Emergent Principles for the Design, Implementation, and Analysis of Cluster-Based Experiments in Social Science

By
THOMAS D. COOK

In experimentally designed research, many good reasons exist for assigning groups or clusters to treatments rather than individuals. This article discusses them. But cluster-level designs face some unique or exacerbated challenges. The article identifies them and offers some principles about them. One emphasizes how statistical power and sample size estimation depend on intraclass correlations, particularly after conditioning on the use of cluster-level covariates. Another stresses assigning experimental units at the lowest level of aggregation possible, provided this does not subtly change the research question. A third emphasizes the utility of minimizing and measuring interunit communication, though neither is easy to achieve. A fourth advises against experiments that are totally black box and so leave program implementation and process unstudied, though such study often makes the research process more salient. The last principle involves the utility of describing treatment heterogeneity and estimating its consequences, though causal conclusions about the heterogeneity will be less well warranted compared to conclusions about the intended treatment, every experiment's major focus.

Keywords: cluster random assignment; cluster level; allocation principle; interventions; unit of assignment; statistical power; treatment contamination; causal chain

Purposes

This article discusses some of the rationales, problems, and solutions associated with randomized experiments that assign groups to treatments rather than individuals. These groups might be entire schools, communities, or work sites, and typically, every individual within such a unit is assigned to the same treatment status, though individuals could also be randomly chosen within groups. Experiments of this form

Thomas D. Cook is a professor in the Institute of Policy Research, Northwestern University. He received his Ph.D. from Stanford University in 1967. His areas of interest include social psychology, social science of human development, evaluation research, and education. He is interested in social science methods for inferring causation, and through this interest he examines

DOI: 10.1177/0002716205275738

are often called group-, cluster- or place-based, since the unit of assignment is a cluster of individuals who share space.

Systematic social forces operate so that neighborhoods tend to vary in their constellation of residents just as schools vary in their profiles of students and teachers. That people fill space in systematic ways is a major finding of demography and sociology and is central to studying the determinants, processes, and consequences of who lives where or who attends a particular type of institution. However, such systematic selection is a problem for other research purposes and leads to most of the problems associated with using cluster random assignment to learn about how a given intervention affects groups of people who share space.

Social scientists have had considerable experience with randomly assigning individuals to treatments and even more with randomly selecting individuals to be in surveys. But experience with the design, implementation, and analysis of cluster-based experiments is more recent and limited. This is unfortunate since social scientists often identify the nature of methodological problems, and some possible solutions to them, in the crucible of experience—by doing a particular type of research and reflecting on it. Since multilevel experiments are rare, the wisdom about how to do them well is less developed than the corresponding wisdom about individual-level experiments. And it is certainly less than the wisdom about implementing random selection in survey research, where seventy years of relevant experience have accumulated on improving some discrete aspect of survey research. This experience even comes from experiments on, say, how to word items, to record responses, to do face-to-face or telephone interviews, or how interviewer and interviewee race or gender should be managed. My main purpose here is to increase the relevant wisdom about successfully implementing cluster-level randomized experiments, even if only at the margin.

While I will consider theoretical work on cluster-based studies in statistics and research design, my main emphasis is on lessons I have learned over the past decade in implementing experiments with schools as the unit of analysis (Cook et al. 1999; Cook, Hunt, and Murphy 2000), in consulting about experiments designed to prevent cardiovascular disease in smaller cities (Farquhar et al. 1990; Blackburn et al. 1984), and in doing work on neighborhood social relations and their effects on family life (Cook, Shagle, and Degirmencioglu 1997; Cook et al. 2002; Furstenberg et al. 1999). In education, many different units of assignment are possible, including districts, schools, grade levels, classrooms, or individuals. In disease prevention and urban improvement, different units are again possible—cities, planners' neighborhoods, census tracts, block groups, or individuals. However, school-based prevention studies assign schools and classrooms, not neighbor-

---

issues in evaluation research, primarily in the areas of education and community health. Works include Quasi-Experimentation Design and Analysis Issues for Field Settings, Qualitative and Quantitative Methods in Evaluation Research *and* The Foundations of Evaluation Theory. *He is a fellow of the American Academy of Arts and Sciences and a Margaret Mead Fellow of the American Academy of Political and Social Science. At Northwestern University, he it the Joan and Serepta Harrison Professor of Ethics and Justice and a professor of sociology.*

hood spaces, to the treatment groups used to explore ways of reducing violence, drug use, or smoking among students.

I want to be clear up front about some limitations to this article. Many of the problems I consider do not have empirically and consensually validated solutions of which I am aware. This forces me to consider incomplete solutions, both contingent ones that work under special circumstances and partial ones that ameliorate a problem without solving it. Of the problems I consider, some are unique to cluster-based designs because they are products of social clustering. Others are germane to experiments that assign individuals but are exacerbated in the group context. Also, I limit the analysis to single experiments that probe whether a causal relationship is plausible with the specific populations, settings, historical periods, manipulations, and measures sampled in that study. Examining the many other factors that moderate such a cause-effect relationship is better accomplished by synthesizing results across many experiments. But space precludes examining the special issues entailed in synthesizing cluster-based experiments rather than individual-level ones, a topic that is even less developed than thinking about single cluster-level experiments.

Why Assign at the Cluster Level?

It is intrinsic to some interventions that they are based on superindividual concepts. For instance, whole school reform seeks to modify the academic and social climate of a school building and to improve teacher practice norms throughout the whole school. Program developers vary in how they want to achieve such goals. But most efforts include giving staff more responsibility for the choices that affect their practice, including a role in setting the school's annual objectives and goals. Professional development is also typically provided to enhance specific practice skills, to build commitment to the reform effort, and to get staff to accept their interdependence with colleagues and parents in furthering students' development. The relevant theories regularly invoke constructs like school governance, culture, climate, norms, teams, and networks, each a concept that cannot be reduced to individual behavior. The hope is even to create a new culture whose norms will affect the behavior of current teachers and students and also future ones.

Interventions can also be designed to target individual staff members and students and yet require a school level of assignment. For instance, program designers can pay special attention to those individuals who function as building-level opinion leaders by virtue of their social networks and power to influence, using these individuals to catalyze building- or classroom-level change. As an explanatory construct, social network is also inherently superindividual.

