Variable Definitions, Data and Access, and Programs for "Teachers, Schools, and Academic Achievement"

by STEVEN G. RIVKIN, ERIC A. HANUSHEK, AND JOHN F. KAIN

This supplement to the paper, "Teachers, Schools, and Academic Achievement," provides corrections to tables in the original papers along with added technical details to the data, variable definitions and estimation. That paper has two major empirical sections: estimation of the variance in teacher quality and estimation of more traditional educational production functions. Both rely upon administrative data for students and teachers from the Texas Education Agency. The microdata are currently available only under special procedures because of federal privacy regulations, but the aggregate data on school outcomes used in the estimation of the variance in teacher quality are available through this site. In addition to a description of the data and variables, this supplement also includes all of the STATA do-files used for data construction and estimation.

KEYWORDS: student achievement, teacher quality, variable definitions, estimation programs, school data

The estimation results reported in the original paper have been corrected because of discovery of some coding errors in the original data on class size and also problem with the original estimation files. While not affecting the conclusions, the corrected tables are included in this supplement (see Econometrica.corrected_tables.pdf).

The estimation in this paper relies upon administrative data compiled by the Texas Education Agency and covering all schools in the State of Texas. Access to these data is covered by federal legislation. The *Family Educational Rights and Privacy Act* (FERPA) (20 U.S.C. § 1232g; 34 CFR Part 99) is a Federal law that protects the privacy of student education records. The law applies to all schools that receive funds under an applicable program of the U.S. Department of Education. Under this law, access to individual student information that could lead to identification of individual students is possible only under restricted circumstances.

The Texas Schools Project of the University of Texas at Dallas has been selected as an Educational Research Center of the State of Texas and in that capacity maintains the microdata for use by qualified researchers who use the data for the purpose of research to improve instruction in Texas schools. Access to those data requires application and acceptance of a research proposal to the Joint Advisory Board (JAB) of the Texas Higher Education Coordinating Board. Procedures for this can be found on the website of the Texas Schools Project (http://www.utdallas.edu/research/tsp-erc/). The Texas Schools Project can currently provide other researchers the aggregate school data used the in estimation of the variance of teacher quality but cannot provide the underlying microdata on student performance except as authorized by the JAB.

This website provides the programs used in the creation of the data sets, in the estimation of the variance in teacher quality, and in the estimation of the educational production functions. These programs are STATA "do-files" that execute the commands to perform the operations that are described. As noted, the data for the do-files in part II of the description below are contained on this website, but the data for the do-files in part III is not currently available.

Texas School Data

The data that are used in this paper come from the data development activity of the UTD Texas Schools Project of the University of Texas at Dallas; see Kain (2001) [http://www.utdallas.edu/research/tsp/Index.htm]. Working with the Texas Education Agency (TEA), this project has combined a number of different data sources to compile an extensive data set on schools, teachers, and students. Demographic information on students and teachers is taken from the PEIMS (Public Education Information Management System), which is TEA's statewide educational data base. Test score results and a limited amount of student demographic information are stored in a separate data base maintained by TEA and must be merged with the student data on the basis of unique student IDs. Data are compiled for all public school students in Texas, allowing us to use the universe of students in the analyses. In this paper all of the information on students comes from the test score data base, and we combine student information from the Texas Assessment of Academic Skills (TAAS) data base with teacher and school information contained in the PEIMs data base for three student cohorts: 3rd through 7th grade test scores for one cohort (4th graders in 1995) and 4th through 7th grade test scores for the other two (4th graders in 1993 and 1994).[1]

Beginning in 1993, the Texas Assessment of Academic Skills (TAAS) was administered each spring to eligible students enrolled in grades three through eight. We focus on test results for mathematics and reading. The bottom one percent of test scores are trimmed from the sample in order to reduce measurement error. Participants in bilingual or special education programs are also excluded from the sample, because of the difficulty in measuring school and teacher characteristics for students who split time between regular classrooms and special programs.

Student data are merged with information on teachers using unique school identifiers. The personnel data provide information on all Texas public school teachers for each year. Experience and highest degree earned are reported, as are the class size, subject, grade, and population served for each class taught. Although the currently available data do not permit linking individual students with specific teachers, the available information is used to construct subject and grade average characteristics for teachers in regular classrooms.

