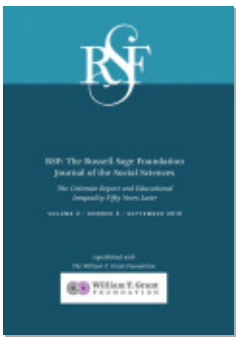
First- and Second-Order Methodological Developments from the Coleman Report



Samuel R. Lucas

RSF: The Russell Sage Foundation Journal of the Social Sciences, Volume 2, Number 5, September 2016, pp. 117-140 (Article)

Published by Russell Sage Foundation

First- and Second-Order Methodological Developments from the Coleman Report

SAMUEL R. LUCAS

Equality of Educational Opportunity *was a watershed for sociological engagement with public policy, yet the questions the project addressed drew attention to several challenging methodological issues. Statistical advances, such as the multilevel model, were important first-order developments from the Coleman Report. Second-order developments, however, may be far less visible but perhaps even more important. Second-order developments of the Coleman Report stem from two sources: (1) social scientists' reactions to proposed resolutions of the statistical challenges that the report navigated, and (2) Coleman's own (perhaps implicit)* theoretical *response to criticisms of such works as* Equality of Educational Opportunity. *Heightened interest in the challenge of identification serves as an example of the former type of second-order effect, whereas "Coleman's boat" (Coleman 1990)—and the social analytics that adopt, among other approaches, simulation strategies of inquiry consistent with Coleman's typology of causal pathways—serves as an example of the latter. First-order developments take the questions as given and see the challenge as a practical, technical issue; second-order developments explicitly or implicitly reassess the question, treating the challenge as epistemological or social-theoretic. Second-order developments therefore may change the game, upsetting or rejecting routine practice at a fundamental level. I contend that as knowledge of second-order developments and their means of practical implementation in analyses diffuses among social analysts, they will prove of far more value than first-order developments to social understanding, sociology, and social policy.*

Equality of Educational Opportunity (Coleman et al. 1966) constituted arguably the peak moment of sociologists' influence on social, economic, and poverty policy as the national leadership turned to the discipline to contribute fundamental research to help address long-standing racial and socioeconomic inequality. Legislators endeavored to inform policy first and advance social science second. To further this aim Coleman and colleagues administered an ambitious, freestanding, multilevel data collection enterprise. All sampling, in-

Samuel R. Lucas is professor in the Department of Sociology at the University of California, Berkeley.

I thank Stephen Morgan, Karl Alexander, Adam Gamoran, Gary Orfield, Jennifer Jennings, Douglas L. Lauen, William A. Darity Jr., and two anonymous reviewers for comments on an earlier draft, as well as Jan Jacobs, Susan Rachel Schacht (posthumously), H. Sorayya Carr, and participants in the Methods and Epistemology of the Social Sciences (MESS) workshop at Berkeley for continuing helpful conversations on methods. All errors and omissions are the fault of the author. Direct correspondence to: Samuel R. Lucas at lucas@berkeley.edu, Sociology Department, University of California–Berkeley, 410 Barrows Hall, #1980, Berkeley, CA 94720-1980.

strument development, data collection, and report writing was completed in two years—an amazing feat.

The amazement only increases once one realizes the massive nature of the undertaking and the resulting analyses in the Coleman Report. Nearly 100 staff members (Pfautz 1967) used mid-1960s computing power, software, and storage technology to estimate, analyze, and report 500,000 correlations and results of up to 2,000 separate regression equations (Crain 1967, 354)! The multitudinous analyses, the importance of the issues, and the availability of the data combined to motivate and facilitate rigorous reassessments and syntheses, chasing behind findings that reverberated through the halls of Congress and judges' chambers (Bowles and Levin 1968, n. 4) and even the Oval Office (Grant 1973). Thus, by any measure, the Coleman Report was a social scientific watershed.

But as social scientists engage in systematic research, they build on both the notable accomplishments and the stubborn limitations of earlier work. Intriguingly, the Coleman Report is as noteworthy for the advances it has inspired as for those it embodied. Considering the future of research, it will be useful to turn our attention to the former, for those may develop into the foundation on which further comprehension will stand.

To make this turn, I briefly identify a few additional noteworthy aspects of the Coleman Report study design. I then relate key critiques of the Coleman Report. Next, I discuss two exemplary first-order responses to the criticisms, after which I describe a design that takes these responses into account, a "neo–Coleman Report I." As analysts did not enter stasis upon the development of these responses, I next turn to criticisms of the first-order responses. These criticisms motivated additional innovative responses, which are related next. Returning again to the concrete, I describe a design for a "neo–Coleman Report II" that takes account of the second-order response. In the penultimate section, I mention several other efforts that bear the marks of Coleman Report and first- and second-order response influence. Concluding remarks follow. I begin with selected relevant aspects of the Coleman Report.

IMPORTANT DESIGN ELEMENTS OF THE COLEMAN REPORT

As is well known, data collection for the Coleman Report was mandated by the Civil Rights Act of 1964 (Coleman et al. 1966, 549). Data were collected from school superintendents, principals, teachers, and students in grades 1, 3, 6, 9, and 12, with an oversample of schools with high proportions of nonwhites. Multiple tests were administered to students, and a test of verbal achievement was administered to teachers (Ehrenberg and Brewer 1995).

The multistage sampling design nested students in probability-sampled schools and -sampled students such that the students sampled represented their peers at the school. Seventy percent of sampled high school principals responded, and 67 percent of sampled high schools delivered student tests and questionnaires; all told, 59 percent of high schools had both types of data. Seventy-four percent of principals for sixth-grade schools returned both principal and student materials (Coleman et al. 1966, 564). Given the massive target sample sizes, these response rates meant that data were obtained on over 3,000 schools and over 625,000 students, an impressive number given that several large urban school districts (Sewell 1967), including Chicago (Havighurst 1967) and Los Angeles (Grant 1973), refused to participate.

Question-specific nonresponse also occurred. Missing data were addressed by substituting the mean of the valid values for respondents with missing responses. The analysis team conducted several tests to assess the impact of both nonparticipation and question-specific nonresponse, finding the bias to be small and unpatterned.

Parenthetically, what has often gone unnoticed is that the investigation team used mixed methods. Raymond W. Mack directed ten case studies of the education of urban nonwhites, while George William Foster directed an analysis of the legal issues involved in desegregation in seven cities (Sewell 1967, 475). These qualitative analyses were later published (see,

for example, Mack 1968), but chapter 7 of the Coleman Report presents highlights of some of those analyses.

The legislation aimed to document resource differences, motivating collection of data on school-level inputs to education. For decades, research had documented racial inequality in number of school days (for example, Norton 1926), teacher qualifications (Norton 1926), per-pupil expenditure (Phillips 1932), teacher-student ratio (Moses 1941), facilities (Moses 1941; Strayer 1949), and curriculum (Strayer 1949; Wallace 1951). Thus, the data collection team focused on such inputs—a task that may have become more difficult with the end of de jure racial segregation. Still, the question of interest was whether nonwhites attended schools with fewer, low-quality, or otherwise substandard resources. Thus, data on these kinds of school factors were obtained because decades of research, while not national in scope, documented racial inequality in these resources.

For all districts, the study team obtained district-level measures on expenditures per student from which school-level expenditures per student were calculated. Data on the size of each school and its library were also obtained. For high schools, data were obtained on whether the school had guidance counselors, a science laboratory, an accelerated curriculum, a comprehensive school curriculum, tracking, and extracurricular activities, as well as the school's location in a city, suburb, town, or rural area. In addition, the study team used aggregate characteristics of the student body to measure the peer environment. The proportion of families owning an encyclopedia, the level of attendance, and the number of student transfers were obtained for all schools. The study team also aggregated elementary school teachers' reports of their perception of the quality of the student body; for high schools, they calculated the average hours of homework and the proportion of students planning to attend college as additional measures of peer environment.

Data were collected on parents' education level, father's occupation, and the presence of several items in the home (television set; telephone; record player, hi-fi, or stereo; refrigerator; dictionary; encyclopedia; automobile; vacuum cleaner; delivered newspaper). Notably, the occupation data seem to have been omitted in indexing students' backgrounds.[1]

The analysis documented high levels of racial segregation. However, it showed general equality in the distribution of school-level resources. Further, in comparing the amount of variance explained by student background and school-level factors, it showed that background factors won the explained variance "horse race." Although students' school peers and some characteristics of teachers explained some of the variance in achievement, the horse-race results were interpreted as indicating that schools had little impact on student achievement.

KEY CRITICISMS OF THE COLEMAN REPORT

Selected Criticisms and Their Implications

The claim that "the schools bring little influence to bear on a child's achievement that is independent of his [*sic*] background and general social context" (Coleman et al. 1966, 325) was perhaps the most famous and controversial statement from the study. But many of the study's findings were controversial, and these controversies motivated microscopic reexamination of and debate about many aspects of the design (see, for example, Crain 1967; Sewell 1967; Bowles and Levin 1968; Dyer 1968; Pettigrew 1968; Wilson 1968; Aigner 1970; Cain and Watts 1970; Hanushek and Kain 1972; Mosteller and Moynihan 1972).

The low response rates were one immediate concern (Nichols 1966; Sewell 1967; Bowles and Levin 1968; Dyer 1968). Although the number of schools and students participating was impressive, central cities were more likely to opt out; thus, metropolitan *samples* were more suburban than metropolitan *populations,* and impoverished schools with high proportions

1. Diane Looker (1989) finds that children are better reporters of their parents' occupation than of their parents' education.

of nonwhites may have been underrepresented. The low response rate and its nonrandom nature threaten the generalizability of the findings and raise the possibility that the variance of school-level variables was artificially deflated, reducing the chance of finding an association between school-level factors, race, and achievement.

Question-specific nonresponse was also a concern (see, for example, Bowles and Levin 1968). Because the solution, mean substitution, lowers the variance of the variable so treated as well as its covariance with other variables, the chance of a discernible association between mean substituted variables and other variables is lowered as well.

