Spurious Relationships from Aggregate Variables in Linear Regression Models

©David J. Armor, Chenna Reddy Cotla, and Thomas Stratmann George Mason University

Abstract

Linear regressions that use aggregated values from a group variable such as a school or a neighborhood are commonplace in the social sciences. This paper uses Monte Carlo methods to demonstrate that aggregated variables produce spurious relationships with other dependent and independent variables in a model even when there are no underlying relationships among those variables. The size of the spurious relationships (or postulated effects) increases as the number of observations per group decreases. A necessary and sufficient condition for eliminating spurious effects is to include the individual version of the aggregated variables in the regression. These results suggest caution when using and interpreting the effects of aggregate variables when the individual versions of those variables are not available.

Introduction

Over the years, aggregate variables in the social sciences have generated numerous methodological discussions regarding their appropriate use and interpretation. One of the earliest methodological issues was the "ecological fallacy" problem, which raised the question of whether correlations among aggregate variables can imply correlations among the underlying individual variables. This issue is of particular concern to political scientists who use aggregate voting data (e.g., King, 1997). A more recent methodological issue is constructing models using both individual and aggregated variables, which raises questions about proper estimation of standard errors as well as the possibility of varying individual-level coefficients within the groups used for the aggregated variable. These issues are addressed in various ways by different disciplines, such as random or mixed effects models in economics (Moulton, 1990; Wooldridge, 2003) and hierarchical (or multi-level) models in the behavioral sciences (Bryk and Raudenbush, 1992).

Conceptually, an aggregate variable is generated to measure a "contextual" effects of some characteristic, such as the demographic composition of a school, or a neighborhood, or of a larger geographic unit such as a city or county. One widely studied contextual effect is the racial composition of a school or a classroom, as for example the effect of racial composition of a school on achievement (Hanushek, Kane, and Rivkin, 2009). Another widely studied contextual effect is the racial composition of neighborhoods. A classic example here the effect of a neighborhood's racial composition on crime and related behaviors (Sampson, Raudenbush, and Earls, 1997).

One of the most common aggregate variables is the socioeconomic composition of a group, such as a neighborhood, a school, or a classroom. In social science studies of academic achievement, as measured by test scores, it is well-established that test scores are strongly influenced by the socioeconomic status (SES) of individual students, as measured by their parents occupation, income, and education (e.g., Loveless, 2012). The higher a student's SES, the higher his or her test scores. Scholars frequently measure SES by a single index comprised of all three of these measures, usually equally weighted.

Likewise, many studies have demonstrated significant contextual effects of school or classroom SES on student test scores. School or classroom SES is obtained by averaging individual student SES for all students in a school or classroom. A good example is a study by Willms (2010) who uses achievement test data from 57 countries. The potential influence of school SES on test scores, which has been replicated in many cross-sectional studies, has provided a major rationale for school economic integration, whereby lower SES students are allocated across schools with the aim of attaining a certain degree of economic balance within each schools (Kahlenberg, 2012).

There has been little or no discussion in the statistics or social science literature of another potential methodological problem, which is the possibility that an aggregated measure of a given attribute has a spurious

relationship with the individual measure of that attribute. Nor is there a discussion that that this spurious relationship gives rise to an additional spurious relationship between the aggregate measure and a third (dependent) variable such as academic achievement. By spurious, we mean that the correlation arises strictly by chance, even though there is no underlying substantive relationship between the aggregate variable and the dependent variable.

Monte Carlo methods are used to demonstrate that significant spurious correlations and regression coefficients arise among aggregate variables and their individual level measures. This phenomenon is demonstrated first using hypothetical data on school SES, individual SES, and achievement test scores,. The phenomenon is then illustrated using real data from two different national data sets These spurious effects are caused by the spurious correlations between individual and aggregated SES and individual and aggregated race. The magnitudes of these spurious effects depend both on the number of schools and the average size of schools (students per school). The paper also demonstrates that the necessary and sufficient conditions for eliminating the spurious effects of aggregate variables is to include their corresponding individual measures in the models.

Methods and Data

This paper examines the linear relationships among several variables. While the dependent variable in the examples and simulations is student achievement as measured by student test scores, the results generalize to other types of dependent variables measured on individuals. The dependent variable is designated as A, and students (individual cases) are indexed by i. One independent variable is individual student socioeconomic status (SES), designated I. Another independent variable will be the aggregate version of SES, defined as the average of I in each school, and it is designated by S. A third independent variable will be an individual student's race B (black vs. nonblack), and the final independent variable the aggregate variable of school percentage black, or P. The number of students is N, the number of schools is m, and the number of students in each school will be a constant n, so that N=m x n. The simulations do not allow the number of students to differ for each school, although small differences should not alter the basic results found here.

All variables are standardized with means of zero and unit variances. Correlations will be represented by r's and regression coefficients will be represented as β's, and they will have subscripts indicating the variables involved. Correlations will be indicated by subscripts for the two variables involved. For example, r_{AS} indicates the correlation between A and S. Regression coefficients will be indicated by β and the subscript indicating the independent variable. So β_I denotes the regression coefficient for the effect of I on A controlling for other variables (A is always the dependent variable of interest).

Monte Carlo simulation methodology is used to allocate N students randomly to different numbers of schools, which can vary from 5 to 50, and then to calculate the correlations (and regression coefficients) that arise among all the variables in a given model. Since the students are allocated randomly, there is no intrinsic reason for a correlation to arise between any two variables. The approach involves calculating the distributions of correlations across the dependent and independent variables that arise by chance in a given dataset. Ideally, correlations across variables would be calculated for all possible allocations of students to schools in a dataset, but this is computationally infeasible except for very small datasets with less than 25 students.

For example, even if the number of students is 30, the number of all possible allocations of students across 3 equally-sized schools would imply enumerating $30C10 \times 20C10 = 5{,}550{,}996{,}791{,}340$ allocations, and computing correlations across variables for each of those 5.5 trillion allocations. The number of possible allocations rises exponentially with the number of students, and the computation time to enumerate exact distributions of correlations becomes prohibitive. The number of students in most achievement studies are in the thousands, thereby ruling out the calculation of exact distributions.

Instead, Monte Carlo simulation methodology is used to approximate the exact distribution of all possible correlations across variables. The distributions of correlations across variables are estimated by considering 100, 000 Monte Carlo samples. Each Monte Carlo sample here represents one random allocation of students across schools, given a particular number of students and a particular number of equally-sized schools. The correlations presented in the results are the mean correlations from the sample of 100,000 correlations for a given number of students and number of schools; that is, they are the expected value of those correlations. If there is no spurious relationship between any two variables, the expected value (or mean) of the correlations from the Monte Carlo samples--for a given number of students and schools--should be 0.

To test the generality of results across different types of data, three datasets are used. The first is a small hypothetical dataset that examines correlations and regressions among A, S, and I. The second dataset is achievement and socioeconomic status data from the 2009 Program for International Student Assessment (PISA) study for the United States sample. In this data set, only the correlations among A, S, and I, (and related regression coefficients) are calculated. The third dataset uses the 2003 U.S. National Assessment of Educational Progress (NAEP) for a single state to conduct simulations for the various correlations and regression coefficients for A, S, P, I, and B.

In the first hypothetical dataset, the values for student achievement (A) and individual student SES (I) were created for 60 students so that the correlation between A and I is approximately .5, a magnitude found in many real datasets. For this dataset, correlations among all variables are calculated for eight different groupings of students in hypothetical schools. The first case is 3 schools with 20 students per school, then 4 schools with 15 students each, then the combinations of 5 and 12, 6 and 10, 10 and 6, 12 and 5, 15 and 4, and finally 20 schools with 3 students per school. For each case, 100,000 random assignments are generated by randomly allocating students across the number of schools given for that particular case. For simplicity and computational tractability, the number of students per school is held constant in each random assignment. For each of these random assignments, S is computed first by averaging I across each school, and then the correlations among S, I, and A are calculated. This step is repeated for each of the 100,000 iterations, and then these 100,000 correlations are averaged to obtain an expected value for each correlation. This process is repeated for each combination of students and schools.

