Educational Evaluation and Policy Analysis
Spring 1995, Vol. 17, No. 1, pp. 113-131

The Challenge of Developing New Educational Indicators

J. Douglas **Willms**
University of British Columbia
and University of Edinburgh
Alan C. Kerckhoff
Duke University

Proposals for national assessments of educational performance provide opportunities to develop valuable social indicators, but they also pose important problems of methodology and interpretation. Researchers in the U.S. have emphasized the importance of analyzing and reporting indicator data at the school level, mainly because much of the variation in schooling outcomes is at ***the school level, and because findings at*** this ***level are particularly relevant to policy and practice. We do not refute this, but contend that indicator data should be analyzed and reported also at state and district levels. We present an analysis of British data that shows there is important variation at higher levels of the system, and that it is possible to examine whether this variation is related to national or district-level policies. The article discusses the issues involved in conducting analyses at higher levels. We identify three kinds of indicators: gross productivity, net productivity, and inequality. Our analysis indicates the degree of interdistrict variation in Great Britain on all of these measures and the extent to which organizational features of the districts' programs help explain that variation. Recommendations are made regarding the necessary features of any educational assessment if it is to produce useful social indicators.***

Moves toward establishing a national assessment of education are generated by many forces. The current state of the economy and the sense of a loss in international economic competition have been important contributors in both Great Britain and the United States. The shift from a manufacturing to a service economy has exposed the limited skills of many in the labor force, and the educational system has been identified as the source of those limitations. In Great Britain, a major concern is the relatively small proportion of 16- to 19-year-olds who are in full-time education or training' ("Teenagers' Charter," 1991). In the United States, invidious international comparisons of academic achievement have also contributed to the sense of educational inadequacy. Not only has the United States fallen back in international economic competition, but the evidence also has been increasingly clear that American youth do not fare very well in international comparisons of academic achievement (McKnight et al., 1987; Suter, 1988).

On March 31, 1994, President Clinton signed into law the Goals 2000: Education America Act. In his transmittal letter to Congress, he stated, "This legislation strives to support States, local communities, schools, businesses and industry, and labor in reinventing our education system so that all Americans can reach internationally competitive standards, and our Nation can reach the National Education Goals" (1993). The legislation established in law the National Education Goals Panel to build public support for the eight National Education Goals, and to monitor progress toward those goals. The legislation also established a council to evalu-

ate standards and assessment systems submitted by state or professional groups. The submission of plans to the council is voluntary.

When Goals 2000 was enacted, nearly every state had some form of monitoring system. Previous efforts to monitor educational performance at the national level had induced many states to develop their own state-level monitoring systems (Selden, 1988). To some extent this was a response by state-level officials who were dissatisfied with the national system, particularly the state-by-state comparisons made for the national "wall chart" (Smith, 1988). The emphasis of assessment at the state level has been to use indicator data to make comparisons among schools, or to diagnose discrepancies between the intended and achieved curriculum (Porter, 1991; Richards, 1988). In South Carolina statewide data have been used to reward schools that are superior in student achievement gains and in student and teacher attendance (Richards & Sheu, 1992).

Researchers and district-level administrators also have emphasized the importance of analyzing and reporting data at the school level. This emphasis can be attributed to at least three factors. One is that analyses examining the effects of factors at higher levels of the system, such as the effects of district expenditures, have generally been unsuccessful (e.g., see Hanushek, 1989). Another factor is that school-site administrators in the United States have a relatively high degree of autonomy (Meyer & Rowan, 1988), and recent reforms have emphasized further devolvement to the local level. The third factor is that researchers who have been discontent with the input-output production function approach to the study of schooling have made a convincing case that what matters most is how teachers and school-site administrators mobilize schooling resources in their particular context (e.g., Fullan, 1992; Hammersley, 1985; Hargreaves & Hopkins, 1991).

We do not decry this shift of emphasis to the school level. However, we maintain that district- and state-level analyses are important also to the understanding of educational policy and practice, and that researchers may have dismissed too early the importance of effects at these higher levels. A number of policies and practices at district and state levels affect the distribution of student outcomes. For example, school choice programs are often implemented at the state or district level, with consequences for schools that win or lose pupils (Plank, Schiller, Schneider, & Coleman, 1992; Witte, 1992). Desegregation programs and the associated litigation are largely district matters. Levels of funding and teachers' salaries, which directly affect pupil-teacher ratios, are generally determined by state- or district-level policies.

The Goals 2000: Education America Act uses the term *community*, which could be interpreted as meaning *school community*, but there are good reasons to interpret it more broadly. The first five of the National Education Goals concern the educational outcomes of *all* American children, and in most respects they address the needs of what Secretary of Education Riley called the "neglected majority (the 75% who don't go on to a four-year college)" (U.S. Department of Education, 1994, p. 1). And success in achieving all eight of the goals requires that all schools meet some minimum standard. But if state or district assessment systems emphasize comparisons among schools in their outcomes, without attention to how outcomes are distributed in communities defined more broadly, they may encourage an unhealthy competition amongst schools that results in their competing for students and teaching resources. So far as this is a zero-sum game, with some schools' gains balanced by other schools' losses, the assessment efforts may be another factor that engenders greater variation among schools in their outcomes. If so, the assessment programs would be counterproductive in enabling communities to achieve the National Education Goals.

Our purpose here is to point to the need for careful consideration of how data at higher levels of the system are analyzed and reported, and to present an example of the kinds of social indicators that can be developed and the ways in which they can be used. Specifically, we: (a) outline some of the kinds of potentially valuable social indicators that can be derived from a state or national assess-

ment, (b) demonstrate the different uses to which indicators can be put and how misleading some indicators can be if used improperly, (c) show how resource levels and the organizational characteristics of districts affect the assessment outcomes, and (d) describe the kinds of information needed if an assessment is to be maximally useful.

Our analysis employs multilevel modeling, a statistical technique that is appropriate for analyzing data that describe hierarchical organizations. (See Bryk & Raudenbush, 1992, and Goldstein, 1987, for a discussion of the methods.) Our approach is not novel; earlier research has demonstrated the need to adequately control for students' prior achievement (Boardman & Murnane, 1980; Lau, 1979) and to take into account the hierarchical structure of the data when estimating the effects of schools (Raudenbush & Bryk, 1986). But despite the improvements in statistical methods, even states such as South Carolina and California, which administer two of the most comprehensive monitoring programs, use statistical methods that do not take account of the multilevel structure of the data. These techniques lead to biased estimates of the effects of schools (Raudenbush & Bryk, 1989; Raudenbush & Willms, in press). South Carolina's system, for example, classifies schools into "comparison bands" based on the percentage of students in the school that met the readiness standard on a test of children's cognitive ability, the percentage of students who were eligible for the free lunch program, and two indicators of school resources (Richards & Sheu, 1992). Schools are then compared within bands on student and teacher attendance and on students' achievement scores adjusted statistically for their previous year's scores. The California system is similar, except that its banding system is more elaborate, and its comparisons do not take account of students' prior academic achievement (California Department of Education, 1992; Fetler, 1989). Because multilevel techniques are more complicated than those typically used in state assessment systems, a secondary purpose of this article is to demonstrate their use in developing indicators following a step-by-step approach, and to identify their advantages. We present an analysis of English and Welsh data as a means of highlighting some central issues and of illustrating both the potential value and possible inadequacies of some of the indicators. The analysis makes clear how important proper design and measurement are to conducting a meaningful educational assessment. We use data from the United Kingdom because they are public, cover the entire national system, and allow for the development of district-level indicators that include controls for prior academic achievement. Some state assessment systems, although not maintained for public use, meet the latter criteria. We expect, therefore, that our example can be transferred easily to the American context.

