common*

Core of Data, but he ought take up such matters with the agencies in question. What he ought *not* to do is what he does: arbitrarily choose some particular set of metropolitan area codes and then declare that codes outside his arbitrary set are obsolete or invalid. He offers no justification for why we should discard the contemporaneous administrative data and use another set of codes instead. If I had arbitrarily picked codes in this manner, he might well have suggested that picking them arbitrarily was wrong.

In any case, the number of districts changing codes over the short period in question is so small that they hardly affect the results. Thus, so long as one sticks to codes that are contemporaneous with the data, the choice does not matter. This is shown in Table 3.

Table 3

Instrumental Variables Estimates of the Coefficient on the Choice Index

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Using the 1987-88	4.41**	4.18**	6.57**	7.84**	5.25*	3.60
metropolitan area codes	(1.99)	(1.79)	(2.44)	(2.18)	(2.88)	(2.29)
Using the 1989-90	4.32**	4.22**	6.43**	7.72**	5.07*	3.49
metropolitan area codes	(1.96)	(1.78)	(2.40)	(2.15)	(2.84)	(2.26)
Using the 1991-92	4.36**	4.32**	6.46**	7.72**	5.02*	3.51
metropolitan area codes	(1.95)	(1.76)	(2.40)	(2.15)	(2.82)	(2.26)

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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

Given that the results are nearly unaffected by whether one uses metropolitan area codes corresponding to the first, second, or third year of NELS, and given that it would be hard to justify using *non-contemporaneous* metropolitan area codes, why does Rothstein discuss this issue? It is an illustration of argument by innuendo: he may think that it is enough simply to imply that the codes are wrong.

F. A Typo

Rothstein also makes much (page 5 and thereafter) of how he had to correct my dataset to incorporate students whose eighth grade school identification codes were missing but whose tenth or twelfth grade codes were not. He implies that the coding mistake was serious. In reality, he is referring to the fact that the single word "update" is missing from lines 52 and 62 of the main do-file on the CD (there are almost 18,000 words in this do-file alone). These lines did not need the word "update" when I wrote them originally but they were indirectly affected by a change I made in coding elsewhere in order to satisfy NCES concerns. (Specifically, NCES was concerned about my using zeros to indicate missing values in some places, owing to the possibility of misinterpretation. When I changed certain zeros to missing values, the word "update" should have been added to lines 52 and 62.) If Rothstein had simply sent me an email though the *AER* asking about the typos, I would have checked, readily agreed and, corrected then. Instead, he frames much of his comment around the typos, repeatedly referring to his corrected dataset as though he had made a dramatic improvement. With a mature view of empirical work, I know that any researcher ought to consider himself fortunate to find only two missing words in 80 pages of code. In any event, the typos were created when modifying the code for NCES redistribution, not in the process of writing the original paper. Thus, his correction is really a correction to the redistribution CD, not Hoxby (2000). Moreover, the typos have little effect on the results.

Table 4 compares the estimates without and with the typos.[9]

[9] All other results in this paper are without the typos, so that readers can see that they were in no way important to the results.

Table 4

Instrumental Variables Estimates of the Coefficient on the Choice Index

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
without typos	4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (2.29)
with typos	4.82** (1.95)	4.41** (1.86)	7.36** (2.50)	7.86** (2.15)	5.88* (2.86)	3.84* (2.30)

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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

Given that correcting the typos makes little difference, why does Rothstein attempt to make them an important issue? Again, it is argument by innuendo rather than evidence.

III. Other Comments

Let us consider some of Rothstein's other comments.

A. Private School Students and the Zipcode Backing-Out Procedure

I am the first to agree that the number and composition of students who attend public schools in a metropolitan area may be endogenous to the absence or presence of competition among public schools in the area. Indeed, one section of Hoxby (2000) is dedicated to showing that the private school attendance rate is affected by the amount of public school competition. It is perfectly reasonable to want to know whether part of the positive effect of competition among public schools on public school students comes through the channel that "better" students are more likely to stay in the public schools when there is competition. In other words, as economists, we can be interested in both the partial equilibrium effect of competition on public school students' achievement and the general equilibrium effect, which includes the effects of students potentially shifting among schools.

In passing, it is worth noting that policy makers are not particularly interested in the partial equilibrium effect. When they ask whether competition improves or undermines public schools, they are usually asking the general equilibrium question.

