1

ECOMETRICS: TOWARD A SCIENCE OF ASSESSING ECOLOGICAL SETTINGS, WITH APPLICATION TO THE SYSTEMATIC SOCIAL OBSERVATION OF NEIGHBORHOODS

*Stephen W. Raudenbush**
Robert J. Sampson†

This paper considers the quantitative assessment of ecological settings such as neighborhoods and schools. Available administrative data typically provide useful but limited information on such settings. We demonstrate how more complete information can be reliably obtained from surveys and observational studies. Survey-based assessments are constructed by aggregating over multiple item responses of multiple informants within each setting. Item and rater inconsistency produce uncertainty about the setting being assessed, with definite implications for research design. Observation-based assessments also have a multilevel error structure. The paper describes measures constructed from

Revision of a paper presented at the annual meeting of the American Society of Criminology, San Diego, November 1997. The research reported here was supported by the Project on Human Development in Chicago Neighborhoods (PHDCN), with funding from the National Institute of Justice and the John D. and Catherine T. MacArthur Foundation. We thank Felton Earls, Albert J. Reiss Jr., and Steven Buka for their essential role in the design of PHDCN, and Richard Congdon for the programming of all computations. Meng-Li Yang checked the derivations and computed all results. We would also like to acknowledge the helpful comments of anonymous reviewers.

*University of Michigan
†University of Chicago

interviews, direct observations, and videotapes of Chicago neighborhoods and illustrates an "ecometric" analysis—a study of bias and random error in neighborhood assessments. Using the observation data as an illustrative example, we present a three-level hierarchical statistical model that identifies sources of error in aggregating across items within face-blocks and in aggregating across face-blocks to larger geographic units such as census tracts. Convergent and divergent validity are evaluated by studying associations between the observational measures and theoretically related measures obtained from the U.S. Census, and a citywide survey of neighborhood residents.

This paper addresses the challenge of assessing the social and physical properties of ecological settings, especially the neighborhood. Most published research on neighborhoods relies on data collected by administrative agencies for other purposes, principally the U.S. decennial census. Measures gleaned from the census typically cover socio-demographic factors such as poverty, family structure, unemployment, and racial composition. Other sources of administrative data often aggregated to the neighborhood level include government crime reports (e.g., the FBI's Uniform Crime Reports), vital health statistics (e.g., rates of infant mortality; suicide), records of social-service agencies (e.g., public assistance caseloads), and school records (e.g., dropout rates; average test scores).

Although much can be gained from these administrative sources, they are not helpful in revealing unofficial behavior (e.g., undetected crime, disorder) and the social-organizational processes that lie behind neighborhood demography. Mayer and Jencks (1989) have argued that if neighborhood effects on social outcomes exist, presumably they are constituted from social processes that involve collective aspects of community life. To date, however, theories emphasizing collective processes such as neighborhood social control and cohesion have rarely been translated into measures that directly tap hypothesized constructs. Common sources of administrative data also poorly capture the physical properties of neighborhoods such as the markings of gang graffiti, the density of liquor stores, and abandoned cars.

A key focus of this paper is on the statistical methods needed to evaluate the quality of such ecological assessments. Two data collection strategies will be considered: (1) the neighborhood survey and (2) the direct observation of physical conditions and social interactions occurring within neighborhoods. We concentrate primarily on the second approach, as it is the more novel of the two and illustrates all of the basic principles involved in assessing reliability and validity. However, we shall briefly review work on the neighborhood survey and compare measures generated from survey work with those derived from direct systematic observations.

1. FROM PSYCHOMETRIC TO "ECOMETRIC" STANDARDS

It is tempting to describe the problem at hand as the need to understand "the psychometric properties of ecological measures." But this awkward phrasing merely reveals the individualistic bias of modern social science, underscoring the need to take ecological assessment seriously as an enterprise that is conceptually distinct from individual-level assessment. Ecological constructs need not be merely the aggregate of individual ones, and thus we seek to understand what we call the "ecometric" rather than psychometric properties of ecological measures. We show that "ecometric assessment," while borrowing tools from the rich tradition of psychometrics, has its own logic.

Moreover, without a coherent strategy for evaluating the quality of ecological assessments, a serious mismatch arises in studies that aim to integrate individual and ecological assessments. The assessment of individual differences, building on decades of psychometric research, employs measures that have withstood rigorous evaluation. This is especially true of measures of cognitive skill and school achievement, but it extends as well to measures of personality and social behavior. These measures have been thoroughly evaluated in many studies; each scale includes many items; ill-performing items have been discarded; and psychometric properties have been found to hold up in many settings. Without comparable standards to evaluate ecological assessments, the search for individual and ecological effects may overemphasize the individual component simply because the well-studied psychometric properties are likely to be superior to the unstudied ecometric ones.

The history of psychometrics is indeed instructive to our case. Beginning in the early years of this century, educational psychologists, statisticians, and others launched a new realm of applied social science destined to have a profound impact on modern society: the assessment of human ability and personality. An enormous demand arose for standardized tests that seemed to offer a meritocratic basis for selecting persons for advanced schooling, for employment, and for specializations within the armed forces. The testing movement that resulted made permanent contributions to statistical methodology, including correlational and factor analysis, and produced a branch of applied statistics called *psychometrics* that has come to dominate thinking about the reliability and validity of measurement in social science.

In contrast, until recently there has been no parallel effort to create a scientific basis for the methodological assessment of human ecological settings such as neighborhoods and schools. While there have been many studies of organizational climate (cf. Pallas 1988), one rarely encounters a rigorous evaluation of the reliability or validity of such measures, nor are standard errors of measurement associated with them. Measures of organizational climate, ironically, have historically been studied psychometrically at the level of the individual respondent rather than "ecometrically" at the level of the organization, even when the analysis used the organization as the unit of analysis in structural models (Sirotnik 1980). As part of a larger study of individual and ecological correlates of social behavior, we are engaged in a multipronged effort to assess neighborhoods as important units in their own right. In approaching the problem of ecometric assessment, we borrow, integrate, and adapt three analytic strategies that are prominent in modern psychometrics: (1) item response modeling, (2) generalizability theory, and (3) factor analysis.

Item response models conceive the probability of a correct response to an item on a test as a function of the ability of the examinee and the difficulty of the item (Lord 1980; Rasch 1980). Assuming all items represent the same ability domain, difficult items will be answered correctly less often than will easy items. Similarly, given the difficulty of the item, more able examinees will obtain a correct response with higher probability than will less able examinees. If the model is sensible, it will generate an interval scale along which every item and every examinee can be located. A visual examination of this "item map" provides useful clues about the construct validity of the test, because one can assess whether the empirically estimated item difficulties conform to cognitive theory regarding the sources of item difficulty. It is also possible to identify misfitting items (e.g., difficult items frequently solved by persons of low ability) and misfitting persons (e.g., able persons who frequently miss easy items). Such analyses form a basis for discarding poor items and assessing the overall quality of the scale. The analysis produces a measure of scale reliability and a standard error of measurement for each examinee (Wright and Stone 1979).

Generalizability theory enables the study of multiple sources of measurement error in an assessment (Cronbach et al. 1972; Brennan and Kane 1979). Suppose, for example, that an examinee is asked to write an essay on Saturday morning and that the essay is rated by a single rater. Possible error sources would be day of week, time of day, the specific task (e.g., the topic chosen for the essay), and the rater. A generalizability study might assess persons on several days of the week and times of day and on varied tasks, with essays read by multiple raters. Such a study would provide not only a summary measure of reliability but also an estimate of the magnitude of each component of error. It would presumably influence future assessments. For example, if tasks and raters produce large error variance, future assessments might require essays on several topics, each to be rated by two raters, thus averaging over task and rater errors, and achieving an acceptable level of reliability. However, the design of future assessments would depend heavily upon their use. For example, if the writing task were used as part of a program evaluation, it might be cheaper to sample more examinees in each comparison group rather than to hire more raters or to require more tasks per examinee. A generalizability study would specify the sample size per group required to achieve a given reliability of the program group mean.

Factor analysis enables a determination of the interrelationships among measures. Often studies collect data on a fairly large number of measured variables. However, these variables may in fact reflect variation in a smaller number of latent variables or factors. Confirmatory factor analysis enables one to test *a priori* hypotheses about the associations between underlying factors and observed variables (Joreskog and Sorbom 1988). Often a factor analysis lays the basis for a parsimonious representation, and this can be particularly important in the case of ecological measurement. Typically the sample size of ecological units is small and the intercorrelations among ecological variables high. Thus a parsimonious representation of variation at the ecological level may be essential for meaningful analysis and interpretation.

With this backdrop in mind, we now turn to the description of two forms of ecological assessment that are not yet standard in social science. We begin with a brief consideration of survey-based measures of ecological settings, where experience has accumulated rapidly in recent years. Building on the survey approach, we then turn to an extended treatment of the more novel technique of systematic social observation.