Assessment Criteria

The primary impetus to establishing monitoring systems at all levels is the belief that keeping score will directly or indirectly generate increased efforts to improve the educational system. National and statewide assessments can serve at least three quite different functions, however, and it is important to distinguish among them. They have very different educational and political implications, and they call for different kinds of information.

At the broadest level, assessment at higher levels provides the basis for a continuing record of progress if the procedure is kept in place over time. This is the most obvious and most frequently cited purpose. Assessment data are also inevitably used to make comparisons among various subunits (states, districts, schools), and such comparisons are almost certain to affect administrative and policy decisions. Finally, assessment provides the basis for evaluating the effectiveness of particular kinds of educational programs. Local districts (in both the United States and Great Britain) operate with varied arrangements, and the programs that characterize the more effective districts are bound to receive widespread attention.

But if, over time, interdistrict and programmatic comparisons are to be made, what should be compared? What are the appropriate assessment criteria? What do we mean by an *effective* program? If we seek improve-

ment over time, how should we define ***improvement?***

Much of the discussion of assessment in both the United States and Great Britain implies that the answer is obvious, but it actually involves many complex issues, only some of which can be dealt with in this article. We will not attempt to deal with issues such as the use of standardized tests *versus* other forms of assessment, or the political and social issues surrounding the use of performance indicators. (See Willms, 1992, for a broader discussion of these issues). Much of the discussion of flaws in the American educational system reflects a concern about the poor showing of American pupils in international comparisons. Those comparisons are based on pupils' scores on standardized achievement tests, and the goal for many is to increase American pupil performance on such tests. That may well be too narrow a focus. A major theme in the literature on social indicators is that the indicators we have usually used are too narrowly concerned with bottom-line variables (e.g., gross national product) and not sufficiently concerned with either the processes involved in generating the bottom line or the subjective aspects of social life (Land, 1983). Thus, in assessing the "success" of our educational system we might well want to include measures of such things as curricular offerings, teacher qualifications, parent participation in school programs, pupil-teacher ratios, pupil assignment practices, or the degree of pupil, teacher, and parent satisfaction with the system. Such indicators may help to explain ***why*** some schools or districts perform better than others, or assess the impact of state or district reforms (Murnane, 1987). They can serve also to describe the educational opportunities provided by schools, independent of the amount of student learning (Porter, 1991). We also might want to broaden the scope of our outcome measures to include nonacademic outcomes such as self-esteem, physical fitness, or the quality of school life. Oakes (1989) suggests that a broader set of indicators may influence educators to place greater emphasis on the wider goals of schooling.

Even if we limit our focus to measures of pupil achievement, however, there are multiple possibilities. For instance, one current debate is between those favoring performance-based measures and those who favor traditional achievement tests (Shepard, 1991; Wolf, Bixby, Glenn, & Gardner, 1991). We will limit our focus to traditional measures of pupil achievement, but the discussion and analysis that follow are relevant to any of the kinds of social indicators that might be produced by a national educational assessment program.

It is clear that the impetus behind establishing a national educational assessment mechanism in both the United States and Great Britain is to generate efforts to improve the educational system. Although not all important social indicators are explicitly normative in purpose or interpretation (Land, 1983), any social indicator generated by an educational assessment program is almost certain to have strong normative meaning. The normative standard establishes an external reference point against which all units can be compared, and provides a basis for saying that one unit is "better" than another. For educational indicators related to achievement, the standard is often the average for the nation or state, and a major index of improvement is increased pupil achievement levels.

Even if the goal is to increase average pupil test performance, however, that goal can be achieved in at least three ways: by increasing test scores at all levels, by increasing only scores in the upper part of the range, or by increasing only those in the bottom part of the range. The first retains the current range of performances whereas the second increases the range and the third decreases it. Would we consider all of those equally desirable "improvements"? An indicator based solely on average scores, whatever the unit involved (nation, state, or district), will fail to make such distinctions.

Because there are bound to be differences among districts in average performance at the outset, one measure of success might be to make districts more equal. However, would that mean equal in actual test performance or equal in the performance of pupils with equal backgrounds and equal prior achievement? If some districts are more ef-

fective in educating some kinds of pupils (e.g., high-ability pupils) and others are more effective in educating other kinds of pupils (e.g., low-ability pupils), would one measure of success be to move districts to being equally effective in educating pupils of all kinds?

Finally, if there is change in performance levels across time, or if there are differences in performance levels across units (states, districts, schools), what significance should we attribute to those differences? Is it part of the assessment process to attempt to account for such differences? If assessment is a normative process, isn't it necessary not only to identify the differences but also to find the reasons for them? If the purpose of the assessment is normative, won't the identification of "successful" structures and programs contribute to system improvement?

One goal of this article is to demonstrate that these questions are not simply academic, and are not beyond the reach of what could be accomplished with a national assessment. We show that districts do vary in all of these ways, and that such variation contributes to the distribution of academic achievement.

We consider three basic kinds of indicators that can be derived from a national educational assessment:[2]

1. *Gross productivity*. This is the average level of performance of pupils in a district, a state, or the country. It is the measure most often used in international comparisons, and the one most often discussed in the political arena. Gross productivity is important because it indicates actual levels of performance at a given stage of schooling. However, it is the least useful measure in internal analyses, and can result in misleading conclusions.

2. *Net productivity*. This is the average level of performance of pupils in a district, state, or the country, adjusted for their characteristics on entry. The phrase *on entry* calls attention to the fact that assessment criteria must specify the part of the educational system that is being evaluated. In our example, it is the secondary school, but it could be elementary school, middle school, higher education, or the entire system.[3] The phrase *their characteristics* calls attention to the significance of specifying which characteristics should be considered. It is rare for analyses of national data to take input characteristics into account, but to the extent they do, socioeconomic-status (SES) measures are most commonly used. We include gender, academic achievement upon entry to secondary school, and SES in our analysis, and we demonstrate that the combination is a much more powerful control than SES alone.[4]

3. *Inequality*. This is the degree to which pupil achievement is associated with pupil characteristics on entry. For instance, do pupils who have low prior academic achievement have greater or smaller performance gains than those with high prior academic achievement? We show that districts do vary on this measure, and we suggest that it is an important assessment index.

In addition to charting district differences in these three assessment criteria, our analysis considers two possible sources of district variation in them. One is the average pupil-teacher ratio (PTR) in the districts, which is a reasonably good indicator of per-pupil investment in education. Much of the early research on the effects of PTR has yielded inconclusive findings (see Hanushek, 1989, for a review); however, the majority of studies have lacked adequate control for pupil intake, and the analyses have not taken account of the hierarchical structure of educational data. The initial findings of a longitudinal study in Tennessee, where pupils were randomly assigned to large and small classes, demonstrate considerable benefits in achievement for smaller classes (Finn & Achilles, 1990). Our second measure is the degree to which the districts offer separate schools for more and less talented students. Research in the United States has emphasized stratification within schools 'through tracking or ability grouping (e.g., see Oakes, Gamoran, & Page, 1992), or the differentiation caused by selective private schooling (Jencks, 1985). In the United Kingdom, the issue of selective schooling within the public sector has been a more salient issue. McPherson and Willms (1987) examined the effects of school reorganization in Scotland following 1965 legislation that abolished selective schooling. They found that comprehensive schooling significantly benefited pupils in

their educational attainment, particularly for pupils from working-class backgrounds. The effects of selective schooling within the public sector may become equally important in the United States when parental choice programs become fully implemented. We do not intend our analysis to be a thorough study of the effects of PTR or selective schooling; our purpose is simply to show how the effects of district-level factors can affect indicators of net productivity and inequality. We discuss further the purpose and meaning of this analysis later in the article.