Another criticism centered on the school-level data collection. Historic research had shown that under dual systems of education, expenditures per white student exceeded the expenditure per nonwhite student (see Irby 1930; Phillips 1932). That same research had documented the great difficulty in obtaining data that would reveal the disparities, because the records released often combined categories in a way that, purposely or not, masked them (Moses 1941). Such data and difficulties suggest that within-district school-to-school differences might have been an important mechanism through which racial inequality in education was maintained. Alas, the data collection team collected district-level expenditure data and allocated an equal amount of funds to each school in the district (Bowles and Levin 1968, 8–9). Coleman (1968, 239) contended that, owing to the state of financial records in U.S. school systems, the survey team's strategy was the only feasible option, and further, that its use simply transformed the interpretation from one of school-to-school differences into one of system-to-system differences. Yet historic evidence indicates that in some jurisdictions those accounting systems had long masked large within-district racial disparities in expenditure. If such masking had continued, the failure to gather truly school-specific expenditure data in the mid-1960s could underlie the finding of small or even nonexistent differences in resources by race and, by restricting the range of the variables and introducing measurement error, may explain why school-level factors of the period appeared to be of little consequence for racial inequality.

Analysts also questioned the effort to draw causal conclusions (see Bowles and Levin 1968). One version of this criticism (for example, Nichols 1966; Dyer 1968) highlighted the difficulty of using cross-sectional data to draw causal conclusions and called for a longitudinal design to enable statistical controls for prior achievement. Such critics often admitted and lamented that the legislatively imposed time frame precluded a longitudinal study.

A second version of the criticism (for example, Dyer 1968) of the effort to draw causal conclusions focused on the horse-race statistical reporting. This criticism noted that the horse-race approach failed to consider the dynamic production of achievement—that is, it failed to consider a process by which student race and socioeconomic background factors affected peer-level and school-level characteristics and treatment of the student (including the possible differentiating behavior of school personnel), which affect student attitudes and school-provided opportunities, which lead to the opportunities, effort, and gains that eventuate in measured achievement. This criticism implied that the achievement process was so much more complex than the model specified that findings from the model were likely to mislead. Coleman (1968) contended that the limited state of knowledge about the achievement process prevented the use of a more comprehensive and dynamic model. Further, to refrain from research until such a model became possible was to force policymakers to act in the absence of understanding.

While Coleman's response rings true, this criticism raised key questions about the stark, implicit input-output model employed in the achievement analyses. In the intervening decades, analysts have focused intently on this issue, in essence heeding both the criticism *and* Coleman's response, by developing and elaborating theories of the education process and the methodological tools needed to analyze that process. Thus, criticism of the implicit input-output model employed has spurred important advances for contemporary analysts and set the stage for additional advances in the decades to come.

Selected Implications of the Critique of the Input-Output Model

One implication of the criticism of the report's input-output model turns on the specification and "horse-race" assessment of the regression model. Focusing on a horse race between socioeconomic background and school-level variables implied that 100 percent of the variance in achievement could be explained by cross-school differences (often cross-system differences in the Coleman Report case, owing to its measurement strategy). This approach was likely to produce what appeared to be low estimates of school-level effects because individual-level variables, such as socioeconomic background, varied within and between schools. Thus, the maximum variance in an individual-level outcome that could be explained by individual-level variables was the sum of all of the variance that occurred within schools and all of the variance that occurred between schools—in other words, 100 percent of the variance in the outcome. In contrast, school-level variables only varied across schools, not within them; thus, school-level variables could only explain the outcome variance that occurred between schools, which was almost always less—and often a great deal less—than 100 percent of the variance in the outcome. Essentially, a horse race of this character compares the distance covered after allowing one horse to run up to 100 percent of the course while constraining the other horse to stop running long before it can cover most of the course. The Coleman Report horse-race comparison thus unequally calibrated the variance explainable at each level, leading to what can be termed the *calibration critique*.

A second implication of the critique of the input-output model is more theoretical. The critique is that in-school processes were mostly ignored (Dyer 1968; Bowles and Levin 1968). For example, teachers' own cognitive levels as an input to students' learning are admittedly important. Yet, depending on grade level, a teacher with great subject matter knowledge but less (or lower-quality) pedagogical knowledge may produce less student achievement than a teacher with less subject matter knowledge but more (or higher-quality) pedagogical knowledge. Such possibilities change the issue from one of determining which teachers know more to assessing both what teachers know and whether and how they use that knowledge in teaching students. The *education process critique* implies that the Coleman Report's findings may be off or unclear in part because the findings are not grounded in solid prior understanding of how students, teachers, and schools actually work.

These two critiques differ in character. The calibration critique contends that the school effect estimates may be right, but that they have been interpreted incorrectly (and probably understated), whereas the education process critique implies that by missing central aspects of the way learning occurs, the school effects estimates are wrong and are likely to be underestimated.

In the decades following publication of the Coleman Report, many analysts attended to either or both of these critiques, and the resulting dialogue pushed the analysis of education —and other socially important phenomena— to new heights of sophistication and capability. That dialogue can be understood as having occurred in stages. In the original stage, which I term the *first-order response,* analysts resolved the calibration and education process critiques in terms outlined in the original debate. This promising stage ended, however, as compelling criticisms of the first-order responses emerged, criticisms that prevented the easy adoption of the solutions produced by the first-order response. Those criticisms pushed scholars further, inspiring critical, reflective engagement with the foundational issues that were perhaps latent within the original critiques. The responses that followed these criticisms form the *second-order response*. For the remainder of this analysis, I turn attention to all three phases of the developing, multistage, somewhat discontinuous dialogue following from the Coleman Report.

EXEMPLARY FIRST-ORDER RESPONSES

The calibration critique calls for development of means to properly partition the variance across individual and contextual levels. The education process critique calls for better understanding of how schools work to produce learning and other outcomes. The multilevel

model was, in part, a response to the calibration critique, while an intensified turn to the differentiating power of in-school and cross-school processes and arrangements was a response to the education process critique.

The Multilevel Model as a Response to the Calibration Critique

The Coleman Report was perhaps the most complex effort to take account of multiple levels of analysis simultaneously, but several analysts of the period were already attempting such work under the label of *contextual analysis* (for example, Bowers 1968). Contextual analysts define contexts by the aggregated values of a single individual-level variable. For example, for contextual analysis schools are placed into categories according to the mean level of parents' education; thus, high schools might be classified as high, medium-high, medium-low, and low parental education contexts. Then analysts assess whether some outcome of interest (such as student test scores) varies by schools' parental education context. Some analysts might compare the power of the education of students' own parents to the power of schools' parental education context.

For several reasons, contextual analysis failed at its appointed task. Robert Hauser (1970a, 1974) persuasively showed that contextual analysis forced researchers to decompose individual outcomes into only two factors: (1) that due to individual-level variation in independent variables, and (2) that due to variation in contexts *defined by an aggregation (mean, proportion) of one other individual-level variable of interest* (for example, the mean level of education in the context). A key basis of the method's failure, therefore, was that it forced analysts to define contexts, which necessarily have multiple traits, in a way that ignored that multiplicity. Further, because contexts are at least somewhat homophilous, an aggregated individual-level characteristic arguably provides an error-corrected version of the individual-level variable (Bowles and Levin 1968). These and other problems in part motivated Hauser's (1970b, 517) call for analysts to "operationalize directly those variables—individual or aggregate—which play a part in the social processes we study." These features imply that a horse race between an individual-level variable and its context-level analog is uninformative as to the relative power of either, whether that race be scored in terms of explained variance or otherwise.

However, some were reluctant to abandon contextual analysis (for example, Farkas 1974), perhaps owing to the undeniable possibility that context matters. Absent an apparatus for simultaneously studying multiple levels, analysts would be forced to adopt one of two ineffective approaches: either (1) reduce all phenomena to the individual level, or (2) aggregate all individual-level factors to a contextual level (Burstein, Linn, and Capell 1978). Thus, contextual analysis joined the alternatives of the period in being unable to satisfy researchers' substantive and theoretical demands.

A more promising line of approach grew in part out of attempts to resolve the reduction/aggregation challenge, an effort that would lead regression coefficients themselves to be conceived as partially random outcomes (see, for example, Zellner 1969; Akkina 1974) and that would eventually produce a method to answer the questions that contextual analysts had sought to address (Boyd and Iverson 1979). William Mason, George Wong, and Barbara Entwisle (1983) note, however, that the breakthrough of Lindley and Smith (1972), built on a Bayesian framework and applying the concept of exchangeability, was not immediately recognized as relevant in part because the statistical, sociological, and economics literatures were pursued largely in isolation. However, advances in computer processing power (Fuchs 2001), software (Bryk et al. 1988), and key didactic works (Bryk and Raudenbush 1992; Hox 1995; Kreft and de Leeuw 1998; Snijders and Bosker 1999; Goldstein 2003; Pinheiro and Bates 2004) eventually brought the model to the wider community.

One can identify at least two types of multilevel models. In *means-as-outcomes models,* the regression equation intercept, b_0, is allowed to vary across contexts. It is an adjusted mean if any other microlevel variables are included in the model. In *slopes-as-outcomes models,* one or more regression slopes for microlevel variables are allowed to vary across contexts.

Appendix A conveys both multilevel models in equation form. An intriguing implication arises upon juxtaposing the means-as-outcomes and slopes-as-outcomes models. In the means-as-outcomes models, macrolevel factors are associated with the outcome only by being associated with context-specific outcome means. Thus, the means-as-outcomes model simply restates the constraints of the original input-output model—that is, school-level variables can only affect between-school differences. However, the slopes-as-outcomes model differs: in that model, macrolevel factors interact (mathematically) with one or more individual-level factors, such that the payoff of one or more individual-level factors depends on the values of macrolevel variables. The major implication of this specification is that *within*-school differences can vary across schools in line with *between*-school factors—that is, the relationships inside the school can be different depending on the value of variables that differ across schools. Thus, the complete inability of cross-school differences to alter within-school effects is no more!

Modelers began by considering interval-level outcomes, but analysts have developed forms of multilevel models for logistic regression (Wong and Mason 1985), ordinal outcomes (Hedeker and Gibbons 1994), time series analysis (Goldstein, Healy, and Rasbash 1994), structural equations (Muthén 1994), event history analysis (Steele, Goldstein, and Browne 2004), and more (for cross-classified data, for example, see Goldstein 1994).

The multilevel model has been used in multiple areas of the social sciences. As the key response to the calibration critique, the multilevel model became an important resource for analysts in multiple areas of inquiry, in part because its application seemed to require few demands. For some, the multilevel model required only a data set that linked individuals to macrolevel contexts (Luke 2004); indeed, even a convenience sample was deemed sufficient (Hox 1995, 1). Still, before the model could be used to deepen our understanding of how schools work, the theoretical and empirical lines that emanated from the Coleman Report needed to mature. It is that development to which I now turn.