2. ASSESSING SURVEY-BASED MEASURES OF ECOLOGICAL SETTINGS

The problem of measuring high school climate provides a useful lead-in to considerations of using survey questionnaires to assess neighborhoods. Raudenbush, Rowan, and Kang (1991) analyzed national survey data yielding questionnaire responses from 15 to 30 teachers in each of about 400 schools. Dimensions of climate included teacher control over the conditions of instruction, teacher collaboration, and administrative support. Multiple Likert scale items tapped each of these constructs. The investigators used a three-level hierarchical statistical model to assess sources of measurement error.

At the first and lowest level of aggregation, item responses within a given scale varied within a teacher around that teacher's "true perception." The source of variation at this level was item inconsistency. At the second level, the "true perceptions" on each scale varied among teachers within a given school around the school's "true score." Here the variation reflected individual variation in perceptions. At the third and highest level of aggregation, school "true scores" varied around a grand mean. This analysis strategy enabled Raudenbush et al. (1991) to estimate (1) the reliability with which teacher perceptions vary; but more importantly, (2) the reliability of the school-level measures of each aspect of climate; and (3) the correlation structure at the teacher level and at the school level among the three climate dimensions.

The analysis just described was in fact a generalizability analysis, laying the groundwork for assessing how adding items to each scale or sampling more teachers per school would increase the reliability of assessment of either persons (teachers) or ecological units (schools). The analysis showed that adding items was far more useful in improving teacher-level reliability than in improving school-level reliability. Viewing teachers as raters of the school, school-level reliability relies principally upon the degree of rater agreement and the number of raters per school. The analysis thus aids in determining the needed sample size of teachers per school to achieve a given school-level reliability on each climate dimension and helps in allocating resources between investing in more data collection per teacher (through more items) or more teachers per school. The analysis also involved a multilevel principal components analysis that revealed the number of reliably varying dimensions of school climate, in addition to, and distinct from, the number of reliably varying dimensions of teacher perceptions. A further extension might have involved multilevel factor analysis (Muthen 1991, 1997).

A similar logic may be applied to the use of interviews to measure social organizational aspects of neighborhoods. Sampson, Raudenbush, and Earls (1997) used a multilevel research design (described below) to construct and evaluate measures of neighborhood social organization. Within each of 343 Chicago neighborhoods, between 20 and 50 households were selected according to a multistage probability sample. The total sample size was 8,782, with a response rate of 75 percent. Within each household, a randomly chosen adult was interviewed concerning conditions and social relationships in the local neighborhood. Sampson et al. (1997) employed a three-level hierarchical model (formally presented in Raudenbush and Sampson [forthcoming]) to investigate the statistical properties of neighborhood measures of social cohesion and informal social control. The analysis yielded estimates of item inconsistency within each scale, interrater agreement on each scale, and an overall estimate of the reliability of measurement of each scale.

This analysis is extended in Table 1, which displays five scales that tap theoretically relevant aspects of the physical and social properties of neighborhoods as perceived by Chicago residents. The table also includes the items composing each scale, the interrater agreement, and the scale reliability at the neighborhood level. Interrater agreement is measured by an intraneighborhood correlation coefficient (ICC)—that is, the ratio of between-neighborhood variance to the sum of between- and within-neighborhood variance, where the variance attributable to item inconsistency has been removed. In essence, these ICCs capture the extent to which assessments of the "ego-defined" neighborhood, as conceived by the individual rater, are correlated within the physical spaces defined *a priori* as neighborhoods.

Table 1 reveals that the ICCs are modest, ranging from .13 for informal social control to .36 for social disorder. Because these correlations are variance ratios, it is clear that in no case does most of the variation in ratings lie between neighborhoods. The relatively modest ICCs are similar to those found in other studies looking at contexts such as schools and even families. Duncan and Raudenbush (1997:10) advise caution in interpreting small ICCs, as effect sizes commonly viewed as large translate into small proportions of variance in individual outcomes explained by neigh-

TABLE 1
Selected Variables from the PHDCN Community Survey (8,782 respondents, 343 neighborhood clusters)

Scale	ICC	Reliability
Social Disorder	.36	.89
Litter		
Graffiti		
Vacant or deserted houses		
Drinking in public		
Selling or using drugs		
Teenagers/adults causing trouble		
Perceived Violence	.25	.82
Fights in which a weapon was used		
Violent arguments between neighbors		
Gang fights		
Sexual assaults		
Robbery		
Social Cohesion	.24	.80
Close-knit neighborhood		
Helpful people		
People get along with each other		
People share the same values		
People can be trusted		
Social Control	.13	.74
Neighbors are willing to do something about:		
children skipping school		
children painting graffiti		
children showing disrespect to adult		
someone being beaten or threatened		
keeping the fire station open		
Neighborhood Decline	.18	.75
Personal safety worse		
Neighborhood looks worse		
People in neighborhood less helpful		
Level of police protection worse		

borhood membership. In fact, neighborhood effect sizes as large as .8 of a standard deviation difference can give rise to an ICC as low as .14. Therefore a small correlation among neighbors does not rule out a large effect size associated with a measured difference between neighborhoods (Duncan and Raudenbush 1997:11).

Although the interrater agreement appears modest, only a moderate sample size of raters per NC is required to achieve reasonably high interrater reliabilities at the neighborhood level. This association between sample size of raters and reliability is graphed in Figure 1, for informal social control (which has the lowest interrater agreement) and social disorder (which has the highest interrater agreement). The curves for the other three measures lie between the two curves in Figure 1 because their interrater agreements are neither as low as that for informal social control nor as high as that for social disorder. It is clear that sampling 20 raters per neighborhood produces interrater reliabilities ranging from .70 to .90 while 40 raters yields reliabilities ranging from .83 to .95. The curves make vividly clear the diminishing returns to investments in raters beyond a given number to yield acceptable reliability.

Further analysis revealed some redundancy among the scales. For example, the correlation between social control and social cohesion, disattenuated for measurement error, was $r = .88$. This result was conceptually sensible. Informal social control taps the extent to which neighbors can be relied upon to intervene to protect the public order. Without some degree of social cohesion, which involves neighbors knowing and trusting

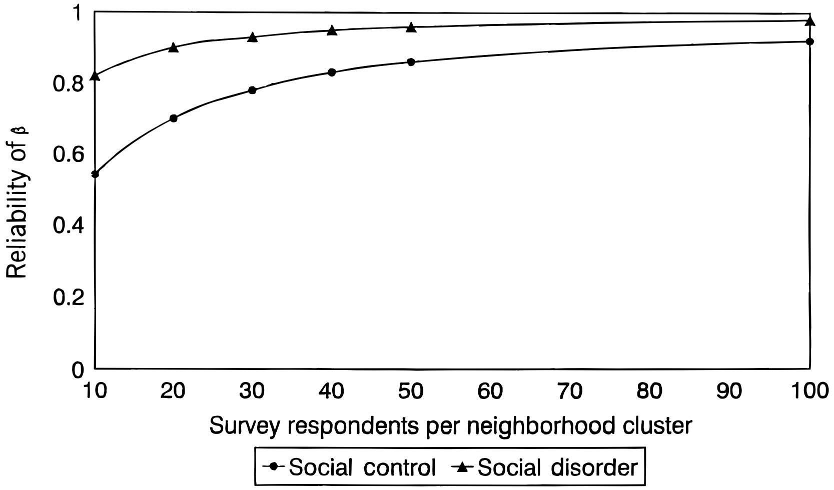

FIGURE 1. Reliabilities of community survey measures of social disorder and social control.

each other and having shared values, informal social control would appear impossible. And the exertion of such informal control would likely enhance social cohesion: people get to know each other by working together for common goals. The two sets of items appeared closely linked to the larger notion of collective efficacy (Sampson et al. 1997). Thus the two measures were combined to create a more parsimonious, reliable, and readily interpretable measure of an ecological construct with strong theoretical connections to crime reduction.

The two examples of ecological assessment just described involved paradigm examples of generalizability theory. In both cases, measurement error was decomposed into variation attributable to item inconsistency and rater inconsistency. This strategy provided the basis for assessing the needed sample size of raters (and of items, though not emphasized in the second example) in future studies. The two cases varied from standard generalizability theory in that multiple measures were simultaneously assessed; in this way, both analyses are easily amenable to multilevel factor analyses. The methodology linking latent variable analysis with multilevel modeling and the assessment of indirect associations (Raudenbush and Sampson forthcoming) can be applied to the sort of nested neighborhood-level designs now appearing in the social sciences (e.g., Elliott et al. 1996; Cook et al. 1997; Brooks-Gunn et al. 1997).

