Educational Evaluation and Policy Analysis
Winter 1999, Vol. 21, No. 4, pp. 385–403

The Puzzling Case of School Resources and Student Achievement

Jens Ludwig
Georgetown University

Laurie J. Bassi
American Society for Training and Development

Whether additional school spending translates to overall improved student learning remains unclear. One explanation for the mixed findings in the literature is the possibility that studies confound the effects of school resources with those of unobserved variables. We show that commonly used "value-added" models are misspecified when estimated using the National Education Longitudinal Study, which raises questions about whether previous value-added studies are unbiased. We also review a more recent literature that uses instrumental variables (IV) methods to address omitted variables bias. Most IV studies suggest that additional resources typically translate to (modest) gains in test scores and that the biases associated with value-added models are large enough to be of practical importance.

The primary objective of education policy in the United States is to increase the amount of student learning that results from the 4% of gross domestic product allocated to public schools each year (U.S. General Accounting Office, 1995). Yet, reliable information about how to consistently improve student learning on a large scale remains elusive (e.g., Barnett, 1996). As a result, policymakers and advocates regularly turn to the one policy lever that is (relatively) straightforward to manipulate: the education budget. Some policy proposals seek to achieve across-the-board increases in school budgets and leave decisions about how to allocate these resources to local administrators. Other proposals seek to ensure that additional funds are spent on specific instructional items, such as the Clinton administration's initiative to devote $11 billion over 7 years to hiring 100,000 additional elementary school teachers (Alvarez, 1999).

The obvious question for education policy is whether across-the-board or targeted increases in education spending will translate into improved student learning. The effects of across-the-board increases in school budgets will obviously depend in large part on how these funds are spent; for example, on average, additional spending on instructional items will presumably generate more student learning than increased spending on administration. The productivity of additional instructional expenditures will in turn depend on a school's organization, internal incentives, and response to the higher resource environment (e.g., Murnane & Levy, 1996). Because of the emphasis on local control over education in the United States and the difficulty of closely monitoring school practices at the micro level, state and national policymakers have typically been able to affect overall education budgets or expenditures on specific items, but have had relatively modest influence over how schools and teachers respond to these changes. As a result, policymakers and researchers alike remain interested in the question of how additional resources affect overall student learning given the current distribution of school practices.

Despite an enormous body of empirical research, there is currently little consensus about whether additional education spending will, on average, improve student test scores, the most commonly used measure of student learning. The literature on whether increases in overall spending or specific expenditure items improve student test scores is surprisingly mixed. Hanushek's (1996) review of the education production function literature noted that "71 percent of the estimated effects [of school spending on student test scores] are statistically insignificant or negative" and concluded that there is "no strong or systematic relationship...between spending and student performance" (p. 56). Others believe that the same research literature provides support for the idea that additional spending will, on average, increase student test scores (Hedges & Greenwald, 1996; Hedges, Laine, & Greenwald, 1994).

This article explores one possible explanation for the mixed results shown in previous studies, namely, that these earlier investigations did not adequately control for omitted variables that are correlated with how students learn and are assigned to different schools and classrooms. Many researchers assume that omitted variables will produce an upward bias in the estimated effects of school resources on student outcomes because more affluent families are likely to choose higher spending districts (Tiebout, 1956) and family socioeconomic status is positively correlated with student learning (Mayer, 1997). If this were the only source of potential bias, then available studies could be interpreted as upper bound estimates for the effects of school resources on student achievement. Yet, it is also possible that administrators or parents target poorly performing schools or students for compensatory resources. In this case, unobserved variables that affect student learning may lead to underestimates of school resource effects (Heckman, Layne-Farrar, & Todd, 1996a, 1996b).

In this article, we use data from the National Education Longitudinal Study of 1988 (NELS) to test whether the estimation approaches used in previous studies can identify the causal effects of school resources on student test scores. We focus on "value-added" models that use a previous period's test score as a control variable to address the problem of omitted variables bias. Value-added models have long been viewed as the "gold standard" in the education production function literature; however, these models will produce unbiased estimates only under certain conditions that can be empirically tested. We find evidence that value-added models estimated with data from the NELS are likely to be biased. Although NELS has several important limitations, the data set provides detailed information on family background and school characteristics for a large, nationally representative sample of students and is arguably one of the best available data sources for estimating education production functions. Evidence that value-added models are not unbiased when applied to the NELS raises questions about the validity of previous value-added studies that rely on less detailed data sets, and perhaps suggests that less weight should be given to the value-added literature in future debates about education policy. At a minimum, it would be useful for future value-added studies to report basic diagnostic information that might illuminate whether investigators' assumptions for unbiased estimation are met.

We also review a more recent empirical literature that may be less familiar to education analysts, one that improves upon previous research by using instrumental variables (IV) methods to control for omitted variables problems. Most of these studies have shown that IV methods produce larger positive effects of school resources on student learning than do value-added or other estimation approaches, which suggests that the biases identified by our NELS empirical work are large enough to be of substantive importance. We conclude by noting that because most of the IV estimates are fairly modest in size, it is not obvious that increased school spending is the most cost-effective way to improve the skills of students. Education production function studies might usefully be combined with cost-effectiveness or cost-benefit analyses to identify the most promising approaches for improving academic and other outcomes among U.S. students.

The article is organized as follows. In the next section, we review the conditions under which value-added models will produce consistent estimates of the effects of school spending on student test scores, after which we review the NELS data set. We then apply various value-added models to the NELS data and use several consistent model misspecification tests to show that the conditions for unbiased estimation have not been met. Finally, we review recent IV studies and discuss implications for research and policy.

The Value-Added Approach

Value-added models that use test scores from previous periods as control variables are typically viewed as the gold standard in studying the effects of school inputs on student achievement. Yet, studies that use this evaluation strategy rely on several assumptions to produce unbiased estimates of the effects of school resources on student learning, as discussed in this section. Later we describe empirical tests of whether these assumptions are met with the NELS data.

To understand the assumptions behind the value-added approach, consider the model of student achievement in Equation 1 from Boardman and Murnane (1979) and Hanushek and Taylor (1990). The achievement test score of student *i* in period *T*, Y_{iT}, is a function of the set of school inputs the child receives in period *T*, S_{iT}, and the set of family characteristics in period *T* that affects how much the child learns at home, F_{iT}. In this model, the child's learning in period *T* is cumulative and is related to school and family inputs in previous years ($t = 1,2,...,T-1$). Because no education data set can hope to capture all of the factors that affect learning, the student's test score in period *T* will also be a function of omitted variables. These omitted variables may be student specific (ε_i), such as innate ability or other omitted student or parent characteristics, or specific to the student and year (ε_{it}), such as unmeasured characteristics of the child's home or school environment that might change over time (e.g., the mental or emotional health of parents or students, student motivation, or the quality of the parents' relationship).

$$Y_{iT} = a_0 + a_{1T} S_{iT} + a_{2T} F_{iT} + \Sigma_t a_{1t} S_{it} + \Sigma_t a_{2t} F_{it} + \varepsilon_i + \varepsilon_{iT} + \Sigma_t \varepsilon_{it} \quad (t = 1,2,...,T-1). \tag{1}$$

Early attempts to estimate the education production function in Equation 1 relied on cross-sectional data, which provided information on test scores, family background, and school inputs for a sample of students at only one point in time. For example, the so-called Coleman Report of 1966 estimated the achievement equation described in Equation 2 (Coleman et al., 1966). If the lagged school and family resources captured by the error term e_{iT} (Equation 3) are uncorrelated with current school inputs, then estimating Equation 2 via ordinary least squares will provide unbiased estimates of the effects of increases in contemporaneous school resources on student achievement. Conversely, if current school resources are perfectly correlated with lagged school inputs (and uncorrelated with lagged family variables and the other individual effects in Equation 3), then the coefficients in a_{1T} reveal the effects of a cumulative increase in school resources in periods $(1,...,T)$ on student achievement in period *T*.

$$Y_{iT} = a_0 + a_{1T} S_{iT} + a_{2T} F_{iT} + e_{iT} \tag{2}$$

$$e_{iT} = \Sigma_t a_{1t} S_{it} + \Sigma_t a_{2t} F_{it} + \varepsilon_i + \varepsilon_{iT} + \Sigma_t \varepsilon_{it} \tag{3}$$

In practice, unmeasured and lagged family variables (F_{it}) and unobserved student characteristics (ε_i) are likely to be correlated with the student's current school resources, and, as a result, ordinary least squares estimates for the parameters (a_{1T}) in Equation 2 will be biased.[1] Of primary concern is the possibility that the decisions that families make about their children's schools or classroom assignments are related in part to unobserved family or student variables. Similarly, school administrators may also assign children to schools or classrooms in part on the basis of prior achievement[2] (which in turn is a function of lagged and unmeasured family and student effects) or other unmeasured student characteristics. In this case, estimates of the effects of school resources from Equation 2 will reflect both the direct effects of school resources and the indirect effects of those family and student aspects that have an impact on achievement and how students are sorted across schools and classrooms.