For the second set of simulations, descriptive statistics show that the PISA math achievement scores for the U.S. are approximately normally distributed with a mean of 485 and sd of 84. The sample size is approximately 5,000. For computational simplicity the I variable was standardized to have a mean of 0 and a standard deviation of 1. An artificial dataset was constructed by drawing a random sample of students from the PISA dataset for each combination of the number of schools and the number of students. All possible combinations of 5, 10, 20, 30, 40, and 50 schools and 5, 10, 20, 30, 40, and 50 students per school were used in the simulations. In this Monte Carlo exercise, the smallest N is 25 (5×5) and the largest is 2500 (50×50). In other words, the size of the total population varies from 25 to 2500, and the total number of schools varies from 5 to 50.

The Monte Carlo simulations procedure is illustrated for the case of 20 schools and 200 students, or 10 students per school. First, a random sample of 200 students is drawn from the PISA dataset, which provides actual values for the A and I variables. This sample is used to construct 100,000 random allocations of 10 students to each of the 20 schools, and for each allocation the S variable is calculated for each school, and then correlations are computed between A, I, and S. The 100, 000 Monte Carlo samples are utilized to approximate the distributions of correlations that occur due to chance. These approximate distributions are then used to assess statistical significance of the average correlation across variables. The approximate distribution of the correlation between any given two variables is used to test if the mean correlation between those variables is significantly different from zero and to compute corresponding simulated probability value.

The third set of simulations uses the U.S. NAEP data for 8th graders in one state and proceeds in a fashion similar to that described for the PISA data, using the NAEP math score as the dependent variable A. The main difference is that two aggregate variables S and P are constructed from their corresponding individual values I and B. As for the PISA data, the A, I, and B variables are based on real data, while S and P are based on arbitrary (random) allocations of individual students to hypothetical schools. The same combinations of number of students and number schools are used. The main difference is that the regression models for A now have two individual independent variables I and B and two aggregate independent variables S and P.

Returning to the PISA data, a fourth example introduces another aggregate variable, average student achievement in a school (D), as an additional independent variable. Some economists have used average classroom achievement as a predictor of individual achievement, contrasting its effect with the effects of clasroom SES and racial composition (Hoxby & Weingrath, 2005; Vigdor and Nechyba, 2004).[1] This model uses I, S, and D as independent variables and A as the dependent variable.

Simulation runs using Hypothetical Data

The hypothetical data in introduced to illustrate how the Monte Carlo simulations work. Simulations are constructed for all 60 students allocated randomly to various combinations of number of schools and number of students per school, from 3 schools with 20 students per school to 20 schools with 3 students per school. Since all 60 students are used in each case, the correlation between A and I is a constant. In this case the correlation is equal to .48, and it is statistically significant.

In this simulation excercise, there are two regression models of interest, as represented in equations 1a and 1b. All models reflect standardized variables with mean zero and unit variance. The first hypothesized model is

(1a) $A = \beta'_S S + e$ (standardized; $\beta'_S = r_{AS}$)

indicating that achievement is caused only by the aggregate school SES, and there is no effect of I. This is a model where causation arises entirely from the contextual (aggregate) variable, which may be unrealistic but is offered here to illustrate a causal model using only the aggregated variable. Since the variables are standardized, the coefficient β'_S is just the correlation between A and S.

The second and more realistic model is

(1b) $A = \beta_S S + \beta_I I + e$ (standardized)

where achievement scores are caused by both school SES and individual SES. Here the coefficient β_S is the effect of S on A controlling for I, and β_I is the effect of I on A controlling for S.

Table 1, columns (1) to (3), show the simulation results for the expected correlations among A, I and S for the various allocations of students to a given number of schools. Columns (4) and (5) show the regression coefficients for model (1b). The entries represent means (or expected values) for 100,000 correlations among A, I and S, as well as the means (or expected values) for 100,000 regression coefficients β_S and β_I. Since S is a constructed variable for each case, from a random allocation of students, it has no intrinsic meaning. That is, for each draw, students are randomly assigned to a particular school, and then S is calculated. Thus, the values for S will be different for each (randomly generated) subsequent draw.

---

[1] These studies use more sophisticated time-series models, compared to the cross-sectional models evaluated here, but the general issue of spurious relationship still applies.

Since the constructed S values are completely arbitrary, any correlation that arises between S and another variable represents a "spurious" relationship. It is simply the expected value of correlations between S and the other variable for all combinations of the 60 individual SES (I) values allocated to a given number of schools.

Column (1) is the expected correlation between the individual student measures of A and I, and since all 60 students are used in each draw, the expected value of the correlation is a constant and is equal to approximately .48. It is also statistically significant at p<.001 for all cases as determined by the number of schools and students per school.

Column (2) shows the expected values of the spurious correlation between I and S, which is also the coefficient β'_S in model (1a). All of these correlations are also statistically significant at p<.001. Using these hypothetical data, the spurious correlation between I and S rises in a nonlinear fashion, starting at .16 for the case 3 schools, 20 students per school and increasing with each additional school until it reaches .56 for the case of 20 schools, 3 students per school. The rate of increase starts at about .04 but slows to less than .02 when the number of schools gets to about 10 or higher.

Column (3) shows the expected value of the correlations between A and S, which are also the coefficients in the simple regression shown in equation (1a). These spurious correlations are smaller than the I & S correlations, and they do not reach statistical significance until there are at least 15 schools with 4 students per school. In that case, the correlation is .23, and the correlation for the next case of 20 schools, 3 students per school, is .27, which is also statistically significant. These are moderate correlations which would lead one to confirm model (1a), that S has a significant effect on A, assuming that one did not have measures of I available to estimate model (1b). Of course, this is due mainly to the fact that a relaively small sample of students is being used in this hypothetical example.

Table 1 Expected Values of Correlations and Regression Coefficients(β) for Test Scores (A) , Individual SES (I), and School SES (S): Hypothetical Data for 60 Students

Number of Schools	Correlations:			Reg. Coefficients	
(Students	(1)	(2)	(3)	(4)	(5)
per school)	r_{AI}	r_{IS}	$r_{AS} = \beta'_S$	β_I	β_S
3 (20)	.4823***	.1637***	.0799	.4821	.0010
4 (15)	.4823***	.2085***	.1011	.4823	.0006
5 (12)	.4823***	.2457***	.1187	.4823	.0003
6 (10)	.4823***	.278***	.1339	.4823	-.0001
10 (6)	.4823***	.3812***	.1841	.4822	.0003
15 (4)	.4823***	.4804***	.2316*	.4823	-.0001
20 (3)	.4823***	.5627***	.2715**	.4821	.0002

* Correlations differ from 0 at p<.05; ** at p<.01; *** at p<.001

Columns 4 and 5 show results from a linear regression of individual level achievement A on individual level socioeconomic status I and school socioeconomic status S, which tests model (1b). The coefficients for

the effect of I on A, controlling for S, are in column 4, and they are all very close to .48, which is the orginal correlation between A and I. The coefficients for S on A, controlling for I, are all virtually zero (to the third decimal place). In other words, by estimating a regression using both the constructed variable S and the individual measure of socioeconomic status I, the spurious effect of S on A is eliminated.