On the other hand, neither of these survey-based examples provided independent or “objective” assessments of the ecological environment based on direct observation. Moreover, instead of a serious item response analysis, in both cases Likert-scale (ordinal) responses were treated as interval-level data. This practice, certainly open to criticism, is widespread and will often cause little trouble when large numbers of item responses are aggregated to produce a scale score. However, a serious item response analysis is not only better grounded theoretically, it also produces information relevant to future scale construction and to interpretation of effect sizes (Wright and Masters 1982). We thus turn to the assessment of direct observational measures, illustrated with the use of item response analysis. The approach is similar to that used in the two survey-based measures described above in that generalizability and factor analysis also come into play.

3. SYSTEMATIC SOCIAL OBSERVATION

Direct observation is fundamental to the advancement of science. With this in mind, more than 25 years ago in an early volume of *Sociological*

Methodology, Albert J. Reiss Jr. (1971) advocated systematic social observation (hereafter, SSO) as a key measurement strategy for a wide variety of social science phenomena. Reiss (1971:4) defined systematic observation to include explicit rules which permit replication. He also argued that the means of observation, whether a person or technology, must be independent of that which is observed. As his main example, Reiss described systematic observations of police-citizen encounters but noted as well the general application to physical conditions and social interactions within neighborhood settings (see also Reiss 1975). In particular, SSO provides measures independent of the perceptions of survey respondents and can tap aspects of the social and physical environment that survey respondents have difficulty describing accurately. The key disadvantage of observational methods in neighborhood research, of course, is that they cannot capture the theoretical constructs that require resident perspectives. Thus, for example, assessing resident perceptions of social cohesion and social control (Sampson et al. 1997) requires survey methods. If researchers rely entirely on observations, there is a danger that they will misinterpret the significance of observable conditions such as physical disorder, building conditions, and land use. Nevertheless, when used in conjunction with survey-based methods, direct observation can provide an independent source of data that can strengthen inferences about neighborhood social organization and its consequences. For example, Sampson and Raudenbush (1998) have tested the association between social control and cohesion, as tapped by survey methods, and neighborhood disorder, as indicated by independent observation. This strategy avoids reliance on resident reported disorder, which would possibly create a “same-source” bias.

Despite the potential of observation for providing quantifiable, objective, and replicable measures of physical and social structure, published examples of systematic social observation at the neighborhood level are relatively infrequent. We believe one of the primary reasons has been methodological uncertainty on how to properly conduct and assess systematic observations. A major exception and an advance in systematic observational study was provided by the research program of Taylor and colleagues in Baltimore (Taylor, Shumaker, and Gottfredson 1985; Taylor, Gottfredson and Brower 1984; Covington and Taylor 1991). Using observations conducted by teams of trained raters walking in the neighborhood, Taylor et al. (1985) assessed 20 percent of the occupied street blocks in 66 Baltimore neighborhoods. They identified two physical dimensions of neighborhoods that stood out empirically: physical decay and nonresidential land use. These two dimensions were reliable in terms of individual-level standards (e.g., Cronbach's alpha and interrater reliability) and were related as expected to independent measures of perceived disorder and fear of crime derived from neighborhood surveys. A smaller-scale version of systematic observation based on interviewer ratings in a neighborhood survey was also used in Taub et al. (1984).

Building on the conceptual framework of Reiss (1971) and the techniques of Taylor and colleagues (1984, 1985), the Project on Human Development in Chicago Neighborhoods initiated in 1995 a combined person-based and video-taped approach to collecting systematic observations of neighborhood social and physical disorder. This substantive focus drew on considerable theory and past research indicating that physical and social disorder provide important environmental cues to residents and potential predators alike (Skogan 1990). After describing the sample design and data collection, we present a generalizable model for how to properly assess such observational techniques.

3.1. *Sample Design*

Chicago's 865 census tracts were first combined into 343 neighborhood clusters (NCs). The overriding consideration in the formation of NCs was that they should be as ecologically meaningful as possible, composed of geographically contiguous census tracts, and internally homogeneous on key census indicators. The resulting ecological units contained about 8000 people, much smaller than Chicago's 77 community areas but large enough to approximate local neighborhoods. Geographic boundaries (for example, railroad tracks, parks, and freeways) and knowledge of Chicago's neighborhoods guided this process.

The 343 NCs in Chicago were then stratified by seven levels of ethnic mix and three levels of SES. Within strata, 80 NCs were sampled with the aim of obtaining a near balanced design, thus eliminating the confounding between ethnic mix and socioeconomic status (SES). However, there were two empty cells (low SES, predominantly European-American; and high SES, predominantly Hispanic). Also, the largest stratum was low SES and predominantly African-American, containing 177 NCs, generally characterized by concentrated poverty, racial segregation, and other forms of disadvantage. The final design randomly sampled four NCs within cells that had at least four, all NCs within cells having fewer than four, with an over sampling of the largest and most disadvantaged cell.

In the first wave of the PHDCN's longitudinal study, approximately 6500 young people have been sampled and assessed within the resulting 80 NCs. Data gathered by means of the systematic observations and neighborhood survey will provide explanatory variables to be used in conjunction with information about individual and family characteristics to account for variation in the developmental trajectories of these young people.

3.2. *Instruments and Data*

Between June and September 1995, observers trained by the National Opinion Research Center (NORC) drove a sport utility vehicle at a rate of five miles per hour down every street within the 80 sample NCs. The composition of the vehicle included a driver, a videographer, and two observers. The unit of recorded observation was the face-block: the block segment on one side of the street. For example, the buildings across the street from one another on any block comprised two separate units of observation. An advantage of this microlevel of coding is that observations can then be pieced together to form higher levels of aggregation desired by theory or as suggested by patterns in the data.

As the NORC team drove down the street, a pair of video recorders, one located on each side of the vehicle, captured social activities and physical features of both face-blocks simultaneously. Also at the same time, the two trained observers—one on each side of the vehicle—recorded their observations onto an observer log for each face-block. Additionally, the observers added commentary when relevant (e.g., about unusual events such as a drug bust) by speaking into the videotape audio. Using these procedures, the SSO team produced Hi-8 videotapes, observer logs, and audiotapes for every face-block in each of the 80 sampled NCs. In all, 23,816 face-blocks were observed and video-recorded for an average of 298 per NC.

NORC collected data on 14 variables in the 23,816 observer logs with an emphasis on land use, traffic, the physical condition of buildings, and evidence of physical disorder. The observer log data were easily transformed into machine readable data files as they were entered on scannable forms. By contrast, because of the expense of first viewing and then coding the videotapes, a random subsample of all face-blocks was selected for coding. Specifically, in those NCs consisting of 150 or fewer face-blocks, all face-blocks were coded. In the remaining face-blocks, sample sizes were calculated to approximate a balanced design as closely as possible in order to maximize statistical power for comparisons of NCs. A total of 15,141 face-blocks were selected for videotape coding, for an average of 189 face-blocks per NC. From the videotapes, 126 variables were coded, including detailed information on physical conditions, housing characteristics, businesses, and social interactions occurring on each face-block (NORC 1995). Coders were trained in multiple sessions, including an intercoder reliability training where 90 face-blocks were independently double coded, differences resolved, and coding procedures revised. Moreover, as a check on quality control, a random 10 percent of all coded face-blocks were recoded by new observers, and the results compared. This test produced over 98 percent agreement (for full details see NORC 1995; Carter et al. 1996).

3.3. *Measures and Scales*

Given the focus of this paper on methodological issues in evaluating ecological measures, we have selected two scales for illustrative analysis. The first is a scale intended to capture the level of physical disorder, represented by items indicating the presence or absence in the street, sidewalk, or gutter of empty beer bottles; cigarettes or cigars; drug paraphernalia; condoms; garbage; abandoned cars; and various types of graffiti. Although some of the scales were measured initially on an ordinal scale, the data behaved essentially as dichotomous items, coded for analysis as 1 = presence and 0 = absence of the indicator of disorder.

Table 2 gives the frequency distribution of the items. The variation in sample size reflects the fact that six of the ten items were taken from the observation log and thus have nearly complete data. The other four variables were derived from the videotapes, and are thus based on the reduced subsample selected for coding. Note that the items behave essentially as one might expect. Less serious indicators of disorder (presence of cigarettes and garbage) arise more frequently than do indicators that might be regarded as more serious (drug paraphernalia and condoms) with the presence of beer bottles arising with moderate frequency. An exception occurs in the case of graffiti: political graffiti is very rare, though not necessarily indicative of severe disorder.