As noted earlier, the direction of bias from unobserved variables could, in principle, be either positive or negative. Poorly measured or unmeasured aspects of family socioeconomic status may lead to an upward bias in the estimated effects of school resources because more affluent parents may be more likely to place their children in resource-intensive environments. Similarly, unmeasured student ability may lead to upward bias in estimates of a_{1T} if schools or parents select more able students into higher resource settings (such as magnet schools or gifted programs). On the other hand, schools or parents may seek additional educational resources for students who are performing below expectations. If unobserved variables that are negatively correlated with student achievement cause students to be placed in higher resource settings, then estimates of a_{1T} will be biased downward.

Because our understanding of how students are assigned to schools and classrooms is quite limited, the direction and magnitude of bias caused by omitted variables are quite difficult to predict. Moreover, the magnitude and direction of bias could, in principle, vary across different measures of school inputs. For example, if affluent parents focus on moving their children into areas with high teacher salaries, but school administrators compensate for poor student performance by moving children into small classes, the effects of teacher salaries may be overstated and the effects of class size understated.

The availability of panel data has enabled researchers to address omitted variables problems by estimating value-added models, as in Equation 4. These models use a previous period's test scores as a covariate to control for unmeasured school, family, and student characteristics ($\Sigma_t a_{1t} S_{it} + \Sigma_t a_{2t} F_{it} + \varepsilon_i + \Sigma_t \varepsilon_{it}$).

$$Y_{iT} = b_0 + b_{1T} S_{iT} + b_{2T} F_{iT} + b_{3T} Y_{iT-1} + v_{iT}. \qquad (4)$$

The value-added model described in Equation 4 will be unbiased only if whatever remains in the error term (v_{iT}) after controlling for a previous period's test score is uncorrelated with current school resources. This condition requires that several strong assumptions be met.

First, the previous period's test score must be a good proxy for the student's prior achievement. To the degree to which previous achievement is measured with error, the lagged test score will be an imperfect proxy for omitted factors that affect both achievement and current school resources, and the estimates from Equation 4 will converge to those in Equation 3 (Meyer, 1992).[3]

Second, all unmeasured and lagged family and student characteristics that affect Y_{iT} and current school or classroom assignments must exhibit their full effects in the pretest score Y_{iT-1}. If, for example, these unobserved background variables have delayed effects, then Y_{iT-1} will be an imperfect proxy for these factors, and the estimates for b_{1T} in Equation 4 may be biased. Similarly, if unobserved student attributes affect the rate at which students learn, including a previous period's test score may not fully control for the effects of these unmeasured variables on Y_{iT}.

Third, the time interval between testing periods must be sufficiently short that there will be no changes in student or family characteristics between testing periods that affect achievement and school resource assignments. This condition may not be met in practice. Many studies use test scores measured 4 years or more in the past (e.g., see Pallas & Alexander, 1983), and even many of the best education data sets, such as NELS and High School and Beyond, test students only at 2-year intervals. As an example of the bias that might be induced by these testing lags, suppose there is some change in a child's household from period $T-1$ to period T, such as increased conflict between parents, that negatively affects the child's achievement and causes parents or school administrators to move the child to a higher resource educational environment. This difficult-to-measure family effect will not be captured by the previous period's test score and will result in downward bias in estimates of b_{1T}.

Finally, whatever random events affect the student's test score in one period (v_{iT}) must be uncorrelated with previous "shocks" (v_{iT-1}) if the value-added model is to be unbiased. For instance, reconsider our example of increased conflict between a child's parents. If this conflict negatively affects student achievement and increases over time, leading to divorce, the unobserved shock to the child's education production function in period $T-1$ will be correlated with the shock in period T. If the errors are serially correlated, then the error term v_{iT} in Equation 4 will be correlated with the lagged test score Y_{iT-1}. As a result, the coefficient estimates for both the lagged test score and the other variables will be biased in a way that will depend on how the previous test score is correlated with the current school resource variables (Hausman, 1983).

While previous studies have typically assumed that the conditions necessary for unbiased estimation with value-added models are met, very little research has been conducted to determine whether this is actually the case. In the empirical work described subsequently, we applied several empirical tests developed in the economics literature to different value-added models fitted to the NELS data. We found evidence to suggest that the conditions for unbiased estimation are violated.

Data

The Department of Education's National Education Longitudinal Study of 1988 surveyed a nationally representative sample of eighth-grade students; follow-up interviews were conducted in 1990 and 1992. The original sample involved a two-stage sampling design, with 1,052 schools selected in the first stage and 26 students per school selected in the second stage.[4] Base-year respondents were selected to participate in follow-up surveys in part on the basis of the number of other base-year NELS participants in the student's school at the time, and dropouts were surveyed as well (U.S. Department of Education, 1994). Since NELS participants were clustered in schools and classrooms, and all of the educational input measures used in this study were measured at the group level (school or classroom), ordinary least squares may underestimate standard errors for regression coefficients (Moulton, 1986). To correct for this problem, we calculated standard errors using the delta method (Kish & Frankel, 1974; Moulton, 1986).

The Department of Education provides weighting variables that account for probabilities of participation in the base-year and follow-up surveys, as well as school administrator and student survey nonresponse rates (U.S. Department of Education, 1994). Our preferred estimates were calculated with these weights, although we also developed unweighted estimates to examine the robustness of our findings.

Descriptive statistics for key variables are presented in Table 1. The outcome measures of inter-

TABLE 1
Descriptive Statistics for NELS Public School Students, 1990

	Sample (N = 13,738)
Male (%)	49.3
Race/ethnicity (%)	
Non-Hispanic White	68.6
African American	12.9
Hispanic	10.2
Asian	3.3
Other	5.0
Household income ($) (%)	
0–10,000	11.5
10,000–20,000	15.8
20,000–35,000	30.8
35,000–50,000	20.9
50,000–75,000	14.8
>75,000	6.0
Mother's education	
Less than high school	16.3
High school	39.3
Some college	22.8
College plus	8.5
Father's education	
Less than high school	16.7
High school	35.0
Some college	21.0
College plus	15.5
Reading achievement (*SD*)	49.4 (10.1)
Math achievement (*SD*)	49.8 (9.7)
School inputs	
Mean student-teacher ratio (*SD*)	16.2 (4.3)
% African American students (*SD*)	13.6 (20.4)
% Hispanic students (*SD*)	9.9 (21.1)
% teachers with more than college degree (*SD*)	51.7 (21.9)
% students in free/reduced-price lunch program (*SD*)	21.6 (20.8)
Lowest full-time teacher salary (1990 $ × 1,000) (*SD*)	20.19 (3.20)
Student class characteristics	
English class size (*SD*)	23.5 (6.2)
Math class size (*SD*)	23.4 (7.0)
English teacher: postcollege education	55.0
Math teacher: postcollege education	52.7
English teacher: total years of experience (*SD*)	16.1 (9.3)
Math teacher: total years of experience (*SD*)	15.3 (10.0)

Note. Values are weighted means.

est were student scores on standardized reading and mathematics achievement tests, administered to students and dropouts in each wave of NELS. The standardized tests administered in previous Department of Education panel studies, such as High School and Beyond, have been criticized as being "too easy" and therefore labeled poor measures of what students learn in high school (Goldberger & Cain, 1982). In contrast, students in NELS were administered different versions of the tests in follow-up years according to base-year performance to minimize these "ceiling" and "floor" effects (U.S. Department of Education, 1994). The Department of Education presents scores on these tests in a way that is scaled to allow for valid across-wave comparisons.

Educational input measures included the school's student-teacher ratio, the proportion of teachers with advanced degrees, and the lowest (and, in some waves, the highest) salary paid to a full-time teacher, all taken from the school administrator survey. Since teacher salaries typically increase with both educational credentials and experience, we used the lowest full-time teacher salary as a proxy for the starting salary within a school. This proxy could potentially confound the effects of teacher experience and pay; for example, a school with a low pay scale whose most junior teacher had 4 years of experience might report a higher minimum salary than a school with a high pay scale and a first-year teacher on staff.[5] This did not appear to be a serious concern, since the average minimum salary for a full-time teacher in the 1990 NELS survey was almost identical to the national mean starting teacher salary for 1990 ($17,750 and $17,900, respectively, in 1987 dollars) (Smith, 1996). Teacher salaries and teacher-student ratios together should capture most of the variation in the instructional resources that students receive.

NELS does not include certain other school-input measures that have been shown to be correlated with student test scores, such as the test scores of the teachers themselves (Ferguson, 1991; Ferguson & Ladd, 1996). NELS also does not include measures of overall per pupil school spending, although estimates of the effects of specific expenditure items may be more useful in highlighting how additional resources should be allocated (Murnane, 1991).