But social communication processes can occur in contexts where classrooms are not selected because of special network or normative links to the rest of the school. Interventions designed to be self-contained within smaller groups like classrooms can still be talked about in other classrooms, with the result that program details can percolate from classes receiving a particular intervention into classes designed

to receive a different intervention. Since this treatment dissemination clouds interpretation of any study results, whenever unplanned treatment dissemination is plausible schools need to be the unit of assignment rather than classrooms or grade levels. In this case, the unit of assignment is chosen to protect against a potential source of bias rather than to support an intervention that is specifically designed to activate a superindividual process like norm creation or dissemination through an existing social network.

Another argument for cluster-based assignment has to do with desired impact. The hope is that individual change will be greater in size, permanence, and generalization if it is achieved through group- rather than individual-level processes. One reason for this hope is that larger units like schools entail reaching more individuals when compared to efforts designed to reach only individuals. Another is that novel norm and network changes can emerge when larger groups are targeted, changes that then come to characterize the setting as a whole, serving not just to maintain change in those originally exposed to an intervention but also to influence the next cohorts who enter a school or community even after the intervention has been removed. A third reason for hope is that intervention with larger aggregates promises change that is stubborn and resistant to counterforces because it is anchored in multiple rather than single influences. It is one thing for a principal to urge teachers to change, and it is another thing for principals, fellow teachers, school boards, and parents to urge the same thing at different times, in different venues and perhaps even modeling the change in different ways. An educational intervention aimed at individual teachers would not reach as many children or entail as many potentially convergent influences for creating and maintaining new individual and group behaviors.

The hope is to reach more individuals, to change local norms and networks, and to link healthy living to multiple source and times.

The same potential advantages hold in public health. An intervention designed to promote a heart healthy lifestyle might affix heart healthy signs to specific commodities in local grocery stores. It might involve blood pressure screening stations at convenient locations. It might include community-wide races and walks to promote exercise as well as outreach activities to teach students about nutrition, exercise, and stress reduction. Even the local media might be solicited to focus on healthy living. The hope is to reach more individuals, to change local norms and networks, and to link healthy living to multiple source and times.

One should also not forget an important political reason for assigning clusters rather than individuals. Assignment to different treatments inevitably creates a source of inequality and potential resentment among those assigned to the less desirable treatment. Depending on the nature of the intervention, these inequalities and resentments can be especially large and pointed if they involve individuals who are in contact with each other. So it gets administrators off unpleasant hooks if, instead of assigning different treatments within the same class and so making inequalities particularly salient, the treatments are instead assigned between classes or—even better—between schools. The assumption is that individuals are less likely to react negatively to differences that are not under their very noses.

The rationales above reflect the loose way experiments are discussed in ordinary language. Discourse centers on testing the causal influence of some named *X*. But experiments test the effects of a contrast between what happened in the intervention group exposed to *X* relative to some other group, often a no-treatment control group. In formal thinking about experiments, the causal agent is always a comparative entity, whereas in public discourse, it tends to be an absolute one. In ordinary language, the treatment is considered to be invariable, as though *X* were implemented at the same strength in all the units. Variation within each treatment group is common, though, some of it due to differences in treatment exposure. So the real experimental question is whether *X* has a marginally greater influence than its comparison *despite* this within-group variation. However, variation in treatment implementation invites questions about the effects of the treatment implemented at its best rather than its average, a question that particularly interests theorists, program designers, and policy makers desperate to find something that works. This "treatment on treated" question is not technically experimental because assignment to implementation levels is not random. Experiments test whether the average difference in implementation between an *X* and comparison group has an effect over and above the noise from within-group variation. This experimental "intent to treat" conception is more nuanced than the ordinary language understanding that an experiment tests the effects of *X*; and it is not a direct test of the effects of *X* when implemented at its best.

Why the *Random* Assignment of Larger Units?

I offer here a structural rationale for assigning clusters at random that alternatively describes the traditional statistical rationale. The starting point is the sociological reality that societies are structured in complex ways. The American educational system includes a federal level within which states are nested and a state level within which districts are nested. Within these districts are schools, and within these schools are grade cohorts. Within these grade cohorts are classrooms, and within these classes are individual students and teachers. So the nesting is hierarchical, multilayered like a Russian matryoshka doll.

However, the nesting is rarely balanced. A student can be with one set of fellow students in English and another in math. Also, the nesting is often multidimen-

sional. For instance, schools are complexly related in space both to neighborhoods and families. Every student lives in a family and neighborhood that has links to schools and that can also affect educational performance either independently of schools or in complex interaction with them (Cook et al. 2002). And the cross-classification of social contexts is itself unbalanced since some children from a particular neighborhood attend one school and others another, the children in one class or school come from many different neighborhoods, and siblings can attend different schools or even live apart in different neighborhoods and homes (Cook 2003). Thus, to state that social structures are hierarchically ordered only scratches the surface of the myriad forms of this ordering.

To add to the complexity, social structures are not fixed in time. Individual students can change tracks (classes) within high school, both within and between school years. They can also leave their school for reasons other than formal graduation. Furthermore, new schools are founded and old ones shut down; schools regularly change their senior personnel; and where they exist, neighborhood schools sometimes redraw their receiving boundaries. Individuals are contextually embedded in ways that are synchronously multilevel, multidimensional, and imbalanced, and these complex relationships vary with time. This is a daunting structure to understand and model.

Fortunately, well-implemented cluster-based random assignment makes fewer demands on the analyst to know and model the structural complexity of society. When schools are randomly assigned in sufficient numbers, concerns about specific state and district confounds should be minimal since any one state or district factor should be equally represented in each of the treatment groups being compared. At most, one needs to describe these larger contexts and speculate about whether the causal results obtained might have varied with them. Also of little concern are school differences in how classrooms are constituted or otherwise used pedagogically and whether school-correlated neighborhood or family factors may have operated as confounds. Again, such factors should be equally distributed across the various treatment groups. Even temporal shifts in structure are only a causal concern if schools (or students) leave the study in patterned ways that differ by treatment condition. This can be easily checked, of course, and it is a consequence of the treatment even if it bedevils the understanding of other treatment consequences. Nor need one worry about imbalanced context crossings since these also should not vary by condition.

Without random assignment, any school-based causal study would have to struggle to rule out the possibility that these various interdependent structural realities function as causal confounds. With correctly implemented random assignment, the threats emanating from how society is organized and changes over time are dealt with by design and not by any ancillary information and assumptions the analyst is forced to use (and measure where possible). If the art of research design is to make the fewest and most transparent assumptions possible whose behavior is also the best known, then for answering a causal question the random assignment of clusters meets this bill better than its quasi-experimental and nonexperimental alternatives.