The second scale is intended to capture direct evidence of social disorder. All items were coded from videotape. They include presence of adults loitering, public drinking, peer gangs, drunken adults, adults fight-

TABLE 2
Frequency Distribution of SSO Item Responses, Face-Block Level

Variable	Category	Frequency
Physical Disorder		
Cigarettes, cigars on street or gutter	no	6815
	yes	16758
Garbage, litter on street or sidewalk	no	11680
	yes	11925
Empty beer bottles visible in street	no	17653
	yes	5870
Tagging graffiti	no	12859
	yes	2252
Graffiti painted over	no	13390
	yes	1721
Gang graffiti	no	14138
	yes	973
Abandoned cars	no	22782
	yes	806
Condoms on sidewalk	no	23331
	yes	231
Needles/syringes on sidewalk	no	23392
	yes	173
Political message graffiti	no	15097
	yes	14
Social Disorder		
Adults loitering or congregating	no	14250
	yes	861
People drinking alcohol	no	15075
	yes	36
Peer group, gang indicators present	no	15091
	yes	20
People intoxicated	no	15093
	yes	18
Adults fighting or hostilely arguing	no	15099
	yes	12
Prostitutes on street	no	15100
	yes	11
People selling drugs	no	15099
	yes	12

ing, prostitutes, and drug sales. In general, indicators of social disorder are present far less frequently than are indicators of physical disorder (see again Table 2). Activities viewed as indicative of serious disorder (prostitution, drug selling, adults fighting) are again especially rare. Indicators that are somewhat less severe are also somewhat less rare (drinking alcohol, presence of peer gangs), though they remain very rare. One item—adults loitering—is the least severe and occurred with much higher frequency than did any other item.

A simple visible inspection of two scales suggests that the physical disorder scale will behave better "ecometrically" than will the social disorder scale. First, it has more items (10 versus 7). Second, and more important, the physical disorder items appear to range widely in severity; several occur with large frequency, several others with modest frequency, and several are comparatively rare. In contrast, the social disorder indicators all occur with extremely rare frequency except one: adults loitering or congregating. The concern is that the social disorder scale will be dominated by this single item. Even that item has a low frequency, so that the overall scale may well lack reliability. In the next section, tools are developed to more formally test these intuitions.

4. A MODEL FOR UNCERTAINTY IN SYSTEMATIC SOCIAL OBSERVATION

Let us now consider how to adapt tools found useful in psychometrics to the problem of evaluating measures of ecological settings, here obtained through systematic social observation of neighborhoods. First, it will be desirable to understand how the items function within each construct and to use this information to build an interval scale for each. In the analogy with ability testing, each face-block is an "examinee," each indicator of disorder is an "item," and a "correct response" occurs when a face-block achieves a "yes" on that item. In this setting, item "difficulty" is the severity of the indicator of disorder, and face-block "ability" is its summary score on the disorder measure.

Second, it is essential to recognize that if the goal is to assess neighborhood clusters (NCs), there will be at least three components of measurement error: (1) item inconsistency within a face-block; (2) face-block variation within NCs, and (3) temporal variation. Temporal variation is an obvious problem in the case of measuring social disorder. The probability of finding adults loitering or drinking or finding peer gangs hanging out, or of seeing prostitution or drug deals will clearly depend on the time of day on which a face-block is observed. Thus it will be necessary to estimate and adjust for time of day. Fortunately, time of day varied substantially within every NC because of the time required to complete the observation. The attempt to model and estimate each component of error variation is consistent with generalizability theory in psychometrics.

Third, the item response model must allow for randomly missing data because only a random sample of the face-blocks yielded data coded from the videotapes. The hierarchical logistic regression model we describe below makes use of all available data.

Fourth, we are interested in the association between the constructs of physical and social disorder, adjusting for measurement error. This is akin to a confirmatory factor analysis in which ten items reflect physical disorder, seven items reflect social disorder, and the aim is to understand the association between physical and social disorder conceived as latent variables or factors.

To achieve these three goals, we formulate a three-level hierarchical logistic regression model. The level-1 units are item responses within face-blocks, the level-2 units are face-blocks, and the level-3 units are NCs.

4.1. *Level*-1 *Model*

The level-1 model represents predictable and random variation among item responses within each face-block. This is a standard one-parameter item response model and might be termed a Rasch model with random effects.[1] However, it will contain two dimensions (physical and social disorder) rather than the single dimension in classical applications of the Rasch model.

Let Y_{ijk} be an indicator taking on a value of unity if indicator i of disorder is found present in face-block j of neighborhood k, with $Y_{ijk} = 0$ if not; and let μ_{ijk} denote the probability $Y_{ijk} = 1$. That is,

$$Y_{ijk} | \mu_{ijk} \sim \text{Bernoulli} \;\; ;$$

$$E(Y_{ijk} | \mu_{ijk}) = \mu_{ijk}, Var(Y_{ijk} | \mu_{ijk}) = \mu_{ijk}(1 - \mu_{ijk}) \;\; . \tag{1}$$

[1] An important advantage of the random effects approach is that data from all face-blocks, even those with a zero on every item, contribute to the analysis. In contrast, a standard fixed effects Rasch analysis would exclude such cases.

As is standard in logistic regression, we define η_{ijk} as the log-odds of this probability. Thus we have

$$\eta_{ijk} = \log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) . \tag{2}$$

The structural model at level 1 accounts for predictable variation within face-blocks across items. It views the log-odds of finding disorder on item i as depending on which aspect of disorder is of interest (physical or social) and which specific item is involved. Let D_{pijk} take on a value of 1 if item i is an indicator of physical disorder, 0 otherwise; and let $D_{sijk} = 1 - D_{pijk}$ similarly indicate whether that item indicates social disorder. Then we have

$$\eta_{ijk} = D_{pijk}\left(\pi_{pjk} + \sum_{m=1}^{9} \alpha_{mjk} X_{mijk}\right) + D_{sijk}\left(\pi_{sjk} + \sum_{m=1}^{6} \delta_{mjk} Z_{mijk}\right) , \tag{3}$$

where

X_{mijk}, $m = 1, \ldots, 9$ are nine dummy variables representing nine of the ten items that measure physical disorder (each taking on a value of 1 or 0);
Z_{mijk}, $m = 1, \ldots, 6$ are six dummy variables representing six of the seven items that measure social disorder.

In fact, we "center" each X and Z around its grand mean. This enables us to assign the following definitions:

π_{pjk} is the adjusted log-odds of finding physical disorder on a "typical item" when observing face-block j of NC k;
π_{sjk} is the adjusted log-odds of finding social disorder on a "typical item" when observing face-block j of NC k;
α_{mjk} reflects the "difficulty" or "severity" level of item m within the physical disorder scale;[2] similarly, δ_{mjk} reflects the "difficulty" or "severity" level of item m within the social disorder scale.

[2] The interpretation of these coefficients as "difficulty" or "severity" requires that they be multiplied by -1.

Using the analogy of educational testing, π_{pjk} and π_{sjk} are the pair of abilities being measured and each α and δ reflects item difficulty. These item difficulties could, in principle, be allowed to vary across face-blocks or NCs; however, in the absence of theory that might predict such variation, they will be held constant in the interest of parsimony. Thus $\alpha_{mjk} = \alpha_m$ and $\delta_{mjk} = \delta_m$ for all j, k. Note that one item within each scale must serve as the "reference item" (it is not represented by a dummy variable). This item is defined to have a difficulty of zero and all other item difficulties are compared to it.

One benefit of explicitly representing the item difficulties in the model is that face-block measures of disorder, π_{pjk} and π_{sjk}, are adjusted for missing data. In the current data set, missing data arise because of the expense of coding the videotapes, leading to the decision to code just a random subsample of face-blocks within NCs. Face-blocks not sampled will have data from the observation log but not the coding log. No bias arises because the coded face-blocks constituted a representative sample of face-blocks in the NC. Nevertheless, controlling the item difficulties enables all of the data collected to be effectively used in the analysis.

4.2. *Level*-2 *Model*

The level-2 model accounts for variation between face-blocks within NCs on latent face-block disorder. Each is predicted by the overall NC level of disorder and the time of day during which the face-block was observed:

$$\pi_{pjk} = \beta_{pk} + \sum_{q=1}^{5} \theta_{pqk}(\text{Time})_{qjk} + u_{pjk}$$

$$\pi_{sjk} = \beta_{sk} + \sum_{q=1}^{5} \theta_{sqk}(\text{Time})_{qjk} + u_{sjk} \quad . \tag{4}$$

$(\text{Time})_{qjk}$ for q = 1, . . . ,5 are five time-of-day indicators (specifically, they indicate 7:00 to 8:59 AM; 9:00 to 10:59 AM; 11:00 AM to 12:59 PM; 1:00 to 2:59 PM; and 3:00 to 4:59 PM, where the omitted group is from 5:00 to 6:59 PM).

θ_{pqk} and θ_{sqk} are regression coefficients that capture the time-of-day effects on observing physical and social disorder within NC k. In principle, these could be allowed to vary over NCs, but for parsimony we shall hold them constant: $\theta_{pqk} = \theta_{pq}$ and $\theta_{sqk} = \theta_{sq}$ for all k. Note that the model allows different time-of-day effects for the social disorder items than for the physical disorder items. Driving this decision is the fact that certain observable social interactions (e.g., adults drinking) are much more likely to occur later in the day than early in the day while physical evidence such as the presence of graffiti should not be so sensitive to time of day. The model can also be elaborated to allow time-of-day effects to vary across items. Thus the "item difficulties" in equation (3)—the α and δ coefficients—could be separately modeled as a function of time of day. We forgo this option to reduce the complexity of the model, particularly in light of the low frequency associated with many of the items (Table 2). β_{pk} and β_{sk} are the "true" scores for NC k on physical and social disorder, respectively, adjusting for time of day.