In the base year and first follow-up, NELS randomly selected two teachers to be surveyed from among each participating student's English, mathematics, science, and history instructors. In the second follow-up, either a mathematics or science teacher was surveyed for students enrolled in at least one math or science course. These teacher surveys provided an opportunity to use information on the student's classroom in a particular subject area, including the teacher's experience and education and overall class size, as it relates to the student's performance on a standardized test for that subject. These classroom-level variables provide more direct measures of the resources that each student receives relative to school-level variables. On the other hand, the student's test score in a particular subject area may be affected by inputs in her or his other classes (for example, reading scores may be affected by the quality of the student's history class). The NELS classroom-level variables will miss these spillover effects because the data set provides classroom information for, at most, two of the four subject areas just described for each student.

Other relevant explanatory variables in the NELS include sociodemographic characteristics of the other students attending each respondent's school (percentage African American students, percentage Hispanic students, percentage eligible for free or reduced-price lunch). The characteristics of a school's student body may influence achievement through "peer effects" (Duncan, 1994; Jencks & Mayer, 1990; Ludwig, Duncan, & Hirschfield, 1999), or by acting as proxies for job characteristics that make the school more or less attractive to talented teachers. Whether such peer or reference group variables should be included in achievement models is controversial, since such groups may be endogenous and thus may introduce new biases (Evans, Oates, & Schwab, 1992). All of the results presented here were estimated both with and without these variables.

The NELS has several important limitations, including the fact that students were tested only at 2-year intervals and classroom-level information is available only for a subset of each student's classes. Nevertheless, NELS has several advantages over data used in previous studies: a large, nationally representative sample of students; detailed information about the family backgrounds and personal characteristics of students available at the student level, with many of these variables taken from parent and school reports (rather than simply from student surveys); a longitudinal survey structure including data on students before they enrolled in high school and on students who dropped out; and numerous measures of educational resources at both the school and classroom levels.

Empirical Results for Value-Added Models

The results of estimating the value-added model in Equation 4 for math and reading scores are shown in Tables 2 and 3. We focus on student test scores from the second wave of the NELS (1990), when most students were in 10th grade. Our focus on 1990 test scores allowed us to estimate a value-added model that controlled for 1988 achievement scores, and very few students in NELS (less than 4%) had dropped out of school by 1990. The set of controls for family background in the model is at least as rich as what is found in other value-added studies, and includes whether the student was living with her or his father or another adult man, family income taken from parent reports (coded as a sequence of dummy variables to allow for nonlinear income effects), and mother's and father's educational attainment (also taken from parent reports and also coded as a sequence of dummy variables to allow for nonlinear effects).

Table 2 presents the estimated effects of classroom-level inputs on student performance in mathematics and reading. The key input measures included the size of the student's math or reading classroom in 1990, teacher experience (in years), teacher experience squared (to allow for nonlinear experience effects, as reported by Murnane, 1975), and whether the teacher had more or less education than a college degree. The model followed Ferguson and Ladd (1996) and allowed for nonlinear class size effects by including spline terms, defined as class size minus 15 if the class size was larger than 15 and zero otherwise, and similarly for a spline cutoff of 30. The effect of a marginal change in classroom size when size exceeded 15 is given by adding the coefficients for class size and spline 15; for changes in classrooms larger than 30 students, we added the coefficients for class size and both splines. (As noted earlier, one drawback of this classroom-input model is that we could not control for the quality of the classroom environments in all of the student's other classes, even though inputs in these classes may also have affected math and reading scores.)

As can be seen in Table 2, the results show the same puzzling pattern of estimated school resource effects as that reported in the previous literature. For example, college education for teachers appears to have a positive effect on student reading scores but a negative effect on math scores. Reductions in class sizes when classrooms are large (more than 30 students) improve test scores in reading but have a very modest negative effect on math scores. Standard *F* tests suggested that the class size variables, and all of the classroom-input variables taken together, are jointly statistically significant.

The results for school-level resource effects on math and reading scores revealed the same pattern, as seen in Table 3. Reductions in student-teacher ratios in schools with high ratios (more than 25 pupils per teacher) appeared to have negative effects on math but positive effects on reading. Teacher salaries had no clear effect on math scores but appeared to have a modest negative effect on reading scores.

Whether these results provide useful information on the actual effects of school or classroom resources depends on whether the conditions for unbiased estimation with these value-added models were met. From our empirical work, using several different statistical specification tests, we concluded that these conditions were not met. The first two specification tests that we used focus on whether the covariates in the value-added model are uncorrelated with the regression error term (as required by the first two conditions for unbiased estimation with the value-added model), while our final test examined whether the error terms are serially uncorrelated.

We first used the test described by Ramsey (1969), which compares the null hypothesis that explanatory variables are uncorrelated with the error term and the alternative hypothesis that the model suffers from omitted variables or misspecified functional form. The intuition behind the test is that model misspecification occurs when the expected value of the regression error term varies according to the value of the explanatory variables. Ramsey hypothesized that the expected value of the error term conditional on the value of the explanatory variables is some nonlinear function of the explanatory variables. We implemented the test by first estimating the value-added regression described in Equation 4 and calculating the predicted value of each student's test scores implied by the regression coefficients. Then we reestimated the value-added model with the square, cube, and quartic of the student's predicted test score included as explanatory variables to examine whether the error term in the value-added regression was re-

TABLE 2
NELS Achievement Regression Coefficients for 1990 Public School Students: Classroom-Level Resources

	Coefficient (*SE*)	
	Reading	Math
Class inputs		
Class size	0.27 (0.07)**	0.16 (0.07)*
Spline 15	−0.27 (0.09)**	−0.19 (0.09)*
Spline 30	−0.11 (0.07)	0.06 (0.04)
Teacher experience	−0.02 (0.03)	0.00 (0.03)
Teacher experience squared	0.00 (0.00)	0.00 (0.01)
Teacher: less than college	1.33 (0.94)	−0.60 (0.83)
Teacher: more than college	0.03 (0.22)	0.40 (0.21)+
School, % Black	0.01 (0.01)	0.00 (0.01)
School, % Hispanic	−0.01 (0.01)	0.01 (0.01)
School, % free lunch	0.01 (0.01)	0.01 (0.01)
Father in household	−0.14 (0.29)	0.51 (0.26)*
Male respondent	−0.23 (0.20)	−0.02 (0.19)
Asian	0.31 (0.42)	0.71 (0.53)
Hispanic	−0.54 (0.36)	−1.13 (0.38)**
Black	−1.26 (0.36)**	−0.66 (0.45)
American Indian	−1.10 (0.58)+	−1.14 (0.48)*
Father's education		
High school	0.35 (0.35)	−0.14 (0.34)
Less than college	1.40 (0.39)**	0.56 (0.36)
College	1.97 (0.44)**	0.88 (0.38)*
College plus	1.87 (0.46)**	1.01 (0.41)*
Don't know	1.02 (0.43)*	0.91 (0.40)*
Mother's education		
High school	0.21 (0.30)	0.62 (0.31)*
Less than college	0.39 (0.39)	0.78 (0.39)*
College	0.75 (0.44)+	0.83 (0.47)+
College plus	0.96 (0.48)*	0.74 (0.50)
Don't know	0.11 (0.41)	−0.01 (0.47)
Family income ($)		
10,000–15,000	0.53 (0.45)	0.67 (0.50)
15,000–20,000	0.90 (0.45)*	0.92 (0.54)+
20,000–25,000	0.44 (0.45)	0.75 (0.50)
25,000–35,000	0.74 (0.42)+	1.14 (0.49)*
35,000–50,000	0.90 (0.42)*	1.61 (0.52)**
50,000–75,000	1.17 (0.45)*	1.25 (0.56)*
>75,000	1.17 (0.54)*	1.58 (0.58)**
Missing	1.44 (0.45)**	0.75 (0.57)
1988 test score	0.89 (0.01)**	0.95 (0.01)**
N	4,839	3,655
R^2	0.65	0.77
Specification test results		
Ramsey test: *F* (*df*)	34.06 (3, 4745)**	18.18 (3, 3561)**
Johnson-McClelland test (normal distribution)	−1.44	−2.73**
Serial correlation (*SE*)		−0.14 (0.04)**

Note. Standard errors are adjusted for the cluster-sample design of the NELS data. The regression models also include a constant term (not shown) and are estimated with the NELS sampling weights. Intercept, period, urban/rural, and regional dummy variables are also included in the model.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

TABLE 3
NELS Achievement Regression Coefficients for 1990 Public School Students: School-Level Resources