Unfortunately, there is no practical method of recovering the partial equilibrium effect. What one can do is estimate *another* general equilibrium effect: the effect of public school competition on all students, not just public school students. Rothstein (page 13) argues that this second general equilibrium effect is an unbiased estimate of the partial equilibrium effect but he is confused: students are still moving between public and private schools. Thus, the public/private mix of instruction is changing, peer effects are potentially changing, and parental support for public schools is potentially changing. The second general equilibrium is neither better nor worse than the first general equilibrium effect; it is just different. In any case, I have always considered the second general equilibrium effect to be interesting, and it is an effect I first estimated and showed in Hoxby (1994). Rothstein (2004) suggests that I should also have estimated this effect in Hoxby (2000), and I am not averse to such a suggestion. The difficulty is a practical one.

In the NELS, there is a unique school code for public school students that allows one to identify a student's geographic location. There are no other explicit geographic identifiers of a student's location within his state. This poses a problem for private school students, who can be traced only to their state using the codes provided. One can attempt to back out a private school student's location, with some error, by using Census variables in the restricted-access NELS that are associated with each school's zipcode. For clarity: students' zipcodes are *not* provided by NELS. NELS merely makes available a few variables derived from the Census that describe the zipcode of student's school, not his residence.[10] By cross-referencing these variables, one can back out a unique school district location for some students. However, this method associates other students with up to nine districts. When one validates the backed-

[10] The Census variables are drawn from 1990 Summary Tape File 3B.

out district codes using actual NCES codes for public schools, one finds that more than a third of backed-out district codes are incorrect or not unique. Table 5, below, shows the percentage of public schools with missing, incorrect, or non-unique "backed-out" codes.

Table 5

	Base Year (8th Grade School)	1st Follow-Up (10th Grade School)	2nd Follow-Up (12th Grade School)
Percentage of public schools for which the zipcode backing-out method produces the a missing, incorrect, or non-unique district code	39%	37%	49%

Clearly, the zipcode-backing out method is error-prone. Another problem with the method is that, unlike public school students, private school students need not live in the immediate neighborhood of their school. The backing-out method associates them with neighborhood variables as though they lived in their school's zipcode. This is problematic for cities where prestigious private schools and Catholic schools are often located, for historical reasons, in neighborhoods that are more urban and mixed-income than the areas from which they draw their students.

In short, the zipcode backing-out method may be the best one can do for a student who has never attended a public school. However, one should not use the backing-out method to infer a location for a student who has attended a public school and thereby revealed the actual location where he resides. Whenever one uses the zipcode backing-out method, one should remember that it adds measurement error to all of the independent variables. This measurement error will predictably attenuate coefficients, so that truly significant effects appear to be insignificant. Adding erroneously measured independent variables may also produce systematic bias if the backing-out method is more erroneous for certain types of students or certain metropolitan areas.

In Table 6, I show what happens to the results when I add private school students to the sample.

In the second row, I add only those private school students whose locations are known with certainty because, at some time, they attended a school with a identifying code. In the third row, I add private school students whose locations can only be approximated using the zipcode backing-out method.

Table 6

Instrumental Variables Estimates of the Coefficient on the Choice Index

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
Public school students only	4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (2.29)
Public school students plus private school students whose locations are known with certainty	4.39** (2.03)	3.91** (1.75)	7.31** (2.65)	8.46** (2.38)	5.73** `(2.86)	4.51** (2.26)
Public school students, plus private school students whose locations are known with certainty, plus private school students whose locations can be approximated using the zipcode back-out method	4.57* (2.42)	2.20 (1.82)	6.68** (2.61)	7.11** (2.15)	4.41* (2.70)	2.53 (2.17)

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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

It is clear that the results are extremely similar with and without private school students whose location is known with certainty. Also, the results remain similar when I add private school students whose location can only be approximated with the zipcode backing-out method. The results that use the backing-out method are, however, consistent with the idea that method introduces attenuation bias. This is what we would expect given the known tendency for students' location assignments to be incorrect when zipcode backing-out is used.