The explicit causal assumptions between A, I, and S can be represented in a simple path model, as shown in Figure 1. The correlations for the case of 20 schools and 3 students per school have been placed on the causal arrows reflecting the spurious relationships for r_{IS} and r_{AS}. In this graph β_I and β_S are are the estimated regression coefficients from model (1). The A-I correlation is always a constant as explained earlier. But the I-S and A-S correlations (.56 and .27, respectively) are both spurious, because they are simply the expected values generated by constucting all possible allocations of 60 students to 20 schools with 3 students per school.

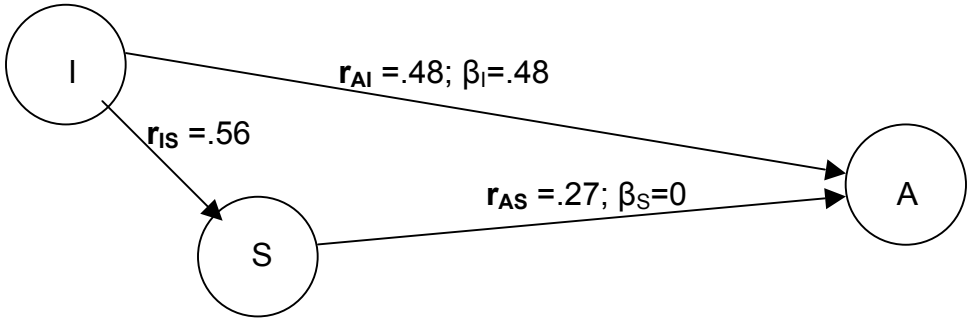

Figure 1 Path Model for A, I, and S showing bivariate correlations and path (beta) coefficients for the case of 20 schools and 3 students per school

The results from estimating the multiple regression (1b) using the correlations shown in Figure 1 (the expected values from the simulation) yields the standardized coefficients $\beta_I = .48$ and $\beta_S = 0$, for the case of 20 schools and 3 students per school. In other words, applying the causal model in Figure 1 eliminates the moderate but spurious relationship between school SES and achievement, showing that S has no effect on A ($\beta_S = 0$) once one controls for I in the regression model. The effect of I on A, controlling for S, is just the original correlation of .48. The same results occur for all combinations of number of schools and number of students per school. The central conclusion here is that the spurious relationships between the aggregate variable S and A is eliminated for all combinations of number of schools and number of students per school, provided that regression is used to estimate the S effect on A controlling for I.

Simulations Using Real Data: The PISA Data

The second set of simulations was carried out using real data, in this case the 2009 PISA math test scores for the United States. The use of real data eliminates any potential distortions stemming from distributional issues in the hypothetical A and I scores, which follow more of a rectangular than a normal distribution.

Because of the results for the hypothetical dataset, it is anticipated that the sizes of the spurious correlations depend on both the size of the school (the number of students per school) and also the total number of schools, with the former having a stronger impact. Accordingly, simulations using all combinations from 5, 10, 20, 30, 40, and 50 schools and the same variation in the numbers of students per school, from 5 to 50, in

increments of five.[2] That is, the smallest simulation sample 25 students with 5 schools and 5 students per school, and the largest was 2500 students with 50 schools and 50 students per school.

In addition, to test the generality of results, simulations were run for two scenarios: (1) two independent variables, I and the constructed S variable (representing school SES); and (2) three independent variables, adding a invidual-level covariate which was a self-rating of a particular learning strategy.[3] In the results below the additional covariate is identified as C.

Scenario 1: Two Independent Variables

Figures 2 and 3 show plots for two of the mean correlations from the simulation runs for the first scenario of one aggregate independent variable S and its individual values I; the correlations plotted in these figures are presented in Table 2.

Figure 2 plots the correlations between the constructed S variable and the I variable, all of which were statistically greater than 0 at $p<.001$. Holding the number of schools constant, the relationship between school size and the school SES-individual SES correlation is approximately inverse quadratic. The spurious correlation decreases but is still modest and statistically significant at about .15 for school sizes of about 40. It is interesting to observe that the average number of students per school in the actual PISA USA data is approximately 34, implying that this international database gives rise to sizeable spurious correlations between S and I. Note that the total number of schools makes a diminishing difference after reaching about 30, regardless of the number of student per school.

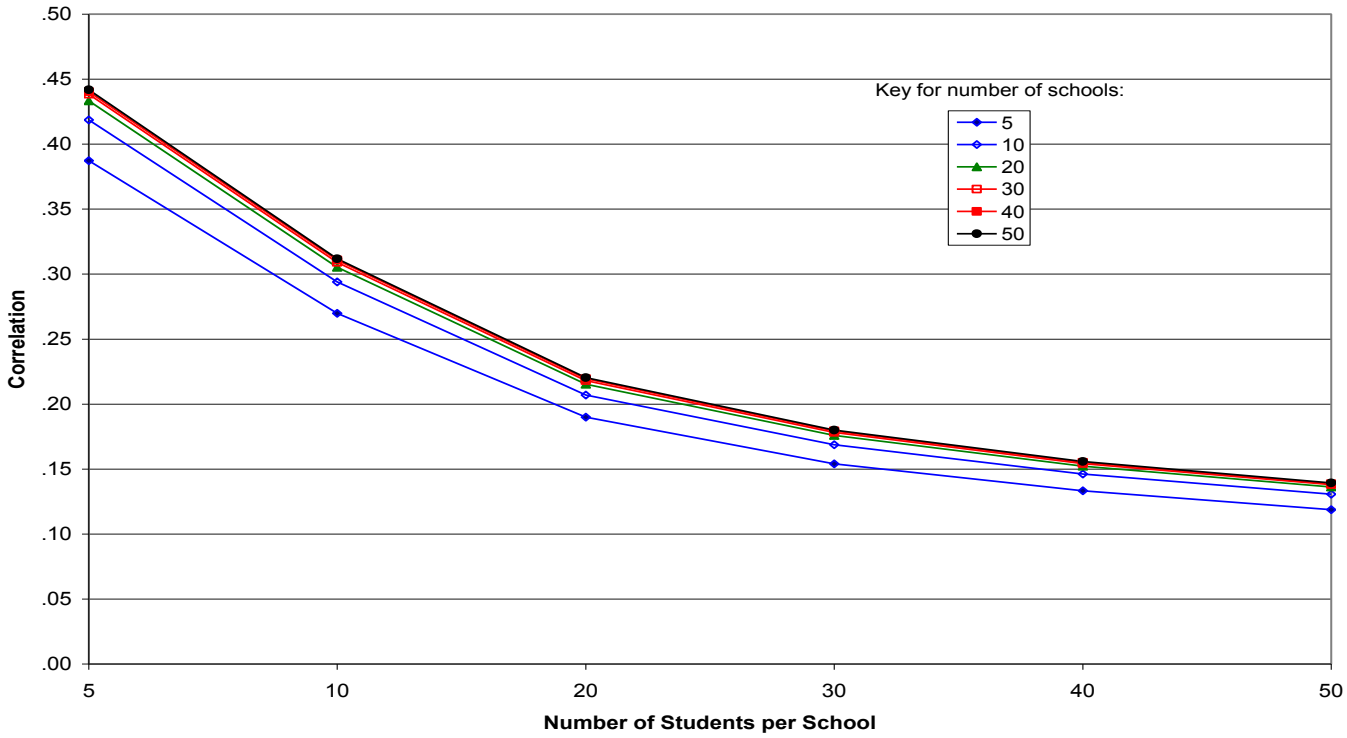

Figure 2 Correlation between School SES (S) and Individual SES (I) by School Size & Number of Schools *(see Table 2)*

[2] Simulations were also run for 60 and 80 schools, but results differed only slightly from the 50 school case.

[3] In the dataset, the variable was named "metasum."