Theorizing and Investigating In-School and Between-School Stratification: Response to the Education Process Critique

The Coleman Report analysis team adopted the input-output model because consensus around a more complex, more faithful-to-the-inner-workings-of-education model did not exist (Coleman 1968, 239–40). Some critics acceded to the proposition that research on the inner workings of schools had not coalesced into, much less provided, a solid enough position to guide the analysis of an endeavor such as the Coleman Report (see, for example, Dyer 1968, 54).

Henry Dyer (1968) used tracking research as an example of an area of study that had been pursued with insufficient coordination and systematicity. The immense number of possible ways of grouping students for multiple subjects or even one subject of study, coupled with a lack of systematicity across the many studies, had harmed analysts' efforts, he argued, to develop consensus understandings of the ubiquitous school practice of grouping.

In a paper titled "Organizational Differentiation of Students and Educational Opportunity," Aage Bøttger Sørensen (1970) provided a major step forward from this state of affairs. Setting aside age-grading—the allocation of students to grades based largely (but not completely) on an age-qualified initiation of formal schooling—Sørensen noted that school systems divide the curriculum horizontally and vertically. Vertical divisions are such that one lesson or course facilitates the next. For example, arithmetic is a precursor to algebra. Horizontal divisions, however, do not build on each other; for example, neither calculus nor matrix algebra is a prerequisite for the other, nor are chemistry and French II prerequisites for each other.

Sørensen identified several dimensions along which systems of curriculum differentiation might vary. For example, systems might allow large, moderate, small, or no incidence of mobility. And the system of allocation to subjects might have wide or narrow scope, with the former meaning that much of a student's day is spent with the same peers. This theoretical development facilitated empirical efforts to determine schools' placement along

multiple dimensions, enabling assessment of the complex possibilities and potential effects of different ways of arranging curriculum differentiation.

Given Sørensen (1970), analysts used multiple analytic strategies, comparing students in different curricular locations (Gamoran 1987), schools with and without tracking (Hoffer 1992), students at different levels of schooling (Hotchkiss and Dorsten 1987; Pallas et al. 1994), and schools with different organizational features (Riehl, Pallas, and Natriello 1999) and studying a multiplicity of outcomes (on college entry, see Rosenbaum 1980; on delinquency, see Wiatrowski et al. 1982; on alienation, see Oakes 1982). For reviews of this extensive literature, see, for example, Gamoran and Berends (1987), Slavin (1990), and Lucas (2008).

If Sørensen (1970) outlined the dimensions within which the black box of schooling is located, other analysts began to open that receptacle to detailed study of the learning process. Identifying classrooms and subunits of classrooms as a primary site of learning, Rebeca Barr and Robert Dreeben (1983) plumbed those sites for the processes underlying their effects while simultaneously tracing how the stage for those effects is set by higher-level organizational factors. Sørensen and Maureen Hallinan (1977) specified a process model of student achievement involving students' ability, effort, and opportunity to learn. Through this model, Sørensen and Hallinan established the theoretical finding that under some conditions good schools (defined as those with many opportunities to learn) increase within-school achievement inequality.

On the basis of such work, analysts learned that the resources that schools supply—especially time and materials—affect teachers and thus differentially affect the learning of students in high- and low-track positions (Gamoran and Dreeben 1986); that class size, another determinant of the time available for each students' learning, matters (Krueger and Whitmore 2002); that racial-ethnic and socioeconomic diversity are associated with track rigidity independent of students' prior profiles of achievement (Lucas and Berends 2002); that high school achievement depends in part on students' placement in the structure of curriculum differentiation and what the structure is (Gamoran 1992); and more.

Alongside these research streams, Coleman, Thomas Hoffer, and Sally Kilgore (1981) analyzed school sector differences, finding that students in public schools performed less well than peers at Catholic and private non-Catholic schools. That analysis can be seen as an effort to establish school effects using the vehicle of school sector. Each sector represented different in-school process *regimes,* such that to compare sectors is to compare arguably coherent sets of in-school processes (Chubb and Moe 1988).

However, this effort sparked serious criticisms (see, for example, Goldberger and Cain 1982), some of which applied to the in-school stratification literature as well; the compelling nature of this criticism signaled the beginning of the end of the heyday of the first-order response. Before turning to those criticisms, I first describe a Coleman Report replication design that takes account of the first-order response.

The Neo–Coleman Report I: A First-Order, Critique-Influenced Coleman Report Replication

Replicating the Coleman Report based on the first-order responses would leave some features of the study intact. Notably, the collection of data on students nested in schools and the collection of cognitive test data on teachers would be maintained. And as before, data to allow study of both school outcomes, such as cognitive achievement, grade retention, graduation, and college entry, and factors within the schooling process, such as educational aspirations and other social-psychological factors, would be collected. But major design changes would also follow.

First, simultaneous collection of data on multiple schooling grades would be maintained but reduced to shift study resources to lower the incidence of missing data. Offering participating districts and schools analyses of key relations of interest (for example, anonymized graphs of the relation of sex-gender, socioeconomic background, race-ethnicity, course levels, teaching strategies, and more with achievement) could aid persuasion efforts. In

addition, follow-up contacts on missing critical items would be used to reduce missing data for key variables (see, for example, Lucas et al. 1987).

Second, the collection of data would be improved for concepts that were originally measured with proxies. For example, instead of collecting district-specific expenditure data and allocating the expenditures equally across schools, actual school-specific expenditure data would be collected. As another example, instead of calculating a proxy for class size by dividing the number of students by the number of teachers, actual class size (and attendance) data would be collected for each class period.

Data would be collected on students' course-taking in sufficient detail to allow students' placement in the national, differentiated curriculum (Lucas 1999). This part of the design necessitates collecting sufficient data on additional levels of nesting (for example, the classroom) to allow assessment of classrooms inside schools and thus facilitate assessment of within-school differences. And because systems of student allocation to courses may matter as well, data sufficient to score schools on Sørensen-dimensions of curricular stratification would also be collected, facilitating cross-school analyses.

Much of schooling occurs inside classrooms, but knowing no more than which students are within which classrooms addresses only the structural aspects of education processes, not the instructional aspects. In order to fully address the education process critique data on instruction are also needed. For example, instead of stopping the measurement of resources at the level of depth reflected in counting the number of library books, analysts would go further, collecting the textbooks and syllabi used in each class and coding the materials to peg the rigor of each student's classes. Such measures would enable analysts to study both within- and between-school inequalities in resource and instruction quality.

Data would be collected on sampled students for two consecutive academic years, making the design longitudinal, to allow analysts to consider achievement growth from a measured, preexisting level. (Data on their teachers each year would also be collected.) To fully analyze growth, multiple data collections would occur each year (for example, in the fall and spring of each academic year of data collection). The longitudinal design would allow analysts to focus on processes occurring in school by arguably washing out occurrences prior to the school year of interest and enabling the removal of summer learning complexities (Heyns 1979; Alexander, Entwisle, & Olson 2007). Such a design would have the added advantage of allowing analysts to consider the role of both transitory and stable socioeconomic background factors (such as transitory income versus "permanent income").

With such data, analysts would use the multilevel model to estimate coefficients for school-, classroom- (or teacher-), and school-level factors on student outcomes. Analysts would not only use school-level factors, such as the rate of track mobility, and classroom-level factors, such as the rigor of textbooks and syllabi, but also compositional variables, such as the mean level of education of the parents of students in the school or classroom. Using such measures, analysts would attempt to estimate context-level effects on student outcomes. Using the multilevel model with such data addresses the calibration critique. Indeed, the model's ability to properly calibrate context-level effects while allowing coefficients for individual-level factors to vary according to macrolevel factors (for example, allowing socioeconomic background effects on achievement to depend on schools' level of track mobility) renders the stark partition unnecessary and often inappropriate.

Finally, analysts would draw on research on missing data (Little 1992) to systematically employ a better approach than mean substitution. Further, analysts would check the robustness of findings against different assumptions for missing data.

Proponents of the first-order critiques imply that a replication along these lines would alter the Coleman Report results. Alas, we may never know whether using this design in the mid-1960s would have produced the original findings. What could be known, however, is which results from the past would replicate upon using the neo–Coleman Report I design

in the contemporary period. Before such an issue is assessed, however, we must recognize that first-order developments were not the final word on the issue. Instead, those developments themselves faced criticisms.

FIRST-ORDER METHODOLOGICAL DEVELOPMENTS CHALLENGED

The Education Process Critique: Questioning the First-Order Response

Many theoretical advances were largely accepted as identifying dimensions or phenomena worthy of study. Yet analysts continued to tighten the focus to discern the key pedagogical factors underlying learning. For example, given the importance of student effort in the production of achievement, David Shernoff and his colleagues (2003) have applied flow theory—a theory of optimal experience—to identify the kinds of pedagogies that maximize student engagement. Martin Nystrand and his colleagues (2003) have studied the emergence of student engagement in the dynamic conditions of classroom dialogue, finding that authentic (teacher) questions, teacher uptake of student contributions, and student questions are important spurs to student engagement. Such work has allowed the visualization of how institutional actors (such as superintendents) and macrolevel factors (for example, socioeconomic residential segregation) might send varying resources (for example, teachers of different skill levels) to different schools, enabling differential teacher-student collaboration in ways that maintain or exacerbate achievement inequality.

However, amid the flurry of empirical research, a long-standing, nagging question became an increasing concern: are outcomes driven by processes allocating students to schools or positions in school, on the one hand, or instead, are outcomes a result of processes occurring inside the schools and the positions to which students are allocated? The question pushed analysts back to consider the determinants of placement and, conditional on those findings, to devise strategies of research to address selection issues that might confound efforts to determine the effects of students' schools or positions in school.

With respect to in-school processes, analysts have found that the higher a student's socioeconomic background (Rosenbaum 1980) or prior achievement (Jones, Vanfossen, and Ensminger 1995), the more likely the student is to enter demanding curricular locations. Findings with respect to race have often been less clear (Garet and DeLany 1988; Mickelson 2001), but later research has reconciled the differences by using the multilevel model to allow black-white differences in probabilities of advanced course-taking to vary across schools. The findings reveal that, net of prior achievement and socioeconomic background, blacks' and whites' probabilities of entering demanding courses differ across schools. Those differences are associated with school diversity, region, and school size (Lucas and Berends 2007). Indeed, the pattern resembles one-for-one substitution of whites for blacks into demanding courses as one traces the graph from less to more diverse schools (see Lucas and Berends 2007, 180, fig. 2).