In an effort to reduce problems associated with measurement error, a number of observations are excluded from the data set. The following paragraphs describe in detail the construction of the variables and the sample selection procedures.

Measurement error in the teacher characteristics is an important issue. In many cases reported teacher experience in one year does not correspond with reported teacher experience for other years. If the experience sequence is valid except for one or two years that do not follow from the others, we correct experience for those years. If experience data are inconsistent for all the years, if there are two consistent patterns, or if correction would impute negative years of

[1]Note that, while we have 3rd grade test information, our analysis begins at 4th grade because of the focus on achievement gains.

experience, no corrections are made. In any case, no teachers are excluded from the final sample on the basis of inconsistent experience data, though the results are not sensitive to their inclusion, possibly because we used discreet experience categories.

The case of average class size is somewhat more complicated. Teachers were asked to report the average class size for each class they taught that was of a different size. Unfortunately, many teachers appear to have reported the total number of students taught per day. This becomes particularly problematic for schools that move from general to subject specific teachers. Consider a school with two fourth grade classes of twenty students in which the two teachers each teach all subjects. If the school switches to math and reading specialists for 5th grade and each teaches one subject for each class, they will report class sizes of forty if they report total number of students served. It will appear that class sizes doubled as students aged, when in fact they remain the same.

In order to reduce problems introduced by measurement, all reported class sizes that fall below 10 or above 25 in 4^{th} grade (35 in higher grades) are set to missing prior to the computation of school averages for each grade. By statute, 4^{th} grade classes are not supposed to exceed 22 students, though some schools receive waivers to provide slightly larger classes. It is our understanding that very few elementary schools in Texas have actual class sizes in later grades that exceed 35 students during this period. Estimates of class size effects increased in magnitude following these exclusions, suggesting that class size was measured with error for these schools.

Analytic Data and Variables

The TAAS data base contains annual files for elementary school students in grades three through eight, and we use information on the aforementioned grades and years. These files contain test results and a small number of demographic variables in addition to year, grade, student and campus identifiers. The variables used in this paper are:

ethnic - five way classification of ethnicity
disadv - eligibility for a free or subsidized lunch.
sex - gender dummy variable
rawred - number correct on the reading test
rawmth - number correct on the mathematics test
rscode - categorical variable indicating the testing conditions on the reading test
mscode - categorical variable indicating the testing conditions on the mathematics test
biling - variable indicating participation in bilingual education
speced - variable indicating participation in special education

The test score sample is restricted to non-special education, non-bilingual students who have valid scores as indicated in the rscode and mscode variables and who do not score in the bottom one percent of the distribution in the subject, year, and grade in which they take the test.

Information on teachers is contained in the PEIMs data base. The teacher data are divided among a small number of annual files based in part on timecards turned in by teachers. Teachers are supposed to turn in a separate card for each class taught. Therefore teachers who teach multiple grades or subjects and are not classified as general teachers should complete multiple timecards. The program masttea.do shows the sources of each of the teacher and classroom variables. The variables used in this paper in addition to a teacher identifier, grade, and year are:

experience - number of years of experience
class size - teacher self reported class size
post-graduate degree - A variable indicating completion of an M.A. or Ph.D.
subject - the subject taught in that class. We keep teachers who report math, English, or general.
population taught - We keep teachers in regular classes.
set - this variable indicates the type of setting. We keep teachers who teach in classroom settings.

STATA do-file Descriptions

The files are divided into three groups. The first creates the teacher data sets used in both empirical parts. The second contains files used in the estimation of the variance of teacher quality. These include files that extract the student information from the raw data, create the necessary variables, combine the data with the appropriate teacher files, and run the regressions. The third group contains the files used in the education production function analysis. These include files that extract the student information from the raw data, create the necessary variables, combine the data with the appropriate teacher files, and run the regressions.

I. Files used to create teacher and school data sets.

1. masttea193198.do - This file reads TEA PEIMS data sets and combines information on teachers into a single file.

2. fixexp.do - This file checks and fixes the teacher experience variable.

3. teavar1.do - This file processes the teacher data and produces four data sets. One contains school average math teacher characteristics by grade and year; one contains school average math and general teacher characteristics by grade and year; one contains school average English teacher characteristics by grade and year; and one contains school average English and general teacher characteristics by grade and year. The file also creates a data set indicating the school/grade/year combinations that have no teachers who teach both math and reading.