The size of the spurious correlations between achievement and school SES are generally smaller, as shown in Figure 3, and they have more variability, particularly after the number of students per school reaches 10 and lower. Their statistical significance also varies (see Table 2). None of the correlations are statistically significant for the case of 5 or 10 schools, regardless of number of students per school. In contrast, all of the correlations are statistically significant for 30 or more schools and all variations in number of students per school. For the case of 20 schools, statistical significance varies.

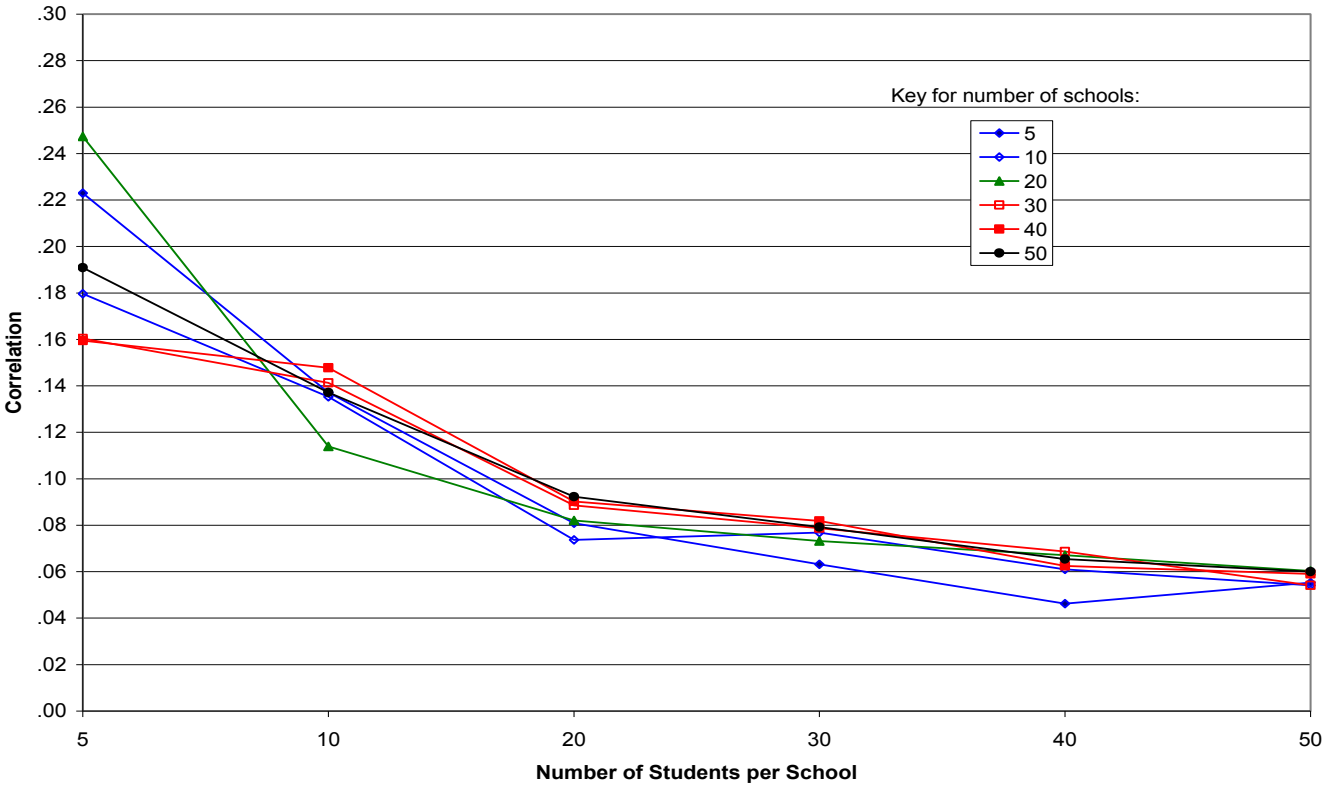

Figure 3 Correlation between School SES and Achievement by School Size and Number of Schools (*see Table 2)*

Table 2 displays the mean correlations, regression coefficients, and significance tests for all variations in the number of schools and number of students per school. In these simulations, since not all students are sampled in each of the 100,000 draws (the largest sample is 2500), the actual A-I correlations vary. The average A-I correlations appear in the first panel of the table. These are not spurious correlations because they correspond to the actual correlations between individual SES and math achievement for a given draw. Although they vary between .58 and .36, most lie in a narrower range near the actual value of the correlation in the full data set, which is .43. All of the A-I correlations are statistically significant at $p<.001$.

The second panel shows the expected values of the A-S correlations, and this set of correlations shows the greatest variation in statistical significance. These are all spurious correlations, because they represent the "constructed" S variable; that is, they are the expected values of an arbitrary creation of S values for a given number of schools and a random allocation of individual SES (I) scores to each of those schools. None of the spurious correlations in the rows for 5 and 10 schools are statistically significant, but all of the correlations are significant in the rows for 30, 40 and 50 schools. In the row for 20 schools, the case of 5 students per school

has a mean correlation of .25 and is significant at p<.01. The only other significant correlations are in the cases of 40 or 50 students per school, where the spurious correlations are .07 and .06, respectively.

Table 2 Expected values of Correlations & Betas among Math Scores, Student SES, and School SES

	Number of	Students per School				
No. of Schools	5	10	20	30	40	50
Student SES-Ach Correlations r_{AI}	(all ***)					
5	.5764	.5112	.4268	.4105	.3447	.4652
10	.4286	.4605	.3552	.4536	.4174	.4140
20	.5700	.3731	.3809	.4172	.4404	.4423
30	.3666	.4562	**.4070**	.4415	.4445	.3927
40	.3627	.4749	.4107	.4568	.4028	.4257
50	.4322	.4398	.4183	.4404	.4200	.4305
School SES-Ach Correlations r_{AS}						
(all p>.05) 5	.2230	.1372	.0809	.0632	.0462	.0553
(all p>.05) 10	.1797	.1353	.0737	.0768	.0610	.0541
20	.2474**	.1139	.0821	.0732	.0671*	.0603*
(all p<.03) 30	.1604	.1413	**.0886**	.0787	.0687	.0541
(all p<.02) 40	.1595	.1478	.0902	.0819	.0625	.0590
(all p<.001) 50	.1909	.1372	.0923	.0792	.0654	.0600
Student SES-School SES Correlations r_{IS}	(all ***)					
5	.3873	.2698	.1899	.1541	.1333	.1188
10	.4186	.2940	.2070	.1688	.1462	.1306
20	.4333	.3053	.2153	.1759	.1523	.1362
30	.4384	.3090	**.2181**	.1782	.1542	.1380
40	.4403	.3108	.2196	.1794	.1552	.1387
50	.4419	.3118	.2203	.1800	.1558	.1393
Standardized Coefficients for β_I	(all ***)					
5	.5766	.5114	.4268	.4105	.3446	.4652
10	.4285	.4605	.3552	.4535	.4174	.4140
20	.5698	.3731	.3808	.4172	.4404	.4423
30	.3668	.4561	**.4071**	.4415	.4445	.3927
40	.3628	.4748	.4108	.4568	.4028	.4257
50	.4322	.4398	.4183	.4404	.4200	.4305
Standardized Coefficients for β_S	(all p>.999)					
5	-.0003	-.0008	-.0002	-.0001	.0002	.0001
10	.0003	-.0002	.0001	.0002	.0000	.0001
20	.0005	.0000	.0000	-.0002	.0000	.0001
30	-.0004	.0004	**-.0002**	.0000	.0001	-.0001
40	-.0002	.0002	.0000	.0000	-.0001	.0000
50	-.0002	.0001	.0001	.0000	.0000	.0001