Findings of differential entry to demanding curricular positions by socioeconomic background and race, even after other determinants (such as achievement) are controlled, have further motivated analysts to attempt to estimate effects of track location that take account of selection on observables. But to purge selection bias from estimates of track effects has required analysts to account not only for observable factors shown to correlate with placement but also for unobservable factors that might nonrandomly allocate students to different tracks.

Analysts have used endogenous switching regression models that can control for selection into tracks on both observable and unobservable factors (Gamoran and Mare 1989; Lucas and Gamoran 2002). Estimates have revealed positive effects of track location on outcomes, indicating that nonrandom allocation to curricular positions does not fully explain prior estimated positive high-track effects. The models address selection bias, but as with all such methods, applications may rely on difficult-to-establish identifying assumptions (Winship and Mare 1992). If those assumptions are faulty, the calculated parameters will not capture the causal effect.

The same conundrum bedevils efforts to estimate school sector effects. The landmark Bryk, Lee, and Holland (1993) study used a mixed-methods approach to investigate the determinants of students' entry into Catholic school and to estimate Catholic school effects. Many aspects of the work are informative and impressive. Yet it is also true that the findings can be explained by selection processes.

For example, key findings conveyed in three figures graph the association between socioeconomic status (SES) and senior-year achievement in public and Catholic schools that resemble: (1) schools with the average social composition of Catholic schools (Bryk, Lee, and Holland (1993, 264–65, fig. 10.6); (2) schools with the average social composition of public schools (fig. 10.7); and (3) schools serving large numbers of disadvantaged students (fig. 10.8). In the first figure, the curve for public schools is lower but steeper than the curve for Catholic schools, a pattern that can be explained by selection on unobservables. Low-SES students with advantaged unobservables (for example, parents especially enabled to find or provide solid educational opportunities for their child) likely have higher achievement than do their unobservably disadvantaged low-SES peers, and they may also be more likely to enter Catholic schools than such peers. The enrollment pattern boosts the achievement of low-SES students in Catholic schools compared to their peers left behind in public school. The Catholic school advantage declines as SES rises, which may occur if selection on achievement-related unobservables declines as SES rises. The boost for lower-SES students in Catholic schools, coupled with less selection on unobservables for higher-SES students, reduces the slope for socioeconomic status in Catholic schools relative to public schools. The same story applies to the second figure, but more mutedly because the comparison is for Catholic and public schools with compositions similar to the average public school.

In the third figure, the curve for Catholic schools is higher and steeper than the curve for public schools. Comparing schools with large numbers of disadvantaged students, unobservables distinguish high-SES students who select into high-disadvantage Catholic schools compared to high-SES students who attend high-disadvantage public schools, leading to higher achievement for the former. Selection on unobservables may be less strong among low-SES students in high-disadvantage schools because the illustrative low-SES parent particularly enabled to find solid educational opportunities for their child may be less able to do so if the relevant Catholic school strongly resembles the nearby public school. But selection may still exist, as reflected in the somewhat higher Catholic school achievement of low-SES students compared to their lower achievement in similar public schools.

Certainly, Bryk, Lee, and Holland (1993) offer plausible interpretations of the patterns that highlight differences in in-school processes. And their mixed-methods analysis does address several criticisms one might articulate. Yet the point of the discussion is not to unequivocally assert that the patterns they document flow from selection; the point is to establish that what appear to be school effects identified through cross-sector comparison could be equally explained as selection effects. And the plausibility of the selection explanation greatly diminishes the utility of the sector comparison strategy for identifying school effects, regardless of whether the method is qualitative or quantitative, unless selection is explicitly addressed.

In this way, in both the effort to consider in-school processes and the effort to look across schools, robust critiques sidelined many empirical elements of the first-order response.

The Calibration Critique: Questioning the Applicability of the Multilevel Model

The multilevel model can be seen as a feasible means to study contexts and thus as an answer to the problems of contextual analysis. Alas, and unbeknownst to many, the multilevel model does not seem to provide the widely applicable escape from many of the problems that have hampered contextual analysis. Problems can be placed into two categories: (1) problems due to common *application* of the multilevel model, and (2) problems *inherent* to the model as a means of decomposing contextual effects.

Application Problems

The application problems are easiest to resolve. One such problem occurs because a high proportion of the research using multilevel models uses secondary data. Analysts collecting their own data might have a chance to ensure sufficient sample size for informative estimation, but the use of secondary data can lead to insufficient level 2 sample size (Bryan and Jenkins 2016), rendering the level 2 estimates unreliable.

Another application problem is that many context analyses capitalize on the higher reliability of aggregated variables, as noted earlier (Bowles and Levin 1968). If we use the mean level of mothers' education in a school as a context-level variable, that variable will be more reliably measured than mother's education is for any single individual student. Such a context-level variable may simply capture variation associated with the individual level that is not captured at the individual level owing to measurement error. A related measurement problem is that inferences are often erroneous because macrolevel factors (for example, the mean level of parents' education in the school) are estimated from the sample rather than known. If analysts treat these estimates as if they are known, standard errors are underestimated (Manski 1995).

An application problem also arises from the common assumption among analysts that processes that allocate entities to contexts can be ignored (Hauser 1974). Ignoring nonrandom selection into contexts, however, can contaminate putative macrolevel causal effects with selection biases. Analysts also often assume that persons know their peers' outcomes, but rarely introduce evidence to support this assumption; the result is yet another application problem (Manski 1995).

A final insidious application problem is that secondary data are often collected with a complex sample design that does not support unbiased estimation of multilevel model parameters. Elsewhere (Lucas 2014), I have proposed the concept of fully multilevel probability (FMP) samples. FMP samples meet three criteria: (1) level 1 entities (for example, students) are probability-sampled to represent the population of level 1 entities (students in the state or nation studied); (2) level 2 entities (for example, elementary schools) are probability-sampled to represent the population of level 2 entities (elementary schools in the state or nation); and (3) level 1 entities (students) are probability-sampled to represent the other level 1 entities (their fellow students) inside their specific level 2 entity (their school). Appendix B provides equations that illustrate the implications of failing on one of those criteria.

Many nationally representative data sets with geocoded data, such as the Panel Study of Income Dynamics (PSID), the General Social Survey (GSS), and the National Longitudinal Study of Adolescent to Adult Health (Add Health), fail to satisfy the criteria of an FMP sample for some research questions. When the criteria are not satisfied, macrolevel parameters are biased to an unknown degree and in an unknown direction (Lucas 2014, 1625–28).

Inherent Problems

The debilitating *inherent* problem of contextual analysis was identification, and as Charles Manski (1995) indicates, the problem remains with the multilevel model. Manski distinguishes: (1) endogenous effects, (2) contextual effects, (3) ecological effects, and (4) correlated individual effects. To fix ideas, consider a cohort of students in several different schools and ask the following question: what factors are associated with a particular student's performance on a standardized achievement test?

If we expect that a student's measured achievement is affected by the measured achievement of other students in the school, then we have posited an *endogenous effect*. Endogenous effects concern the impact of the behavior of others with respect to some phenomenon on a particular person's behavior with respect to the *same* phenomenon; for example, the academic performance of a student's peers may affect that student's academic performance.

In contrast, if we posit that students in schools with a high rate of delinquency will be less likely to have high achievement scores, we have posited a *contextual effect*. In doing so, we are concerned with the question of whether a

person's performance varies with the distribution of *other* characteristics in the reference group. Manski (1995) notes that often analysts posit an endogenous effect but investigate several contextual effects (see, for example, Crane 1991). Confusing endogenous and contextual effects impedes effective theorizing of the mechanisms through which contexts might have their effects.

If we expect that schools composed of a large proportion of delinquent students will have other negative characteristics, such as poor facilities or authoritarian discipline practices, then we have posited that students in the same school face similar institutional environments that may account for their similar behavioral response (in this case, their academic performance). Manski regards such effects as *ecological effects*.

Finally, if we believe that students in similar schools share similar unobserved individual characteristics, such as levels of curiosity, and that these individual-level attributes account for their similar behavior, then we have posited *correlated individual effects*. Appendix C develops the difficulty with estimating all four types of effects.

In response to old-style contextual analysis, Hauser (1970a) shows that analysts may often conflate endogenous, contextual, and correlated individual effects. Alas, the same problem can hound multilevel modelers, because, regardless of whether one uses 1960s-era contextual analysis or twenty-first-century multilevel models, there are many ways the analysis can go awry. For example, one may conflate context and correlated individual effects by positing an indefensible individual-level model, inadvertently leaving more unexplained variance that may be captured by context-level variables. That unexplained variance would partially reflect unmeasured individual-level determinants that are correlated across individuals and may be incorrectly assigned to context-level variables. Thus, Hauser contends that the individual-level model must be a defensible baseline.

Manski (1993) demonstrates that even with the introduction of controversial assumptions, contextual and endogenous effects cannot be distinguished. Moreover, even the determination of whether such effects exist is fraught with peril; Manski (1993, 35–36) shows four common scenarios in which important conditions for these results are violated.

The many difficulties that Hauser and Manski identify culminate in a resounding critique of the search for context-level effects even by means of the multilevel model. The inherent problems present serious impediments to obtaining informative estimates of targeted parameters. The application problems escalate the difficulty but have far more tractable solutions. Still, taken together, these complexities establish that the challenge of estimating school effects remains daunting despite popularization of the multilevel model.

Critique of the Neo–Coleman Report I

Key aspects of the ideal first-order replication of the Coleman Report are implicitly criticized, especially the specification of the multilevel model. Data collection for the neo–Coleman Report I would have led several context-level factors to be measured by aggregating student-level variables, running afoul of the aggregation critique discussed earlier. The longitudinal data collection would be seen as helpful but probably insufficient to identify causal effects of context. And the lack of effort to address selection into contexts creates further problems. The existence of information on students' placement in the stratified curriculum would be valuable but, again, insufficient owing to the prospect of the analysis being complicated by unobservable determinants of selection into curricula.

Thus, whether the neo–Coleman Report I design would reproduce the 1966 findings in 1966 or later, or would not do so, the findings would be regarded as indeterminate. More research would be indicated.