	Coefficient (*SE*)	
	Reading	Math
School inputs		
Student-teacher ratio	−0.17 (0.19)	−0.47 (0.31)
Spline 10	0.13 (0.20)	0.52 (0.32)
Spline 25	0.14 (0.06)*	−0.12 (0.09)
Low teacher salary	0.00 (0.03)	−0.05 (0.04)
% Teachers: more than college	0.02 (0.40)	−0.02 (0.56)
% Black students	0.01 (0.01)	−0.00 (0.01)
% Hispanic students	0.01 (0.01)	−0.00 (0.01)
% Students in free lunch program	−0.01 (0.01)	−0.01 (0.01)
Father in household	0.21 (0.17)	0.32 (0.22)
Male respondent	0.07 (0.13)	−0.31 (0.19)
Asian	1.16 (0.33)**	0.24 (0.48)
Hispanic	−0.35 (0.29)	−0.82 (0.35)*
Black	−1.10 (0.31)**	−1.11 (0.39)**
American Indian	−0.70 (0.31)*	−0.50 (0.43)
Father's education		
High school	0.66 (0.24)**	1.04 (0.35)**
Less than college	0.72 (0.26)**	1.39 (0.36)**
College	0.82 (0.29)**	1.87 (0.38)**
College plus	1.04 (0.28)**	1.75 (0.39)**
Don't know	1.12 (0.28)**	0.88 (0.41)*
Mother's education		
High school	0.65 (0.24)**	−0.01 (0.38)
Less than college	0.82 (0.27)**	0.71 (0.42)$^{+}$
College	1.39 (0.32)**	0.53 (0.42)
College plus	1.35 (0.35)**	1.00 (0.44)*
Don't know	−0.09 (0.35)	0.08 (0.52)
Family income ($)		
10,000–15,000	0.34 (0.32)	0.05 (0.43)
15,000–20,000	0.82 (0.34)*	0.40 (0.45)
20,000–25,000	0.67 (0.34)*	0.03 (0.52)
25,000–35,000	0.85 (0.31)**	0.47 (0.4)
35,000–50,000	0.99 (0.31)**	0.37 (0.41)
50,000–75,000	0.72 (0.34)*	0.64 (0.43)
>75,000	0.51 (0.39)	0.42 (0.56)
Missing	0.54 (0.43)	0.71 (0.54)
1988 test score	0.96 (0.01)**	0.90 (0.01)**
N	7,837	7,845
R^2	0.77	0.65
Specification test results		
Ramsey test: *F* (*df*)	40.42 (3, 7795)**	53.48 (3, 7803)**
Johnson-McClelland test (normal distribution)	−2.12*	−3.08**
Serial correlation (*SE*)	−0.14 (0.02)**	−0.33 (0.03)**

Note. Standard errors are adjusted for the cluster-sample design of the NELS data. The regression models also include a constant term (not shown) and are estimated with the NELS sampling weights. Intercept, period, urban/rural, and regional dummy variables are also included in the model.
$^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$.

lated to these nonlinear functions of the school and family variables. The null hypothesis of no model misspecification would be rejected if an F statistic suggested that the squared, cubed, and quartic terms of the predicted test score were jointly significant (Ramsey & Schmidt, 1976).

As seen in the bottom rows of Tables 2 and 3, the Ramsey test rejected the null hypothesis of no misspecification at the 1% level for each of the models, regardless of whether inputs were measured at the classroom or school level. While the Ramsey test is not informative about the exact nature of the misspecification, the results nonetheless caution against taking the school effect estimates literally.

A second test for correlation between the explanatory variables in Equation 4 and the model's error term was derived from Johnson and McClelland (1997). Their test is based on the observation that if two random variables are independent, then the joint probability distribution will equal the product of the marginal densities; that is, $P(AB) = P(A) \times P(B)$. In terms of our value-added models, we defined the two probabilities as (a) the probability that the regression error is small and (b) the probability that the test score value predicted from the regression coefficients (which is a linear function of the explanatory variables) is small. Johnson and McClelland outlined a formal way of testing the proposition that these two events are independent; if the events are not independent, then the explanatory variables and the regression error term are correlated, and the underlying value-added model does not produce unbiased estimates for the effects of school resources on student test scores. As seen in the bottom rows of Tables 2 and 3, the Johnson-McClelland test rejected three of the four models at the 5% significance level.[6]

We also examined, using the test described in MacKinnon (1992), the necessary condition that the error terms in Equation 4 are serially uncorrelated. (We used the MacKinnon test because the usual Durban-Watson statistic is not accurate when a lagged dependent variable is included as an explanatory variable, as is the case with our value-added model.) We calculated the test using all three waves of the NELS data, estimating value-added models for achievement gains from 1988–1990 and 1990–1992 for every member of the sample, and retaining the regression residuals from the 1988–1990 achievement equation. We then reestimated the 1990–1992 value-added model with the 1988–1990 regression residual included as an explanatory variable. The null hypothesis of no serial correlation in the value-added error terms is rejected if the coefficient estimate for the lagged regression residual is statistically significant.

As seen in the last rows of Tables 2 and 3, the null hypothesis of no serial correlation was rejected at the 1% level in every equation. It is interesting to note that the serial correlation term was negative in each case, which implies that students who "underperform" in terms of what the regression model predicts in one period tend, on average, to "overperform" in the next period. This pattern could be explained by measurement error involving the NELS test scores, or by the efforts of parents or school administrators to provide compensatory resources or attention in response to a student's poor previous performance (or, alternatively, by parents or administrators becoming complacent as a result of previous success).

Since there is some variability in how previous studies specified Equation 4, we reestimated the value-added model in a variety of different ways. For example, we reestimated the model in log-log form to allow for additional nonlinearities and interaction terms between family and school variables. We also reestimated the model without peer group characteristics and without the NELS sampling weights. In each case, the coefficients for the classroom- or school-level input measures exhibited the same mixed pattern observed in Tables 2 and 3. More important, our specification tests consistently rejected the null hypothesis of no model misspecification for each of these alternative models.

The ultimate questions of interest are whether the biases detected by our model specification tests are large enough to matter for the purposes of education policy evaluation and whether unbiased estimates show that school resources affect student test scores. We address these points in the next section by reviewing several recent studies that used IV methods to address the problem of correlation between school or classroom resources and omitted variables. In most cases, the IV estimates revealed statistically significant, positive effects of school resources on student achievement. Most of the studies also showed larger IV estimates than those derived by applying ordinary least squares to the same data sources, which provides additional support for our concern that the biases found with value-added models in the NELS data may afflict other value-added studies as well. Moreover, the study findings suggest that these biases are large enough to substantially distort value-added estimates of school resource effects.

Instrumental Variables Studies

In this section, we review the results of studies using IV methods to address the omitted variables problems that appear to plague many value-added studies. As outlined in Table 4, most of these IV studies suggest that increases in school resources lead, on average, to increased student learning. Most IV studies also suggest that omitted variables bias estimates of school resource effects downward (toward zero) and that the magnitude of this bias is large enough to be relevant for evaluation and policy. We begin by reviewing the IV procedure and the key assumptions of unbiased estimation, and then discuss each of the IV studies described in Table 4 in some detail.

Instrumental Variables Estimation

IV methods seek to identify sources of variation in school resources that are unrelated to unobserved characteristics of students or their families and do not affect student learning other than by influencing measurable resources. The most common method of estimating IV models is through two-stage least squares analyses. For simplicity, assume that S_{iT}, the set of school resource measures of interest, contains only one variable.[7] The first stage of the two-stage least squares analysis (Equation 5) consisted of regressing S_{iT} against the other control variables (family and student variables [F_{iT}] and, if available, previous test scores [Y_{iT-1}]), as well as a set of instrumental variables, Z_{iT}. The coefficient estimates from

TABLE 4
Instrumental Variables Studies of Effects of School Resources on Student Outcomes

Study	Sample	Instrumental variable	Findings
Angrist and Lavy (1997)	Israeli classrooms in 1991 (2,053 for 4th grade & 2,024 for 5th grade) and 1992 (2,162 for 3rd grade)	Variation in school enrollment interacted with Israeli class size rules	Class size effects on math and reading scores for 5th graders, more modest effects on reading for 4th graders, no measurable effects for 3rd graders
Krueger (1997)	11,600 K–3 students in Tennessee STAR experiment, 1985–1986	Random assignment to class size "treatments"	Effects on test scores from initial assignment to small classes, modest gains thereafter
Hoxby (1998)	Connecticut school data for cohorts born 1965–1987	Population variation	No effects of class size on student test scores
Ludwig (1999)	9,500 students in National Education Longitudinal Study, 1988–1992[a]	State and federal aid interacted with school poverty rate	Modest effects of student-teacher ratio and teacher salary on math scores, less clearly for reading
Figlio (1997)	5,600 students in National Education Longitudinal Study, 1988–1990[b]	State property tax limitations	Tax limitations reduce school spending and student test scores
Card and Payne (1998)	Annual samples of 100,000 SAT takers, 1978–1992	School finance reforms	Modest positive effects of additional resources on SAT scores
Cullen (1997)	5,000 Texas schools, 1993–1995	Variation in school special education enrollment	Positive effect of school spending on student test scores

[a] All public school students in NELS who participated in each of the three survey waves between 1988 and 1992.
[b] 5,600 public school students in NELS who did not change schools between 1988 and 1990 and for whom there were no missing values with the covariates of interest.