Sample, Measures, and Method

Our analysis uses data from a longitudinal study of a British birth cohort. The National Child Development Study (NCDS) followed a birth cohort of all the babies born in England, Scotland, and Wales in the first week of March 1958. (See Kerckhoff, 1990, for further details.) Data were collected when the children were 7, 11, 16, 20, and 23 years of age. The members of the cohort, their parents, their teachers, and school administrators were the sources of the data.

The data used here are primarily relevant to the secondary school education of this cohort. The assessment measures are reading and mathematics achievement tests administered at age 16 and the level of secondary school examinations' passed (data collected from the schools when the cohort members were 20 years old). Measures of social and academic background come from the age 11 sweep. Our SES measure is an index based on weighted standardized measures of father's occupation, father's and mother's education, and number of siblings.[6] Academic achievement at time of entry into secondary school was measured by reading and mathematics tests administered at age 11.

Although much of the literature on educational structures uses the school as the unit of analysis, a national assessment will necessarily focus on larger units, those based on divisions of political and legal responsibility. In the United States, the individual states are highly visible units for summary reporting. In both Great Britain and the United States, however, school districts have considerable autonomy and are the source of much of the variation in organizational structure and potential program innovation. In addition to measures of the characteristics and achievements of individual pupils, therefore, our analysis uses data on Local Education Authorities (LEAs) as units of analysis. Our analysis is based on the 148 LEAs in England and Wales.' The LEAs varied widely in both the nature of the pupil populations they served and the way they organized the schools the pupils attended. The two measures of district-level features we use in the analysis are the pupil-teacher ratio in the LEA and the percentage of the LEA pupils who attended schools with selective intakes (i.e., grammar and secondary modern schools). These were based on statistics reported by the Department of Education and Science (1974).

The advantages of this data set include the large number of districts covered, the amount of detailed longitudinal information on individuals, and the availability of district-level data on resources and organizational structure. No American data set has all of these features (see Note 1). As will become clear as we present the analysis, however, the data set has limitations that any adequate national assessment will need to overcome. One of the chief limitations is the size of the within-district samples, which could be overcome easily with a different sampling strategy.

We use a hierarchical linear model (HLM) to analyze the data. The strength of a hierarchical analysis is that it estimates statistics for each unit of a hierarchical structure using data for that unit, while borrowing strength from the information available on all units. In this case, we estimate the indices of gross productivity, net productivity, and inequality for each district. For each of these indices we estimate: (a) their average across all districts, (b) the extent that they vary among districts, and (c) the accuracy with which they are estimated. We also ask whether their variation can be explained by district-level features. The HLM analysis is conducted in three stages. We discuss each stage as part of the presentation of results that follow.

Gross Productivity

The simplest indicator of the effectiveness of the educational system is its gross productivity. Measures of this type are reported in Table 1, which also indicates the intercorrelations among the major variables in our analysis. Without a point of reference, however, such measures have little meaning.

Two points of reference can be envisaged for external assessment, although they are not available in the present instance. One is based on some statement of goals for the system, what pupils "should" achieve. The English and Welsh are moving rapidly in the direction of establishing such "attainment targets," stated in terms of achievement in specific subjects (English, mathematics, science, etc.; Department of Education and Science, 1989). In the United States, there has also been movement in this direction, especially with respect to mathematics achievement (National Council of Teachers of Mathematics, 1989, 1991). A second reference point could be based on prior years' average performance, but that is feasible only after an assessment procedure has been in effect for at least 2 years. Because our data do not relate to the new British attainment targets, and because there is no prior record of achievements, the average scores reported in Table 1 are not useful in either of these kinds of comparisons.

The average performance of the schooling system can serve as a point of reference for internal comparisons, however. For example, we might say that the average mathematics score for a particular LEA was 1.3 items higher than the national average. To assist in the interpretation of this type of comparison, in the analyses that follow we have standardized the measures of reading and mathematics achievement, based on the means and standard deviations for the national sample, as was done for the measure of examination attainment (see Note 4). This allows us to express differences among LEAs in all three types of achievement as *effect sizes*, that is, as fractions of a standard deviation. Theeffect size metric is useful because it allows us to compare the magnitude of differences among units with our estimates of the effects of policy interventions (e.g., reduction in PTR), or with the effect sizes'associated with specific types of educational interventions (e.g., see Hattie, 1992; Levin, Glass, & Meister, 1984).

It is apparent from the correlations among the three achievement measures (reading, mathematics, and examination attainment) that they are not measuring exactly the same kinds of achievement, and assessment interpretations may differ depending on which is used as the criterion of achievement. Thus, if multiple criteria of achievement are used, it will be important to examine the extent to which they lead to the same conclusions.'

The other variables included in Table 1 are indicators of the pupils' backgrounds-their gender, SES, and prior academic achievement . As would be expected, there are significant correlations between most of these background variables and the measures of academic achievement in secondary school. The correlations of the three outcome measures with the age 11 reading and mathemat-

TABLE 1
Means, Standard Deviations, and Product-Moment Correlations for the Pupil-Level Variables

Variable	*M*	*SD*	Product-moment correlation						
			1	2	3	4	5	6	7
1. Reading (age 16)	25.470	6.890	1.00						
2. Mathematics (age 16)	12.641	6.803	.64	1.00					
3. Examination attainment	0.000	1.000	.63	.68	1.00				
4. Reading (age 11)	16.061	6.208	.79	.63	.61	1.00			
5. Mathematics (age 11)	16.691	10.256	.69	.76	.64	.74	1.00		
6. Socioeconomic status	0.000	1.000	.38	.40	.44	.39	.39	1.00	
7. Sex	0.501	.500	(.00)	.09	−.05	(−.01)	(.02)	(−.02)	1.00

Note. Correlations were estimated with pairwise deletion of missing data. Total N = 8,826; minimum N = 6,839. Correlations in parentheses were not statistically significant at the .001 level.

ics test scores tend to be higher (ranging from 0.61 to 0.79) than their correlations with SES (which range from 0.38 to 0.44), a fact that becomes important in our later analysis. The correlations with gender are quite small, and in most cases are not statistically significant.

Table 2 presents an analysis of the variation in gross productivity among LEAs. The results are derived from what is called a *null model*, which is the first stage of our HLM analysis.' The analysis estimates the mean outcome score for each LEA and then estimates the average of the LEA means." The reported averages are all close to zero because of the standardization of the measures.

The analysis also estimates the variance (and standard deviation) of the LEA means for each outcome measure and tests whether it is statistically significant. This is the most important part of Table 2 for our purposes. The variation among LEAs that we observe might be a result of measurement and sampling error only. In that event, further analyses that attempted to uncover the processes underlying LEA differences would be futile. The HLM analysis tests (with a χ^2 test) whether the "true" variation (i.e., variation independent of measurement and sampling error) is statistically different than zero. For our data, the estimated standard deviations of the LEA means range from .137 to .171, and are statistically significant. This amount of variation is also substantively significant. Figure 1 displays the variation among LEA means, expressed as effect sizes, with box plots" for each outcome measure. The range in LEA means is greater than 0.4 for reading and attainment, and nearly 0.6 for mathematics. Effect sizes of this magnitude are bigger than those associated with many educational interventions. We conclude there is large variation among LEAs in their gross productivity.