Principles of Improved Cluster-Level Design

Principle 1: Know the size of unconditional and conditional intraclass correlations, what determines them, and how they affect statistical power and hence sample size estimation

The systematic ways in which individuals cluster within space means that the observations they provide for data analysis are rarely independent. The degree of nonindependence is indexed by the intraclass correlation (ICC). In its unconditional form, this measures what fraction of the total variance in a variable lies between higher-order units and how much is due to individual differences within them, including error. In its conditional form, it indexes what fraction of the variance is between higher-order units after statistical adjustments have been made for other variables. When the ICC is zero, observations are independent and statistical power and statistical tests can be computed using traditional, individual-level methods. But in my experience, few unconditional ICCs are zero, and inventories of ICCs for a variety of different types of clusters and variables show this across many fields. Such systematic clustering creates problems of two kinds.

The first concerns statistical power and hence sample size estimation. The needed sample size of clustered units is sensitive to the magnitude of ICC, and even very small correlations can increase the sample size needed (Raudenbush and Bryk 2002). Moreover, in most (but not all) circumstances, the number of clusters affects power much more than the number of units within a cluster. So cluster-level experiments should be designed to minimize the ICC by using legitimate statistical adjustments. To understand this, consider academic achievement where, at the school level, unconditional ICC values on nationally available tests usually vary between .05 and .15 (Raudenbush 1997). The rule of thumb has developed from these values that school sample sizes of forty to fifty are usually needed to attain the statistical power traditionally needed to test intercept differences in balanced school-level experiments with two treatment conditions. This is a large number of schools, given the usual financial costs and especially if several school districts have to be sampled, forcing researchers to struggle with different school boards, time schedules, institutional review board procedures, and the like.

Individual- or cluster-level covariates are useful in this context because statistical power and sample size depend on the conditional rather than the unconditional ICC and covariates can reduce the unconditional ICC. The more the covariates do this, the more the study approximates an individual-level analysis. One type of covariate stands out for its ability to minimize the conditional ICC, a cluster-level covariate that is highly correlated with the outcome. Consider school-level achievement means from before and after an intervention and the correlation between them. Being larger units, schools are very reliably measured because individual differences are averaged out and this high reliability will increase correlations. Both measures are on the same test, probably measured in the same way, thus also increasing the correlation between the covariate and outcome. Indeed,

one-year school-level achievement correlations typically range between .70 and .90, and correlations at the higher end of this range can reduce the conditional ICC to close to zero and so require fewer schools. Gargani and Cook (2005) showed this using two reading measures a year apart that were correlated .85, that reduced the ICC from .11 to .02 and that, in a balanced two-group design, required only 22 schools to detect an effect as small as .20, with alpha = .05 and beta = .80. Also relevant is that, when the conditional ICC is so close to zero, within-cluster sample sizes and the number of repeat measures then count more for power compared to when the ICC is higher. So very powerful cluster-level covariates are important for reducing the number of clusters needed and hence for getting more experiments done on the same fixed budget.

Knowing unconditional and especially conditional ICC values and their determinants is vital for the design of multilevel experiments and for analyzing the data from them. And they are important in their own right as descriptions of the extent to which spatially separated units vary on a given attribute.

Nonetheless, the ideal experiment selects more clustered units than the minimum necessary. This is to protect against the attrition of clusters. When this happens, the loss in power can be considerable in small-*N* designs. And cluster attrition does sometimes happen, as when new principals are appointed during a study and want nothing to do with their predecessors' initiatives. The technical ideal in this situation is to keep their schools in the measurement framework even if not in the treatment framework. Retaining them respects the original treatment assignment and intent-to-treat conception and permits unbiased estimates, though these are likely to be smaller because schools with only partial treatment exposure are included in the treatment group. But keeping schools that have dropped out of treatment in the measurement framework is not always possible, and so sample sizes need to be calculated with some slack, slack that can be financially burdensome when entire schools or communities are under study.

But consider the alternative by examining the community heart healthy studies of the 1970s and 1980s. The financial cost of organizing entire communities of at

least forty thousand inhabitants for five years or more meant that the largest study randomly assigned but six communities, three to each treatment condition (Blackburn et al. 1984). Realizing this was a hopelessly underpowered study, the researchers sought to bolster it by measuring many individual-level covariates and up to nine years of annual outcome measurement, but the drag from so few communities was still too great. The next wave of studies was more focused, on, say, just smoking prevention, and the sample size of communities was increased to about twenty spread over two conditions, Project COMMIT having eleven pairs of matched settings (Gail et al. 1992). Since then, the sample size requirements have continued to creep up, with the median target being thirty-five, a number that Eldridge et al. (2004) bemoaned as too low in light of the often binary nature of outcomes in public health. In whole school reform, the experiments have not been much bigger, with the two Comer evaluations having twenty-three and nineteen schools, each equally distributed over two conditions (Cook, Hunt, and Murphy 2000; Cook et al. 2002). While these last studies also utilized school-level covariates, their correlation with the outcome was not at the level that Gargani and Cook (2005) showed were needed.

There are other things that can also be done to increase statistical power at the cluster level when powerful covariates are not available. Increasing sample size is obviously one. The most interesting version of this is when the costs of adding treatment units far exceeds that of adding control ones and so controls are oversampled, as in a third Comer study that had thirty-six schools sampled 2:1. However, the power increment associated with design imbalance depends on the harmonic mean of the two treatments (Raudenbush 1997), making the total effect less than if the thirty-six schools had been equally assigned to treatment and control. But still, power does increase over having twenty-four schools equally balanced. To add clusters, it is also sometimes possible to sample fewer individuals within a school, using the resources saved to add more schools. However, it is usually much more expensive to add a single school than to add many students within a school that is already in the research design. So this option usually adds few schools. But when studying achievement with school sample sizes between twenty and forty, even a small change in the number of schools can make a big difference to power. So the option should not always be rejected out of hand.

The second problem that follows from nonindependent, clustered observations is that statistical tests of the hypothesized causal relationship will be biased unless the ICC is taken into account. If it is not, standard errors will be based on the (inappropriate) large number of units within clusters rather than on the smaller number of clusters. Since standard errors will be inappropriately small, effect sizes will be too large and causal conclusions will be too often positive. This is not a problem for researchers who rely on magnitude estimates that are independent of standard errors. But the vast majority of social scientists use statistical tests, and they have a problem because their hypothesis tests are too liberal.