Rothstein (2004) claims that adding private school students changes all of the results, substantially. This is not true, as we have seen. What accounts, then, for Rothstein's claim? He gets very different results because he switches to assigning *all* students, not just private school students, to locations based on the zipcode backing-out method. (The fact that he switches assignment procedures entirely is easy to miss, but see paragraph 2 of page 13 of Rothstein's paper.) That is, even for all of the public school students for whom Rothstein has an exact district code, he uses an error-prone assignment procedure. It is hardly surprising that, having added so much measurement error, he obtains statistically insignificant results. If he were to use the zipcode backing-out method for only those private school students who do not have a valid school code, he would obtain the results in Table 6 above.

One can only wonder why Rothstein uses the zipcode backing-out method to assign error-prone geographic codes to students for whom he has a valid school code. There is no possible justification for the procedure.

As an aside, no one should be surprised by the smallness of the difference between the first and second general equilibrium effects, which are shown in Table 6 above. (The difference shown in the table also reflects attenuation bias, so the true difference is probably smaller.) Suppose that public school competition reduces private school enrollment by a substantial amount, like a fifth.[11] Because private school students make up only about ten percent of students in the U.S., only about two percent of students would be shifting schools. This two percent of students would have to be very special to raise achievement significantly among public school students.

B. Misspecification of the Income Variable

Rothstein claims (page 6) that the results of Hoxby (2000) should be robust to substituting, as a covariate, the natural log of the mean income of districts in a metropolitan area for the mean of the

[11] This change corresponds to the effect of a two standard deviation increase in the index of choice among public school districts.

districts' log mean incomes. This claim is strictly wrong. The two variables in question are not the same. The substantial section on aggregation in Hoxby (2000) goes to some pains to discuss how important it is to include a metropolitan-level mean of any district-level covariate that is included in the regression. If this is not done, the estimating equation is misspecified. (See Section VII of Hoxby (2000), especially page 1217.) Thus, the mean of the districts' log mean incomes *must* be included to avoid misspecification. The variable is *not* interchangeable with the log of the mean income of districts.

The question of how aggregation affected the specification was scrutinized carefully by the *AER*'s editor (Epple) and referees. Indeed, much of the correspondence on Hoxby (2000) was directly on the point discussed above: how the inclusion of aggregate variables was needed in the estimating equation so that misspecification could be avoided. Readers will notice that the section on aggregation in Hoxby (2000) is long --disproportionate to the length of the paper. This length was occasioned by the fact that the issue was important to the editor and referees.

Given the care with which aggregation and its implications are discussed in the paper, one can only wonder why Rothstein would knowingly misspecify an aggregate variable and then claim that the results ought to be robust to the misspecification. At a minimum, he ought to be held to the standard to which I was held when publishing the paper. His departure from this standard may be an example of specification searching with the intent of finding different results (as, of course, one is likely to do when one misspecifies an relationship).

C. Counting Navigable Streams

Rothstein (page 8) suggests that it was dubious and subjective to measure the larger rivers from the United States Geological Survey quadrangle maps of the United States. This is the first time that I have heard that measuring an extremely accurate map with a ruler is subjective. Moreover, the quadrangle maps are the underlying source of the information contained in the electronic dataset known as the Geographic Names Information System (GNIS). The difference between the maps and the GNIS is that the latter contains only a tiny fraction of the information contained on the maps. Data do not become more reliable simply because only a fraction are coded and the resulting, sparse set of variables is put in electronic form. When it comes to geographic information, which is easy to portray visually but hard to convey in a few variables, the maps are much more informative than the electronic data in GNIS. Rothstein seems to feel that there are deep structural ambiguities in maps, but anything he does not like about the maps is carried over to GNIS.

Let us back up and remember why I spent weeks with rulers in the map room measuring rivers on quadrangle maps. I first proposed using streams to instrument for the number of school districts because I noticed, on maps, that district boundaries often *were* streams. Later, I learned that early American laws and even Maimonides' Rule 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' traveling to school.

The unfortunate thing about streams as instruments, however, is that some large ones are navigable and may be (or have been) important channels for commerce. 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. Would the city exist if it were not for the confluence of the Allegheny, Monongahela and Ohio? In other words, large navigable rivers might generate certain types of commerce, which might attract certain industries and workers, who might live in densely populated central municipalities, which might monopolize much of the student population in a metropolitan area.