Equation 5 were then used to calculate the level of school inputs predicted for each student by the covariates on the right-hand side of Equation 5; we denote this predicted value by $\hat{S}_{iT}$. The second stage of the procedure (Equation 6) consisted of substituting the predicted for the actual value of the student's school resources in the achievement equation.[8]

$$S_{iT} = c_0 + c_1 F_{iT} + c_2 Y_{iT-1} + c_3 Z_{iT} + v_{iT}. \quad (5)$$

$$Y_{iT} = d_0 + d_1 \hat{S}_{iT} + d_2 F_{iT} + d_3 Y_{iT-1} + u_{iT}. \quad (6)$$

The intuition behind the IV model is straightforward. The key empirical concern with the value-added model in Equation 4 stems from possible correlation between the regression error term (v_{iT}) and the school resource variables (S_{iT}). The two-stage least squares procedure isolates the variation in school resources (S_{iT}) that is due to variation in the instruments (Z_{iT}). If the instruments are uncorrelated with the error term in the achievement equation, the predicted value of $\hat{S}_{iT}$ will be uncorrelated with the error term, and the coefficient estimate for the effects of school resources on student achievement (d_1) will be unbiased.

Thus, the three necessary conditions for unbiased IV estimation are as follows. First, variation in the instruments (Z_{iT}) must be related to variation in the school inputs of interest (S_{iT}). Put differently, the instruments must be correlated with school resources even after we control for the other family and student variables in Equation 5. Second, the instruments cannot have a direct causal effect on student achievement other than by affecting the level of school inputs for the student. Third, the instruments cannot be correlated with whatever unobserved factors affect student achievement (see Angrist, Imbens, & Rubin, 1996, for an excellent discussion). If the instrumental variables are only weakly correlated with school resources, even a slight violation of the second or third conditions can lead to substantial bias in the resulting estimates (Bound, Jaeger, & Baker, 1995).

The second and third conditions are similar in a statistical sense, in that both require the instrument (Z_{iT}) to be uncorrelated with the error term for student achievement (u_{iT}). Yet distinguishing between the two conditions may be useful for researchers because many candidate instruments will arguably not have direct effects on achievement, but may be correlated with unobserved factors that affect student test scores. Consider, for example, the idea of using a student's region of residence (South, West, Midwest, or Northeast) as an instrument. There has traditionally been variation in school spending across regions of the United States, with spending in the South, in particular, lagging behind that of other regions (Armor, 1972). Moreover, the physical location of a child's family in a specific quadrant of the country is unlikely to have a direct causal effect on academic achievement. Thus, region seems to satisfy the first two conditions for valid instruments just described.

However, region is unlikely to meet the third condition for valid IV estimation, since there appear to be systematic population differences across regions of the country that are not well measured in most education data sets. For example, adults (including teachers) in the South have historically had lower test scores on average than those in other regions of the country (Armor, 1972), and most education data sets do not include test scores for teachers or parents, even though the cognitive skills of both groups are related to student achievement (Ferguson, 1991; Ferguson & Ladd, 1996; Phillips, Brooks-Gunn, Duncan, Klebanov, & Crane, 1998). Similarly, whatever cultural factors cause the South to have higher rates of violence than other regions (Butterfield, 1995) might plausibly be related to academic outcomes as well.

In sum, IV estimation requires that the analyst identify a variable that is correlated with the school- or classroom-level inputs a student receives but otherwise does not affect the student's achievement. One advantage of the IV method is that researchers are forced to identify the source of variation in school resources that is being related to variation in student test scores, and to argue why this source of variation meets the conditions just described. The validity of the instruments can then be judged according to basic statistical diagnostics,[9] social science theory, and common sense.

Instrumental Variables Studies of School Resources and Student Achievement

This section reviews the current IV literature on the effects of school resources on student achievement. Our review focuses on those IV studies in which the instruments are most likely to be valid.[10] Some readers may take exception with the instruments used in any individual IV study; taken together, however, these studies provide stronger evidence for a positive relationship between school inputs and student test scores than the previous value-added literature.

Angrist and Lavy (1997) estimated the effects of class size on the test scores of Israeli fourth and fifth graders in 1991 and third graders in 1992. As instruments, they used variation in school enrollments combined with a religious rule, dating back to the 12th century, that governs class sizes. When a school's classroom size is close to the cutoff specified by the rule, modest changes in school enrollment can trigger sharp reductions in average class sizes as schools add teachers to keep classrooms below the limit. Angrist and Lavy found that reductions in class size of approximately 35% (about 10.5 students per class) imply statistically significant increases in student test scores of 0.10 to 0.20 standard deviations for fifth graders, gains about half that size for fourth graders, and little gain for third graders. The authors attributed the lack of effect for third graders to across-the-board changes in instruction and testing conditions that occurred in Israel in 1992 in response to public dissatisfaction with the 1991 test results, which in turn served to reduce the variance in student test scores. Angrist and Lavy also found that ordinary least squares estimates either understated the effects of class size on test scores relative to IV models or produced estimated effects that were of the opposite sign.

Angrist and Lavy's estimates will be valid so long as parents and school administrators do not behave strategically to affect whether a given child's school is just over or under the class size cutoff. In practice, such strategic behavior is probably difficult for both parents and teachers, given the uncertainty about both the school's enrollment in a given year and how other parents will react to particular enrollment or class size patterns.[11]

The magnitudes of the Angrist and Lavy class size estimates are roughly the same as those of the estimates from the Tennessee STAR experiment (Finn & Achilles, 1990; Krueger, 1997). The well-known STAR experiment overcame the evaluation problem stemming from correlation between student characteristics and educational resources by randomly assigning 11,600 students in kindergarten through third grade to small classes, regular classes, or regular classes with a teacher aide. While it may be surprising to some readers to think of the STAR results as IV estimates, social science experiments such as Tennessee STAR in fact provide the ideal opportunity for IV estimation. The reason is that the variable indicating the treatment group to which the student is randomly assigned is by construction a valid instrument, one that is correlated with student class sizes (given the nature of the experiment) but uncorrelated with unobserved student variables that affect achievement (since assignment to classroom size treatment groups is random).

Analysis of the Tennessee STAR experiment reveals sizable effects on achievement during the first year that students were assigned to smaller classrooms, but the effects of each additional year in a small classroom were, at most, one quarter as large (Krueger, 1997). One implication of this finding is that estimates from value-added models, which focus on changes in test scores from one year to the next, will miss the initial gain from small classes if students are observed during subsequent years spent in such classrooms. At first glance, the Angrist and Lavy findings that the effects of smaller classes are larger for fourth and fifth graders than for third graders seem inconsistent with the STAR results showing that achievement gains are largest for kindergarten students and first graders (when most students were assigned to smaller classes). However, the Angrist and Lavy study relied on cross-sectional data for Israeli students, which did not allow determination of how long students had been in a classroom of a given size. As a result, it is not clear whether the Angrist and Lavy achievement estimates represent one-time or cumulative effects.[12]

In contrast to the Angrist and Lavy study and the Tennessee STAR experiment, Hoxby (1998) examined the effects of class sizes on fourth-, sixth-, and eighth-grade test scores in Connecticut school districts and found little evidence for class size effects on student learning. Her instrument was derived from departures from general population trends in Connecticut towns (which correspond to individual school districts in that state). Departures from population trends, in turn, affect class sizes through fluctuations in school enrollments. Angrist and Lavy (1997) hypothesized that the statistically insignificant effects of class size found in Hoxby's study may have been due in part to the already-low average class size in Connecticut (19 students, as compared with an average of more than 23 students per class in the nation as a whole [see Table 1]).

Ludwig (1999) examined the effects of student-teacher ratios and teacher salaries on test scores using the nationally representative NELS data set (described earlier). As instruments, he used changes in U.S. federal and state education spending over the 1988 to 1992 period. The share of total education spending covered by federal and state sources varies across states and, because such aid tends to target schools serving poor students, within states as well. Ludwig found that reductions in student-teacher ratios and increases in teacher salaries produce statistically significant increases in math scores but have less clear effects on reading. The estimated effects of student-teacher ratios on math scores were similar in magnitude to those from Angrist and Lavy and the STAR experiment.

Changes in the federal education assistance provided to a state will be a valid instrument if patterns of federal school spending are not motivated by the performance of students in the state. Some evidence to support this assumption comes from Mayer's (1991) study of federal defense appropriations, which suggests a very indirect relationship between local citizen preferences and congressional legislation, as well as substantial time lags associated with both the process through which Congress approves funds and the actual obligation of funds once approval is granted. Changes in state aid may be a valid instrument because the reductions in aid in many states during the 1988–1992 period were driven in part by court-mandated state spending on prisons (partly in response to prison overcrowding) and federally mandated expenditures on Medicaid (U.S. General Accounting Office, 1995) rather than by factors related to the outcomes of the state's students.