The random effects u_{pjk}, u_{sjk} are assumed to be bivariate normally distributed with zero means, variances τ_{pp} and τ_{ss}, and covariance τ_{ps}. The variances will be large when face-blocks vary greatly within NCs on their levels of disorder.

4.3. *Level*-3 *Model*

The third and final level of the model describes variation between NCs, the key units of measurement, on physical and social disorder. We have simply

$$\beta_{pk} = \gamma_p + \upsilon_{pk}$$

$$\beta_{sk} = \gamma_s + \upsilon_{sk} \tag{5}$$

where γ_p and γ_s are the grand mean levels of physical and social disorder in Chicago neighborhoods and the random effects υ_{pk} and υ_{sk} are assumed to be bivariate normally distributed with zero means, variances ω_{pp} and ω_{ss}, and covariance ω_{ps}. The variances will be large when NCs vary greatly on their levels of disorder.

Estimation. Combining equations (2)–(5), our task is to estimate the nonlinear mixed model

$$E(Y_{ijk} | \mu_{ijk}) = \mathrm{Prob}(Y_{ijk} = 1 | \mu_{ijk}) = \mu_{ijk} = (1 + \exp\{-\eta_{ijk}\})^{-1} \tag{6}$$

with

$$\eta_{ij} = D_{pijk}\left(\gamma_p + \sum_{q=1}^{5} \theta_{pq}(\text{Time})_{qjk} + \sum_{m=1}^{9} \alpha_m X_{mijk} + u_{pjk} + \upsilon_{pk}\right)$$

$$+ D_{sijk}\left(\gamma_s + \sum_{q=1}^{5} \theta_{sq}(\text{Time})_{qjk} + \sum_{m=1}^{6} \delta_m Z_{mijk} + u_{sjk} + \upsilon_{sk}\right) \ . \qquad (7)$$

For purposes of illustration in the pages to follow, all model parameters were estimated simultaneously by penalized quasi-likelihood or "PQL" (Breslow and Clayton 1993) using an algorithm described in detail by Raudenbush (1995) and implemented in Version 4 of the HLM program (Bryk et al. 1996). The advantages and disadvantages of this approach relative to alternative approaches are discussed in Appendix A. That appendix also provides a sensitivity analysis based on a better approximation to maximum-likelihood estimates.

4.4. *Measurement Properties to Be Estimated*

The three-level hierarchical logistic regression model described above can be viewed as an item response model embedded within a hierarchical structure in which the secondary units of measurement, the face-blocks, are nested within the units of primary interest, the NCs. It extends the usual item response model also in allowing for multiple characteristics to be measured—in this case, physical social and physical disorder, rather than a single, unidimensional trait—and in allowing for randomly missing responses.

Fitting the model produces considerable information of interest in assessing the quality of the measures. The item difficulties have been mentioned above and their use in creating and interpreting a scale will be illustrated in the next section. Other key quantities are described below.

Intra-NC Correlations. The variance estimates within and between NCs yield an estimated "intra-NC correlation" on each measure that expresses the consistency of disorder across face-blocks. Consider the physical disorder items. If we substitute equation (5) into equation (4), we have a combined model for π_{pjk}, the latent trait being measured for face-block j of neighborhood k:

$$\pi_{pjk} = \gamma_p + \sum_{q=1}^{5} \theta_{pq}(\text{Time})_{qjk} + u_{pjk} + \upsilon_{pk} \ . \qquad (8)$$

This leads to the following definition of the intra-NC correlation for physical disorder:

$$\rho_{NCp} = \text{Corr}(\pi_{pjk}, \pi_{pj'k}) = \frac{\text{Cov}(\pi_{pjk}, \pi_{pj'k})}{[\text{Var}(\pi_{pjk}) * \text{Var}(\pi_{pj'k})]^{1/2}}$$

$$= \frac{\omega_{pp}}{\omega_{pp} + \tau_{pp}} \ . \tag{9}$$

Here face-block *pjk* and face-block *pj′k* are two different face-blocks within the *k*th NC. The intra-NC correlation for social disorder is, of course, analogous. The intra-NC correlation in equation (9) represents the proportion of variation in the true latent traits that lies between NCs. By definition, such variation excludes item inconsistency. By conceiving η_{ijk} as a latent variable following a logistic distribution, it is also possible to define an intra–face-block correlation and an alternative intra-NC correlation that would incorporate item inconsistency (see Gibbons and Hedeker 1997:1533).[3] Large intra-NC correlations imply that face-blocks within NCs are comparatively similar and that NCs vary considerably.

NC-level Reliabilities. Closely related to the intra-NC correlation is the internal consistency reliability of NC measurement. It depends on the intra-NC correlation but also on the number of face-blocks sampled, the number of items per scale, and the item difficulties. An approximation to the reliability for NC *k*, in the case of physical disorder, is given by

$$\lambda_{pk} = \frac{\text{Var}(\beta_{pk})}{\text{Var}(\hat{\beta}_{pk})} \approx \frac{\omega_{pp}}{\omega_{pp} * \dfrac{\tau_{pp}}{J_k} + \dfrac{1}{n_k J_k \omega_k}} \ , \tag{10}$$

where

λ_{pk} is the internal consistency of the physical disorder measure for NC *k*;
n_k is the average number of items per face-block in NC *k* ($n_k = 10$ if videotapes for all face-blocks in that NC are coded);
J_k = the number of face-blocks sampled within NC *k*;
w_k is the average within NC *k* of $\mu_{ijk}(1 - \mu_{ijk})$ on physical disorder items.

[3] The latent trait π_{pjk} is what we seek to measure more and more accurately as we add items to the scale, and the intra-NC correlation indexes the relative importance of NC variation and face-block variation within NCs on this trait. This is different from the intra-NC correlation on a measure based on a fixed number of items.

This conception of internal consistency can be motivated as follows. Suppose we use only the data from NC k to estimate β_{pk} and we regard that estimate as our measure of β_{pk}, the true level of physical disorder in NC k. Equation (10) is then the proportion of the variance in the estimates that is attributable to variance in the trait of interest; it is also the correlation between two such estimates derived from independent random samples of face-blocks. This approach to measurement reliability in an ecological setting is a direct extension of the approach used by Raudenbush et al. (1991) to measure school climate. While they used a three-level linear model, we extend that methodology to a three-level logistic model for dichotomous item responses. Appendix B provides the details.

Inspection of equation (10) reveals that reliability will be high when (1) the between-NC variance ω_{pp} is large relative to the within-NC variance τ_{pp}; (2) when the number of items in scale n_k is large; (3) when the number of face-blocks sampled, J_k, is large; and (4) when the typical probability of finding an aspect of disorder in a given face-block—that is μ_{ijk}—is near .50, at which point w_k achieves its maximum.

Face-block Reliability. It may be desirable to measure disorder at a lower level of geographic analysis, indeed, at the face-block level. Such measures could be assigned to individuals in a longitudinal study—for example, by geocoding their addresses. Reliability at the face-block level is given by

$$\lambda_{pjk} = \frac{\omega_{pp} + \tau_{pp}}{\omega_{pp} + \tau_{pp} + \dfrac{1}{n_{jk} w_{jk}}} \tag{11}$$

and will depend heavily on the number of items and the value of w_{jk},[4] the average of $\mu_{ijk}(1 - \mu_{ijk})$ within face-block jk. Here n_{jk} is the number of items assessed in that face-block.

[4] Equation (11) gives an internal consistency measure for discriminating among face-blocks in different NCs. An internal consistency measure for discriminating among face-blocks within the same NC is

$$\lambda_{\text{within}\, pjk} = \frac{\tau_{pp}}{\tau_{pp} + \dfrac{1}{n_{jk} w_{jk}}}.$$

Interscale Correlation. Of obvious interest is the correlation between physical and social disorder. This correlation can be estimated at the NC or face-block level. At the NC level, we have

$$\mathrm{Corr}(\beta_{pk}, \beta_{sk}) = \frac{\omega_{ps}}{(\omega_{pp} + \omega_{ss})^{1/2}} \tag{12}$$

while at the face-block level we have

$$\mathrm{Corr}(\pi_{pjk}, \pi_{sjk}) = \frac{\tau_{ps} + \omega_{ps}}{[(\tau_{pp} + \omega_{pp}) * (\tau_{ss} + \omega_{ss})]^{1/2}} \ . \tag{13}$$

We illustrate application of these ideas in the next section.

5. RESULTS

Tables 3–6 provide the model fitting results. The two scales behave quite differently, as expected.