Another problem with larger streams is that a single channel of water can have multiple names if it is long and, even worse, the number of names can be endogenous to the number of jurisdictions through which a stream runs. An example that may be known to readers familiar with the Boston area is the Concord/Sudbury River, an important "working river" in its day. When Henry Thoreau famously wrote about the river, it was known in its entirety as the Concord River.[12] Sometime between 1852 and 1906, the part of the river that runs through the town of Sudbury came to be called the Sudbury River and the part that runs through Concord came to be called the Concord River, but it is a single water course.

In short, although all streams are created by nature, small streams are more credible instruments than are large navigable rivers. The presence of many small streams indicates that an area is generally "watery" and dense in natural barriers. The count of large navigable rivers does indicates natural barriers but it may also indicate omitted factors that cause districts to be densely populated or to break apart. In short, I realized that I needed to measure the smaller and larger streams separately so that they could enter the first-stage equation separately and so that I could use only the smaller streams variable as the excluded instrument if necessary. The correlation between the two types of streams is only 0.41.

Examining the maps, I recognized that the GNIS would be problematic for measuring the number of large, navigable streams. The GNIS is only a list of names and a few geographic coordinates. It does not record width, which is crucial to assessing navigability. In short, the use of visuals and not merely names is important for counting navigable streams.

When I undertook to measure the larger streams, I knew that I was undertaking a very time-consuming task that had an uncertain pay off, but I believed that accurate measurement was often the difference between better and worse empirical research. I believe this even more strongly today and do not regret my weeks in the map room with rulers. There is *no* finding that can or even should be robust to wrong measurement.

Often, social scientists complain about the unwillingness of fellow researchers to measure the variable they really mean to measure (in this case, larger streams) rather than use a crude, but readily available proxy for it. Rothstein (2004) seems to have the reverse complaint and holds up readily available electronic data as necessarily better than the visual data it partially summarizes.

[12] Thoreau (1849).

In any case, as emphasized in Hoxby (2000), it is the smaller streams, the complement of the larger streams, that provide explanatory power as instruments in the first-stage regression. One can, and perhaps should, leave out the larger streams variable as an instrument because it is somewhat suspect owing to endogenous commerce. I do this in Hoxby (2000). One indication that the larger streams variable may have an endogeneity problem is that its coefficient in the first-stage regression is smaller than the coefficient of the smaller streams variable. This suggests that larger streams produce offsetting effects --for instance, more natural barriers but also a tendency toward the establishment of a densely populated center city district. The bottom line is that there is no reason to expect the larger streams and smaller streams variables to have the same coefficient, as Rothstein implicitly claims when he asserts that adding them together is the correct thing to do. (He does not make an explicit argument but relies on assertion.) It appears that Rothstein is promoting not just less-informed measurement but also the destruction of the important information contained in the differentiation between streams that are more and less suitable for commercial use.[13]

Rothstein (page 9) makes an extraordinary argument that it really does not matter how one measures an instrumental variable so long as it remains correlated with the potentially endogenous variable. This is incorrect: some measures will violate the second instrumental variables condition (no correlation between the instrument and the error term in the limit).

Rothstein (page 9) argues that I was wrong to use GNIS data that assigns streams to counties based on their source and that I should have used an alternative GNIS dataset that assigns a stream to any county through which it flows. Here again, he relies somewhat on innuendo by not mentioning the fact

[13] In first stage regression that includes all of the metropolitan areas, both streams variable have a positive coefficient. Rothstein notes that in the subset (approximately 60 percent) of metropolitan areas included in certain regressions in which NELS achievement is the dependent variable, the larger streams variable sometimes has the wrong sign. This may be evidence that the larger streams variable is related to commerce in at least some metropolitan areas. It would be a reason to rely only on the smaller streams variable, not for discarding both streams variables or --even worse-- combining them as Rothstein does.

that the alternative GNIS dataset did not exist when I began writing the paper. (Rothstein is equally critical when I do and do not employ a newly available dataset for the reproduction CD.) In any case, I have no objection to the alternative dataset, which strikes me as a good thing. Use of the alternative data which makes very little difference to the results, however, because the vast majority (93.5 percent) of streams flow only in one metropolitan area. As shown in Table 7, the two datasets produce estimates that are extremely similar.