* Correlations differ from 0 at p<.05; ** at p<.01; *** at p<.001

The third panel of I-S correlations are all statistically significant, and they are fairly sizeable for the cases of 5 and 10 students per school (over .4 and about .3, respectively). Note that after about 20 schools, there is very little variation as more schools are added. The cases of 30 students per school, which is similar to the number of students per school drawn in many national surveys including the PISA USA survey, yields spurious

correlations on the order of .18 or greater after the number of schools reach 20. In the full PISA data set, the actual I-S correlation is .55, and the actual A-S correlation is .42. The actual I-S correlation of .52 does not seem so large if it is understood that for samples of this size, a random allocation of students to schools produced an expected spurious correlation of nearly .2

The mathematical models for this data are the same as (1a) and (1b), and the second panel corresponds to the coefficients for model (1a) if one hypothesized that only school SES had an effect on achievement. If one postulates model (1b) instead, then the bottom two panels of Table 2 show the regression coefficients of student SES (β_I) and school SES (β_S) on math scores for all combinations of the number of schools and students per school. All of the β_I coefficients are significant at $p<.001$, and they are virtually identical to the original A-I correlations (to the third decimal place). In contrast, all of the β_S coefficients are equal to zero to the third decimal place. In other words, the spurious correlation between school SES and math scores is eliminated by controlling for individual SES.

Figure 4 shows a path model constructed for the case of 30 schools and 20 students per school (N equals 600 students). This case generates a statistically significant spurious correlation of .22 between I and S, and a statistically significant spurious correlations of .09 between S and A. The correlations of .41 between I and A is the actual average correlation for simulation draws. The beta (or path) coefficients for the regression return .41 and 0 for the effects of I and S on A.

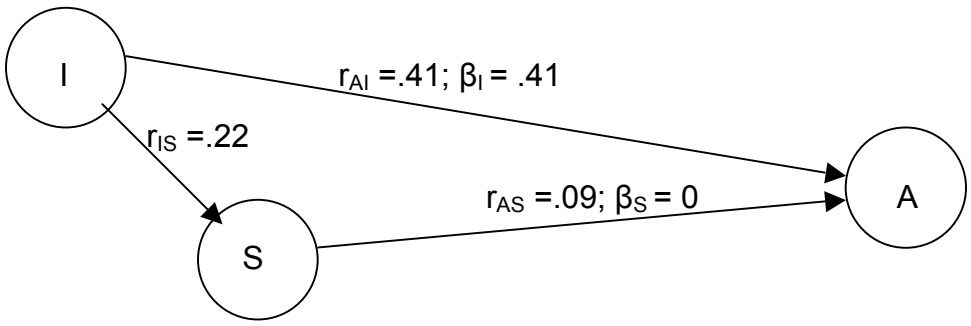

Figure 4 Path Model for A, I, and S Using PISA Data for Case of 30 Schools and 20 Students per School

Similar results occur for all of the possible regressions run on the average correlations generated for the simulations using differing number of schools and students per school. Although this paper does not develop a mathematical proof, as long as the effects of S on A are estimated after controlling for I, the resulting coefficients are zero. In other words, controlling for the individual-level variable eliminates the spurious effect of the aggregate variable in model (1a). One could also say it eliminates the spurious correlations between I and S, and A and S, in that those correlations come from random allocations of students to schools and thus do not reveal substantive relationships.

Scenario 2: Adding a Covariate

In this scenario an individual-level covariate is added to the model to test whether a more complex causal model has any impact on the regression results given the spurious correlations generated by the constructed S scores. Equation (2) shows the mathematical model assumed for the case of three independent variables:

(2) $A = \beta_I I + \beta_S S + \beta_C C + e$ (standardized)

Table 3 presents the correlations among C and A, C and I, and C and S. The same random starting seed used in Scenario 1 was also used for the Scenario 2 simulation runs, so the correlations among A, I, and S are virtually identical to those shown in Table 2; there are some minor variations due to rounding precision.

The C-A and C-I correlations represent estimated relationships among these individual-level variables, as influenced by the various sample sizes. The actual correlations in the full dataset are .32 and .14, respectively, and mean correlations are very close to these values when both the number of schools and the number of students per school reach 50 (N = 2500 students). These correlations are statistically significant for all variations of number of schools and number of students per school.

Table 3 Expected Values of Correlations among Three Independent Variables

	No. of Students per School					
No. of Schools	5	10	20	30	40	50
C-A CORRELATIONS (all ***)						
5	.2407	.3291	.1827	.4616	.3784	.3288
10	.1958	.4096	.3074	.3216	.3006	.3832
20	.3220	.3848	.3607	.4346	.3577	.3026
30	.3012	.2615	**.3551**	.3209	.3366	.3264
40	.2015	.4089	.3154	.3092	.3497	.3367
50	.2272	.3228	.3188	.3237	.3347	.3200
C-I CORRELATIONS (all ***)						
5	.2054	.3854	-.0010	.1313	.0595	.1476
10	-.0093	.3003	.1159	.1291	.1323	.1628
20	.1886	.0569	.1773	.1703	.1637	.1192
30	-.0346	.2365	**.1594**	.1780	.1427	.1636
40	.0342	.1680	.1669	.1408	.1427	.1382
50	.0876	.0928	.1268	.1309	.1577	.1444
C-S CORRELATIONS (all p>.15)						
5	.0792	.1043	-.0005	.0205	.0082	.0173
10	-.0044	.0882	.0242	.0215	.0194	.0213
20	.0813	.0174	.0382	.0303	.0250	.0164
30	-.0155	.0729	**.0348**	.0316	.0220	.0224
40	.0151	.0523	.0365	.0253	.0223	.0193
50	.0388	.0289	.0279	.0234	.0245	.0201

* Correlations differ from 0 at p<.05; ** at p<.01; *** at p<.001

The spurious correlations between C and the constructed S variable are generally small, and none of them are statistically significant. The largest is .09, which occurs for the case of 10 schools and 10 students per school (N=100). But this spurious correlation is not statistically significant. The actual C-S correlation in the full USA PISA data set is .13, which is quite modest. This is good news, because it eliminates the possibility of a significant indirect effect of S on A operating through C.

The regression coefficients for equation (2) are presented in Table 4. The standardized coefficients for the constructed S variable on achievement are zero for all iterations of number of schools & students, just as they were in the two variable case. So, the spurious correlation between I and S generates a null coefficient when A is regressed on I, S, and C; adding an individual-level covariate does not change the result for the

constructed variable. Again, the only way to eliminate the spurious correlation between I and A is to conduct a multiple regression analysis, with both S and I as independent variables.

Table 4 Expected Values of Standardized Coefficients for Model (2)

No. of Schools		Number of Students per School 5	10	20	30	40	50
STANDARDIZED COEFFICIENTS FOR I							
(all ***)	5	.5505	.4516	.4269	.3560	.3233	.4259
	10	.4302	.3709	.3240	.4191	.3844	.3612
	20	.5283	.3524	.3272	.3534	.3924	.4121
	30	.3775	.4177	**.3596**	.3970	.4047	.3486
	40	.3562	.4180	.3683	.4216	.3602	.3866
	50	.4153	.4135	.3841	.4050	.3766	.3924
STANDARDIZED COEFFICIENTS FOR S							
(all p>.999)	5	-.0007	-.0006	.0002	.0003	.0001	-.0001
	10	.0007	.0000	-.0003	-.0002	-.0003	.0001
	20	-.0005	-.0001	-.0001	-.0001	.0000	-.0001
	30	-.0001	.0000	**.0000**	.0001	.0001	.0000
	40	.0002	.0000	.0002	.0000	.0000	-.0001
	50	.0003	-.0001	.0000	-.0001	-.0001	.0000
STANDARDIZED COEFFICIENTS FOR C							
(all*** except 5/5 at p<.02)	5	.1277	.1551	.1832	.4149	.3591	.2660
	10	.1998	.2982	.2698	.2675	.2498	.3244
	20	.2224	.3647	.3027	.3744	.2934	.2535
	30	.3143	.1627	**.2978**	.2502	.2789	.2693
	40	.1893	.3387	.2539	.2498	.2982	.2833
	50	.1908	.2844	.2701	.2707	.2753	.2634

In contrast, the I and C coefficients are not equal to their correlations with A, as was the case of I when it was the only independent variable. Because both of these variables are correlated with A, and also with each other (albeit modestly; see the fifth panel in Table 3), their regression coefficients are reduced somewhat from the correlations. Considering the case of 30 schools and 20 students per school, the I coefficient β_I is .36 versus its correlation of .42, and likewise the C coefficient β_C is .30 versus its correlation of .36. All of the coefficients for I and C are statistically significant at $p<.001$ (with one minor exception).