5.1. *Item Severity*

In Table 3 items with negative coefficients have low probabilities of occurrence and thus are rarer and, presumably, more “difficult” or “severe” than are items with positive coefficients. Thus, in the physical disorder scale, the presence of cigarettes or cigars and garbage on the street or sidewalk, along with the presence of empty beer bottles, are comparatively less severe than the presence of gang graffiti, abandoned cars, condoms, or drug paraphernalia (needles and syringes). Thus item severity conforms to intuitive expectations. The exception is political graffiti, which is exceptionally rare yet not generally regarded as especially severe. A nice feature of the physical disorder scale is that the item severities vary substantially, a feature of a “well-behaved” scale.

In contrast, all of the severities in the social disorder scale are clumped at the severe end except for the item indicating adults loitering or congregating. This pattern reflects the low frequency of the social disorder indicators apparent in Table 2 and discussed earlier. Although the item severities are not well separated, their ordering does correspond to theoretical expectation, with adults loitering and drinking alcohol being less severe than adults fighting, prostitution, or drug sales.

TABLE 3
Model Fitting Results: Item Difficulty at Face-Block Level

Item	Coefficient	SE
Physical Disorder		
Intercept	−2.215	0.225
Cigarettes, cigars on street or gutter	3.456	0.032
Garbage, litter on street or sidewalk	2.431	0.031
Empty beer bottles visible in street	1.126	0.032
Tagging graffiti	0.338	0.036
Graffiti painted over	(0)	(reference item)
Gang graffiti	−0.667	0.043
Abandoned cars	−1.297	0.046
Condoms on sidewalk	−2.569	0.071
Needles/syringes on sidewalk	−2.893	0.082
Political message graffiti	−5.028	0.269
Social Disorder		
Intercept	−7.017	(0.153)
Adults loitering or congregating	3.884	(0.227)
People drinking alcohol	0.590	(0.280)
Peer group, gang indicators present	(0)	(reference item)
People intoxicated	−0.106	(0.325)
Adults fighting or hostilely arguing	−0.512	(0.366)
Prostitutes on street	−0.599	(0.376)
People selling drugs	−0.696	(0.388)

5.2. *Scale Construction*

The item maps are displayed graphically in Figures 2 and 3. The horizontal axis gives scale scores and the vertical axis gives the frequency of NC's. The figures include the list of items that compose the scale; distances between items represent differences in item difficulty. Note the spread of item difficulties in the case of physical disorder (Figure 2) and the clumping in the case of social disorder (Figure 3). NC scale scores are in the same metric as are item severities, and the figure suggests that these are nearly unimodal and symmetric in distribution. Note that this "nice distribution" is defined on the logit scale on which the NC scores are measured. Indeed, the construction of such a scale is a key goal of the item response analysis. Differences between NCs in their disorder scores can be interpreted unambiguously as expected differences in the log-odds of finding disorder on a typical item in the scale. The resulting scale is thus mean-

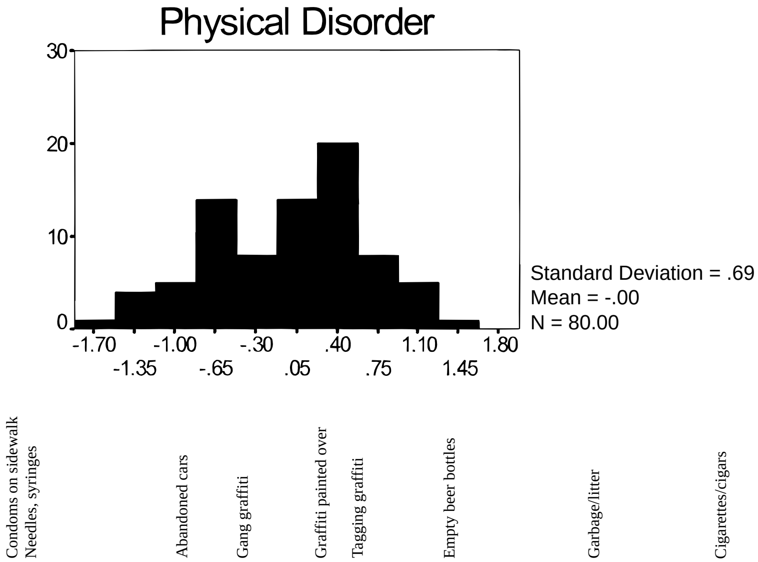

FIGURE 2. Item map for physical disorder.

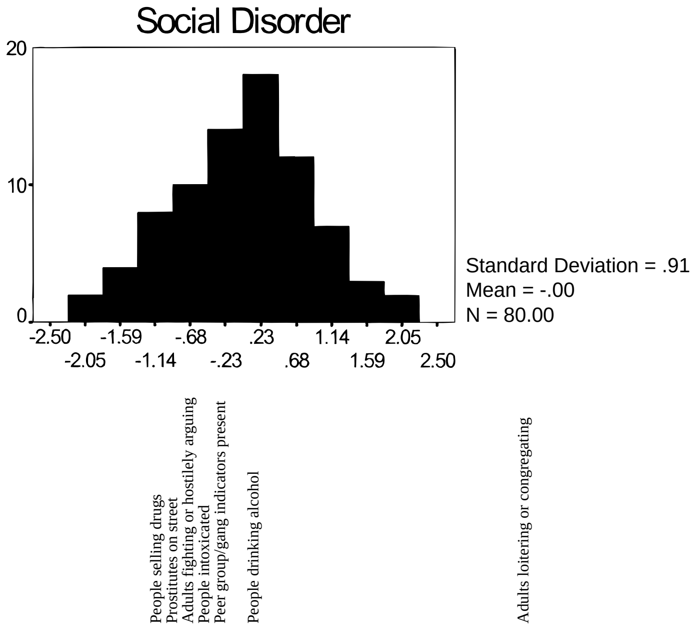

FIGURE 3. Item map for social disorder.

ingfully interpretable and arguably a linear (interval) scale appropriate for analysis via standard linear models (Rasch 1980; Wright and Masters 1982; Lord 1980).

5.3. *Time of Day*

Table 4 provides estimates of the effects of time of day. One would expect social interactions in public view to occur with relatively little frequency early in the morning and more frequently later on. This would presumably be true of those social interactions indicative of disorder as well, and that is what the results show. Note that there is a near linear positive trend in time for social disorder with coefficients of (−0.766, −0.715, −0.363, −0.057, −0.134, and 0.000) as the day progresses. No such trend is apparent in the case of physical disorder. All other model estimates are adjusted for any time-of-day effects.

5.4. *Variance-Covariance Components and Related Quantities*: *Physical Disorders*

Estimation of the variance-covariance components (Table 5) provides the necessary data to compute useful indicators of data quality (Table 6). For

TABLE 4
Model Fitting Results: Time-of-Day Effects at NC Level (N = 80)

Item	Coefficient	SE
Physical Disorder		
7:00–8:59	0.213	0.043
9:00–10:59	0.036	0.031
11:00–12:59	0.057	0.036
1:00–2:59	0.073	0.040
3:00–4:59	0.020	0.033
5:00–6:59	(0)	(reference time)
Social Disorder		
7:00–8:59	−0.766	0.180
9:00–10:59	−0.715	0.115
11:00–12:59	−0.363	0.137
1:00–2:59	−0.057	0.129
3:00–4:59	−0.134	0.107
5:00–6:59	(0)	(reference time)

TABLE 5
Model Fitting Results: Variance-Covariance Components

	Variance-Covariance Component	Estimate	SE
(a) Between face-blocks within NCs	Variance of physical disorder	0.734	0.019
	Variance of social disorder[a]	—	—
	Covariance[b]	—	—
(b) Between NCs	Variance of physical disorder	0.475	0.076
	Variance of social disorder	0.981	0.184
	Covariance	0.394	0.096

[a,b] Variance and covariance were constrained to zero.

physical disorder, the estimated variance between face-blocks is 0.734, while the estimated variance between NCs is 0.475. Thus the estimated ICC for physical disorder (see equation (9) and note 3) is 0.475/(0.475 + 0.734) = 0.39. Thus about 39 percent of the variation in the physical disorder of face-blocks is estimated to be between NCs. This fact, when combined with the typical frequency of "yes" responses (Table 2) and the large number of face-blocks per NC (equation 10), yields a high average reliability of 0.98 (Table 6) at the NC level. Thus the data enable us to distinguish among NCs with high reliability. The reliability for distinguishing among face-blocks within NCs is estimated to be much lower, at 0.36. This reflects the dependence of the reliability at the face-block level on the number of items. More items would be required to increase this reliability.