SECOND-ORDER METHODOLOGICAL RESPONSES

The criticisms already described set the stage for contemporary second-order responses to the Coleman Report. Those responses are necessary because, as Coleman implied, if solutions to the problems excavated remain elu-

sive, our understanding will simply be limited and policy decisions will be made anyway.

Second-Order Responses to the Calibration Critique

The implications of the criticisms range from limiting to dire. On the limiting side, better sample and questionnaire design and the use of, in principle, macrolevel measures instead of aggregated microlevel factors to measure macrolevel phenomena could readily resolve some of the application limitations. Other application problems, such as the problem of accounting for the process of selection into contexts, are more challenging. Even more challenging yet are efforts to resolve the identification problems inherent in the multilevel model. Indeed, in this area the implications are dire. Advances forward must, as Manski notes, come at the cost of introducing a priori information. There are several ways of introducing such information.

For example, one approach to identifying endogenous and contextual effects is to posit a lag structure. One defense of this alternative is the claim that individual-level factors act contemporaneously but contextual factors act with lag.[2] Although the lag assumption is defensible in general, making it requires not only that one have access to information about the prior performance of others but also (and most important) that one can defend the particular time lag employed. If the time lag employed is simply asserted, then identification again rests on a strong and perhaps unjustified assumption.

In addition, some circumstances conspire to render this strategy ineffective. As Manski notes, the assumption requires our observation of the system in disequilibrium. If the system is in equilibrium, then the lagged value of the context-level variable is a linear function of the other constituent parts of the equation in the same way that a contemporaneous value is, vitiating the identifying power of the time-lag assumption.

A second approach to resolving the identification problem is to restrict some of the parameters; for example, one might assume no endogenous effect and, conditional on this assumption, estimate the contextual effect. This is a common way of identifying effects in the individual-level case. However, one limitation to following this approach is that our knowledge of context-level factors is much less developed than our knowledge of individual-level factors. Thus, an assumption to constrain some types of effects to zero is correspondingly stronger.

These approaches to identifying context-level effects are potentially useful. A general theme of the analysis is that researchers should present an explicit identification analysis that precedes any empirical analysis. In the case of context-level effects, the observations made here set the stage for any empirical analysis attempting to search for contextual effects. In the kind of identification analysis that Manski proposes, the researcher would make plain the assumptions used to identify causal effects. The identification analysis would therefore allow other researchers to determine for themselves whether they accept or reject the assumptions and, in so doing, accept or reject the findings of the analysis.

The introduction of a priori information is a pathway to sustaining the model, but it can come at a high cost. As Manski (1995) argues, often the prior information one can introduce is the very kind of information about which analysts disagree. Thus, the results produced conditional on acceptance of the priors can simply push the debate back one step instead of forward.

A tool that might help break the impasse is needed. Such a tool will make assumptions visible and allow analysts to discern testable implications of their causal hypotheses, if such implications exist. In the contemporary period, such tools are provided by the graphical causal model (GCM) and directed acyclic graphs (DAGs) (Pearl 2010; Elwert 2013).

The ability to determine testable implications, conditional on the posited causal struc-

2. To defend the time-lag assumption for standardized test scores, consider that the student is alone during any given test administration. Peers' collective performance on the same test cannot have an effect; however, peers' prior achievement can have an effect.

ture, opens the door to a step-by-step bootstrapping operation by which, over time, causal inferences become tested and certified by cautious and systematic analysis of observational data. Because the posited structure is a precursor to any hypothesis, it is obvious that, if there is disagreement on the set of linkages needed for the causal theory, progress will depend on positing multiple plausible causal structures to determine their testable implications, followed by evaluating those implications with appropriate data. Just as obviously, such a bootstrapping operation may take time.

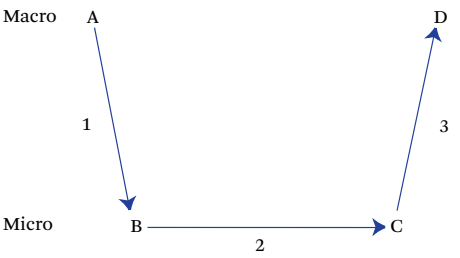
Second-Order Responses to the Education Process Critique

For DAGs to become more than another rote tool applied quickly and mechanically in the process of churning out text for citation counters to tally, the causal structures posited must be theory-laden (Horan 1978)—that is, they must be based on well-reasoned theoretical positions. Although in many cases observational data have been incredibly suggestive, the selection critique still prevents complete dissolution of analysts' uncertainty as to the mechanisms underlying the parameters estimated. What is needed is a breakthrough that brings into awareness the deeper structure of any claims about individuals in context, as well as an apparatus that can serve as a laboratory for studying the specificities of the claims that flow from that awareness.

Intriguingly, Coleman (1990) provides a candidate resource for the needed structural breakthrough. In *Foundations of Social Theory,* Coleman literally sketches the relation of individuals and the contexts in which they are nested, vis-à-vis any given feature of the latter. The contribution can be interpreted as another in a series of Coleman's own responses to the dialogue that transpired after the Coleman Report. In the public- and private-school work, he makes additional efforts to discern school effects. Here, however, Coleman takes a different tack: the very nature of what it means to talk about contexts having effects becomes the object of study.

Coleman's contribution is suggested by *Coleman's boat,* replicated in figure 1. In one interpretation, nodes A and D represent the macro level (for example, the school), while nodes B and C represent the micro level (individuals). Any effect of A on D goes through the interaction of individuals in nodes B and C. Further, from interaction between B and C can emerge macrolevel phenomenon D. So, for example, institutional effects are traced by arrow 1. Yet the system is dynamic, such that the individual recipients of that treatment interact with each other at node B, possibly producing endogenous or contextual effects, which would be reflected in node C. Arrow 2 traces the causal path from interaction at node B to microlevel outcomes at node C, outcomes that may be produced by endogenous or contextual processes. Finally, the result of their interaction may have implications for the context level in node D. The causal force of microlevel actors on the macro level is traced by arrow 3.

Figure 1. Coleman's Boat



Source: Coleman 1990.

The framework points analysts toward discerning the microfoundations on and through which all social entities (institutions, norms, extra-individual structures) are based—that is, the mechanisms through which they activate their effects. As they do so, social analysts need to recognize that action can flow from a complex combination of individuals' desires, beliefs, and opportunities (Hedström 2005), and that part of the social analyst's job is to attend to the real motivations of individuals even as analysts may build stark models to test specific, contained mechanisms that might explain patterns and features of the social world. Yet accessing the motivations of real individuals is almost impossible, for multiple technical reasons, including our inability to fully access our own cognition, the tendency of causal claims for the same event to change over time with additional experience and wisdom, the

impossibility of resolving the fundamental problem of causal inference on the basis of one case, and finally, the Heraclitian same-river-twice phenomenon (Plato 1921, 402a). Thus, techniques such as agent-based modeling (ABM) (for example, Schelling 1971; Bruch and Mare 2006)—with which analysts can explore whether observed patterns will emerge given certain theoretical claims—hold promise.

However, the concept of mechanisms must not devolve into metaphor, for that would endanger all of the gains in understanding preceding analysts have struggled to develop. By mechanisms what is meant are concrete, identifiable processes, entities, and sets of behaviors. Instead of metaphor, it would be more accurate to view mechanisms as catalytic ingredients of causal recipes for social outcomes —as long as analysts remain epistemological stochasticists regardless of their ontological views (Lucas and Szatrowski 2014, 23–27).

WHAT IS TO BE DONE? HARVESTING THE INSIGHTS OF SECOND-ORDER DEVELOPMENTS

Once we bring together the second-order developments from the calibration and education process critiques, we are left with a perhaps more difficult task. However, we are also left with more possibilities for completing the task than perhaps ever before. As we recognize the threats to proper causal inference (Morgan and Winship 2007), the process of research becomes longer, as should the products of that research as well. Indeed, a subsection on identification (possibly referencing DAGs) should become a routine part of the methods section of articles and the methods appendices of books. Papers and books should grow in length as their numbers decline, for if they remain as constrained and common as now—given the task before us and the need to describe it in sufficient detail for expert readers to evaluate and modify in later work—the counsel to follow is already likely impossible.

The Neo–Coleman Report II: A Second-Order Critique-Influenced Coleman Report Replication

Some changes in the neo–Coleman Report I would be needed to replicate the Coleman Report on the basis of the second-order critique. First, while acknowledging Coleman's (1968) claim that lack of knowledge about the achievement process in 1966 prevented the use of a more comprehensive and dynamic model, the neo–Coleman Report II would be built on the recognition that every analysis requires an implicit or explicit model of the production of its outcome. Ideally, analysts would select a model to guide data collection and analysis, such that the model should be evident prior to data collection.

Thus, in an effort to ensure that the analysis produces results that address the questions of interest while limiting the threats to proper inference, the neo–Coleman Report II would begin by using tools generated by the second-order response, such as DAGs and ABMs, in a complex process of theoretical analysis. In one part of the analysis, researchers would specify a DAG, based largely on previous research, and analyze it to determine which variables to control or not control in models as well as to discern the testable implications of the DAG. In another part of the analysis, they would use ABMs to assess whether context-level patterns can emerge from the posited individual-level factors; if not, then, given the Coleman's boat understanding, they would know that something is missing from the analysis. Further, ABMs might be helpful when a DAG reveals that there are no testable implications of a posited causal structure. Through such theoretical work, analysts will be better placed to identify variables of promise to measure and to determine whether methods as simple as (nonparametric) comparison of group-specific means, methods as complex as marginal structural models (Robins 1999; Robins, Hernán, and Brumback 2000; Sharkey and Elwert 2011), or methods somewhere in the middle are necessary to estimate the causal effects of interest.

After such efforts, data collection would begin. In addition to data collected as described for the neo–Coleman Report I, multiple other data collection methods that collectively capture the processes of interest would be implemented. For example, standardized classroom observations (Nystrand and Gamoran 1991) as well as novel data collection strategies that can capture interaction, such as experience sam-

pling (ES) (for example, Csikszentmihalyi 1990), would be used.[3] Use of such data collection strategies would allow measurement of the quality of pedagogic dialogue specifically, and the pedagogies in general, to which students are exposed, a necessary addition to fully address the education process critique.