Cullen (1997) used school-level data from Texas for 1993–1995 and found that increases in school spending improve student achievement. Her instrument was derived from natural variations across areas in the number of students with physical disabilities. Since additional school spending on physically disabled students tends to "crowd out" spending on other students, the number of physically disabled students will be negatively correlated with regular education spending. Cullen's study differed from the IV studies described earlier in that the latter tested the hypothesis that increases in specific instructional spending items translate into additional student learning; in contrast, Cullen examined the joint hypothesis that schools spend additional funds on expenditure items that are relevant for student learning and are then, in turn, able to translate additional spending on these items into increased student learning. She found that IV estimates of the effects of school spending on student test scores were up to five times as large as ordinary least squares estimates.

Cullen's IV procedure assumed that parents of physically disabled students do not choose their residential location on the basis of the quality of local school services for disabled students.[13] If parents of disabled students move to school districts with higher quality services for disabled students, the nature of the bias will depend on why some schools have better programs for disabled students than other schools. Suppose, for example, that local voters who are unusually concerned about education choose to support programs for disabled children. If local voter concern about education translates into a school system of parents who are unusually involved in their children's schooling, then areas with high-quality programs for disabled students will have more students with disabilities (because of self-selection) and lower regular education spending (because of crowd out). Cullen's estimates would then understate the effects of spending on test scores (because regular education spending in this case is negatively correlated with unobserved measures of parents' involvement in their children's education).

Finally, Figlio (1997) used NELS data to present what can be thought of as indirect IV estimates. He first showed that state property tax limitations serve to limit school spending. He then regressed student test scores against indicators of whether a student's state has a property tax limitation (rather than against direct measures of the student's school resources). Figlio found that tax limits are associated with lower student test scores. Similarly, Card and Payne (1998) examined the relationship between school finance reforms and inequality in school spending and student SAT scores. They found evidence for a positive but weak effect of additional school spending on student test scores. Both studies provide useful information on school resource effects if state tax limits and school finance reforms are unrelated to other changes in state policies or population characteristics that may affect educational achievement.

These IV studies reveal the variety of approaches that can be used to identify the effects of school resources. The instruments in any IV study are always open to question, with the exception of those used in the Tennessee STAR study, which are valid by virtue of the program's randomized experimental design. The studies described here all used instruments that are plausibly valid and, taken together, provide some evidence that additional resources may improve student achievement. The findings also suggest that the biases revealed with value-added models applied to the NELS data are apparently large enough to matter, since ordinary least squares analysis typically produces smaller (or even negative) estimates than IV estimates obtained from the same data.

Discussion

To date, the question of how additional resources affect student learning has largely been answered by meta-analyses of the conflicting value-added literature. The results presented here are consistent with the idea that the mixed results shown in previous studies may be due to omitted variables problems.

We have shown that value-added models cannot produce unbiased estimates in the NELS data set, which is true across a wide range of choices about model specification, functional form, and sample weighting decisions. Given the large number of data sets and model specifications that have been used to estimate value-added models in the literature, even an exhaustive application of different models to a single data set such as NELS cannot conclusively prove that all previous studies were biased. Moreover, the NELS has several important limitations; for example, classroom-level information is available for only a subset of each student's classes and teachers, and students were tested only at 2-year intervals. Nevertheless, the NELS is arguably among the best micro data sets available for this application, with rich student, family, school, and even classroom information provided at the student level, taken from surveys of parents and school administrators as well as students. Our findings cannot demonstrate that every previous value-added study was biased, but they do suggest that perhaps less weight should be given to the value-added literature in future debates over education policy. At the very least, future value-added research should provide the results of the basic diagnostic tests used here, which might reveal whether the conditions for unbiased estimation are met in the particular data set being used.

We have also shown that more recent studies using IV methods provide somewhat stronger support than the value-added literature for the idea that additional resources translate, on average, to improved student learning. While the instruments used in IV studies are always open to question, the similarity between estimates from the Tennessee STAR experiment and several of the IV studies provides some support for the idea that many of these instruments are valid. On the other hand, several of the IV studies suggest weak or no effects of school resources on certain educational outcomes. Taking the full set of IV results together with our NELS results, we conclude that the biases we have identified with the value-added studies are large enough to be of substantive significance and typically lead to understated school resource effects.

While the recent IV literature on the effects of school spending is far from definitive, the IV technique may be a useful tool to overcome the omitted variables problems that may plague value-added methods. The IV method relies on identifying variables that are correlated with the educational resources that students receive but are otherwise uncorrelated with student achievement. Creative application of IV estimation may help researchers identify the average effects of school resources on achievement and the conditions under which resources can be used most effectively.

We end by noting that even definitive evidence showing that additional school expenditures raise student test scores is not sufficient to conclude that increasing education budgets or expenditures on specific instructional items is good public policy. If the objective of government policy is to obtain the highest level of student learning for a given level of expenditure, cost-effectiveness analysis should be used to determine which instructional expenditures (such as reductions in class sizes versus increases in teacher salaries) are able to achieve a given increase in student test scores at the lowest possible cost. Moreover, other government interventions that compete with public schools for resources may affect test scores as well as other socially desirable outcomes such as increased future earnings, decreased criminal involvement, and improved health. Determining which programs produce the largest gains to society for a given expenditure will require that all outcomes be measured in a common metric that can be compared with costs (e.g., dollars). Combining unbiased estimates of school resource effects with cost-benefit analysis could help guide policy decisions about allocating resources across competing uses such as school funding, expansions in the coverage or intensity of early childhood interventions, parenting classes or other efforts to enhance the cognitive stimulation that children receive within the home, and dropout prevention programs (Donohue & Siegelman, 1998; Jencks & Phillips, 1998; Levin, 1989).

Notes

We thank Duncan Chaplin, Charles Clotfelter, Jeffrey Conte, Paul Harrison, Christopher Jencks, Richard Murnane, Steve Pischke, Eric Ralph, Leslie Whittington, and seminar participants at the Bureau of Labor Statistics, the University of Chicago, Georgetown University, Northern Illinois University, and the 1997 Association for Public Policy Analysis and Management meetings for valuable comments. We also thank Heath Einstein, Brian Komar, Marisa De La Cruz, and Theresa Luhm for valuable assistance. This paper was written in part while the first author was visiting scholar at the Northwestern University/University of Chicago Joint Center for Poverty Research and National Academy of Education/Spencer Foundation postdoctoral fellow in education research. Any errors in fact or interpretation are our own.

[1]Goldhaber and Brewer (1997) showed that correlation between any school-specific unobserved variables and school resource measures (S_{iT}) is unlikely to produce substantial bias in estimates of the effects of school resources. This finding does not mean that unobserved family- or student-level variables are unimportant sources of bias in estimating school resource effects.

[2]For example, schools frequently assign students to remedial or advanced classes on the basis of prior achievement, and both classroom size and teacher attributes may systematically vary across classes by academic level.

[3]Meyer (1992) discussed two strategies for addressing measurement error in previous test scores. The first method adjusts the results using a measure of the variance of the measurement error. The second method is two-stage least squares; in Meyer's application examining the effects of math courses on math achievement, he used previous math courses as instrumental variables to predict the previous period's test scores.

[4]Excluded from the NELS sample in 1988 were students with mental handicaps, physical or emotional problems, and inadequate command of the English language. In most cases, 24 of the 26 students per school included in NELS were randomly sampled, while the other 2 students were selected from among Hispanic and Asian Islander students (U.S. Department of Education, 1994).

[5]Thanks to Richard Murnane for pointing out this potential problem.

[6]As a minor technical point, the Johnson-McClelland test assumes that the regression error terms are identically and independently distributed. This assumption is technically not met with the NELS data because the cluster sample design may have introduced some correlation in the error terms of students within the same school. In practice, this nonindependence has little effect. We examined the sensitivity of our calculations to violations of the error-independence assumption by randomly selecting only one student per NELS school. We found that the Johnson-McClelland test statistics differed from those reported in Tables 2 and 3 only because of the reduction in sample size.

[7]Also of interest is the case in which there are multiple school input variables of interest that may be correlated with the unobserved variables that affect achievement. In this case, the procedure is the same as described here, but predicted values are calculated separately for each of the school input variables in first-stage regressions. The key requirement for the two-stage least squares analysis is that the number of instrumental variables in Z_{iT} must at least equal the number of school input variables for which predicted values are calculated.

[8]As a minor technical point, the standard errors obtained from estimating the second regression must be adjusted to account for the inclusion of the predicted value as a covariate. The adjustment consists of dividing the standard errors obtained from estimating Equation 6 using the actual rather than predicted school resource variable by the standard errors obtained from estimating Equation 6 with the predicted rather than actual school resources (e.g., see Gujarati, 1988, or Greene, 1993).

[9]Bound, Jaeger, and Baker (1995) suggested that readers report the partial R^2 and F statistics to describe how much explanatory power the instrumental variables have in the first-stage regression. As for the second and third conditions for valid IV estimation, if we assume that the effects of additional resources are equal for all students (i.e., a "constant treatment effect") and if the analyst has more instrumental variables than potentially confounded school input variables that must be measured, then these conditions can be tested by essentially regressing the residuals from the second-stage regression against the instruments (see Hausman, 1983, or Newey, 1985, for details). If the treatment effect is assumed to vary across students, interpretation of the IV estimate and the Hausman/Newey diagnostic tests becomes complicated (see Angrist, Imbens, & Rubin, 1996).