TABLE 6
Some Measurement Properties

Property	Physical Disorder	Social Disorder
Intra-NC correlations	.39	—
Between NC reliability (average)	.98	.84
Between face-block reliability (average)	.36	—

Level	Correlation
Inter-scale Correlation at NC Level	.58

5.5. *Variance-Covariance Components and Related Quantities*: *Social Disorder*

The social disorder scale behaves quite differently. The point estimate of the variance within NCs for social disorder is zero. This result does not imply that face-blocks within NCs are homogeneous. Rather, the result appears to reflect the extreme rarity of "yes" responses of most social disorder items (Table 2). The data simply are too sparse at the face-block level to facilitate stable estimation of variance between face-blocks within NCs. Yet the variation between NCs is quite substantial ($\hat{\omega}_{SS} = 0.981$), leading to a respectable NC-level reliability estimate of 0.84. Although the frequency of indicators of social disorder is rare at the face-block level, when we aggregate over the many face-blocks within an NC, we are able to achieve a respectable between-NC reliability.

5.6. *Implications for Research Design*

Applying the logic of generalizability analysis, we can use our data to inform the design of new research. Figure 4 plots the expected NC-level reliability of the two scales as a function of the number of face-blocks

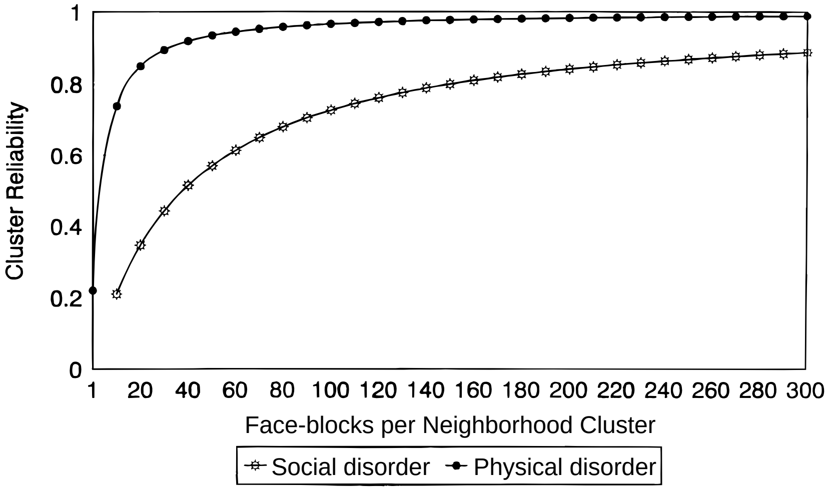

FIGURE 4. NC reliability as a function of face-blocks sampled (holding constant the number of items).

sampled. For physical disorder, there appears to be little point in sampling more than 80 to 100 face-blocks per NC if the sole aim is to obtain reasonable NC reliability. The same is not true for social disorder. More face-blocks are required for adequate reliability in measuring social disorder (as compared to physical disorder).

Physical disorder results provide good news for the next analyses of the PHDCN data. It is clear that physical disorder can be reliably measured at much lower levels of aggregation than the NC. Thus it is feasible to construct physical disorder measures at the level of the block group or census tract,[5] creating a measure that is more proximal geographically to the longitudinal cohort subjects of PHDCN than is the NC. Our results are less encouraging about the measurement of social disorder at lower levels of aggregation because of the low frequency of “yes” responses on most social disorder items.

5.7. *Dimensionality*

The correlation between physical disorder and social disorder is estimated to be .58 (Table 6). This is the estimated correlation of the two latent variables and is therefore automatically adjusted for measurement error in each. The implication is that physical and social disorder as conceived here are quite strongly related, although not so strongly as to be viewed as a single dimension. There is reason to pursue sound measures of each construct separately. This aspect of the analysis has parallels with confirmatory factor analysis. Here we have hypothesized multiple indicators for each of two traits. The data suggest that the two traits are highly related, but not entirely confounded. In a typical confirmatory factor analysis, factor loadings would vary while the variance of latent traits would be constrained. Here we impose equal loadings while allowing latent trait variances to vary.

5.8. *Convergent and Divergent Validity*

Key tests of validity of measurement involve assessing correlations with theoretically related constructs measured independently. Convergent va-

[5] Indeed, subsequent analysis showed that the reliabilities of the physical disorder at the census tract level were nearly identical to those at the level of the NC even though there are two to three tracts per NC.

lidity implies that theoretically linked measures ought to correlate highly. Divergent validity implies that correlations should be smaller with variables that are not clearly linked theoretically. Table 7 gives some evidence about physical and social disorder.

Observed physical disorder is correlated highly with those constructs measured in both the community survey (described earlier) and independent sources (census data, official police records) most theoretically linked to it. Thus we see that a substantial correlation of $r = .71$ emerges with perceptions of social disorder as measured in the community survey. SSO physical disorder also has a moderately strong correlation with the survey measures of social cohesion and social control ($r = -.62$ and $r = -.55$, respectively), in the direction expected. Further tests not shown indicate that physical disorder also correlates less strongly with those survey-derived constructs for which it has a weaker theoretical connection (e.g., anonymity, intergenerational ties, organizational density).

Turning next to construct validation with other independent sources, physical disorder is strongly related to census measures of concentrated poverty ($r = .64$), and less strongly with residential stability ($r = -.25$) and immigrant concentration ($r = .36$; see Sampson et al. 1997 for further details of these census-based factors). As expected, the observed physical disorder is significantly higher in neighborhoods characterized by poverty and instability, and in areas undergoing ethnic transition. Furthermore,

TABLE 7
Correlations of Systematic Social Observation Scales with Theoretically Related Variables from the U.S. Census and Community Survey

	SSO:	
	Physical Disorder	Social Disorder
I. Community Survey		
Social disorder	.71	.65
Social cohesion	−.62	−.55
Social control	−.55	−.56
II. U.S. Census		
Concentrated poverty	.64	.54
Residential stability	−.25	−.34
Immigrant concentration	.36	.21
III. Violence and Crime		
Perceived violence	.54	.59
Crime victimization	.32	.33

physical disorder measured in the SSO is positively and significantly correlated with survey perceptions of violence (r = .54) and aggregated reports of victimization (r = .32). These patterns conform to extant theory on urban disorder and crime (Skogan 1990; Taylor et al. 1985; Sampson and Groves 1989; Taub et al. 1984).

A similar pattern of correlations appears with respect to social disorder. In some cases the magnitude of correlations is a bit smaller than those involving physical disorder, which may reflect the less sanguine behavior of the social disorder scale. Nonetheless, the SSO measure of social disorder has quite robust relationships with theoretically linked constructs—again whether derived from the neighborhood survey or census. Taken together, then, the multiple sources of data provide independent evidence of both the convergent and divergent validity of SSO measures of disorder.

6. FINAL REMARKS

As interest in the social sciences turns increasingly to the integration of individual, family, and neighborhood processes, a potential mismatch arises in the quality of measures. Standing behind individual measurement are decades of psychometric research, producing measures that often have excellent statistical properties. In contrast, much less is known about measures of ecological settings such as neighborhoods and schools, and the methodology needed to evaluate these measures is in its infancy. The aim of this paper has been to move toward a science of ecological assessment by integrating and adapting tools from psychometrics to improve the quality of “ecometric” measures. We have used systematic social observation (Reiss 1971; 1975) linked to neighborhood surveys as a case study in this effort. The SSO is an especially important case, given the potential utility of videotaping as an observational strategy in the study of neighborhoods and other collectivities.

The two measures selected—physical and social disorder—behaved sufficiently differently statistically to be useful in clarifying how ideas from item response modeling, generalizability theory, and confirmatory factor analysis can be integrated to better understand the process of measuring ecological units. In the future, we plan to construct additional scales from the systematic social observation data, including land use, housing quality, traffic, advertising of tobacco and alcohol, recreational opportunities, and type of commercial district. The approaches described here can be used to evaluate the quality of measures and to build improved scales for use in the study of neighborhoods and human development. A crucial question is the causal link between crime and disorder, which is being addressed elsewhere (Sampson and Raudenbush forthcoming).

While our ecometric analysis borrows from standard psychometric techniques, it also integrates and otherwise extends them. Thus our random effects item response model is embedded in a three-level hierarchical regression model, enabling estimation of time-of-day effects and of variability within and between face-blocks. In this way, the item response analysis is formally incorporated into a generalizability analysis. Moreover, at the second level (between face-blocks) and third level (between NCs), we estimate the variance and covariance of the two latent variables (physical and social disorder), thus constructing a simple but multilevel confirmatory factor analysis. The resulting three-level hierarchical logistic regression model allows for randomly missing data at the level of items and uses all available information, even from those face-blocks having zero incidence on all disorder indicators.

The logic of ecological assessment and resulting multilevel error structure will generally prescribe such combinations and extensions of standard psychometric techniques. Thus a natural model for survey-based measures of settings (neighborhoods or schools) would have an ordinal response model at level 1 (between items within respondents) and would add two higher levels of variation: between respondents within settings and between settings. Such a model would be a three-level ordinal regression model.

Another extension to the approach sketched here would be to take into account spatial autocorrelation. In this paper, neighborhood clusters have been treated as independent.[6] Ongoing work will build spatial associations into the models presented here. We expect information about spatial dependence to reduce standard errors of measurement, possibly substantially, and to make it possible to obtain reasonable measures of neighborhood ecology even for persons residing in areas sparsely assessed by direct observation. In the meantime, the results of the present analysis suggest that the survey and SSO approaches have considerable promise for the reliable and valid assessment of neighborhood-level social processes.