Because estimating context-level effects can be simplified if disequilibrium conditions exist, efforts would be made to allow assessment of whether an equilibrium on relevant matters exists for the cohorts and years of study. To facilitate assessment of this issue, some perhaps less extensive data could be collected on multiple adjacent cohorts over multiple years, a task that might force further reduction in the number of grades studied.

To address the aggregation critique specifically, and the general problem of proxy measures in general, analysts would theorize phenomena of interest at their level of existence and ensure that data are collected at the same level. For example, to ascertain whether students in poor schools fare poorly, one would have to measure the sum total of school resources available—public funds as well as any private or endowment funds provided by the wider community—rather than calculate the mean levels of parents' status characteristics in the school.

The Second-Order Response as Determinant of the Contemporary Context

Through the first- and second-order response, the Coleman Report's influence extended far beyond the analysis of inequality in education. Intensified development of the concept of social capital (see, for example, Coleman 1988), more nuanced neighborhood effects research (Sharkey and Elwert 2011), experiment-based efforts to assess discrimination (Pager 2003), the employment of natural experiment data where possible (Heckman and Payner 1989), and close study of intrinsically interesting cases for causal insight (Vaughan 1996) are all, in part, responses to rising awareness of the challenge of estimating effects of interest, an awareness that is perhaps the most general and widespread effect of the responses to the Coleman Report. Intriguingly, it is just such context effects on outcomes that key responses to the Coleman Report have shown are a challenge to establish.

Notably, however, such work, as well as the neo–Coleman Report II, suggests that analysts can be empowered rather than paralyzed by the second-order critiques and developments that flow therefrom. On the basis of those developments, useful steps in research can be identified. The belief that one can estimate causal effects simply from data, without imposing a set of assumptions that make estimation possible and meaningful, has been shown to be in error. Thus, if the aim is to estimate causal effects, the first advice is to conduct theoretical analyses. DAGs and ABMs are useful resources, but as Sørensen and Hallinan (1977) and many others demonstrate (for example, Breen and Goldthorpe 1997; Lucas 2009), just developing and manipulating the theoretical equations implied by various claims can be illuminating and helpful.

Once such theoretical work has created enough clarity for informative empirical analyses to proceed, the next step is to seek out or collect data sufficient to the task. In the case of the multilevel model, data should be used or collected in such a way that it is a fully multilevel probability sample. It is helpful to follow Hauser's advice and measure the factors of interest at the levels at which they occur—and to never aggregate level 1 variables to produce alleged level 2 variables. Parameters of interest should be explicitly identified.

This counsel is where the second-order developments leave us—cautious, committed to a less rote process of evidence generation, and hopeful if not optimistic.

CONCLUDING REMARKS

The Coleman Report, despite its groundbreaking nature, had limitations. Fortunately for those who seek to understand the implications of that report, many of those limitations were

3. In ES designs, persons are given beepers that beep at random times. When beeped, the person is to record the requested information about their activities, environment, and state of mind. ES designs were used to establish the concept of flow and its relevance for optimal experience (Csikszentmihalyi 1990).

revealed in the immediate aftermath of the study, owing to the institutional structures and goodwill of the researchers, who made data available and engaged the dialogue forthrightly.

Afterwards, analysts took the trail-blazing nature of the Coleman Report as an inspiration to intensify their efforts to inform our understanding of schools and inequality as well as the methodologies and difficulties of such research. Promising responses to the two criticisms of focus were developed out of those intensified efforts. But key aspects of those responses themselves contained hidden flaws that undermined their success. Many of those flaws were exposed in the transition out of the period of first-order response.

Intriguingly, Coleman's later work in analytic sociology, most notably Coleman's boat, contributed key resources to the second-order response. The second-order developments have been characterized by additional complexity and tools that, if used well, require a measured, cautious, step-by-step process of evidence generation.

Running through the entire dialogue has been a concern with the grounds of causal inference and a desire for analyses to be based in a plausible process understanding of schooling. At this stage of knowledge, we may finally be poised to resolve the critiques originally raised many decades ago and satisfy the desire for solid causal conclusions grounded in and contributing to a developing, accurate understanding of how schools work.

APPENDIX A

The Multilevel Model in Equation Form

A multi-equation specification is perhaps the clearest way to convey the distinct features of the multilevel model. In the following multi-equation specification for an interval-level dependent variable, one individual-level (i) equation contains an outcome variable (Y) and several determinants (X's). If a coefficient (α) for a given X is allowed to vary over J macrolevel units, then an equation at the macro level may contain coefficients (λ's) for macrolevel factors (Z's) that may partially determine α_j. So, for example, equations A1–A5 describe a two-level means-as-outcomes model:

$$\text{(A1)}\ Y_{ij} = \alpha_{0j} + \alpha_1 X_{1ij} + \alpha_2 X_{2ij} + \alpha_3 X_{3ij} + \varepsilon_{ij}$$

$$\text{(A2)}\ \alpha_{0j} = \lambda_{00} + \lambda_{10} Z_{1j} + \lambda_{20} Z_{2j} + \delta_{0j}$$

$$\text{(A3)}\ \alpha_1 = \lambda_{01}$$

$$\text{(A4)}\ \alpha_2 = \lambda_{02}$$

$$\text{(A5)}\ \alpha_3 = \lambda_{03}$$

$$\varepsilon_{ij} \sim N(0,\sigma^2);\ \delta_{kj} \sim N(0, \mathbf{T});\ \rho^{\varepsilon\delta} = 0$$

Equation A1 specifies the level 1 equation, whereas equations A2–A5 specify the macrolevel (or level 2) equations. ε_{ij} and δ_{0j} are individual- and macrolevel errors with variance σ^2 and variance-covariance matrix $\mathbf{T}$, respectively. The level 1 coefficient for the intercept varies across macrolevel units, while the other level 1 coefficients do not vary. In equation A1, the variation in α_{0j} is partially associated with macrolevel variables Z_1 and Z_2.

Similarly, equations A6–A10 constitute a slopes-as-outcomes model; for clarity, I switch to β's and γ's for these equations:

$$\text{(A6)}\ Y_{ij} = \beta_0 + \beta_{1j} X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \varepsilon_{ij}$$

$$\text{(A7)}\ \beta_0 = \gamma_{00}$$

$$\text{(A8)}\ \beta_{1j} = \gamma_{01} + \gamma_{11} Z_{1j} + \gamma_{21} Z_{2j} + \delta_{1j}$$

$$\text{(A9)}\ \beta_2 = \gamma_{02}$$

$$\text{(A10)}\ \beta_3 = \gamma_{03}$$

$$\varepsilon_{ij} \sim N(0,\sigma^2);\ \delta_{kj} \sim N(0, \mathbf{T});\ \rho^{\varepsilon\delta} = 0$$

In equations A1–A10, note that context-level variation is measured by context-specific coefficients, α_{0j} in equation A1 and β_{1j} in equation A6—that is, the coefficients with j subscripts. Thus, context-level variation can only be explained by macrolevel variables (Z_1 and Z_2) and the error terms associated with that coefficient (δ_{0j} or δ_{1j}). The model partitions the variance across levels and estimates more appropriate standard errors for macrolevel coefficients.

An equivalent specification of the model writes it all as one equation, as in equation A11, which combines equations A1–A5, and equation A12, which combines equations A6–A10:

$$(A11)\ Y_{ij} = \lambda_{00} + \lambda_{10}Z_{1j} + \lambda_{20}Z_{2j} + \delta_{0j} + \lambda_{01}X_{1ij} + \lambda_{02}X_{2ij} + \lambda_{03}X_{3ij} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N(0,\sigma^2);\ \delta_{kj} \sim N(0,T);\ \rho^{\varepsilon\delta} = 0$$

$$(A12)\ Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{11}X_{1ij}Z_{1j} + \gamma_{21}X_{1ij}Z_{2j} + \delta_{1j} + \delta_{1j}X_{1ij} + \gamma_{02}X_{2ij} + \gamma_{03}X_{3ij} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N(0,\sigma^2);\ \delta_{kj} \sim N(0,T);\ \rho^{\varepsilon\delta} = 0$$

APPENDIX B

Why Context-Unrepresentative Sampling Fails: An Illustration of the Need for Fully Multilevel Probability Sampling for the Multilevel Model

The discussion here (drawn from Lucas 2014) uses population parameters (for example, α's) rather than sample entities (a's) to underscore that at issue is parameter identification, not simply estimation efficiency.

Consider a probability sample designed to represent a nation with sampled individuals lodged in contexts. Owing to the context-unrepresentative sample design, each context j is composed of two groups of persons: (1) sampling-reachable (r), and (2) sampling-unreachable (u). The proportion of r (p) and u ($1-p$) varies across contexts. By definition, persons' allocation to group r or u is not random. Further, the determinants of assignment are unknown, and the allocation process may vary across contexts. Thus, groups r and u differ in unknown yet systematic ways, such that group r provides no retrievable information on the parameters for group u without heavy assumptions.

Given this design, true context-specific population parameters are actually mixtures:

$$(B1)\ \alpha_{0j} = p_j\,\alpha_{0j,r} + (1 - p_j)\,\alpha_{0j,u}$$

Using the multilevel model (MLM) with such data treats $\alpha_{0j,r}$ as if it is α_{0j}. Expressed as a function of the true population parameter, in reality:

$$(B2)\ \alpha_{0j,r} = (p_j\alpha_{0j,u} - \alpha_{0j,u} + \alpha_{0j}) / p_j$$

which is not in general equal to α_{0j}. Using $\alpha_{0j,r}$ as if it is α_{0j} is mistaken, for:

$$(B3)\ \alpha_{0j} - \alpha_{0j,r} = p_j\alpha_{0j,r} + \alpha_{0j,u} - p_j\alpha_{0j,u} - \alpha_{0j,r}$$

which is not in general zero. Equation B3 indicates that it will be difficult to establish the magnitude and sign of the difference between $\alpha_{0j,r}$ and α_{0j}. First, to identify magnitude and sign requires information about the unreachable subpopulation in each context. By definition, one has no such information. Second, the unknown bias varies by context as a function of p_j, $\alpha_{0j,r}$, and $\alpha_{0j,u}$, such that large context-specific biases may exist even if the average bias is zero.