Table 7

Instrumental Variables Estimates of the Coefficient on the Choice Index

	8th Grade		10th Grade		12th Grade	
	Reading	Math	Reading	Math	Reading	Math
GNIS dataset (USGS website)	4.41** (1.99)	4.18** (1.79)	6.57** (2.44)	7.84** (2.18)	5.25* (2.88)	3.60 (2.29)
Alternative GNIS dataset (Dataware (1999))	4.08** (1.89)	3.91** (1.77)	6.14** (2.32)	7.36** (2.13)	5.24* (2.86)	3.75 (2.34)

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. Standard errors are in parentheses. ** (*) iindicates statistical significance at 5% (10%). The results are from the regressions for the "basic instrumental variables" regressions on the redistribution CD. The specification corresponds to Table 4 of Hoxby (2000).

Before leaving this section, it is worth noting that Rothstein (page 9) claims that there are some metropolitan areas in which there are more large streams than there are total streams. This is incorrect, and I have no idea how Rothstein could have come to this conclusion. The redistribution CD has all of the raw data and code. I can only conclude that Rothstein is making coding mistakes. These mistakes may affect all of his assertions about the streams variables. While I could, in theory, re-do and proofread everything that Rothstein has done, imposing such a duty on authors would allow the least objective replicators to absorb much of the time of authors who do original work.

D. Lagged Districts As Instruments for Current Districts

Rothstein (2003) argues that the number of districts in a metropolitan area in 1942 is a better instrumental variable for the index of choice (based on 1990 districts) than are the streams variables.

Of course, the number of districts in 1942 *is* highly correlated with the number of districts in 1990. The correlation is 0.76 in the log(number of districts). See Table 8. Indeed, 15% of metropolitan areas have no change whatsoever in their number of districts between 1942 and 1990, and 25% of metropolitan areas have their number of districts change by less than 3%. The fact that 1942 districts are very similar to 1990 districts does not make them a valid instrument. Indeed, it suggests strongly that they are not a valid instrument –that the sources of endogeneity that make it necessary to instrument for the 1990 choice index are operating in the 1942 districts as well. Alesina, Baqir, and Hoxby (2004) demonstrate that much of the endogenous formation of school districts played out before 1942, an important period of district consolidation.

Table 8

Share of MSAs/PMSAs with *No Change* in Number of Districts from 1942 to 1990	15%
Share of MSAs/PMSAs with 3% or Smaller Change in Number of Districts from 1942 to 1990	25%
Correlation between ln(1942 Districts) & ln(1990 Districts)	0.76

Notes:
I use the 1990 definitions of MSA/PMSAs and map 1942 districts into them. The correlation that uses the natural log of districts (0.076) is the closest available indicator of the correlation one would get with the 1942 and 1990 choice indices. A 1942 choice index cannot be computed.

Sources:
United States Department of Commerce (1944). United States Department of Education, National Center for Education Statistics. *The Common Core of Data, 1989-90.*

I can think of no argument for 1942 districts being a good instrument. By using them, one is instrumenting with a lagged observation of a variable that has a high degree of long-term serial correlation. It is known that lagged variables are invalid instruments for endogenous variables that are highly serially correlated over the relevant lag length. Such instruments are not orthogonal to the error term owing to the serial correlation. Of course, given the serial correlation, using 1942 districts as an instrument produces results similar to ordinary least squares results.

If I had thought that something as simple as the number of lagged districts was a credible instrument, I would have used them in Hoxby (2000) and avoided a great deal of onerous work. In any case, I do not believe that the editors or referees for Hoxby (2000) would have countenanced such a non-credible instrumental variable, even if I had been willing to use it myself. It is difficult to know what could motivate Rothstein to suggest 1942 districts as an instrument, apart from a desire to obtain results different from those in Hoxby (2000).

E. Measures of Achievement Based on a Self-Selected Sample

Rothstein (2003) argues that it is better to use SAT scores as a measure of student achievement than to use a nationally representative survey like the NELS. This is incorrect: a representative sample is much to be preferred. There are multiple, serious problems with using SAT scores instead of test scores from a representative sample:

First, students self-select to take the SAT based on their achievement, which is unobserved for those who do not take the test. There is no correction that can fully account for self-selection on the basis of a latent dependent variable. Second, in most states, only a minority of students take the SAT, and the percentage is as low as 3percent. See Table 9. There are no statistical techniques that allow an econometrician to deduce what is happening in 97 percent of a distribution from looking at the top 3percent of it.