The squares of the standard deviations in Table 2 are estimates of the "true" variances in LEA mean scores. These range from .019 to .029, indicating that only 2 or 3% of the total variance is among LEAs, the rest being within LEAs. This is a paradox that this kind of analysis uncovers: ***Although only a small percentage of the total variance is attributable to variation among districts, the differences between the highest and lowest scoring districts are significant both statistically and substantively***.

The third line in Table 2 reports the reliability with which we could estimate gross productivity for each LEA. The reliability of an estimate depends on the accuracy with which we can estimate an indicator (inaccuracies stem from measurement and sampling error), and the magnitude of the "true" variance of the indicator. The low reliabilities for our estimates of gross productivity reflect the fact that our estimates of the means are based on relatively small samples in many of the individual LEAs. We restricted our analysis to LEAs that had data describing at least 15 pupils, but samples of at least 100 pupils per district would be more appropriate. While the small within-LEA samples have the effect of reducing the possibility of finding statistically significant effects, the effects reported here and in later tables are all adjusted to take this limitation into account. Consistently larger individual LEA samples would have produced more reliable estimates of the individual LEAs means, but the estimates of the amount of variation in LEA means would have been about the same.

The primary conclusion to be drawn from this simple HLM analysis is that there was

TABLE 2
Statistics Describing Gross Productivity of 127 English and Welsh Local Education Authorities

Indicator	Reading	Mathematics	Examination attainment
Gross productivity			
M of LEA means	−.024	−.032	−.028
SD of LEA means	.150**	.171**	.137**
Reliability	.482	.539	,470

• ☒ < .01.

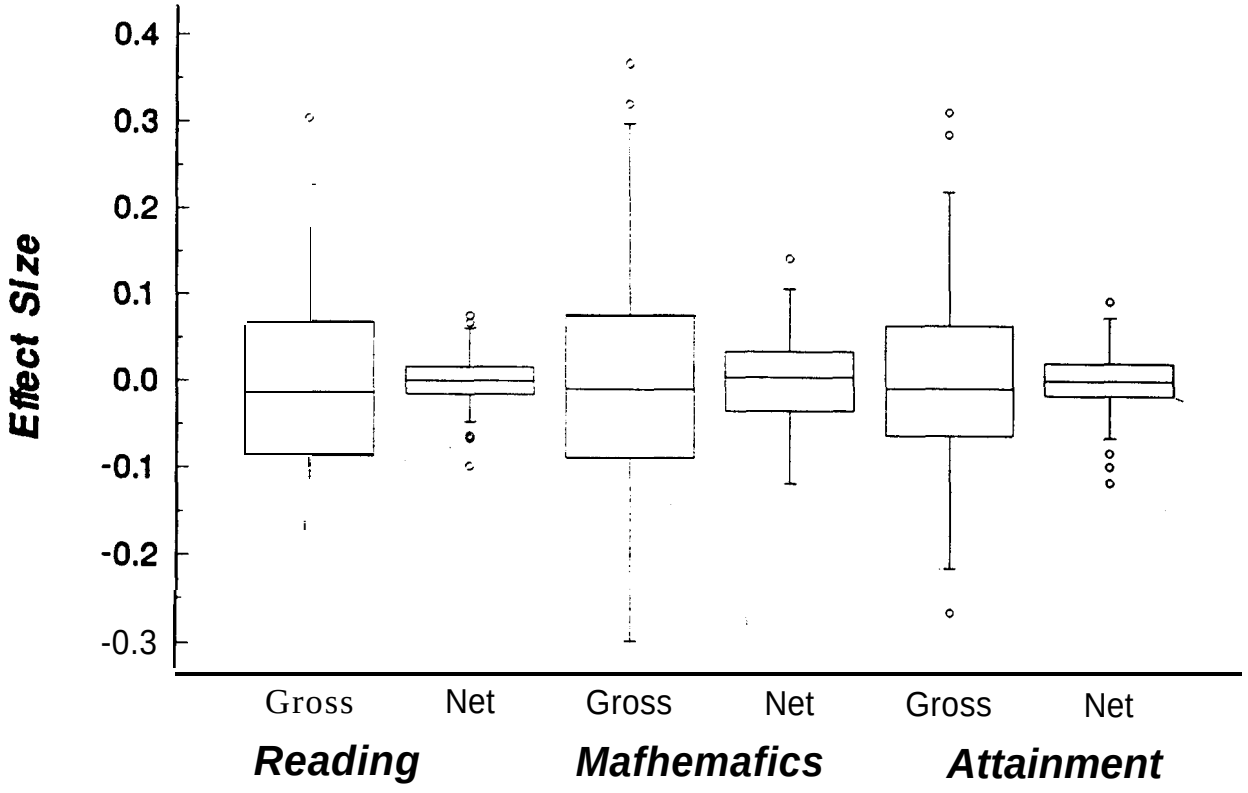

FIGURE 1. *Box plots indicating variability of gross and net productivity*

substantial variation in the gross productivity among the LEAs, even after taking account of measurement and sampling error. This provides the basis for investigating the reasons for that variation. Most reports of national assessments would probably present such measures of gross productivity as if they were meaningful in themselves. But that would imply that the differences resulted from the educational systems of the districts, and it would ignore the differences in the kinds of pupils enrolled in the districts' schools. In the next two analyses we show that much of the variation observed here stems from differences in pupil intake among the LEAs.

A Controlled Measure of Productivity

The second stage of our analysis introduces background measures.'* We begin by examining the effects of SES and gender. The multilevel analysis computes a separate regression of outcome scores on SES and gender for each LEA and produces expected scores for a pupil[13] with nationally average SES. We could refer to these adjusted means as measures of *net productivity*, but we find that SES and gender are inadequate indicators of pupil intake. We therefore reserve the term *net productivity* for adjusted means that also adjust for prior academic achievement.

The top section of Table 3 reports the averages of the within-LEA adjusted means as well as the standard deviations of those means and the reliability of the estimates. The second and third sections of the table report parallel statistics for the relationships of gender and SES with the outcome measures.

In the top section of the table, the average background-adjusted means of all LEAs are again close to zero for the three outcomes, as would be expected because of standardization of the outcome measures. However, the standard deviations of the adjusted means (ranging from .058 to .095) are considerably smaller than the standard deviations for the unadjusted means (which range from .137 to .171). The adjustments for sex and SES have reduced the amount of variation in productivity among the LEAs for all outcomes, but proportionally the greatest reduction is for examination attainment and the least is for reading. Because the variation in adjusted means is smaller, and because of the relatively small sample sizes used to estimate the adjusted means for some LEAs, the re-

TABLE 3
Statistics Describing Background-Adjusted Indicators for 127 English and Welsh Local Education Authorities

Indicator	Reading	Mathematics	Examination Attainment
Background-adjusted means			
M of LEA adjusted means	−.012	−.018	−.005
SD of LEA adjusted means	.094**	.095**	.058*
Reliability	.302	.306	.179
Gender gap (female minus male)			
M gap across LEAs	−.015	−.193**	.093**
SD of LEA gender gap	.035	.018	.063
Reliability	.020	.006	.070
Outcome&ES relationship			
M relationship across LEAs	.378**	.391**	.436**
SD of relationship	.054	.029	.060
Reliability	.137	.049	.179

● *p < .01. ● p < .05.

liability of these controlled measures is also considerably lower than for gross productivity. Therefore, it is more difficult to distinguish among particular LEAs in their background-adjusted outcome scores than it is in their unadjusted scores.