Sensitivity to the downward bias in standard errors in multilevel designs has increased over the past decade. Researchers are more and more using texts and software programs that explicitly model dependencies in multilevel contexts (e.g.,

Raudenbush and Bryk 2002). So ignorance of ICCs and of their relevance for valid statistical tests is a declining problem. It is still a problem, though, in interpreting past research where published reports failed to take data dependencies into account—for example, in studies of peer tutoring that involve groups of two or more students and so require a group-level unit of analysis and in other forms of peer research where many ICC values are in the .20 to .35 range (Cook, Deng, and Morgano 2005) and so much higher than in school or neighborhood research. Unfortunately, sensitivity to clustering is rare in both of these research traditions, though probably increasing. Knowing unconditional and especially conditional ICC values and their determinants is vital for the design of multilevel experiments and for analyzing the data from them. And they are important in their own right as descriptions of the extent to which spatially separated units vary on a given attribute.

Principle 2: Assign units to treatments at the lowest level of aggregation possible, as long as this does not change the research question

On many practical grounds that I later elaborate, the results from cluster-based randomized experiments tend to be more difficult to interpret than the results from experiments with individuals. Thus, it is imperative to ask whether a causal question about an aggregate process absolutely requires assigning clusters rather than individuals. This is actually a special case of an even broader principle—that units should be assigned to treatment at the lowest level possible in the hierarchy from school districts, schools, grade levels, classrooms, and individuals or in the hierarchy from states, cities, city planning areas, census tracts, census blocks, households, and individuals.

To understand why lower levels of assignment are more desirable ceteris paribus, begin by considering quantitative research on neighborhood effects (Wilson 1987). As explicated by Jencks and Mayer (1990), the dominant research question is, Do neighborhoods affect children or adults for reasons that are more than the sum of individual differences? Of course, neighborhoods inevitably differ in composition, in the average of their individual difference profiles. But Jencks and Mayer's framing equates neighborhood effects with "emergent properties," with social processes that emerge from social interaction and that involve superindividual explanatory concepts like norms, networks, cultures, or institutions. In this formulation, summing individual differences cannot suffice to create a neighborhood effect because it does not speak to a collective process or structure.

A specific research strategy is associated with this "emergent properties" conception of a neighborhood effect. It depends on correlating neighborhood attributes selected for their theoretical interest with a particular outcome that has been statistically adjusted for all the measured individual difference attributes that vary with it and with the neighborhood differences under analysis. These individual differences are typically measured from surveys. The individual outcome data also often come from surveys but sometimes from administrative records too. And the

geo-coded neighborhood attributes come either from the decennial census or from individual survey responses about neighborhood conditions aggregated across residents to the neighborhood level.

This strategy has revealed neighborhood "effects" that have been interpreted as modest (Brooks-Gunn, Duncan, and Aber 1997). The implication is that the unadjusted neighborhood differences one observes in daily life and descriptive statistics reflect who lives where and not what happens to them where they live. That is, they would have behaved similarly had they lived in affluent or in impoverished neighborhoods, suggesting that place does not matter much to individual welfare. Another possibility, though, is the methodology generating this momentous conclusion has serious limitations. So some scholars went a step further and sought to examine how spontaneous *changes* in one's neighborhood were related to *changes* in individual outcomes, again after statistically controlling for many measured and neighborhood-correlated individual differences. But this new strategy also failed to show large and systematic neighborhood effects and still leaves sophisticated readers wondering whether all selection effects have been controlled and whether the treatment contrast is not unduly restricted since most spontaneous moves are to neighborhoods whose attributes are not very different from the neighborhoods one just left. What to do, then, given that it is not really feasible to randomly assign whole neighborhoods to some kind of dramatic economic, social, and psychological upgrading, which analysts do not know how to do well in theory, let alone in practice?

The decision was made to do a randomized experiment on the topic, but assigning smaller intact households rather than larger neighborhoods. If I can oversimplify the experiment, called Moving to Opportunity (MTO) (Orr et al. 2003), it took volunteer families living in the inner part of five cities and randomly assigned them either to staying there or to getting housing vouchers and other forms of assistance permitting them to live in the suburbs. The causal question was, How does moving to a neighborhood with greater resources change the behavior of family members? Selection does not now seem to be a problem; nor does restriction on the independent variable; and the research question is still about the superindividual concept of a neighborhood. Yet families are randomly assigned to treatments, not neighborhoods.

Now carry through in greater detail the thought experiment begun earlier in which neighborhoods are the unit of treatment assignment, not households. Poor neighborhoods might be randomly assigned to improvements in the types of material, political, social, and institutional resources that are thought to cause better family outcomes in the suburbs. As desirable as upgrading inner-city neighborhoods might be, the prospects for it are not bright. It would be extremely expensive to sample the number of neighborhoods required for an experiment. It would also require a broad social buy-in that would not be easy to achieve since upgrading some city neighborhoods but not others will likely engender a bruising and high-profile public discussion that many governments would prefer to avoid. Anyway, city planners do not know how to design and implement such simultaneous multidimensional upgrading. And materially upgrading existing communities does not

answer quite the same question as moving individuals to the suburbs since only the latter requires a radical change in the neighborhood's social composition. Such change is central to the conception of MTO, but it need not happen when upgrading existing poor neighborhoods.

As expensive as the multicity MTO was, its research costs were probably less than the combined costs of its many nonexperimental predecessors. MTO also fulfilled better a core function of causal research—maximizing the contrast tested. If one compares the poverty level of the receiving suburbs in the experiment with the poverty level of the neighborhoods spontaneously moved to in prior nonexperiments, the difference is very large. In science, experiments have traditionally sought, in Francis Bacon's words, "to twist nature by the tail," not to mirror it as some policy analysts want them to do. The large treatment contrast suggests again that MTO answers a different question from its predecessors. It describes what radical, large neighborhood changes *can* do to individuals and families—the traditional efficacy question of public health research (Flay 1986). It does not answer questions about what neighborhood change routinely *does* to individuals and families within the range of change that typically takes place—the traditional effectiveness question.