Third, students of affluent, well-educated parents are the most likely to take the SAT in any state, and such families are disproportionately *not* affected by choice among public schools. Affluent, well-educated parents can always choose private schools for their children and are the most able to move into exclusive neighborhoods or gain admissions for their children at selective public schools (such as the "exam" schools in major cities). By focusing on SAT takers, one is focusing on precisely the students least likely to be "treated" by the existence of more or less public school choice.

Table 9
The Representativeness of SAT Takers, By State*

State	SAT Takers as a Percentage of	Combined Average SAT
	Relevant Enrollment**	Verbal & Math Scores
Alabama	7%	984
Alaska	37%	914
Arizona	20%	942
Arkansas	6%	981
California	33%	903
Colorado	26%	969
Connecticut	73%	901
Delaware	60%	903
District of Columbia	55%	850
Florida	36%	884
Georgia	49%	844
Hawaii	52%	885
Idaho	13%	968
Illinois	16%	994
Indiana	54%	867
Iowa	5%	1088
Kansas	6%	1040
Kentucky	9%	994
Louisiana	8%	993
Maine	55%	886
Maryland	60%	908
Massachusetts	67%	900
Michigan	12%	968
Minnesota	13%	1019
Mississippi	3%	996
Missouri	11%	995
Montana	19%	987
Nebraska	10%	1030
Nevada	18%	921
New Hampshire	64%	928
New Jersey	72%	891
New Mexico	11%	1007
New York	63%	882
North Carolina	51%	841
North Dakota	6%	1069
Ohio	22%	949
Oklahoma	8%	1001

Table 9
The Representativeness of SAT Takers, By State*

State	SAT Takers as a Percentage of Relevant Enrollment**	Combined Average SAT Verbal & Math Scores
Oregon	40%	923
Pennsylvania	63%	883
Rhode Island	58%	883
South Carolina	51%	834
South Dakota	5%	1061
Tennessee	10%	1008
Texas	38%	874
Utah	4%	1031
Vermont	60%	897
Virginia	59%	895
Washington	36%	923
West Virginia	14%	933
Wisconsin	11%	1019
Wyoming	12%	977

* For students who are in the same cohort as the typical NELS student (11th graders in 1990-91).
** Public school grade 11 enrollment in 1990-91 plus private high school graduates in the 1991-92 school year. (Data on private school enrollment by grade is not available.)

Sources:
SAT scores and number of test takers: The College Board (2004)
Public school enrollment in grade 11: United States Department of Education, National Center for Education Statistics, *Digest of Education Statistics, 1992*. Table 42.
Private high school graduates 1991-92: United States Department of Education, National Center for Education Statistics, *Private School Universe Survey, 1993-94*. Table 17.

Fourth, the percentage of students who take the SAT varies extraordinarily among states, and the variation is far from random (see Table 9). The states with the top SAT-taking rates are Connecticut, New Jersey, and Massachusetts. They obviously differ on many socioeconomic and demographic dimensions from the states with the bottom SAT-taking rates: Mississippi, Utah, South Dakota, and so on. Thus, the sample selection bias introduced by using SAT scores is not only serious but varies systemically with metropolitan area characteristics. All of the coefficients will be biased by sample selection, including the coefficient on the choice index.

Fifth, there is a strong negative correlation of -0.84 between a state's SAT-taking rate and its combined average SAT score. (The correlation between the two columns in Table 9 is -0.84.) The strong negative correlation makes it obvious that self-selection into SAT-taking is a serious problem. Rothstein (2003) claims that controlling for the SAT-taking rate is a sufficient correction for sample selection bias. This claim is strangely out of touch with the large econometric literature on self-selection.[14] If Rothstein's claim were valid, we could presumably easily control for many serious self-selection problems simply by controlling for the rate at which people self-select into a treatment. We could solve wage bias that results from people self-selecting into the labor force just by controlling for the percentage who participate. We could solve bias in estimating the rate of return to college just by controlling for the percentage of people who attend college.

In short, there is no valid reason to use SAT scores in preference to a nationally representative measure of achievement.