The middle section of Table 3 shows that the differences between the sexes vary according to the outcome being measured. There is no significant gender difference in reading scores, but the males scored significantly higher on the mathematics test (an effect size of nearly 0.2), and the females scored significantly higher on the examination attainment scale (an effect size of nearly 0.1). The middle line of this section shows the standard deviations of the gender gap across LEAs. For example, whereas the *average* gender gap in examination attainment was .093, the observed gap was not the same in every LEA. The standard deviation of the gender gaps was .063. However, not one of the standard deviations of the gender gaps is statistically significant. These results suggest that no LEAs were more effective for one sex than the other in reading, and that the male advantage in mathematics and the female advantage in attainment was relatively constant across LEAs. The reliability here is very low for all three outcomes, which means that it is difficult to reliably distinguish between LEAs that may have been more or less effective for one sex than the other.

The bottom section of Table 3 shows that there are consistently significant associations between SES and the outcomemeasures, the strongest association being between SES and examination attainment. This general result is consistent with many previous studies." As the second line of this section shows, however, the variation among the LEAs in this association is not significant for any of the three outcome measures. This indicates that either the LEAs do not differ in their effectiveness in teaching high- or low- SES pupils, or that the analysis is not powerful enough to detect such differences. Here again the reliability is very low, which also indicates that it is very difficult to distinguish between LEAs according to the strength of these associations between SES and the outcome measures. We consider the magnitude of the estimated variation in outcome&ES slopes (*SD*s ranging from .029 to .060) to be large in substantive terms, and therefore conclude that a more powerful analysis is required to achieve good indicators of inequality. We repeated the educational attainment analysis for the 23 LEAs that had data for at least 100 pupils, and found the variation in outcome slopes to be about the same (SD = .058), but with the larger sample sizes the variation was statistically significant ($p < .01$).

Net Productivity and Inequality

In Table 4 we add measures of the pupils' prior academic achievement to the pupil-

TABLE 4
Statistics Describing Net Productivity and Inequality for 127 English and Welsh Local Education Authorities

Indicator	Reading	Mathematics	Examination attainment
Net productivity			
M of LEA effects	−.020	−.038	−.010
SD of LEA effects	.057**	.081**	.059**
Reliability	.225	.341	,223
Inequality (gender gap, female minus male)			
Average gap	−.014	−.158**	.088**
Inequality (outcome/SES relationship)			
Average relationship	.142**	.171**	.275**
Inequality (outcome/prior achievement relationship)			
M relationship across LEAs	.622**	.589***	.418**
SD of relationship	.116**	.051	.073**
Reliability	.485	.189	.290

● *p < .01. *p < .05.

level model. These too are an indication of the LEAs' varied intakes. These measures were also standardized on the national sample to facilitate interpretation of the regression results. We consider this to be a reasonably well-specified pupil-level model, and we doubt whether additional control variables would alter the results appreciably." As expected, the prior academic achievement measures have very strong relationships with the outcome measures, much stronger than the relationships between SES and the outcome measures. [16]

One of the problems with any longitudinal assessment is the attrition of participants during the assessment period. In estimating either gross or net productivity, we would expect biased comparisons if attrition rates varied among districts, and if those dropping out of the study were a nonrandom sample (e.g., predominantly low-achieving pupils). In many cases, these circumstances could prevail. The NCDS expended considerable resources to follow subjects as they moved from one location to another. Approximately 88% of the pupils who completed the age 11 reading and mathematics tests were still in the study 5 years later and completed the age 16 tests. Thus, attrition may be less of a problem here than in a state assessment that did not attempt to follow pupils longitudinally. But even in assessments where there is attrition that is nonrandom and differs across districts, estimates of net productivity would be less biased than estimates of gross productivity because net productivity includes an adjustment for pupil background. (The bias would depend on whether those who dropped out of the study differed in their outcomes, given their set of background characteristics.) Attrition could also affect estimates of inequality: For example, a district that lost a relatively large proportion of either high- or low-ability pupils would likely have a shallower outcome/prior achievement slope, indicating less inequality than the true value.

The first section of Table 4 reports the results for the estimates of net productivity. The average net productivity across all LEAs is again close to zero because of the standardization of the outcome measures. The standard deviations of net productivity range from .057 to .081. These are less than half the size of the standard deviations of the estimates of gross productivity (shown in Table 2) and for reading and mathematics are smaller than those of the controlled measures of productivity (shown in Table 3). However, they are still statistically significant. Three of the box plots in Figure 1 display the distribution of net productivity of the LEAs as effect sizes. The full range for LEA effects for reading and attainment are slightly greater than 0.1, and for mathematics it is about 0.2. Some analysts would not consider effect sizes of this magnitude to be large, but arguably if an inexpensive intervention boosted the

achievement scores for all pupils in a district by .1 to .2 of a standard deviation, we would consider the intervention worthwhile.

The reliability of the estimates of net productivity are quite low (ranging from .223 to .341) indicating that, although there is significant variation among the LEAs in net productivity, it is difficult to distinguish reliably among the LEAs on that basis. The box plots indicate that half of the LEAs have effects that are very tightly clustered around the national mean (plus or minus .02 for reading and attainment, and plus or minus .04 for mathematics). We conclude therefore that the practice of "ranking" LEAs in their net productivity can be very misleading, for the difference between LEAs at the 25th and 75th percentiles may be quite small in substantive terms.

The second section of Table 4 shows the estimates of the gender gap across LEAs. There are no statistically significant sex differences in age 16 reading scores, but, as in Table 3, there are sex differences in both age 16 mathematics scores and examination attainment scores. On average, males score approximately .16 of a standard deviation above females on the age 16 mathematics test and females outperform males by approximately .09 of a standard deviation on the examination attainment measure. As noted above, the sizes of these gender gaps did not vary significantly among LEAs.[17]

The third and fourth sections of Table 4 report the analyses of the relationships between the outcomes and SES and prior academic achievement. These estimates of inequality give the expected change in the outcome measure for a one standard deviation increase in SES or prior academic achievement, with other factors in the model held constant. For example, the average outcome&ES relationship across all LEAs for reading was .142, which means that a one SD increase in SES (other factors held constant) was associated with a .142 SD increase in age 16 reading scores. Overall, the outcomes are much more strongly associated with prior academic achievement than they are with SES. The associations between outcome measures and prior reading achievement vary significantly across LEAs for both the reading test scores and the examination attainment scores. In all cases, the reliability of this indicator of inequality is rather low, but it is consistently higher than in the analysis in Table 3. In contrast to the analysis reported in Table 3, there is clear evidence here of significant variation among the LEAs in the measures of inequality (their effects on pupils with different prior academic achievements)—at least in reading and examination attainment.

Sources of Variation in Net Productivity and Inequality

The final stage of our analysis considered whether the variance among LEAs in their net productivity or their inequality could be explained with LEA-level variables. We examined the effects of five variables. Three of these can be viewed as contextual variables and two as organizational or structural variables. The three contextual measures describe the pupil composition of the LEAs. These are the mean reading and mathematics scores and the mean level of SES of each of the LEAs.

One of the organizational variables is the PTR in the LEA. This is both a measure of an organizational feature of the school system and an indication of the per-pupil investment in education. The other organizational measure is the percentage of pupils in the LEA who attended selective schools. At the time these data were collected (the mid-1970s), the British secondary school systems were undergoing reorganization. The state-supported system in England and Wales had historically been stratified into three types of schools: grammar schools for the academically talented, technical and secondary modern schools for all others. At this time, comprehensive schools were being introduced that enrolled pupils who would have previously attended any of these types of selective schools. The LEAs varied widely in the rate at which they had made the change. Our measure is the proportion of the pupils in the LEA still attending selective schools.[18]

We conducted analyses of the effects of these five LEA-level variables on net productivity and inequality, using the hierarchical linear model." None of the contextual mea-

sures was a statistically significant predictor of net productivity, so we limited the final analysis to the two organizational variables—pupil-teacher ratio and percentage in selective schools. The results of that analysis are presented in the Appendix for the reader interested in the technical details, but only their general outline needs to be discussed here.