The above discussion speaks to the rationales for MTO and to the advantages that occur from having many lower-order cases for assignment rather than fewer higher-order ones. It has nothing to say about how well MTO was implemented and the implications of variation in such implementation. As a matter of record, almost 50 percent of the treatment families were not in the neighborhood assigned to them after five years. This does not preclude an intent-to-treat analysis since most families were kept in the measurement frame. So one could learn what effect the program had on those who were assigned to a suburb as opposed to a question about the effects it had just on those who stayed in their assigned suburb or who finished up in a suburb "demographically like" the assigned one. Analysts now have promising procedures for providing unbiased estimates of this last type of "treatment on treated" question using the original random assignment as an instrumental valuable (Angrist, Imbrens, and Rubin 1996). But the method is a large sample test and so not feasible with a small number of large-sized clusters. However, it is more feasible with larger numbers of small-sized clusters like the families in MTO, and this type of statistical analysis was actually used in Orr et al. (2003). So an added advantage of smaller clusters is the increased chance to analyze both causal questions: about the intent to treat and the effects of treatment on the treated.

In designing an experiment about clustered entities like neighborhoods, one needs to ask what is the lowest unit at which clustering is necessary, though recourse to lower-order units can sometimes subtly shift the nature of the question. Much more can be gained by assignment at the individual or household level (as with MTO), in contrast to the block group or census tract or city planners' level. By the same principle, when an educational intervention can be assigned to districts, schools, classrooms, or individual teachers and students, the lower level is to be preferred. The same holds if a coronary disease prevention intervention can be assigned to cities, neighborhoods, blocks, families, or individuals. The main advan-

tages are for statistical power, control over implementation, and a greater chance to learn about both the intent to treat and treatment on treated questions.

Principle 3: Minimize and measure interunit communication, though neither will be easy when schools or neighborhoods are the unit of assignment

Whatever the unit of assignment, interunit communication is a problem because it can modify the planned treatment contrast. Cook and Campbell (1979) have enumerated many different processes and consequences associated with interunit communication in experiments. Most likely is that comparison cases will borrow from treatment ones and so reduce the achieved contrast, thereby increasing the chances of incorrectly failing to reject the null hypothesis. However, it is also possible for control cases to become resentfully demoralized at not receiving the treatment, leading them to do less well than otherwise. Then, a false positive conclusion can result if the controls have done worse and not because the experimentals have done better.

While these processes also operate at the individual level, my speculation is that they are exacerbated at the group level because of the larger numbers of individuals involved in toto—given the sample size requirements for and within groups—and also because group-level interventions are likely to be more socially salient and discussed than individual-level ones.

The best way to prevent interunit communication about treatment details is to assign units at such a social remove that they hardly communicate, and if they do, it is not about treatment details. Assigning treatment and control status to individual students or teachers within the same school means they can potentially compare experiences more readily than if intact schools were randomly assigned. If it is true that teachers generally communicate less about treatment details between grade levels than within them, then within-school treatments should be assigned across grades rather than within them. This suggests the utility of randomly assigning schools so that, say, all second-grade teachers are in one treatment and all the third-grade teachers are in another, while the opposite is the case in the other set of randomly assigned schools. This within-school design has obvious advantages over a complete between-school design, especially as regards the number of schools needed and hence financial cost. But interpreting the results depends heavily on teachers not sharing experiences across grades within a school. They are less likely to do so if a manipulation is not salient; but low-profile interventions are not that common in education, especially if academic achievement is the principal outcome and if teachers are called upon to implement the intervention. So I think that treatment contamination is a *potential* major problem in cluster-based experiments and serves as a rationale for often assigning entire schools to treatments rather than grade levels or classrooms within a school.

If such advantages accrue as the unit of assignment increases in size, why stop with aggregates like schools or cities? Why not go to even higher levels, to the district in the case of education or to the state or region in the case of cardiac health?

The brief answer is that increasing the size of the unit increases financial costs, on one hand, and also adversely affects the trade-off between an intervention's intensity and coverage, on the other. Although larger units usually entail greater coverage and less treatment contamination, they can often reduce treatment fidelity because monitoring by research staff tends to decline the more geographically dispersed units are. Also, most professionals in education and health know more about targeting their work to single identified individuals or small groups like classrooms than they know about targeting whole schools and entire communities. Moreover, teachers are probably motivated more by calls for change from their immediate colleagues as opposed to senior staff in school district headquarters. Since clusters are composed of individuals, the fear is that a treatment's average fidelity will drop off as unit size increases and that variation in implementation will increase both within schools and communities and across the group of schools and communities constituting a treatment or control group.

[T]eachers are probably motivated more by calls for change from their immediate colleagues as opposed to senior staff in school district headquarters.

Of course, it is not always possible to know in advance if there will be interunit communication or whether it will be of enough power to imperil causal conclusions. Most interventions have a theory (or logic model) specifying the causally potent components of the intervention, and it is exposure to these components that should be measured in assessments of interunit communication. Thus, Cook, Hunt, and Murphy (2000) showed that interunit communication took place in only three of their eleven control schools. In one instance, a principal from one condition was married to a teacher from another. In another school, a control group teacher was the daughter of a senior official at the program central office and she invited her father to come speak about his work. And in the final case, a control principal read up on the intervention by himself and tried to implement some of its particulars. In addition, the program coordinator in the district did some districtwide professional development and mentioned aspects of the program to teachers from control schools. So interunit communication took place but was restricted to a minority of schools. Moreover, the control schools never had the salaried program staff in them who were responsible for coordinating the program in treatment schools. Nor did any control teachers go to within-district or out-of-

town retreats at the developers' headquarters to learn about program implementation. So measures of what was implemented in the control schools showed few of the specific structural changes the program design called for. To record some interunit communication does not mean that it is widespread or involves central details from the program theory.

But why should one ever incur the risk of treatment contamination? Why not simply add more units deliberately chosen so as to be more physically remote from each other? One way to do this would be by randomly assigning district schools to a second- and third-grade intervention or no intervention at all instead of randomly assigning them to have the intervention in their second or third grade, as in the earlier example. Another way would be to select schools for intervention that are in different school districts rather than all coming from the same district. In the first case, the main hypothesis is now a between-school one rather than a within-school one and so will require more schools. The second case entails schools from many districts, and so they will be more different from each other and more of them will be required for the same level of statistical power. Dealing with possible contamination by increasing physical distance will very often entail greater expense and logistical headaches. Just try, for instance, getting permission to do research from many school districts, going through each one's institutional review board with its unique requirements! So the hope is to assign grades or teachers within schools rather than entire schools.