The use of $\alpha_{0j,r}$ for α_{0j} causes further problems, for equation A2 becomes:

$$(B4)\ \alpha_{0j,r} = (p_j\alpha_{0j,u} - \alpha_{0j,u} + \alpha_{0j}) / p_j = \lambda^*_{00} + \lambda^*_{10}Z_{1j} + \lambda^*_{20}Z_{2j} + \delta^*_{0j}.$$

For equation B4 to produce the sought-after level 2 population parameter:

$$(B5)\ \lambda^*_{00} = \lambda_{00}$$

$$(B6)\ \lambda^*_{10} = \lambda_{10}$$

$$(B7)\ \lambda^*_{20} = \lambda_{20}$$

must be true. But there is little reason to believe that equations B5 through B7 are true, and if they are false, it will be difficult to recover λ_{00}, $\lambda_{10,}$ and λ_{20} from the model for $\alpha_{0j,r}$.

One of two possible conditions can make r sufficient for estimating α_{0j} unbiasedly. First, if all $p_j = 1.00$, then there is no problem. Of course, if all $p_j = 1.00$, then one has context-representative probability sampling.

Failing this condition, however, one may justify the MLM by assuming:

$$(B8)\ \alpha_{0j,r} = \alpha_{0j,u}$$

If equation B8 holds, then there is no problem with using only those in group r to estimate the population parameter(s). There is, however, little reason to suspect that equation B8 will hold in general. Thus, those using the MLM must either use context-representative probability samples ($p_j = 1.00$) or explain why they believe equation B8 holds for the parameters of interest that vary across contexts.

For establishment of the other criteria for an FMP sample, see Lucas (2014).

APPENDIX C

Identification Problems with Estimating Contextual Effects

Equations C1–C3 are drawn directly from Manski (1993); for fuller discussion of these equations and their interrelation, see that work. Manski (1993, equations 42–43) proposes the following equations as a way to distinguish four different effects conceptually:

$$(C1)\ y = \alpha + \beta\, E(y|x) + E(z|x)'\gamma + x'\delta_1 + z'\eta + u$$

$$(C2)\ E(u|x,z) = x'\delta_2$$

where x stands for a set of variables that identify the contexts (for example, a set of dummy variables), z stands for a set of individual-level explanatory variables, β signifies the endogenous effect, γ signifies the contextual effect, $E(y|x)$ stands for the context-specific mean of y, $E(z|x)$ represents the context-specific mean of z, δ_1 signifies the ecological effect, η reflects the effect of z on y, and u captures the error in equation C1. Because $E(u|x,z) \neq 0$ in general, δ_2 reflects the effects of similarity of unobserved attributes for persons in the same context—that is, the correlated individual-level effects. Substituting equation C2 into equation C1 and rearranging the terms reveals that the two equations imply:

$$(C3)\ E(y|x,z) = \alpha + \beta\, E(y|x) + E(z|x)'\gamma + x'(\delta_1 + \delta_2) + z'\eta.$$

Given equation C3, researchers will encounter major difficulty identifying peer effects.

REFERENCES

Aigner, Dennis J. 1970. "A Comment on Problems in Making Inferences from the Coleman Report." *American Sociological Review* 35(2): 249–52.

Akkina, K. R. 1974. "Application of Random Coefficient Regression Models to the Aggregation Problem." *Econometrica* 42(2): 369–75.

Alexander, Karl L., Doris R. Entwisle, and Linda Steffel Olson. 2007. "Lasting Consequences of the Summer Learning Gap." *American Sociological Review* 72(2): 167–80.

Barr, Rebecca, and Robert Dreeben, with Nonglak Wiratchai. 1983. *How Schools Work*. Chicago: University of Chicago Press.

Bowers, William J. 1968. "Normative Constraints on Deviant Behavior in the College Context." *Sociometry* 31(4): 370–85.

Bowles, Samuel, and Henry M. Levin. 1968. "The Determinants of Scholastic Achievement: An Appraisal of Some Recent Evidence." *Journal of Human Resources* 3(1): 3–24.

Boyd, Lawrence H., Jr., and Gudmund R. Iversen. 1979. *Contextual Analysis: Concepts and Statistical Techniques.* Belmont, Calif.: Wadsworth.

Breen, Richard, and John H. Goldthorpe. 1997. "Explaining Educational Differentials: Towards a Formal Rational Action Theory." *Rationality and Society* 9(3): 275–305.

Bruch, Elizabeth E., and Robert D. Mare. 2006. "Neighborhood Choice and Neighborhood Change." *American Journal of Sociology* 112(3): 667–709.

Bryan, Mark L., and Stephen P. Jenkins. 2016. "Multilevel Modelling of Country Effects: A Cautionary Tale." *European Sociological Review* 32(1): 3–22.

Bryk, Anthony S., Valerie E. Lee, and Peter B. Holland. 1993. *Catholic Schools and the Common Good.* Cambridge, Mass.: Harvard University Press.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, Calif.: Sage Publications.

Bryk, Anthony S., Stephen W. Raudenbush, M. Seltzer, and Richard T. Congdon. 1988. *An Introduction to HLM: Computer Program and User's Guide.* 2nd ed. Chicago: University of Chicago Department of Education.

Burstein, Leigh, Robert L. Linn, and Frank J. Capell. 1978. "Analyzing Multilevel Data in the Presence of Heterogenous Within-Class Regressions." *Journal of Educational Statistics* 3(4): 347–83.

Cain, Glen G., and Harold W. Watts. 1970. "Problems in Making Policy Inferences from the Coleman Report." *American Sociological Review* 35(2): 228–42.

Chubb, John E., and Terry M. Moe. 1988. "Politics, Markets, and the Organization of Schools." *American Political Science Review* 82(4): 1065–87.

Coleman, James S. 1968. "Equality of Educational Opportunity: Reply to Bowles and Levin." *Journal of Human Resources* 3(2): 237–46.

———. 1988. "Social Capital in the Creation of Hu-

man Capital." *American Journal of Sociology* 94: S95–120.

———. 1990. *Foundations of Social Theory.* Cambridge, Mass.: Harvard University Press.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederick D. Weinfeld, and Robert L. York. 1966. *Equality of Educational Opportunity.* Washington: U.S. Department of Health, Education, and Welfare, Office of Education.

Coleman, James, Thomas Hoffer, and Sally Kilgore. 1981. *Public and Private Schools: Report to the National Center for Education Statistics.* Chicago: National Opinion Research Center.

Crain, Robert L. 1967. "Review of *Equality of Educational Opportunity*." *American Journal of Sociology* 73(3): 354–56.

Crane, Jonathan. 1991. "The Epidemic Theory of Ghettos and Neighborhood Effects on Dropping Out and Teenage Childbearing." *American Journal of Sociology* 96(5): 1226–59.

Csikszentmihalyi, Mihalyi. 1990. *Flow: The Psychology of Optimal Experience.* New York: Harper & Row.

Dyer, Henry. 1968. "School Factors and Equal Educational Opportunity." *Harvard Educational Review* 38(1): 38–56.

Ehrenberg, Ronald G., and Dominic J. Brewer. 1995. "Did Teachers' Verbal Ability and Race Matter in the 1960s? *Coleman* Revisited." *Economics of Education Review* 14(1): 1–21.

Elwert, Felix. 2013. "Graphical Causal Models." In *Handbook of Causal Analysis for Social Research,* edited by Stephen L. Morgan. New York: Springer.

Farkas, George. 1974. "Specifications, Residuals, and Context Effects." *Sociological Methods and Research* 2(3): 333–63.

Fuchs, Ira H. 2001. "Prospects and Possibilities of the Digital Age." *Proceedings of the American Philosophical Society* 145(1): 45–53.

Gamoran, Adam. 1987. "The Stratification of High School Learning Opportunities." *Sociology of Education* 60(3): 135–55.

———. 1992. "The Variable Effects of High School Tracking." *American Sociological Review* 57(6): 812–28.

Gamoran, Adam, and Mark Berends. 1987. "The Effects of Stratification in Secondary Schools: Synthesis of Survey and Ethnographic Research." *Review of Educational Research* 57(4): 415–35.

Gamoran, Adam, and Robert Dreeben. 1986. "Coupling and Control in Educational Organizations." *Administrative Science Quarterly* 31(4): 612–32.

Gamoran, Adam, and Robert D. Mare. 1989. "Secondary School Tracking and Educational Equality: Compensation, Reinforcement, or Neutrality?" *American Journal of Sociology* 94(5): 1146–83.

Garet, Michael S., and Brian DeLany. 1988. "Students, Courses, and Stratification." *Sociology of Education* 61(2): 61–77.

Goldberger, Arthur S., and Glen G. Cain. 1982. "The Causal Analysis of Cognitive Outcomes in the Coleman, Hoffer, and Kilgore Report." *Sociology of Education* 55(2–3): 103–22.

Goldstein, Harvey. 1994. "Multilevel Cross-Classified Models." *Sociological Methods and Research* 22(3): 364–75.

———. 2003. *Multilevel Statistical Models.* 3rd ed. London: Arnold.

Goldstein, Harvey, M. J. R. Healy, and Jon Rasbash. 1994. "Multilevel Time Series Models with Application to Repeated Measures Data." *Statistics in Medicine* 13(16): 1643–55.

Grant, Gerald. 1973. "Shaping Social Policy: The Politics of the Coleman Report." *Teachers College Record* 75(1): 17–54.

Hanushek, Eric A., and John F. Kain. 1972. "On the Value of 'Equality of Educational Opportunity' as a Guide to Public Policy." In *On Equality of Educational Opportunity,* edited by Frederick Mosteller and Daniel Patrick Moynihan. New York: Random House.

Hauser, Robert M. 1970a. "Context and Consex: A Cautionary Tale." *American Journal of Sociology* 75(4): 645–64.

———. 1970b. "The Author Replies." *American Journal of Sociology* 76(3): 517–20.

———. 1974. "Contextual Analysis Revisited." *Sociological Methods and Research* 2(3): 365–75.

Havighurst, Robert J. 1967. Review. *Journal of the American Statistical Association* 62: 1071–73.

Heckman, James J., and Brook S. Payner. 1989. "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina." *American Economic Review* 79(1): 138–77.

Hedeker, Donald, and Robert D. Gibbons. 1994. "A Random-Effects Ordinal Regression Model for Multilevel Analysis." *Biometrics* 50(4): 933–44.

Hedström, Peter. 2005. *Dissecting the Social: On the Principles of Analytic Sociology.* New York: Cambridge University Press.

Heyns, Barbara. 1979. *Summer Learning and the Effects of Schooling.* New York: Academic Press.