The percentage of pupils in selective schools had a small, but statistically significant, effect on net productivity in the mathematics and examination attainment analyses. For instance, the results indicated that the expected score of an average pupil in a fully selective system (which includes only grammar and secondary modern schools) would be about 10% of a standard deviation higher than a comparable pupil attending school in a fully comprehensive LEA. An effect of that size is enough to alter a pupil's rank from the 50th to the 53rd percentile. The effects of percent selective on inequality were even weaker, although for mathematics and attainment they suggest that there was greater inequality in LEAs with high proportions of their pupils in selective schools.

The effect of PTR was relatively consistent across the outcome measures. In all cases, net productivity was lower the higher the PTR, although the effect was statistically significant only for the mathematics test scores. For mathematics, a one-pupil decrease in PTR was associated with an increase of 3.2 percent of a standard deviation in average mathematics scores. PTR did not have a significant association with the measures of inequality.

The LEA-level organizational measures thus do help to explain some of the variations in net productivity and inequality among the LEAs, although their effectiveness was quite limited, in part because of small sample sizes within LEAs.

Conclusion

We have used the NCDS data to show how national data on pupil achievement might be used to generate useful social indicators. In doing so, we have shown some of the possible pitfalls of such an enterprise and some of the necessary characteristics of any assessment if it is to be fully effective. Rather than emphasize the specific results of our analysis, we focus this section on their general meaning.

1. Most of the need for a national assessment emphasizes what we have called measures of gross productivity. Such measures may indeed be useful for some purposes, but their usefulness is limited. Attempts to measure the adequacy of the skills of the entire population (or any particular age cohort) might use criteria based on gross productivity. International comparisons of labor force competence or time series records of national change over time can be based on such measures. However, internal interdistrict comparisons based on gross productivity measures are bound to distort the relative effectiveness of the districts' schools because those measures do not take variations in intake into account. Gross productivity measures exaggerate the amount of among-district variation, and they lead to unwarranted and misleading comparisons.

2. The usual methods of accounting for interdistrict variation in intake involve controlling for such factors as SES and gender. We have shown that those controls do, indeed, sharply reduce the amount of interdistrict variation in productivity. However, when prior academic achievement is used as an additional indicator of district intake variation, the results are quite different. The three measures together (SES, sex, and prior achievement) reduce the variation in productivity very sharply, and the rate of reduction is essentially the same for all three achievement measures. There is a huge difference between gross and net productivity, measured in this way. In our analysis, the amount of variation in net productivity among LEAs is only one-fifth the amount of variation in gross productivity. The variation in productivity is still significant, but it is much less glaring, and any interdistrict comparisons that are made will be fairer to educators.

3. Almost all of the attention in discussions of assessment is focused on productivity measures. Other measures are possible, however, and may be equally worthy of interest. One of these is the kind of measure we have called inequality. To what extent are the districts equally effective in teaching differ-

ent kinds of pupils-boys and girls, high- and low-ability pupils, pupils from high- and low-status families? Inequality is an important index of district differences, and it is also a potentially valuable national social indicator. Our analysis suggests that reliable indicators of inequality are more difficult to achieve than reliable estimates of net productivity. Stronger designs that employ longitudinal data are required.

4. It is possible to be misled in assessing the overall importance of among-district variation. We have become too accustomed to judging the importance of an association in terms of the amount of variance explained. Only small proportions of the overall variance in achievements will generally lie among districts; most of the variation is typically among schools within districts, and among pupils within schools. But that does not preclude the among-district differences from being both statistically and substantively significant. We have shown that even though districts do not vary greatly in net productivity and inequality, that variation contributes significantly to the overall distributions of achievements. Moreover, even when district-level variation is small, it is possible to identify districts that are statistically significant outliers in their net productivity or inequality. Further study of these districts could enhance an evaluation program.

5. Collecting national assessment data provides an opportunity not only to measure the effectiveness of our educational system but also to evaluate the importance of various structural and programmatic features of the system that vary across districts. Our attempts to explain the LEA variation in net productivity and inequality were not highly successful, but the results were at least suggestive of the kinds of analysis that can be done in seeking greater understanding of the observed among-district variation. Our data did not include information on school-level features, nor could we identify the school attended by each child. Given that a factor such as pupil-teacher ratio may vary substantially among schools within LEAS, as well as between LEAs, a more powerful test of its effect could be achieved with a three-level (pupils, schools, LEAs) analysis. Our results also varied for the different achievement measures, suggesting that any program feature may have different effects on different assessment criteria.

6. One of the most important lessons to be learned from the analysis we have presented has to do with the design of the data collection in any national assessment. The reliability levels in our analysis were regretfully low. This was because of the nature of the sample, based on a birth cohort. The number of NCDS pupils in any given LEA varied widely, and the inclusion of many LEAs with very small numbers posed serious analytic problems. This leads to the obvious conclusion that all districts included in an assessment must collect data from large numbers of pupils. Yet, we have also shown that potentially misleading conclusions can be reached unless the analysis carefully controls for the nature of pupil intake. That calls for kinds of data not always considered as part of the assessment process. In particular, it calls for measures of prior achievement. Adequate assessment design will require careful balancing between the need for large samples and adequate intake data.

7. A dilemma facing educational analysts attempting to monitor performance is the requirement to produce meaningful and easily interpreted indicators and yet describe some of the complexity inherent in all schooling systems. Assessment systems that categorize schools into comparison bands and rank-order schools are appealing because of their simplicity. But the estimates of school performance garnered by these techniques can be biased, and rank orders can be misleading because even a small change in the performance of a school can dramatically alter its rank order. ***We are calling for the use of more complex techniques which not only produce more accurate estimates o f school performance, but also describe the diversity in schooling outcomes across a school district or larger community, indicate the extent of inequality among subpopulations, and provide a means of explaining diversity among schools or districts in terms of school processes.***

The function of statistical techniques is to uncover complexity and help explain it. An

indicator is simply a statistic used to convey the complexity. It derives its meaning from its variation within a sample, its trend over time, or from a comparison to some standard. In our analysis we used effect ***sizes*** as the metric for describing gross and net productivity. This is a useful metric for an audience of researchers and educational policymakers, but is probably not the best metric to convey results to parents. It may be that transforming effect sizes into scores that ranged from 0 to 100 would enable the indicators to be interpreted more easily by some audiences. But we cannot presume that parents, teachers, or school administrators are incapable of understanding indicators of diversity, inequality, or explanation. If we do, then we must also presume that we are unable to inform them in ways that will improve the schooling system.

APPENDIX A

HLM Results for Model-Examining Relationships Between Net Productivity and Policy-Related Variables

	Reading		Mathematics		Examination attainment	
Fixed effects	Estimate	SE	Estimate	SE	Estimate	SE
Average within-LEA equation:						
Intercept (grand mean)	−.038*	.017	−.074**	.018	−.045	.017
Reading (age 12)	.501**	.023	.088**	.010	.193**	.019
Mathematics (age 12)	.174**	.012	.513**	.019	.311**	.012
Socioeconomic status	.121**	.009	.165**	.009	.235**	.009
Sex (female = 1)	−.004	.017	−.172**	.017	.106**	.017
Effects of LEA-level variables:						
Percent selective on intercept	.00051	.00039	.00107*	.00042	.00112*	.00038
Pupil-teacher ratio on intercept	−.01772	.01335	−.03183*	.01375	−.02397	.01263
Percent selective on reading slopes	−.00003	.00046			.00044	.00037
Percent selective on mathematics slopes			.00035	.00037		
Random effects	Estimate	χ^2	Estimate	χ^2	Estimate	χ^2
Estimates of residual parameter variance						
Adjusted LEA means	.00413**	192.88	.00608**	204.69	.00300**	184.66
Reading slopes	.01173**	1047.24			.00319**	169.28
Mathematics slopes			.00293	145.54		
Model statistics						
Reliability of estimates:						
Adjusted LEA means		.267		.329		.210
Reading slopes		.464				.219
Mathematics slopes				.204		
Proportion of variance explained:						
Pupil-level		.482		.468		.387
LEA-level		.815		.792		.840

Note. Degrees of freedom for the χ^2 tests were as follows: adjusted means (124). reading or mathematics slopes (125). *$p < .05$. **$p < .01$.