But moving to the lower level leads to a dilemma that the following example highlights. Currently under way is an educational experiment on whole school reform that centrally features a literacy curriculum with grade-specific materials. In this study, schools are randomly assigned to getting a grade-specific curriculum in one grade, with another grade serving as a control. So all schools get one version of the same theoretical intervention, but it randomly varies as to which one they get. Under these conditions, how plausible is it to assume that teachers at the control grade level are unaware of what their language arts' colleagues are doing in different classes? In my judgment, some interunit communication about the intervention seems likely. After all, the comparison is with what fellow language arts teachers in the other grades are doing, not with what PE or sixth-grade teachers are doing. And since most of the control teachers voted for the new curriculum as a precondition for it entering their school, will they not be curious about it?

The researchers will doubtless measure interunit communication about treatment particulars, though this is difficult to do well; and they will likely construct arguments using these data about how likely such communication was to have biased results. But these arguments are bound to sound defensive or incomplete, and readers will likely vary in how plausible they judge them to be. Although the within-school strategy has considerable statistical advantages, it is a disadvantage that one cannot judge in advance how much serious interunit communication is likely and how convincing measures of such communication will be in arguments about the extent to which treatment contamination is a problem. Still, the measurement of such interunit communication is taking place, and that is all to the good.

The researchers' problem could have been avoided in the first place, albeit at extra cost, by assigning intact schools to the treatments so that, within each school, all the relevant grades received the intervention. The investigators tried this initially, inviting schools that desired the intervention to participate as treatment or control schools depending on the coin toss. But nearly all refused. Since wholesale refusal has not been common in the brief history of experiments on school reform, it is instructive to ask why it occurred this time.

I assume that acceptance of the invitation to random assignment is more likely when those soliciting the assignment are genuinely unsure about the efficacy of the intervention and state this as the major rationale for assigning at random. In the example under analysis, the program developer was the solicitor, and the schools his staff solicited were those that had already agreed to the intervention. In this circumstance, can one readily imagine the program developer asserting that an experiment was needed because the program of his design for which they had already signed up was not demonstrably effective? To make this argument would also have meant abjuring the results of his own prior quasi-experimental research that he had used for years to market the program to schools and funders. Can one easily imagine the developer using the best arguments for randomization, particularly since the schools solicited for the experiment had already voted to accept the program. Was it now being intimated after the fact that their vote only entitled them to be in a lottery to be in the program as opposed to being in the program proper? Were they given a guarantee of getting into the program in future years? What was said about what changes they could and could not make while serving as a control group? Were they asked to introduce nothing new in reading, or were they free to do whatever they wanted after learning they did not get into the treatment group? Schools are continuously evolving organizations and can never be asked to entomb themselves in usual practice. The implication of all this is that program developers should never be in charge of soliciting participation to be in an experiment. Their role is to advocate for their program, not to be brokers of the best approximation to the truth about their own program's effectiveness.

In their evaluation of *Sesame Street*'s second viewing season, Bogatz and Ball (1972) devised an innovative strategy that exemplifies the flexible thinking required to reduce interunit communication when assigning higher-order aggregates to treatments. At the time, the show was available in some markets only on cable. So the evaluators went to poorer neighborhoods without cable, and in families with a child of target age, they offered to provide cable to the home on a random assignment basis, together with books, toys, and games designed to raise the profile of *Sesame Street*. The fear was that children from control-group homes without cable would visit their friends nearby and ask to see *Sesame Street*. So Bogatz and Ball decided that all the families with a child of target age living within a block were to be assigned, thus making the unit of assignment a street block and not a home. And to reduce visiting across contiguous blocks, the sampling strategy excluded all blocks that touched on a block that was to be randomly assigned to the experimental or control condition. This created a cordon sanitaire around the study blocks. While it did not rule out all visiting between treatment and control

homes, it surely reduced it and exemplifies what it takes to reduce this threat that will often not be completely ruled out.

It is obvious that the second and third principles are at odds because selecting the lowest unit for assignment usually increases the likelihood of the very communication between treatments that threatens interpretation. The connection is not inevitable, of course. But selecting units that are distant can increase costs and choosing interventions that are nonsalient—and hence not talked about—can decrease treatment impact. So, the solutions themselves entail other tradeoffs. Indeed, they are only the most striking of a large and complicated network of tradeoffs that currently make the design of cluster-level experiments as much an art as a science—notwithstanding the reality of scientific features like statistical power calculation and data analysis through hierarchical linear modeling. As the years of experience doing experiments accumulate, we will gain more reliable data and insights into these tradeoffs and will doubtless make better choices than today. But today, underinformed guesses will be required about the likelihood of treatment-correlated communication, about how many of the treatment or control units will be affected by such communication, about the communication being about theoretically and empirically significant components of what are usually complex multidimensional treatment packages, and about the quality with which such communication can be measured. Depending on the underinformed judgments that inevitably have to be made, a smaller or larger unit will be selected. At this early moment in the history of cluster-level experiments, traditional scientific conservatism suggests larger units—even though this strains budgets. But resources will force many researchers into the riskier choice of smaller units. So, some scholars are needed who will track—as well and as honestly as possible—just what amount and type of inter-unit communication occurs about treatment particulars.

Principle 4: Avoid black box experiments despite their policy relevance; instead explore implementation and causal mediating processes

When process measurement is made within an experiment, the conclusions cannot be generalized beyond contexts that include such measurement. Yet social policy rarely stipulates such measurement as part of an ongoing program. So if research is to have high fidelity to policy, it should be of a black box type, shorn of all unnecessary measurement. Science assigns measurement a different priority; for assessments of treatment implementation, unit characteristics and causal mediation are important for explaining why a cause-effect relationship occurs (or is not observed). This measurement-related trade-off between policy and science needs affects every experiment, whatever the unit of assignment.

But the trade-off is particularly acute when experimenting with clusters. In part, this is because the sources of variation are particularly complex in multilevel experiments. If schools are assigned to condition, there will be both general and time-varying implementation differences between schools, between grade levels within schools, and between classrooms within grade levels, not to speak of between

teachers and between students within classrooms. Moreover, experimenter-based attempts to enhance implementation quality will nearly always be less effective when dealing with all these levels simultaneously than when dealing only with, say, teachers. It is from the many sources of implementation variance and the presumption of diminished ability to ensure quality implementation that the need springs to check that the planned intervention occurs at an intensity that is considered "adequate" for bringing about effects. And it is impossible to do this without measuring those processes that occur between manipulating the independent variable and measuring the outcome.

In cluster-based experiments, treatment heterogeneity is the norm, not the exception; and it is likely greater than in most individual-level experiments.