The causal assumptions of this four variable model can also be represented in a path model, which is shown in Figure 5. The correlations and regression coefficients are represented for the case of 30 schools and 20 students per school, as was done for Figure 4. C is also an endogenous (dependent) variable, being influenced by both I and S, so a separate regression has to be run to calculate the coefficients, and we denote these coefficients with an asterisk (*). In this case, since there are only two independent variables, one of which is the constructed S variable, the regression coefficient for S (on C) is zero and the coefficient for I (on C) is the same as its bivariate correlation with C.

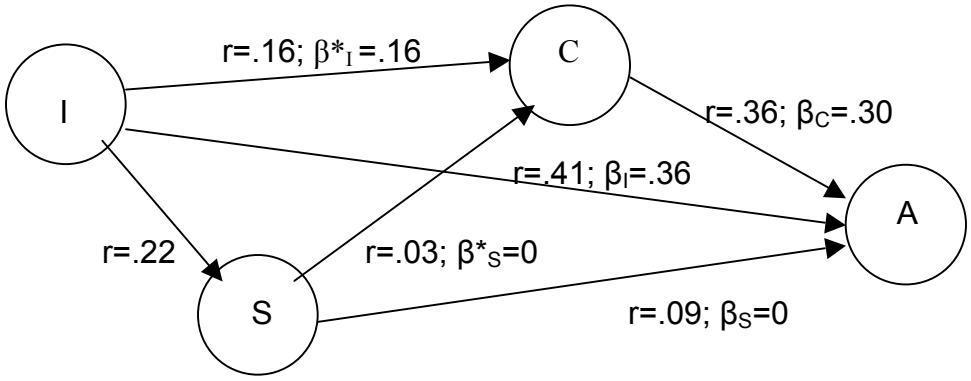

Figure 5 Path Model for Thre Independent Variables for Case of 30 Schools and 20 Students per School
(Note: β's are coefficients for regression of C on I and S)*

Simulations using the NAEP Data

The third set of simulations uses 2003 NAEP achievement data to illustrate a situation of multiple aggregate variables, in this case school SES (S) and school percent black (P). Aside from testing the generality of results when a different set of real data are used, these simulation tests whether significant spurious correlations occur when multiple aggregate variables are used, and also test for the conditions under which these spurious correlations can be eliminated in a linear regression.

The dataset includes 8th grade math scores for about 4,000 students in a single state. The first model considered assumes that A is influenced only by the aggregate variables S and P, and then a second model is estimated where A is influenced by I, S, B, and P. In the full dataset, the correlation between S and A is .37 and the correlation between P and A is -.35. The correlation between S and P is just below .5. The average number of students per school is about 30.

Model (3): Two Aggregate Independent Variables

Table 5 shows the expected values correlations and standardized coefficients for a model where A is a linear function of the aggregate variables, school SES and school percent black, as defined by equation (3).

(3) $A = \beta'_S S + \beta'_P P + e$ (standardized coefficients)

The pattern of the spurious correlations between S and A in the NAEP data is very similar to that observed for the S-A correlations in the PISA data. The correlations increase sharply as the number of students per school declines, reaching highs of in the range of .2 to .25 for 5 students per school--and similar values across different numbers of schools. At the other end, the correlations decline and flatten out at about 50 students per school. The pattern for the P-A correlations is very similar, although the sign of the correlations are negative and the correlations are very slightly weaker (by about .01). In contrast, the number of schools has very little impact on the level of the correlation, but it does affect significance tests. The spurious correlations are all statistically significant after the number of schools reach 20 or 30, regardless of the number of students per school.

Table 5 Expected Values of Correlations & Betas among Math Scores, School SES, and School Percent Black

		Number of Students per School					
No. of Schools		5	10	20	30	40	50
S-A CORRELATIONS							
	5	.2422	.1295	.0849	.0571	.0664	.0523
	10	.1680	.1631	.1061	.0868	.0675	.0616
(all p<.05)	20	.2187	.1525	.1095	.0823	.0731	.0727
(all p<.01)	30	.2294	.1475	.1023	.0875	.0742	.0691
(all p<.01)	40	.2021	.1572	.1062	.0809	.0786	.0696
(all p<.001)	50	.2561	.1586	.1101	.0880	.0740	.0679
P-A CORRELATIONS							
	5	-.2116	-.1136	-.0740	-.0605	-.0460	-.0468
	10	-.1957	-.1308	-.0882	-.0762	-.0687	-.0550
	20	-.1822	-.1320	-.0922	-.0731	-.0671	-.0569
(all p<.05)	30	-.2107	-.1305	-.0957	-.0810	-.0634	-.0594
(all p<.01)	40	-.2359	-.1336	-.0809	-.0790	-.0690	-.0612
(all p<.01)	50	-.1844	-.1482	-.1007	-.0792	-.0672	-.0617
S-P CORRELATIONS							
	5	-.3568	-.1485	-.0600	-.1389	-.1451	-.2629
	10	-.2449	-.3653	-.2431	-.2257	-.2138	-.2364
	20	-.3422	-.3519	-.1767	-.2291	-.2101	-.2445
	30	-.2256	-.1704	-.2645	-.2646	-.2506	-.2365
	40	-.2289	-.2464	-.1789	-.2200	-.2482	-.2471
(all p≤ .1)	50	-.2226	-.3401	-.2186	-.2175	-.2281	-.2563
STANDARDIZED COEFFICIENTS FOR S							
	5	.1885	.1126	.0801	.0489	.0606	.0422
	10	.1261	.1324	.0893	.0728	.0548	.0511
(all p<.05)	20	.1766	.1207	.0959	.0689	.0615	.0624
(all p<.05)	30	.1912	.1288	.0825	.0709	.0621	.0582
(all p<.01)	40	.1559	.1321	.0947	.0666	.0654	.0579
(all p<.01)	50	.2261	.1222	.0924	.0742	.0619	.0556
STANDARDIZED COEFFICIENTS FOR P							
	5	-.1359	-.0944	-.0688	-.0527	-.0363	-.0343
	10	-.1629	-.0804	-.0654	-.0589	-.0564	-.0422
	20	-.1206	-.0886	-.0748	-.0569	-.0539	-.0413
(all p<.05)	30	-.1670	-.1082	-.0736	-.0619	-.0476	-.0454
(all p<.01)	40	-.1998	-.1008	-.0638	-.0642	-.0526	-.0467
(all p<.01)	50	-.1337	-.1063	-.0803	-.0629	-.0530	-.0474

The standardized coefficients reveal a pattern similar to the correlations, with values increasing in magnitude as the number of students per school declines. For the coefficients, however, the flattening out starts when the number of students per school reaches 30 for the S effect and reaches 40 for the P effect. Likewise, the number of schools has little effect on magnitudes but does affect significance levels. All S coefficients are statistically significant (at $p < .05$) after the number of schools reach 20, even when the number of students per

school is only 5 (which means N = 100). For the P coefficient, the $p < .05$ significance starts when the number of schools reaches about 30 (N = 150). These coefficients are statistically significant despite the facts that the coefficients are small and that the sample sizes are also relatively small.