IV. Conclusions

Rothstein, in both his 2003 and 2004 comments, consistently makes decisions that cannot withstand scrutiny. His decisions would have run afoul of editors, referees, and most seminar audiences with whom I dealt. Why, especially when he has the benefit of hindsight, careful explanations, and nearly all of the data work done for him, does he make bad decisions repeatedly? It may be that he has not internalized the purpose of replication and was determined to generate results that would contradict those in Hoxby (2000) at whatever cost. It should surprise no one that if a person makes a determination to change data and specifications until a result disappears, he will eventually succeed, particularly if he gives himself latitude to misspecify equations in ways that are known to be wrong and to substitute error-prone data for relatively error-free data. Ironically, such behavior is actually facilitated by an author

[14] It is impossible to do justice to this literature in a note, but any review should include Heckman's corpus of work (a good survey is Heckman 1987) and the recent literature on causal inference (see, for instance, Angrist, Imbens, and Rubin 1996). A recent overall survey is Moffitt (1999)

making code and raw data (as opposed to just an extract) available.

Rothstein uses innuendo to suggest that things are important when they are not and to justify procedures for which he makes no explicit argument. Allowing a commentator to rely on innuendo sets a double standard that is biased against original research. An original researcher must make an argument for whatever he does and must convince a good many people (editors and referees of course, but also his readership) that his arguments are right. An original researcher must demonstrate that something is important before he gets to publish it.

People who purport to be replicators can impose serious time costs on original researchers, so it is important for economics to develop and maintain a culture of *scientific* replication if the activity is not be productivity-reducing when it has the potential to be productivity-enhancing. Rothstein's (2003) and (2004) comments are not replications in the scientific sense and are without merit.

References

Alesina, Alberto, Reza Baqir, and Caroline M. Hoxby. 2004. "Political Jurisdictions in Heterogeneous Communities," Journal of Political Economy, 112.2.

Angrist, Joshua, Guido Imbens, and Donald Rubin. (1996) "Identification of Causal Effects Using Instrumental Variables," Journal of the American Statistical Association, 91.434 (June), 444-55.

College Board, Office of Public Affairs. 2004. SAT Scores by State. Typescript. New York: The College Board.

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

Heckman James J. (1987), Selection Bias and Self-Selection, in P. Newman, M. Milgate and J. Eatwell (eds.), The New Palgrave - A Dictionary of Economics, Macmillan.

Hoxby, Caroline M. 1994. "Do Private Schools Provide Competition for Public Schools?" NBER Working Paper no. 4978.

Hoxby, Caroline M. 2000. "Does Competition Among Public Schools Benefit Students and Taxpayers?" American Economic Review, 90.5.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press, vol. 1, 221–223.

Kenny, Lawrence, and Amy B. Schmidt. 1994. "The Decline in the Number of School Districts in the U.S.: 1950 - 1980," Public Choice.

Moffitt, Robert A. 1999. “New Developments in Econometric Methods for Labor Market Analysis.” In Handbook of Labor Economics, Chapter 24, Volume 3A, 1367-1397.

Rogers, W. H. 1993. Regression standard errors in clustered samples. Stata Technical Bulletin 13: 19–23. Reprinted in Stata Technical Bulletin Reprints, vol. 3, 88–94.

Rothstein, Jesse. 2003. "Does Competition among Public Schools Benefit Students and Taxpayers? Comment." Typescript.

Rothstein, Jesse. 2004. "Does Competition among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000)." Typescript.

Thoreau, Henry David. 1849. A Week on the Concord and Merrimack Rivers. Boston, Cambridge: J. Munroe and Company. New York: G. P. Putnam.

United States Department of Education, National Center for Education Statistics. 1994. *Digest of Education Statistics, 1992*. Washington DC: United States Department of Education.

United States Department of Education, National Center for Education Statistics. 1996. *Private School Universe Survey, 1993-94*. Washington DC: United States Department of Education.

United States Department of Education, National Center for Education Statistics. 2004. *The Common Core of Data*. 1987-88, 1989-90, and 1991-92 editions. Electronic data. Washington DC: United States Department of Education.

United States Department of Commerce, Bureau of the Census. 1944. Governmental Units of the United States 1942. Washington DC: United States Government Printing Office.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48: 817–830.

Wooldridge, J. M. 2002. Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge, Massachusetts.