A second rationale for measuring implementation processes is to probe the substantive theory through which influence is supposed to be transmitted from the intervention to the outcome. Such theory is usually laid out as a time-flow chart with boxes and arrows that illustrate the expected pathways of influence. Measuring the elements in this logic model is always important but is particularly problematic in multilevel studies because the theory tends to be more complex and more multidisciplinary as one moves through the various organizational levels to finish up at the individual. Also, the time lines might be more unclear as one struggles with causal paths both within and between levels. After all, one needs theories, not just of what happens at the individual level, but also at the classroom and individual levels at different times—a necessary but daunting task.

Experiments that fail to reject the null hypothesis are unhelpful if one cannot assess where the planned sequence of change broke down. And in experiments that do reject the null hypothesis, it is useful for both theory and practice to learn why the effect came about. One might discover, for instance, that classrooms are the major contributors to change and that schools merely provide the physical shell within which classrooms are found and through which children change. Knowing this simplifies the policy and theory implications of an experiment, reducing the role of the school level in this case and simplifying which program elements to emphasize to achieve successful transfer to other locations.

Heterogeneity of implementation usually increases with the size of clusters. Some community-based experiments have targeted cities of forty to one hundred

thousand inhabitants. In one case, after five years of implementation targeted at physicians, physician associations, hospitals, grocery stores, schools, adult voluntary associations, the media, and athletic clubs, about half the inhabitants could not recognize the Minnesota Heart Health Project's logo or remember having come in direct contact with the program (Blackburn et al. 1984). Moreover, many city inhabitants did not immediately need the program services since they already led heart healthy lives. The implication is that mobilizing large communities around heart healthy living is difficult, given competing influences in the lives of many local people and their associations. So the National Heart, Lung, Blood Institute halted this kind of study, preferring to fund preventive work with individuals from populations known to be predisposed to poor cardiac health to whom treatment materials could be better targeted because of their presumed homogeneity, at least relative to whole cities where implementation quality was a real issue, as was targeting those truly in need.

Measures of intervening processes are used in three main ways. One is to describe their relationship to the randomly manipulated independent variable. In this kind of analysis, the intervening variable is an outcome in its own right. One can and should ask how the treatment affected a given intervening process, and answers to questions like this will be unbiased. Second, if the intervening measures are made at several time points they can be used as probes of the hypothesis about the temporal sequencing of change—what changes first and then later. Third, the measures can be used to test the entire causal chain. However, a selection problem arises when testing how one intervening variable affects another. Exposure to the first intervening variable in the causal chain is not random. As noted earlier, statisticians are now working on ways to deal with this using the random assignment as an instrumental variable (Angrist, Imbrens, and Rubin 1996) but assuming large samples that are not relevant to experiments with few clusters, each of large size. Nor are they relevant to models with multiple intervening variables changing at different times. Nonetheless, I think it still worth empirically probing each link in the postulated causal chain as long as the results are presented cautiously and with less certainty than the intent-to-treat results.

Nowhere is this tentative and dangerous analysis more needed than when large-sized clusters are involved. Then, the causal chain tends to be longer, increasing the chances of incorrectly specifying and inadequately measuring any one causal link. Moreover, effects later in the causal chain are likely to be more weakly related to the manipulation because they depend on more prior causal relationships being true when they are, almost certainly and at best, only probabilistically related to each other. Multiplying out all these probabilities highlights how tenuous is the molar causal link between the intervention and the outcome. This is why measurement of causal processes is a necessity at this early stage in the evolution of experience with cluster-based experiments, even if such measurement does go against the policy need to have experiments re-create the conditions of practical policy implementation that rarely include measuring intervening processes.

Principle 5: Explicitly acknowledge the great heterogeneity in treatment implementation in cluster-level designs and analyze its likely consequences even though these causal conclusions will be inferior to conclusions about the intended treatment

Descriptive studies of the outcomes most used in neighborhood and school research show few ICC values greater than .25—even in national studies and after corrections for unreliability (Cook, Shagle, and Degirmencioglu 1997). This implies that neighborhoods and schools are internally heterogeneous; the same is even true at the lower level of classrooms or street blocks through internal homogeneity is understandably greater at these lower levels.

Yet group-level interventions are often introduced into such heterogeneous settings under the implicit assumption that the intervention is equally meaningful and relevant to all residents or students. They are generally not, of course, as was seen with the heart healthy experiments in very large communities. It is naïve to believe that a treatment will be implemented in standard fashion across all the units within a treatment group. Nonstandardization is the reality both from one cluster to the next in a given treatment group and also, as I am discussing here, from one cluster to another within the same treatment group. In cluster-based experiments, treatment heterogeneity is the norm, not the exception; and it is likely greater than in most individual-level experiments.

The situation is exacerbated where local reinvention of the intervention is encouraged. For instance, the School Development Program (Comer 1988) specifies some procedures that should be followed within each school, but the school governance committee is encouraged to decide for itself which goals the school should pursue and how these goals might be realized. As a result, the program differs considerably from one school to the next within a district. To add to the complexity, district officials can add program factors at their more central level. Thus, the program in Prince George's County emphasizes mental health issues (Cook, Hunt, and Murphy 2000); the program in Chicago began emphasizing parent involvement and then switched to stimulating reading (Cook et al. 2002); and the program in Detroit emphasized coordination with outside consultants trying to stimulate reading and math. I surmise that local reinvention is more likely with clusters than individuals because clustered units are usually more physically dispersed, more used to exercising collective responsibility and tailoring materials to their needs and, anyway, researchers can control them even less than they can control individuals. But there is more to it than this. Advantages of ownership and maintenance occur if collectives are actively engaged in an intervention and generate marginal increments that make it more responsive to local circumstances.

This heterogeneity of treatment implementation problem is reduced when an intervention is linked to a detailed and well-articulated protocol and when homogenizing conditions are set for entry into a study. If 80 percent of the teachers have to agree to an intervention before it can be introduced into a school, this weeds out sites with minimal or highly variable interest in implementing a planned change including those where there is active dispute about the intervention's value. The

flip side of this, of course, is that generalization is only warranted to such primed places. Another useful strategy is to monitor implementation at the earliest treatment stages and to use this feedback to tighten up on those aspects of implementation that are within the researchers' control. And of course, implementation should always be carefully and exhaustively measured.