Regarding the magnitude of the significant but spurious regression coefficients, they are larger for S than for P, likely because the original correlations between S and A are larger than those between P and A. The S coefficients are near .2 for 5 students per school and near .1 for 20 students per school. The P coefficients are about .02 less for the same number of schools and students per school.

While these might be small effect sizes for much economics research, they can represent important effects in many social science fields, particularly education research. In educational assessment research, when studying the effects of various school programs or policies on achievement outcomes, standardized effects often fall in the range of .1 to .2. That is to say, providing they are statistically significant, many education policy experts would see a standardized program effect of .1 or .2 as evidence in favor of the program, even though it is an admittedly small effect (Hill, et al, 2008; Lipsey, et al, 2012). Even standardized effects on the order of .05 can be reported as supporting a policy, provided they are statistically significant.

Although the larger spurious effects occur when the number of students per school is 10 or 20, these numbers are not unusual when national or regional samples of schools are drawn, or when the grouping variable is classrooms instead of schools. For example, in the PISA and NAEP datasets used for the simulations, the average sample size per school is about 30 students. In education studies of the contextual effects of classrooms, classroom sizes of 15 to 30 students would be the most common. Average class sizes in many school districts are around 20 students.

Model (4): Two Aggregate Variable Plus the Corresponding Individual Variables

The final model using the NAEP data shows that the spurious effects of the aggregate variables S and P are eliminated when their individual versions, I and B, are included in the simulations. The linear regression model is shown as a fourth equation:

$$(4) \qquad A = \beta_I I + \beta_B B + \beta_S S + \beta_P P + e \qquad \text{(standardized)}^4$$

The expected values for the standardized regression coefficients for model (4), as produced by the Monte Carlo simulations, are shown in Table 6.[5] The estimated coefficients for I in the first panel and B in the second panel are all statistically significant at $p < .001$. The magnitudes differ somewhat from sample to sample, but they are all within a modest range of the "true" value of .41 based on a regression for the full NAEP sample (see Table 7). Indeed, the utility of the Monte Carlo approach is demonstrated by the fact that even very small samples on the order of 50 or 100 students generate estimates of β_I that are close to its true value; e.g., 5 schools and 10 students per school show $\beta_I = .42$. The same is true for the estimated values for β_B

[4] It is understood that the coefficients for S and P in model (4) can be different than those in model (3), even though the same β symbols are used.

[5] The full set of simulation correlations that go into model (4) are available from the authors.

In contrast, the values for the S and P coefficients are now zero to the third decimal point. In other words, when the individual variables I and B are included in the simulations, the spurious coefficients for the aggregate variables S and P in model (3) (as found in the simulation in Table 5)) are eliminated.

Table 6 Mean Simulation Betas for the Full Model in Equation (4)

	Number of Students per School					
No. of Schools	5	10	20	30	40	50
STANDARDIZED COEFFICIENTS FOR I	(all p < .001)					
5	.4859	.4208	.4225	.3149	.4522	.3551
10	.3005	.4511	.4322	.4323	.3748	.3907
20	.4079	.3964	.4446	.3926	.4037	.4588
30	.4367	.4168	.3781	.3971	.4019	.4218
40	.3545	.4248	.4305	.3715	.4213	.4176
50	.5121	.3920	.4189	.4123	.3970	.3991
80	.4104	.4618	.4183	.4087	.3990	.4146
STANDARDIZED COEFFICIENTS FOR B	(all p < .001)					
5	-.3526	-.3503	-.3617	-.3458	-.2710	-.2875
10	-.3909	-.2738	-.3157	-.3472	-.3854	-.3244
20	-.2787	-.2896	-.3480	-.3235	-.3535	-.3027
30	-.3812	-.3503	-.3368	-.3473	-.3087	-.3295
40	-.4532	-.3237	-.2910	-.3586	-.3385	-.3367
50	-.3026	-.3403	-.3638	-.3496	-.3399	-.3397
80	-.4175	-.2978	-.3205	-.3272	-.3314	-.3330
STANDARIZED COEFFICIENTS FOR S	(all p > .99)					
5	.0001	-.0012	-.0003	-.0002	.0005	-.0003
10	-.0001	.0008	-.0001	-.0003	.0000	.0000
20	-.0002	-.0003	.0001	.0000	.0000	.0000
30	-.0001	.0002	-.0002	.0001	.0001	.0000
40	-.0003	-.0003	.0000	.0000	.0001	.0001
50	.0002	.0000	.0001	-.0001	-.0001	.0000
80	.0000	-.0001	.0000	-.0001	.0000	.0000
STANDARDIZED COEFFICENTS FOR P	(all p > .99)					
5	.0011	.0003	.0006	.0000	-.0001	-.0002
10	.0000	.0008	-.0004	.0002	-.0001	-.0002
20	.0001	-.0001	-.0001	.0000	-.0002	-.0001
30	-.0004	-.0002	.0000	.0001	.0000	.0000
40	.0001	-.0002	.0001	.0000	.0002	.0001
50	.0001	-.0001	.0000	-.0001	-.0001	.0000
80	-.0003	.0000	.0000	-.0001	.0000	.0000

It should be noted that the large-sample estimate for the I and B coefficients in Table 6, which are .41 and -.34, respectively, are virtually identical to the coefficients obtained in the actual regression of I and B on A using the full data set and excluding the aggregate variables. This result is demonstrated in Table 7, which shows various linear regressions involving the individual and aggregate variables in the full NAEP data set.

The first regression in the table includes only the two individual measures, and the standardized regression coefficients (betas) for individual SES (I) and black (B) are .41 and -.34.

Table 7 Actual Regressions for I, B, S, and P in the NAEP Data

Variable	Coefficient	Robust SE	P	Beta
INDIVIDUAL VARIABLES ONLY				
Individual SES (I)	14.90	0.56	<.001	0.41
Black (B)	-25.56	1.23	<.001	-0.34
Constant	289.45	0.79	<.001	
R-squared	0.35			
AGGREGATE VARIABLES ONLY				
School SES (S)	20.27	1.56	<.001	0.27
School Per Cent Black (P)	-30.22	3.28	<.001	-0.22
Constant	290.80	1.31	<.001	
R-squared	0.18			
ALL VARIABLES				
Individual SES (I)	13.19	0.56	<.001	0.36
School SES (S)	7.08	1.63	<.001	0.09
Black (B)	-21.99	1.06	<.001	-0.29
School Percent Black (P)	-8.04	3.40	0.02	-0.06
Constant	290.67	1.30	<.001	
R-squared	0.36			

Note: for all regressions, number of students is 4134, number of schools is 144

The other regression results are also worth noting. The regression for Aggregate Variables Only in the first panel shows substantial standardized coefficients for both school SES and school percent black, .27 and -.22 respectively, and these are much larger than the spurious effects found in the simulations for S and P in model (3). However, when all four variables are included in the regression, the coefficient for S is reduced to .09 and the coefficient for P is reduced to -.06, although both remain statistically significant.

Model (5): Introducing School Achievement (D) in the PISA Data

In some discussions of peer or contextual effects of school and classroom composition, aggregate achievement has been introduced, and in at least one study of classroom peer effects, average peer achievement was shown to be more important than either SES or racial composition of the classroom (Hoxby and Wiengrath, 2005). The question here is whether average school achievement also gives rise to spurious effects.