Hoffer, Thomas B. 1992. "Middle School Ability Grouping and Student Achievement in Science and Mathematics." *Educational Evaluation and Policy Analysis* 14(3): 205–27.

Horan, Patrick M. 1978. "Is Status Attainment Research Atheoretical?" *American Sociological Review* 43(4): 534–41.

Hotchkiss, Lawrence, and Linda E. Dorsten. 1987. "Curriculum Effects on Early Post–High School Outcomes." *Research in the Sociology of Education and Socialization* 7(1): 191–219.

Hox, Joop. 1995. *Applied Multilevel Analysis.* Amsterdam: TT-Publikaties.

Irby, Nolan Meaders. 1930. "A Program for the Equalization of Educational Opportunities in the State of Arkansas." PhD diss., George Peabody College for Teachers, Nashville, Tenn.

Jones, James D., Beth E. Vanfossen, and Margaret E. Ensminger. 1995. "Individual and Organizational Predictors of High School Track Placement." *Sociology of Education* 68(4): 287–300.

Kreft, Ita, and Jan de Leeuw. 1998. *Introducing Multilevel Modeling.* Thousand Oaks, Calif.: Sage Publications.

Krueger, Alan B., and Diane M. Whitmore. 2002. "Would Smaller Classes Help Close the Black-White Achievement Gap?" In *Bridging the Achievement Gap,* edited by John E. Chubb and Tom Loveless. Washington, D.C.: Brookings Institution Press.

Lindley, D. V., and A. F. M. Smith. 1972. "Bayes Estimates for the Linear Model" (with discussion). *Journal of the Royal Statistical Society, Series B* 34(1): 1—41.

Little, Roderick J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87(420): 1227–37.

Looker, E. Diane. 1989. "Accuracy of Proxy Reports of Parental Status Characteristics." *Sociology of Education* 62(4): 257–76.

Lucas, Samuel Roundfield. 1999. *Tracking Inequality: Stratification and Mobility in American Schools.* New York: Teachers College Press.

———. 2008. "School Tracking." In *Encyclopedia of the Life Course and Human Development,* edited by Deborah Carr. Farmington Hills, Mich.: Macmillan Reference, USA.

———. 2009. "Stratification Theory, Socioeconomic Background, and Educational Attainment: A Formal Analysis." *Rationality and Society* 21(4): 459–511.

———. 2014. "An Inconvenient Dataset: Bias and Inappropriate Inference with the Multilevel Model." *Quality and Quantity* 48(3): 1619–49.

Lucas, Samuel R., and Mark Berends. 2002. "Sociodemographic Diversity, Correlated Achievement, and de Facto Tracking." *Sociology of Education* 75(4): 328–48.

———. 2007. "Race and Track Location in U.S. Public Schools." *Research in Social Stratification and Mobility* 25(3): 169–87.

Lucas, Samuel R., and Adam Gamoran. 2002. "Tracking and the Achievement Gap." In *Bridging the Achievement Gap,* edited by John E. Chubb and Tom Loveless. Washington, D.C.: Brookings Institution Press.

Lucas, Samuel R., Steven Ingels, Harrison Greene, and Louise Little. 1987. "Student Data Collection." In Steven Ingels et al., *Field Test Report: National Education Longitudinal Study of 1988 (NELS:88),* prepared for U.S. Department of Education, Center for Education Statistics.

Lucas, Samuel R., and Alisa Szatrowski. 2014. "Qualitative Comparative Analysis in Critical Perspective." *Sociological Methodology* 44(1): 1–79.

Luke, Douglas A. 2004. *Multilevel Modeling.* Thousand Oaks, Calif.: Sage Publications.

Mack, Raymond W. 1968. *Our Children's Burden: Studies of Desegregation in Nine American Communities.* New York: Random House.

Manski, Charles F. 1993. "Identification Problems in the Social Sciences." *Sociological Methodology* 23(1): 1–56.

———. 1995. *Identification Problems in the Social Sciences.* Cambridge, Mass.: Harvard University Press.

Mason, William M., George Y. Wong, and Barbara Entwisle. 1983. "Contextual Analysis Through the Multilevel Linear Model." *Sociological Methodology* 14(2): 72–103.

Mickelson, Roslyn Arlin. 2001. "Subverting Swann: First- and Second-Generation Segregation in Charlotte-Mecklenberg Schools." *American Educational Research Journal* 38(2): 215–52.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research.* New York: Cambridge University Press.

Moses, Earl R. 1941. "Indices of Inequalities in a Dual

System of Education." *Journal of Negro Education* 10(2): 239–44.

Mosteller, Frederick, and Daniel P. Moynihan. 1972. *On Equality of Educational Opportunity.* New York: Random House.

Muthén, Bengt O. 1994. "Multilevel Covariance Structure Analysis." *Sociological Methods and Research* 22(3): 376–98.

Nichols, Robert C. 1966. "Schools and the Disadvantaged." *Science* 154(3754): 1312–14.

Norton, John K. 1926. *The Ability of the States to Support Education.* Washington, D.C.: National Educational Association.

Nystrand, Martin, and Adam Gamoran. 1991. "Instructional Discourse, Student Engagement, and Literature Achievement." *Research in the Teaching of English* 25(3): 261–90.

Nystrand, Martin, Lawrence L. Wu, Adam Gamoran, Susie Zeiser, and Daniel A. Long. 2003. "Questions in Time: Investigating the Structure and Dynamics of Unfolding Classroom Discourse." *Discourse Processes* 35(2): 135–98.

Oakes, Jeannie. 1982. "Classroom Social Relationships: Exploring the Bowles and Gintis Hypothesis." *Sociology of Education* 55(4): 197–212.

Pager, Devah. 2003. "The Mark of a Criminal Record." *American Journal of Sociology* 108(5): 937–75.

Pallas, Aaron M., Doris R. Entwisle, Karl L. Alexander, and M. Francis Stluka. 1994. "Ability-Group Effects: Instructional, Social, or Institutional?" *Sociology of Education* 67(1): 27–46.

Pearl, Judea. 2010. "The Foundations of Causal Inference." *Sociological Methodology* 40(1): 75–149.

Pettigrew, Thomas F. 1968. "Race and Equal Educational Opportunity." *Harvard Educational Review* 38(1): 66–76.

Pfautz, Harold W. 1967. Review. *American Sociological Review* 32(3): 481–83.

Phillips, Myrtle R. 1932. "Financial Support." *Journal of Negro Education* 1(2): 108–36.

Plato. 1921. *Cratylus.* In *Plato in Twelve Volumes,* vol. 12, translated by Harold N. Fowler. Cambridge, Mass.: Harvard University Press; London: William Heinemann Ltd. Available at: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0172%3Atext%3DCrat.%3Asection%3D402a (accessed June 28, 2016).

Pinheiro, José C., and Douglas M. Bates. 2004. *Mixed-Effects Models in S and S-PLUS.* New York: Springer.

Riehl, Carolyn, Aaron M. Pallas, and Gary Natriello. 1999. "Rites and Wrongs: Institutional Explanations for the Student Course-Scheduling Process in Urban High Schools." *American Journal of Education* 107(2): 116–54.

Robins, James M. 1999. "Association, Causation, and Marginal Structural Models." *Synthese* 121(1): 151–79.

Robins, James M., Miguel Angel Hernán, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5): 550–60.

Rosenbaum, James E. 1980. "Track Misperceptions and Frustrated College Plans: An Analysis of the Effects of Tracks and Track Perceptions in the National Longitudinal Survey." *Sociology of Education* 53(2): 74–88.

Schelling, Thomas C. 1971. "Dynamic Models of Segregation." *Journal of Mathematical Sociology* 1(2): 143–86.

Sewell, William H. 1967. "Review of *Equality of Educational Opportunity*." *American Sociological Review* 32: 475–79.

Sharkey, Patrick, and Felix Elwert. 2011. "The Legacy of Disadvantage: Multigenerational Neighborhood Effects on Cognitive Ability." *American Journal of Sociology* 116(6): 1934–81.

Shernoff, David J., Mihaly Csikszentmihalyi, Barbara Schneider, and Elisa Steele Shernoff. 2003. "Student Engagement in High School Classrooms from the Perspective of Flow Theory." *School Psychology Quarterly* 18(2): 158–76.

Slavin, Robert E. 1990. "Achievement Effects of Ability Grouping in Secondary Schools: A Best-Evidence Synthesis." *Review of Educational Research* 60(3): 471–99.

Snijders, Tom, and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Thousand Oaks, Calif.: Sage Publications.

Sørensen, Aage Bøttger. 1970. "Organizational Differentiation of Students and Educational Opportunity." *Sociology of Education* 43(4): 355–76.

Sørensen, Aage B., and Maureen Hallinan. 1977. "A Reconceptualization of School Effects." *Sociology of Education* 50(4): 273–89.

Steele, Fiona, Harvey Goldstein, and William Browne. 2004. "A General Multistate Competing Risks Model for Event History Data, with an Application to a Study of Contraceptive Use Dynamics." *Journal of Statistical Modelling* 4(2): 145–59.

Strayer, George. 1949. *The Report of a Survey of the Public Schools of the District of Columbia.* Washington: U.S. Government Printing Office.

Vaughan, Diane. 1996. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA.* Chicago: University of Chicago Press.

Wallace, Elsie Hill. 1951. "A Study of Negro Elementary Education in North Alabama." *Journal of Negro Education* 20(1): 39–49.

Wiatrowski, Michael D., Stephen Hansell, Charles R. Massey, and David L. Wilson. 1982. "Curriculum Tracking and Delinquency." *American Sociological Review* 47(1): 151–60.

Wilson, Alan. 1968. "Social Class and Equal Educational Opportunity." *Harvard Educational Review* 38(1): 77–84.

Winship, Christopher, and Robert D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18: 327–50.

Wong, George Y., and William M. Mason. 1985. "The Hierarchical Logistic Regression Model for Multilevel Analysis." *Journal of the American Statistical Association* 80(391): 513–24.

Zellner, Arnold. 1969. "On the Aggregation Problem: A New Approach to a Troublesome Problem." In *Economic Models, Estimation, and Risk Programming: Essays in Honor of Gerhard Tintner,* Lecture Notes 15, edited by K. A. Fox, J. K. Sengupta, and G. V. L. Narasimham. New York: Springer-Verlag.