The classical statistical response to incomplete and variable implementation at any level is to increase the sample size of units. This should be done where feasible, notwithstanding the many practical difficulties in finding willing clusters, in funding implementation at these sites, and in maintaining implementation quality as sites are added. I also think that a statistical strategy should be followed that is often considered to be ill advised. In addition to the necessary intent-to-treat analysis, a subsidiary analysis of the treatment on the treated should be done that uses measures of implementation from each unit at each level to provide quality of implementation scores. These should then be used in a quasi-experimental analysis of how variation in implementation quality affects the major outcomes. Since the analysis is biased, it is worth analyzing the data under several different assumptions about the direction of bias, and conclusions should be stated with less confidence than those from the unbiased intent-to-treat analysis. It strikes me as unproductively puritan to collect but not use implementation data when implementation quality so obviously varies from person to person within schools or neighborhoods and, particularly, between schools and neighborhoods assigned to the same treatment. The question of the treatment's effectiveness at its best is centrally important to many actors, including theorists, program developers, and some policy makers. It seems odd to neglect it completely in favor of a question about the effects of a treatment at its average that is conditional on the particular comparison groups chosen and on the level of uncontrolled background noise.

Conclusion

All things being equal, it is much more difficult to do an experiment that simultaneously deals with units at two or more levels—usually the individual and some higher-order level like a school or community. Social scientists have not yet acquired much experience in implementing cluster-based experiments, and the knowledge to be gained by emulating medicine's more evolved tradition of multisite randomized clinical trials is limited. In that tradition, individuals are assigned to treatments from within sites rather than between sites so that sites serve to increase power and generalization and not to be units of assignment per se. Social scientists will have to learn most of their lessons about cluster-based experiments in the crucible of their own experience. Solutions will also have to come from reflections on such experiences as well as from developments in statistical theory.

I have tried here to synthesize some of that experience, both as I have learned it in my own cluster-based experiments and in considering the work of others. The major problems are few, but quite messy. Cumulatively, they probably seem daunt-

ing to anyone who has not done experiments or has only done them with individuals. And they are indeed daunting. I make no pretence to having perfect solutions that can be plugged in across all types of clustering and all types of independent and dependent variables. But I do claim to have explicated the problems as I experienced them, and that is a modest start.

References

Angrist, J. D., G. W. Imbrens, and D. B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91:444-55.

Blackburn, H., R. Luepker, F. G. Kline, N. Bracht, R. Carlaw, D. Jacos, M. Mittelmark, L. Stauffer, and H. L. Taylor. 1984. The Minnesota Heart Health Program: A research and demonstration project in cardiovascular disease prevention. In *Behavioral health*, ed. J. D. Matarazzo, S. Weiss, J. A. Herd, N. E. Mller, and S. M. Weiss. New York: Wiley.

Bogatz, G. A., and S. Ball. 1972. *The impact of "Sesame Street" on children's first school experience*. Princeton, NJ: Educational Testing Service.

Brooks-Gunn, J., G. J. Duncan, and J. L. Aber, eds. 1997. *Neighborhood poverty: Context and consequences for children*. Vol. 1. New York: Russell Sage Foundation.

Comer, J. P. 1988. Educating poor minority children. *Scientific American* 259:42-48.

Cook, T. D. 2003. The rationale for studying multiple contexts simultaneously. *Addiction* 98 (Suppl. 1): 151-55.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand-McNally.

Cook, T. D., Y. Y. Deng, and E. Morgano. 2005. Peer group effects in early adolescence: The role of friends' average grade point average. Manuscript, Northwestern University, Evanston, IL.

Cook, T. D., F. Habib, M. Phillips, R. A. Settersten, S. C. Shagle, and S. M. Degirmencioglu. 1999. Comer's School Development Program in Prince George's County: A theory-based evaluation. *American Educational Research Journal* 36 (3): 543-97.

Cook, T. D., M. Herman, M. Phillips, and R. J. Setterston Jr. 2002. Some ways in which neighborhoods, nuclear families, friendship groups and schools jointly affect changes in early adolescent development. *Child Development* 73 (4): 1283-1309.

Cook, T. D., H. D. Hunt, and R. F. Murphy. 2000. Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal* 37 (2): 535-97.

Cook, T. D., S. C. Shagle, and S. M. Degirmencioglu. 1997. Capturing social process for testing mediational models of neighborhood effects. In *Neighborhood poverty: Context and consequences for children*, vol. 2, ed. J. Brooks-Gunn, G. J. Duncan, and J. L. Aber. New York: Russell Sage Foundation.

Eldridge, S. M., D. Ashby, G. S. Feder, A. R. Rudnicka, and O. C. Ukoumunne. 2004. Lessons for cluster randomized trials in the twenty-first century: A systematic review of trials in primary care. *Clinical Trials* 1:80-90.

Farquhar, J. W., S. P. Fortmann, J. A. Flora, C. B. Taylor, W. L. Haskell, P. T. Williams, N. Maccoby, and P. D. Wood. 1990. The Stanford Five-City Project: Effects of community-wide education on cardiovascular disease risk factors. *Journal of the American Medical Association* 26:359-65.

Flay, B. R. 1986. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine* 15:451-74.

Furstenberg, F. F., Jr., T. D. Cook, J. Eccles, G. H. Elder, and A. Sameroff. 1999. *Managing to make it: Urban families in high-risk neighborhoods*. Chicago: University of Chicago Press.

Gail, M. H., D. P. Byar, T. F. Pechacek, and D. K. Corle. 1992. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clinical Trials* 13 (1): 6-21

Gargani, J., and T. D. Cook. 2005. How many schools? The limits of the conventional wisdom about sample size requirements for cluster randomized trials. Manuscript, Graduate School of Education, University of California, Berkeley.

Jencks, C., and S. Mayer. 1990. The social consequences of growing up in a poor neighborhood. In *Inner city poverty in the United States*, ed. L. Lynn. Washington, DC: National Academy Press.

Orr, Larry, Judith D. Feins, Robin Jacob, Erik Beecroft, Lisa Sanbonmatsu, Lawrence F. Katz, Jeffrey B. Liebman, and Jeffrey R. Kling. 2003. *Moving to Opportunity: Interim impacts evaluation*. Cambridge, MA: Abt Associates.

Raudenbush, S. W. 1997. Statistical analysis and optimal design for cluster randomized design. *Psychological Methods* 2:173-85.

Raudenbush, S. W., and A. S. Bryk. 2002. *Heirarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Wilson, W. J. 1987. *The truly disadvantaged: The inner city, the underclass, and public policy*. Chicago: University of Chicago Press.