[10]Three other IV studies providing evidence that additional resources improve student test scores are not included in our review because of concerns with the instruments. Akerhielm (1995) used NELS data and total school enrollment as an instrument, although school enrollment does not appear to be a valid instrument (as noted by Angrist & Lavy, 1997). Boozer and Rouse (1995) also used the NELS to estimate IV models with state special education policies as instruments, although overidentification tests reject the null hypothesis that these instruments are valid. Ferguson and Ladd (1996) applied IV models to district-level data from Alabama and found evidence to suggest that increases in spending produce substantial gains in achievement. However, their district-level instruments (per capita income and property value per student) may have affected student achievement both directly (since they reflected the average family background of the students) and indirectly

(through possible peer or neighborhood effects).

[11]For example, consider a family whose child is in a school with an enrollment level such that class sizes are just below the cutoff that triggers reductions to much smaller classes. If the parents believe that other parents will respond to this situation by moving to other schools with smaller classes, our hypothetical family will stay in place since the departure of other families will reduce classes in the current school. On the other hand, if the parents believe that other families will stay despite the current enrollment situation, our family may be more likely to leave in search of a school with smaller classes.

[12]For example, consider a case in which each student's class size is identical during every year she or he is in school. In this case, students who were in small classes during fifth grade were also in small K–4 classes. The observed test score differences in fifth grade in our example would be equally consistent with a one-time small-class gain, as in STAR, or more modest cumulative gains. Since the class sizes of students are unlikely to be perfectly correlated across years of schooling, many other combinations of one-time and cumulative effects are also possible.

[13]Thanks to an anonymous referee for this observation.

References

Akerhielm, K. (1995). Does class size matter? *Economics of Education Review, 14,* 229–241.

Alvarez, L. (1999, March 9). Senate education bill stalls despite its bipartisan support. *New York Times*, p. A16.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–455.

Angrist, J. D., & Lavy, V. (1997). *Using Miamonides' rule to estimate the effect of class size on scholastic achievement* (Working Paper 5888). Cambridge, MA: National Bureau of Economic Research.

Armor, D. J. (1972). School and family effects on Black and White achievement: A reexamination of the USOE data. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 168–229). New York: Vintage Books.

Barnett, W. S. (1996). Economics of school reform: Three promising models. In H. F. Ladd (Ed.), *Holding schools accountable* (pp. 299–326). Washington, DC: Brookings Institution.

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education, 52,* 113–121.

Boozer, M., & Rouse, C. (1995). *Intraschool variation in class size: Patterns and implications* (Working Paper 5144). Cambridge, MA: National Bureau of Economic Research.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association, 90,* 443–450.

Butterfield, F. (1995). *All God's children: The Bosket family and the American tradition of violence*. New York: Avon Books.

Card, D., & Payne, A. A. (1998). *School finance reform, the distribution of school spending, and the distribution of SAT scores* (Working Paper 6766). Cambridge, MA: National Bureau of Economic Research.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cullen, J. B. (1997). *New evidence on the impact of differential expenditures on student performance* (Working Paper). Ann Arbor: University of Michigan.

Donohue, J. J., & Siegelman, P. (1998). Allocating resources among prisons and social programs in the battle against crime. *Journal of Legal Studies, 27,* 1–43.

Duncan, G. J. (1994). Families and neighbors as sources of disadvantage in the schooling decisions of White and Black adolescents. *American Journal of Education, 103,* 20–53.

Evans, W., Oates, W., & Schwab, R. (1992). Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy, 100*, 966–991.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28,* 465–497.

Ferguson, R. F., & Ladd, H. F. (1996). How and why money matters: An analysis of Alabama schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 265–298). Washington, DC: Brookings Institution.

Figlio, D. (1997). Did the 'tax revolt' reduce school performance? *Journal of Public Economics, 65,* 245–269.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size. *American Educational Research Journal, 27,* 557–577.

Goldberger, A. S., & Cain, G. (1982). The causal analysis of cognitive outcomes in the Coleman, Hoffer and Kilgore Report. *Sociology of Education, 55,* 103–122.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources, 32,* 505–523.

Greene, W. H. (1993). *Econometric analysis* (2nd ed.). New York: Macmillan.

Gujarati, D. N. (1988). *Basic econometrics* (2nd ed.). New York: McGraw-Hill.

Hanushek, E. A. (1996). School resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 43–73). Washington, DC: Brookings Institution.

Hanushek, E. A., & Taylor, L. L. (1990). Alternative assessments of the performance of schools. *Journal of Human Resources, 25,* 179–201.

Hausman, J. A. (1983). Specification and estimation of simultaneous equation models. In Z. Griliches & M. C. Intriligator (Eds.), *Handbook of econometrics* (Vol. 1, pp. 391–448). Amsterdam: North-Holland.

Heckman, J. J., Layne-Farrar, A., & Todd, P. (1996a). Human capital pricing equations with an application to estimating the effect of schooling quality on earnings. *Review of Economics and Statistics, 78,* 562–610.

Heckman, J. J., Layne-Farrar, A., & Todd, P. (1996b). Does measured school quality really matter? An examination of the earnings-quality relationship. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 192–290). Washington, DC: Brookings Institution.

Hedges, L., & Greenwald, R. (1996). Have times changed? The relation between school resources and student performance. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 74–92). Washington, DC: Brookings Institution.

Hedges, L., Laine, R., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.

Hoxby, C. M. (1998). *The effects of class size and composition on student achievement: New evidence from natural population variation* (Working Paper). Cambridge, MA: Harvard University.

Jencks, C., & Mayer, S. (1990). The social consequences of growing up in a poor neighborhood. In M. McGeary & L. Lynn (Eds.), *Inner city poverty in the United States* (pp. 111–186). Washington, DC: National Academy Press.

Jencks, C., & Phillips, M. (1998). The Black-White test score gap: An introduction. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 1–54). Washington, DC: Brookings Institution.

Johnson, D., & McClelland, R. (1997). Nonparametric tests for the independence of regressors and disturbances as specification tests. *Review of Economics and Statistics, 79,* 335–340.

Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B, 1,* 1–37.

Krueger, A. B. (1997, May). *Experimental estimates of education production functions.* Paper presented at the annual meeting of the Society of Labor Economists, Washington, DC.

Levin, H. M. (1989). Economics of investment in educationally disadvantaged students. *American Economic Review, 79*(2), 52–56.

Ludwig, J. (1999). *School spending and student achievement: New evidence from longitudinal data* (Working Paper). Washington, DC: Georgetown University.

Ludwig, J., Duncan, G. J., & Hirschfield, P. (1999). *Urban poverty and juvenile crime: Evidence from a randomized housing-mobility experiment* (Working Paper). Washington, DC: Georgetown University.

MacKinnon, J. (1992). Model specification tests and artificial regressions. *Journal of Economic Literature, 30,* 102–146.

Mayer, K. R. (1991). *The political economy of defense spending.* New Haven: Yale Press.

Mayer, S. (1997). *What money can't buy.* Cambridge, MA: Harvard University Press.

Meyer, R. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency* (Discussion Paper 966-92). Madison: Institute for Research on Poverty, University of Wisconsin.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics, 32,* 385–397.

Murnane, R. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R. (1991). *Who will teach? Policies that matter.* Cambridge, MA: Harvard University Press.

Murnane, R., & Levy, F. (1996). Evidence from fifteen schools in Austin, Texas. In G. Burtless (Ed.), *Does money matter? The effect of school resources on student achievement and adult success* (pp. 93–96). Washington, DC: Brookings Institution.

Pallas, A. M., & Alexander, K. L. (1983). Sex differences in qualitative SAT performance: New evidence on the differential coursework hypothesis. *American Educational Research Journal, 20,* 165–182.

Phillips, M., Brooks-Gunn, J., Duncan, G. J., Klebanov, P., & Crane, J. (1998). Family background, parenting practices, and the Black-White test score gap. In C. Jencks & M. Phillips (Eds.), *The Black-White test score gap* (pp. 103–148). Washington, DC: Brookings Institution.

Ramsey, J. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B, 31,* 350–371.

Ramsey, J., & Schmidt, P. (1976). Some further results on the use of OLS and BLUS residuals in specification error tests. *Journal of the American Statistical Association, 71,* 389–390.

Smith, T. M. (1996). *The condition of education 1996.* Washington, DC: U.S. Department of Education.

Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64,* 416–424.

U.S. Department of Education, National Center for Education Statistics. (1994). *NELS:88 second follow-up user's guide, student sample.* Washington, DC: U.S. Government Printing Office.

U.S. General Accounting Office. (1995). *School finance: Trends in U.S. education spending*. Washington, DC: U.S. Government Printing Office.