The approach is similar to models (3) and (4), except in this case the second aggregate variable is average school achievement (D), and there is no second individual variable because individual achievement (A) is the dependent variable. Two models are hypothesized:

(5a) $A = \beta'_S S + \beta'_D D + e$ (standardized)

(5b) $A = \beta_I I + \beta_S S + \beta_D D + e$ (standardized)

In the full PISA data, the A-D correlation is .55 compared to the A-S correlation of .42, and the correlation between S and D is rather high at .77. The simulation results for model (5a) were surprising, initially, because the average β'_D was always equal to the A-D correlation, and the average β'_S was always 0.

Some analysis revealed that cov(A,S)=cov(D,S) and also that cov(A,D)=var(D). These equivalences lead to the unexpected results of $\beta_D = r_{AD}$ and $\beta_S = 0$

Model 5(b) produced more equivalences, namely that cov(A,S)=cov(D,S)=cov(I,D) and cov(I,S)=var(S).[6] Because of these additional constraints, the coefficient for D was always equal to the A-D correlation and the coefficient for S was always negative, despite the fact that the A-S correlation is a large positive value. Accordingly, instead of showing simulation results, Table 8 shows the correlation and regression results calculated for the full dataset.

Table 8 Actual Correlation and Regression Results for Models (5a) and (5b), Full Sample[a]

CORRELATIONS, VARIANCES, AND COVARIANCES				
	A	I	S	D
Achievement (A)	7066.8	33.2	18.0	2148.9
Individual SES (I)	0.4304	0.8430	0.2534	18.0
School SES (S)	0.4259	0.5483	0.2534	18.0
School Ach (D)	0.5514	0.4235	0.7723	2148.9
REGRESSION RESULTS FOR MODELS (5a) AND (5b)				
	Coefficient	Robust SE	P	Beta
MODEL (5a)				
School SES (S)	0.000	3.096	<.001	0.0000
School Ach (D)	1.000	0.034	<.001	0.5514
Constant	0.000	16.035	<.001	
R-squared	0.304			
MODEL (5b)				
Individual SES (I)	25.774	1.237	<.001	0.2815
School SES (S)	-25.774	3.218	<.001	-0.1543
School Ach (D)	1.000	0.032	<.001	0.5514
Constant	0.000	15.385	<.001	
R-squared	0.359			

a N = 5018 for all analyses

In the first panel of Table 8, the correlations among all variables are shown below the diagonal, the variances are shown in the diagonal, and the covariances are shown above the diagonal. Color coding is used to help identify the equivalences. The second panel shows regression results for the two models. For Model (5a), the regression yields 0 for the A-S coefficient and the A-D coefficient (.5514) is equal to the A-D correlation. Interestingly, in model (5b), the various covariance and variance equivalences yield the same coefficient for school achievement, and the unstandardized coefficients for I and S are equal but of opposite sign.

The relationships shown for Model (5b) attenuate as additional predictor variables are added to the regression equation, provided they have some correlation with the other variables in the model. The reason is because the covariance and variance equivalences have less direct impact as more covariance terms are added. For example, if the variable metasum (C) is added to model (5b), the D coefficient is no longer equal to the A-D correlation and the I coefficient is no longer the S coefficient with opposite sign.

[6] More detailed mathematical derivations for these equivalences are available from the authors.

Discussion and Conclusions

This paper has investigated the relationships among variables which are measured at the individual level and also at an aggregated level using a grouping variable such as a school, a classroom, a neighborhood, and so forth. Aggregate variables have become commonplace in many social science studies as measures of "contextual effects." Characteristics such as race, socioeconomic status (SES), and academic ability are hypothesized to have effects on various outcomes above and beyond the effects of the individual measures themselves.

Using data from the PISA and NAEP national studies, Monte Carlo simulations were undertaken to determine if constructed aggregate variables can produce spurious relationships with a dependent variable which is measured at the individual level. The term "constructed" means that a given aggregation of socioeconomic status or race for a group of persons is not based on the real distribution of observations in real groups, but rather the group is created by randomized draws of observations from all possible permutations of those observations. If the expected value of a given correlation (or a regression coefficient) between an aggregate variable and an individual variable is significantly greater than 0, the term "spurious" is used to signify that it does not have any intrinsic meaning, because the group composition is entirely arbitrary and random.

The Monte Carlo simulations found both correlations and regression coefficients for constructed aggregate variables that were significantly greater than 0. Linear regression coefficients that include only aggregated independent variables had expected values that were significantly greater than zero, even though the aggregated variable was created entirely by random draws. The size of the spurious effects increase as the number of observations per group decrease. These spurious effects are quite large when the number of observation per group is about five, which is smaller than most grouping variables studied in the social sciences. The spurious effects are still sizeable, however, when the numbers of observations per group is between 10 and 20. Most studies of the contextual effects of classrooms have groupings of this size.

When the individual-level version of the aggregated variable is included in the regression, the effect(s) of the constructed aggregated variable(s) becomes exactly zero (within rounding error), regardless of the size of the original spurious relationship; that is, even when the number of cases per group is very small. This appears to be true for both simple models and more complex models with multiple aggregate variables. In other words, a necessary and sufficient condition to eliminate spurious effects of aggregate variables is to include their exact individual-level variable in the model.

The special case of models using the dependent variable as an aggregate independent variable led to a different and unexpected result. A simple example with individual and aggregate SES as independent variables and aggregate achievement as an independent variable caused the aggregate SES measure to have a significant negative effect, even though it had a strong positive effect in models without the aggregate achievement measure.

The results of these simulation exercises means that caution must be used when interpreting correlation and regression coefficients for aggregate variables. If the individual level of these variables are not available for a regression analysis, significant spurious effects can arise, and these effects can be fairly large if the aggregated groups are on the order of 10 to 20 cases per group.

REFERENCES

Bryk and Raudenbush (1992). *Hierarchical Linear Models*. Sage Publications, Newbury Park, C

Hanushek, Eric A., John F. Kain, Steven G. Rivkin (2009). "New Evidence about *Brown v. Boardof Education*: The Complex Effects of School Racial Composition on Achievement" *Journal of Labor Economics,* (27:349-383)

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey (2008). "Empirical Benchmarks for Interpreting Effect Sizes in Research," *Child Development Perspectives* (2:172-177)

Hoxby, Caroline (2005) & Gretchen Wiengrath. Taking Race Out of the Equation:School Reassigmnet and the Structure of Peer Effects available on the world wide web at www.economics.harvard.edu/faculty/hoxby/papers/hoxbyweingarth_taking_race.pdf

Kahlenberg, Richard D. (ed) (2012). *The Future of School Integration: Socioeconomic Diversity as an Education Reform Strategy* Washington D.C.: The Century Foundation

King, Gary (1997**).** *A Solution to the Ecological Inference Problem:Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press

Lipsey, Mark W., Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick (2012). *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*. U.S. Department of Education, Institute for Education Science

Loveless, Tom (2012). *Heow Well Are American Students Learning?* Washington, DC: Brookings Institution

Moulton, Brent R. (1990). "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units" *The Review of Economics and Statistics* (72:334-338)

Sampson, Robert J.; Stephen W. Raudenbush, and Felton Earls (1997). Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy *Science* 277:918-924

Vigdor, Jacob and Thomas Nechyba (2004). Peer Effects in Elementary School: Learning from 'Apparent' Random Assignment, unpublished manuscript, Duke University and NBER

Willms, J. Douglas (2010). School Composition and Contextual Effects on Student Outcomes, *Teachers College Record,* 112(4):1137-1162

Wooldridge , Jeffrey M. (2003). "Cluster-Sample Methods in Applied Econometrics" *The American Economic Review* (93: 133-138)