4. teavar2.do - This file creates separate teacher data sets by grade, subject and year.

5. turnover.do - This file produces a data set that contains rates of teacher turnover for math teachers, English teachers, general teachers, math and general teachers combined, and reading and general teachers combined.

6. turnov3.do - This file produces information on the transitions of new teachers, dividing schools on the basis of those transitions in a particular grade and year. This is used to learn more about experience effects on achievement.

7. addturn.do - This file produces a data set of turnover variables that can be used as regressors in the education production function analysis.

8. suprinc.do - This file produces a data set with variables indicating principal and superintendant transitions.

II. Files used in the estimation of the variance in teacher quality

1. dec493c.do; dec494c.do; dec495c.do - These three files use the TEA TAAS files to create the data sets for the 4th grade 1993 cohort, 4th grade 1994 cohort, and 4th grade 1995 cohort used in the semiparametric analysis of the variance in the quality of mathematics instruction.

2. rdec493c.do; rdec494c.do; rdec495c.do - These three files use the TEA TAAS files to create the data sets for the 4th grade 1993 cohort, 4th grade 1994 cohort, and 4th grade 1995 cohort used in the semiparametric analysis of the variance in the quality of reading instruction.

3. colldatv.do - This file combines the teacher turnover data sets and the student achievement data sets into two files for each subject, aggregating the student data by campus, grade and year. It calculates the squared difference in average gains between adjacent cohorts and produces a data set in which each observation is the difference between adjacent cohorts. For each subject one file includes all students and the other includes only students who remain in the same school for consecutive years.

4. covregv.do; covregss.do; covregsu.do - These three files estimate the regressions used to produce estimates of the variance in the quality of instruction in mathematics and reading. covregv.do uses all schools, covregsu.do uses a sample restricted to schools with no teachers in both English and mathematics, and covregss uses a sample restricted to schools with a single teacher in the grade for the relevant subject.

5. tqualdat.do - This file produces a data set with information on school average student gains and teacher turnover for the large Texas district for which there are student/teacher matches.

6. difexit20.do; difexit40.do; difexit60.do; and difexit30tail.do - These four files estimate the effects of non-random attrition on the estimates of the variance in teacher quality using a sample of students in the large district for which students can be matched with teachers. The number refers to the number of quality intervals used in the calculations, and tail indicates that the top and bottom one percent of students in terms of test score gains are included in the calculations.

III. Files used in the education production function analysis

1. ren493c.do; ren494c.do; ren495c.do - These three files use the TEA TAAS files to create the data sets for the 4th grade 1993 cohort, 4th grade 1994 cohort, and 4th grade 1995 cohort used in the education production function analysis.

2. studteam.do; studtear.do - These two files combine the student and teacher data sets and create the data sets used in the education production function regressions. Separate data sets are created by subject, year, and grade

3. ifixeff.do; ifixeffr.do - These files combine the student/teacher data sets, generate some additional variables, save data sets for the more complicated fixed effects analysis, and run OLS and student fixed effect regression models. Descriptive characteristics also come from these files.

5. mathregs.do; readregs.do - These files run student and school, school by year, and school by year and school by grade fixed effects models including models that estimate separate effects by subsidized lunch eligibility.

6. expreg.do - This file estimates the returns to experience for different samples in order to learn more about the sources of the achievement/experience relationship.

7. corsm.do; corsr.do - These files combine the teacher/student data sets, create school math and reading gains for each grade, and produce Pearson Correlation Coefficients for adjacent grades and years.

IV. Files used to create Table 2

1. teavar1ss.do - This file is a slightly modified version of teavar1.do that makes it clear which grades have only a single teacher in a subject.

2. teavar2ss.do; studteamss.do; studtearss.do; - These files are identical to the same files without the ss suffix but use data sets created by teavar1ss.do rather than teavar1.do.

3. compch.do compchr.do - These files combine the teacher/student data sets in order to run analyses of variance for math and reading achievement over a sample of schools with a single teacher in the relevant subject.

Department of Economics, University of Illinois at Chicago, Chicago, IL 60607;

sgrivkin@uic.edu,

Hoover Institution, Stanford University, Stanford, CA 94305, U.S.A.; hanushek@stanford.edu;

http://www.hanushek.net,

and

University of Texas at Dallas (deceased).