Notes

The research was supported by the U.S. Spencer Foundation, the Canadian Social Sciences and Humanities Research Council, and the Centre for Educational Sociology (University of Edinburgh), a Designated Research Centre funded by the Economic and Social Research Council (C00280004). The analyses and conclusions do not necessarily reflect the views or policies of these or any other organizations.

'For instance, although about 90% of 17–18-year-olds are in school in the United States, fewer than 40% are in Great Britain.

[2] These terms and their general conceptual meaning are adapted from Gamoran and Mare (1989).

[3] Although some form of analysis might be used to assess the effectiveness of the entire educational system, it would necessarily go well beyond the kinds of analysis discussed and presented here. Even an analysis that bridges more than one stage in the educational career raises many additional issues about appropriate measures and the relationship between one-stage and multistage measures. This is especially true when trend data are being assembled and there are historical changes in the proportions of pupils who make the transition from one stage to the next. (See Mare, 1981).

[4] The importance of controlling for both SES and prior academic achievement when estimating school effects is well known; for example, see Jencks' (1985) review of the studies comparing public and private schools in the United States. Previous research has shown also the dramatic decrease in variability of school mean scores, before and after controlling for SES and prior achievement (see Willms, 1992, Chapter 5). One would expect prior performance and SES to be important controls also when examining variation in district mean scores, but there have been no analyses that show the extent of the reduction in variability.

'There is no secondary school "graduation" in Great Britain. At the time these data were collected, secondary school pupils could take various kinds and levels of examinations between the ages of 16 and 19. These included the academically demanding General Certificate of Education examinations (GCEs), offered at ordinary and advanced levels (O-levels and A-levels), and the less rigorous Certificate of Secondary Education examinations (CSEs). Our examinations measure is an ordinal scale that ranges from 0 (no examinations passed) to 9 (2 + A-levels). We scaled the ordinal measure on a logit distribution, as recommended by Mosteller and Tukey (1977), and standardized it on the entire sample.

[6] The SES index is the first principal component of the four indicators. The factor loadings were .74447 for father's occupation, .82305 for father's education, .75499 for mother's education, and .32829 for number of siblings. The eigenvalue was 1.90944.

'Data from Scotland were also included in the data set, but they are not used in this analysis. The Scottish education system is organized differently from the Welsh and English systems, and it uses somewhat different secondary school examinations. Also, the Scottish education authorities were much more homogeneous in their adoption of comprehensive schools than were those in England and Wales (see McPherson & Willms, 1987). For these cross-sectional data, the Scottish education authorities contribute little to our understanding of between-district variation related to school type.

[8] Willms and Raudenbush (1989) present a model for assessing the stability of outcomes across subject areas. Longford (1991) provides a multilevel test for determining whether comparisons among groups (such as the LEAs studied here) on a set of multivariate outcomes can distinguish between individual outcomes, or whether there is only a single dimension.

[9] In the first stage, a null model was fitted for each outcome measure. The null model is identical to a random effects analysis of variance, which partitions the variance into within- and among-group components. It is called a null model because there are no pupil- or district-level factors used to explain variation in pupils' achievement. The model provides estimates of the mean score for each district, and estimates of the variance of these means among districts. The remaining variance is among pupils within districts.

[10] The average is a "precision-weighted mean," which is the average of the LEA means, weighted by the precision with which they were estimated. LEAs with large sample sizes figure more heavily in this calculation.

[11] Box plots summarize graphically the distribution of a set of scores. The median is shown by the horizontal line within the box, and the 75th and 25th percentiles are shown by the top and bottom of the box. The upper and lower vertical lines extend from the box to the maximum and minimum values, respectively, except for outliers, which are shown as individual points. Outliers are values that are more than 1.5 times the interquartile range away from the upper or lower quartiles. See McGill, Tukey, and Larsen (1978) for a full discussion of box plots.

[12] In the second stage of the HLM analysis we included pupil-level background factors (SES, prior achievement, and gender) in the model. The analysis provides estimates of background-adjusted means (net productivity) for each district, and an estimate of the extent to which these means vary among districts. The analysis also provides estimates of the regression coefficients relating achievement to background factors (inequality) for each district, and estimates of the extent to which these coefficients vary among districts. We conducted the analysis with and without prior achievement in the model. For each analysis we report the average background-adjusted mean, the variation of background-adjusted means among districts, the average of the within-district regression coefficients, and the variation of these coefficients among districts.

[13] In these analyses we "centered" sex on the entire sample. Therefore, the intercepts of the within-LEA regressions can be thought of as estimates of net productivity for each LEA, if they had a balance of males and females similar to the national average.

[14] For example, pupil-level correlations between SES and achievement test scores using the 1982 U.S. High School and Beyond Study data were .33 for reading and .34 for general mathematics (Willms, 1983, Table 4-2). The correlations between SES and the 1980 national O-Grade results for Scotland were .39 for English, .41 for mathematics, and .47 for overall attainment (Willms, 1986). Madaus, Kellaghan, Rakow, and King (1979) reported correlations for Ireland ranging from .33 to .43 for standardized tests of Irish, English, and mathematics.

[15] In other analyses we expanded our input controls by including teachers' ratings of pupil behavior at age 11. The measure was statistically significant, but its inclusion did not appreciably affect the proportions of variance explained at either the pupil or LEA levels. Besides, it is not realistic to imagine that any national assessment would include such variables.

[16] The results for the null model indicated that about 3 to 4% of the total pupil-level variance was among LEAs, and 96 to 97% was within LEAs. Of this within-LEA variance, about 14 to 19% was explained by SES and gender, whereas between 41 and 53% was explained by SES, gender, and the measures of prior achievement.

"In these regressions we constrained the outcome/sex slopes to be parallel, as there was no statistically significant variation in these slopes across LEAs. Also, we found that if the SES, age 11 reading, and mathematics slopes were both allowed to vary among LEAs, collinearity problems prevented estimation of the model. Therefore, we constrained the age 11 mathematics slopes to be parallel in the regressions for reading and examination attainment, the age 11 reading slopes to be parallel in the regression for mathematics, and the SES slopes to be parallel for all three regressions. Therefore, our indicators of inequality that vary among LEAs include only the outcome/prior achievement slopes.

[18] The analog of this organizational feature in the United States might be the proportion of pupils attending "magnet" schools or schools with specialized functions (science education, the arts, vocational training, etc).

[19] The third stage of the HLM analysis attempts to explain variation among districts in their background-adjusted means (net productivity) and in their achievement/background relationships (inequality). We attempt to explain variation among districts in these indicators with the percentage of pupils attending selective schools and the pupil-teacher ratio. The multilevel technique provides estimates of the effects of these district-level factors and their standard errors.