Authors

JENS LUDWIG is an assistant professor of public policy, Georgetown Public Policy Institute, Georgetown University, 3600 N Street, NW, Suite 200, Washington, DC 20007. He specializes in education policy, juvenile delinquency, and neighborhood and peer group effects.

LAURIE J. BASSI is Vice President, American Society for Training and Development, 1640 King Street, Box 1443, Alexandria, VA 22313. She specializes in employer-provided education and training, measurement methodology, and return-on-investment analysis.

Manuscript received March 30, 1999
Revision received July 9, 1999
Accepted July 28, 1999

LINKED CITATIONS

You have printed the following article:

The Puzzling Case of School Resources and Student Achievement
Jens Ludwig; Laurie J. Bassi
Educational Evaluation and Policy Analysis, Vol. 21, No. 4. (Winter, 1999), pp. 385-403.
Stable URL:
http://links.jstor.org/sici?sici=0162-3737%28199924%2921%3A4%3C385%3ATPCOSR%3E2.0.CO%3B2-8

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

Notes

[1] **Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity**
Dan D. Goldhaber; Dominic J. Brewer
The Journal of Human Resources, Vol. 32, No. 3. (Summer, 1997), pp. 505-523.
Stable URL:
http://links.jstor.org/sici?sici=0022-166X%28199722%2932%3A3%3C505%3AWDSATS%3E2.0.CO%3B2-%23

[9] **Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak**
John Bound; David A. Jaeger; Regina M. Baker
Journal of the American Statistical Association, Vol. 90, No. 430. (Jun., 1995), pp. 443-450.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199506%2990%3A430%3C443%3APWIVEW%3E2.0.CO%3B2-U

[9] **Identification of Causal Effects Using Instrumental Variables**
Joshua D. Angrist; Guido W. Imbens; Donald B. Rubin
Journal of the American Statistical Association, Vol. 91, No. 434. (Jun., 1996), pp. 444-455.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199606%2991%3A434%3C444%3AIOCEUI%3E2.0.CO%3B2-O

References

NOTE: *The reference numbering from the original has been maintained in this citation list.*

LINKED CITATIONS

Identification of Causal Effects Using Instrumental Variables
Joshua D. Angrist; Guido W. Imbens; Donald B. Rubin
Journal of the American Statistical Association, Vol. 91, No. 434. (Jun., 1996), pp. 444-455.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199606%2991%3A434%3C444%3AIOCEUI%3E2.0.CO%3B2-O

Using Panel Data to Improve Estimates of the Determinants of Educational Achievement
Anthony E. Boardman; Richard J. Murnane
Sociology of Education, Vol. 52, No. 2. (Apr., 1979), pp. 113-121.
Stable URL:
http://links.jstor.org/sici?sici=0038-0407%28197904%2952%3A2%3C113%3AUPDTIE%3E2.0.CO%3B2-X

Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak
John Bound; David A. Jaeger; Regina M. Baker
Journal of the American Statistical Association, Vol. 90, No. 430. (Jun., 1995), pp. 443-450.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199506%2990%3A430%3C443%3APWIVEW%3E2.0.CO%3B2-U

Allocating Resources among Prisons and Social Programs in the Battle against Crime
John J. Donohue III; Peter Siegelman
The Journal of Legal Studies, Vol. 27, No. 1. (Jan., 1998), pp. 1-43.
Stable URL:
http://links.jstor.org/sici?sici=0047-2530%28199801%2927%3A1%3C1%3AARAPAS%3E2.0.CO%3B2-5

Families and Neighbors as Sources of Disadvantage in the Schooling Decisions of White and Black Adolescents
Greg J. Duncan
American Journal of Education, Vol. 103, No. 1. (Nov., 1994), pp. 20-53.
Stable URL:
http://links.jstor.org/sici?sici=0195-6744%28199411%29103%3A1%3C20%3AFANASO%3E2.0.CO%3B2-I

Measuring Peer Group Effects: A Study of Teenage Behavior
William N. Evans; Wallace E. Oates; Robert M. Schwab
The Journal of Political Economy, Vol. 100, No. 5. (Oct., 1992), pp. 966-991.
Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28199210%29100%3A5%3C966%3AMPGEAS%3E2.0.CO%3B2-A

LINKED CITATIONS

Answers and Questions about Class Size: A Statewide Experiment
Jeremy D. Finn; Charles M. Achilles
American Educational Research Journal, Vol. 27, No. 3. (Autumn, 1990), pp. 557-577.
Stable URL:
http://links.jstor.org/sici?sici=0002-8312%28199023%2927%3A3%3C557%3AAAQACS%3E2.0.CO%3B2-4

The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer and Kilgore Report
Arthur S. Goldberger; Glen G. Cain
Sociology of Education, Vol. 55, No. 2/3. (Apr. - Jul., 1982), pp. 103-122.
Stable URL:
http://links.jstor.org/sici?sici=0038-0407%28198204%2F07%2955%3A2%2F3%3C103%3ATCAOCO%3E2.0.CO%3B2-W

Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity
Dan D. Goldhaber; Dominic J. Brewer
The Journal of Human Resources, Vol. 32, No. 3. (Summer, 1997), pp. 505-523.
Stable URL:
http://links.jstor.org/sici?sici=0022-166X%28199722%2932%3A3%3C505%3AWDSATS%3E2.0.CO%3B2-%23

Alternative Assessments of the Performance of Schools: Measurement of State Variations in Achievement
Eric A. Hanushek; Lori L. Taylor
The Journal of Human Resources, Vol. 25, No. 2. (Spring, 1990), pp. 179-201.
Stable URL:
http://links.jstor.org/sici?sici=0022-166X%28199021%2925%3A2%3C179%3AAAOTPO%3E2.0.CO%3B2-Y

Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings
James Heckman; Anne Layne-Farrar; Petra Todd
The Review of Economics and Statistics, Vol. 78, No. 4. (Nov., 1996), pp. 562-610.
Stable URL:
http://links.jstor.org/sici?sici=0034-6535%28199611%2978%3A4%3C562%3AHCPEWA%3E2.0.CO%3B2-B

NOTE: *The reference numbering from the original has been maintained in this citation list.*

LINKED CITATIONS

An Exchange: Part I: Does Money Matter? A Meta-Analysis of Studies of the Effects of Differential School Inputs on Student Outcomes
Larry V. Hedges; Richard D. Laine; Rob Greenwald
Educational Researcher, Vol. 23, No. 3. (Apr., 1994), pp. 5-14.
Stable URL:
http://links.jstor.org/sici?sici=0013-189X%28199404%2923%3A3%3C5%3AAEPIDM%3E2.0.CO%3B2-U

Nonparametric Tests for the Independence of Regressors and Disturbances as Specification Tests
David Johnson; Robert McClelland
The Review of Economics and Statistics, Vol. 79, No. 2. (May, 1997), pp. 335-340.
Stable URL:
http://links.jstor.org/sici?sici=0034-6535%28199705%2979%3A2%3C335%3ANTFTIO%3E2.0.CO%3B2-Y

Economics of Investment in Educationally Disadvantaged Students
Henry M. Levin
The American Economic Review, Vol. 79, No. 2, Papers and Proceedings of the Hundred and First Annual Meeting of the American Economic Association. (May, 1989), pp. 52-56.
Stable URL:
http://links.jstor.org/sici?sici=0002-8282%28198905%2979%3A2%3C52%3AEOIIED%3E2.0.CO%3B2-C

Model Specification Tests and Artificial Regressions
James G. MacKinnon
Journal of Economic Literature, Vol. 30, No. 1. (Mar., 1992), pp. 102-146.
Stable URL:
http://links.jstor.org/sici?sici=0022-0515%28199203%2930%3A1%3C102%3AMSTAAR%3E2.0.CO%3B2-K

Sex Differences in Quantitative SAT Performance: New Evidence on the Differential Coursework Hypothesis
Aaron M. Pallas; Karl L. Alexander
American Educational Research Journal, Vol. 20, No. 2. (Summer, 1983), pp. 165-182.
Stable URL:
http://links.jstor.org/sici?sici=0002-8312%28198322%2920%3A2%3C165%3ASDIQSP%3E2.0.CO%3B2-A

LINKED CITATIONS

Some Further Results on the Use of OLS and BLUS Residuals in Specification Error Tests
James B. Ramsey; Peter Schmidt
Journal of the American Statistical Association, Vol. 71, No. 354. (Jun., 1976), pp. 389-390.
Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197606%2971%3A354%3C389%3ASFROTU%3E2.0.CO%3B2-M

A Pure Theory of Local Expenditures
Charles M. Tiebout
The Journal of Political Economy, Vol. 64, No. 5. (Oct., 1956), pp. 416-424.
Stable URL:
http://links.jstor.org/sici?sici=0022-3808%28195610%2964%3A5%3C416%3AAPTOLE%3E2.0.CO%3B2-P

NOTE: *The reference numbering from the original has been maintained in this citation list.*