References

Boardman, A. E., & Murnane, R. J. (1980). Using panel data to improve estimates of the determinants of educational achievement. ***Sociology of Education,*** 53, 113-121.

Bryk, A. S., & Raudenbush, **S.** W. (1992). ***Hierarchical linear models for social and behavioral research: Applications and data analysis merhods.*** Newbury Park, CA: Sage.

California Department of Education. (1992). ***California middle grades performance report summary.*** Program Evaluation and Research Division.

Department of Education and Science. (1974). ***Schools: Vol.*** 1. London: HMSO.

Department of Education and Science. (1989). ***National curriculum: From policy to practice.*** London: HMSO .

Fetler, M. (1989). ***A methodforthe construction of differentiated school norms.*** Unpublished manuscript, California State Department of Education.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. ***American Educational Research Journal, 27, 557-575.***

Fullan, M. (1992). ***Successful school improvement: The implementation perspective and beyond.*** Buckingham: Open University Press.

Gamoran, A., & Mare, R. D. (1989). Secondary school tracking and educational inequality:

Compensation, reinforcement, or neutrality? ***American Journal of Sociology, 94***, 1146–1183.

Goldstein, H. (1987). ***Multilevel models in educational and social research.*** New York: Oxford University Press.

Hammersley, M. (1985). From ethnography to theory: A programme and paradigm in sociology of education. ***Sociology, 19***(2), ***244-259.***

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. ***Educational Researcher, 18***(4), ***45-62.***

Hargreaves, D., & Hopkins, D. (1991). ***The empowered school: The management and practice of developmental planning.*** London: Cassell.

Hattie, J. A. (1992). Measuring the effects of schooling. ***Australian Journal of Education, 36,*** 5-13.

Jencks, C. (1985). How much do high school students learn? ***Sociology of Education, 58,*** 128-135.

Kerckhoff, A. C. (1990). Getting ***started: Transition to adulthood in Great Britain.*** Boulder, CO: Westview Press.

Land, K. C. (1983). Social indicators. In R. H. Turner & J. F. Short, Jr. (Eds.), ***Annual Review of Sociology, Vol. 9*** (pp. l-26). Palo Alto, CA: Annual Reviews.

Lau, L. J. (1979). Educational production functions. In D. M. Windham (Ed.), ***Economic dimensions of education*** (pp. 33-69). Washington, DC: National Academy of Education.

Levin, H. M., Glass, G. V., & Meister, G. R. (1984). ***Cost-effectiveness of four educational interventions*** (Project Report No. 84-A11). Stanford, CA: Stanford University, Institute for Research on Educational Finance and Governance.

Longford, N. T. (1991). Searching for multivariate outcomes in education. In S. W. Raudenbush & J. D. Willms (Eds.), ***Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective*** (pp. 115-130). San Diego: Academic Press.

Madaus, G. F., Kellaghan, T., Rakow, E. A., & King, D. J. (1979). The sensitivity of measures of school effectiveness. ***Harvard Educational Review, 49, 207-230.***

Mare, R. D. (1981). Change and stability in educational stratification. ***American Sociological Review, 46, 72-87.***

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. ***The American Statistician, 32***(1), 12-16.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. J. (1987). ***The underachieving curriculum: Assessing U. S. school mathematics from an international perspective.*** Champaign, IL: Stipes.

McPherson, A. F., & Willms, J. D. (1987). Equalisation and improvement: Some effects of comprehensive reorganisation in Scotland. ***Sociology, 21***(4), ***509-539.***

Meyer, J. W., & Rowan, B. (1988). The structure of educational organizations. In A. Westoby (Ed.), ***Culture and power in educational organizations*** (pp. 87-112). Philadelphia: Open University Press.

Mosteller, F., & Tukey, J. W. (1977). ***Data analysis and regression.*** Reading, MA: Addison-Wesley.

Murnane, R. J. (1987). Improving education indicators and economic indicators: The same problems? ***Educational Evaluation and Policy Analysis,*** 9, 101-116.

National Commission on Excellence in Education (1983). ***A nation at risk: The imperative for educational reform.*** Washington, DC: U.S. GPO.

National Council of Teachers of Mathematics. (1989). ***Curriculum and evaluation standards for school mathematics.*** Reston, VA: NCTM.

National Council of Teachers of Mathematics. (1991). ***Professional standards for teaching mathematics.*** Reston, VA: Author.

Oakes, J. (1989). What educational indicators? The case for assessing school context. ***Educational Evaluation and Policy Analysis, 11***, 181-199.

Oakes, J., Gamoran, A., & Page, R.N. (1992). Curriculum differentiation: Opportunities, outcomes, and meanings. In P. W. Jackson (Ed.), ***Handbook of research on curriculum*** (pp. 570-608). Washington, DC: American Educational Research Association.

Plank, S., Schiller, K., Schneider, B., & Coleman, J. (1993). Effects of choice in education. In E. Rasell & R. Rothstein (Eds.), ***School choice: Examining the evidence.*** Washington, DC: Economic Policy Institute.

Porter, A. C. (1991). Creating a system of school process indicators. ***Educational Evaluation and Policy Analysis, 13,*** 13-29.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. ***Sociology of Education, 59,*** 1-17.

Raudenbush, S., & Bryk, A. S. (1989). Quantitative models for estimating teacher and school effectiveness. In D. R. Bock (Ed.), ***Multilevel analysis of educational data. New*** York: Academic Press.

Raudenbush, S. W., & Willms, J. D. (in press). The estimation of school effects. ***Journal of Educational and Behavioral Statistics.***

Richards, C. E. (1988). Indicators and three types of educational monitoring systems: Implica-

tions for design. ***Phi Delta Kappan, 69***(2), ***495-499.***

Richards, C. E., & Sheu, T. M. (1992). The South Carolina school incentive reward program: A policy analysis. ***Economics of Education Review, 11***(1), 71-86.

Selden, R. W. (1988). Missing data: A progress report from the states. ***Phi Delta Kappan, 69***(7), 492-494.

Shepard, L. (1991). Interview on assessment issues with Lorrie Shepard. ***Educational Researcher, 20***(2), 21-23, 27.

Smith, M. S. (1988). Educational indicators. ***Phi Delta Kappan, 69***(7), 487-491.

Suter, L. E. (1988, April). ***Selected indicators of education of the United States and Japan.*** Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA.

Teenager's charter. (1991, May 18). ***The Economist, 63.***

U.S. Department of Education. (1994). ***Community Update, No.*** 11.

Willms, J. D. (1983). ***Achievement outcomes in public and private high schools.*** Doctoral Dissertation, Stanford University.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. ***American Sociological Review, 51***, 224-241.

Willms, J. D. (1992). ***Monitoring school performance: A*** guide ***for educators.*** Lewes: Falmer.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical model for estimating school effects and their stability. ***Journal of Educational Measurement, 26,*** 209–232.

Witte, J. (1992). Public subsidies for private schools: What we know and how to proceed. ***Educational Policy, 6***(2), ***206-227.***

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), ***Review of Research in Education: Vol. 17*** (pp. 31-74). Washington, DC: American Educational Research Association.

Authors

J. DOUGLAS WILLMS is a Professor at the Center for Policy Studies in Education, 2125 Main Hall, UBC Campus, Vancouver, BC V6T 1Z4, Canada, and a Senior Research Fellow at the Center for Educational Sociology, University of Edinburgh. His area of specialization is sociology of education.

ALAN C. KERCKHOFF is Professor Emeritus of the Sociology Department of Duke University, Box 90088, Durham, NC 27706. His specialization is sociology of education,

Received February 2, 1993
Revisions received August 17, 1993, and July 29, 1994
Accepted August 16, 1994