[6] This assumption is not entirely implausible in the case of the SSO, which involves a probability sample of 80 NCs from among 343 NCs in Chicago. Many of the sample NCs are not contiguous to other sample NCs. Nevertheless, a more complete treatment would model spatial dependence between NCs.

APPENDIX A

Estimation by Penalized Quasi-likelihood “PQL” has several advantages. First, computations are fast, an important consideration for the data in this article, with 23,816 level-2 units and over 300,000 level-1 responses. Second, convergence is remarkably reliable, even when the probability of success is close to zero (Yang 1998). Third, the methodology is currently widely accessible.[7] However, Breslow and Lin (1995) have found that PQL can produce variance components estimates that are substantially negatively biased when the true variance component is large. PQL is based on a linearization of the model—that is, a Taylor-series expansion of μ_{ijk} in equation (6) around the approximate posterior modes of the random effects. An alternative approach, which Breslow and Clayton (1993) have termed MQL, expands μ_{ijk} around 0—that is, in a MacLaurin series for the random effects, produces even greater bias (see Rodriguez and Goldman 1995). The asymptotic bias is eliminated when the model is estimated by maximum likelihood (ML) or by Bayesian methods. These approaches, however, require difficult integrations: the random effects must be integrated to evaluate the likelihood. Hedeker and Gibbons (1994) developed excellent approximations to ML estimates in the case of two levels by using Gauss-Hermite quadrature for numerical integration. Gibbons and Hedeker (1997) extended this approach to three-level models with a single random effect at level 3. Raudenbush and Yang (1998) developed a high-order LaPlace approximation to the integral that produced results comparable to those with quadrature based on 20-30 quadrature points (generally regarded as a large number of points and therefore yielding a good approximation to the likelihood). The LaPlace approach is computationally remarkably efficient. These results are based on Yang’s (1998) dissertation. Bayesian estimation can be implemented by the Gibbs sampler (Zeger and Karim 1991) and is implemented in the most recent version of MLWiN (Goldstein et al.1998). The Bayesian approach provides perhaps the most elegant solution, in that all inferences fully take into account the uncertainty about the variance components. However, this solution appears infeasible computationally given the size of the data set at hand and the complexity of the model.

To conduct an analysis of the sensitivity of results arising from bias associated with PQL, we settled on the higher-order LaPlace strategy. The

[7] For example, programs HLM, MLWin (Goldstein et al. 1998), and Proc Mixed (Littel et al. 1996) provide PQL or closely related estimation algorithms).

general theory, with application to binary response data, is described in detail by Raudenbush and Yang (1998). The task is first to integrate the random effects from the joint density of the data and random effects, in effect, a binomial-normal mixture. The integrand is represented by a sixth-order Taylor series expansion around the maximizer of this joint density. The integral then can be seen as equivalent to the expectation taken over a multivariate normal distribution of the third-order and higher-order terms. It is therefore possible to evaluate the integral and maximize it using a Fisher-scoring algorithm.

A comparison of results from PQL and the sixth-order LaPlace ("LaPlace 6"), yielded the results shown in Table A.1.

We conclude from these results that estimates of item difficulty and time-of-day effects are essentially insensitive to choice of estimation procedure. (Note that the magnitudes of the estimated time-of-day effects are a bit larger under LaPlace 6, as expected given the bias toward zero of PQL estimates.) Inferences about the ecometric properties of the physical disorder scale are also insensitive (ICC of 0.39 versus 0.32 for LaPlace 6, reliability of 0.98 for both PQL and LaPlace 6). And the interscale correlation estimates are similar, 0.58 versus 0.52. Inferences about ecometric properties of the social disorder scale are more sensitive (reliability of 0.84 for PQL versus 0.70 for LaPlace 6). LaPlace standard errors are generally somewhat smaller for physical disorder items and larger for social disorder items than are the corresponding PQL standard errors. While we are cautiously optimistic that PQL will produce reasonable inferences about ecometric properties, we expect that much better approximations to likelihood-based inference will rapidly become available to researchers over the next couple of years. We recommend use of these better approximations as they become available.[8]

APPENDIX B

To construct the reliability of NC measures, suppose we conceive the level-1 model (equation 3) as a generalized linear model. In matrix notation, we have

$$\eta_{jk} = D_{jk}\pi_{jk} + X_{jk}\alpha \ , \tag{B1}$$

[8] The LaPlace 6 algorithm used in computing this sensitivity analysis is available upon request from the first author (rauden@umich.edu).

TABLE A.1
Sensitivity Analysis Based on Alternative Estimation Approach

	PQL Point Estimates (standard errors)	LaPlace 6 Point Estimates (standard errors)
Item Difficulties: Physical Disorder		
Intercept	−2.215 (0.225)	−2.044 (0.313)
Cigarettes, cigars	3.456 (0.032)	3.689 (0.018)
Garbage, litter	2.431 (0.031)	2.541 (0.015)
Empty beer bottles	1.126 (0.032)	1.097 (0.014)
Tagging graffiti	0.338 (0.036)	0.432 (0.020)
Gang graffiti	−0.667 (0.043)	−0.663 (0.022)
Abandoned cars	−1.297 (0.046)	−1.346 (0.024)
Condoms	−2.569 (0.071)	−2.816 (0.053)
Needles/syringes	−2.893 (0.082)	−3.028 (0.076)
Political graffiti	−5.028 (0.269)	−4.823 (0.299)
Item Difficulties: Social Disorder		
Intercept	−7.017 (0.153)	−7.156 (0.124)
Adults loitering	3.884 (0.227)	3.968 (0.272)
People drinking alcohol	0.590 (0.280)	0.996 (0.382)
People intoxicated	−0.106 (0.325)	0.386 (0.392)
Adults fighting, arguing	−0.512 (0.366)	0.147 (0.527)
Prostitutes	−0.599 (0.376)	0.126 (0.400)
Selling drugs	−0.696 (0.388)	−0.106 (0.505)
Time of Day: Physical Disorder		
7:00–8:59	0.213 (0.043)	0.298 (0.032)

(*Table continues*)

TABLE A.1
Continued.

	PQL Point Estimates (standard errors)	LaPlace 6 Point Estimates (standard errors)
Time of Day: Physical Disorder (Continued)		
9:00–10:59	0.036	0.082
	(0.031)	(0.019)
11:00–12:59	0.057	0.055
	(0.036)	(0.021)
1:00–2:59	0.073	0.141
	(0.040)	(0.025)
3:00–4:59	0.020	0.039
	(0.033)	(0.020)
Time of Day: Social Disorder		
7:00–8:59	−0.766	−1.066
	(0.180)	(0.173)
9:00–10:59	−0.715	−0.923
	(0.115)	(0.091)
11:00–12:59	−0.363	−0.314
	(0.137)	(0.116)
1:00–2:59	−0.057	−0.091
	(0.129)	(0.110)
3:00–4:59	−0.134	−0.121
	(0.107)	(0.101)
Variance-Covariance Components		
$\hat{\tau}_{\mathrm{pp}}$	0.734	1.070
$\hat{\tau}_{\mathrm{ss}}$	0	0
$\hat{\tau}_{\mathrm{ps}}$	0	0
$\hat{\omega}_{\mathrm{pp}}$	0.475	0.500
$\hat{\omega}_{\mathrm{ss}}$	0.981	0.515
$\hat{\omega}_{\mathrm{ps}}$	0.394	0.266
Corr ($\beta_{\mathrm{s}}, \beta_{\mathrm{p}}$)	0.58	0.52
ICC physical	0.39	0.32
NC Reliab physical	0.98	0.98
NC Reliab social	0.84	0.70

where η_{jk} is the n_{jk} by 1 vector consisting of elements η_{ijk}, D_{jk} is an n_{jk} by 2 matrix of indicators (the first column for physical disorder, the second column for social disorder), π_{jk} is a 2 by 1 vector of coefficients; X_{jk} is the n_{jk} by 15 matrix of dummy variables for items; and α is the 15 by 1 vector of item difficulties (assumed equal across all NCs and assumed known). Then, applying maximum-likelihood estimation via iteratively reweighted least squares, we find the approximate variance-covariance matrix of estimates $\hat{\pi}_{jk}$ (McCullagh and Nelder 1989:119, eq. 4.18) to be

$$V_{jk} = \mathrm{Var}(\hat{\pi}_{jk}) = (D_{jk}^T W_{jk} D_{jk})^{-1} \ , \tag{B2}$$

where W_{jk} is a diagonal n_{jk} by n_{jk} matrix with entries $\mu_{ijk}(1 - \mu_{ijk})$. Given the structure of D_{jk} and the diagonal nature of W_{jk}, it is easy to see that V_{jk} is a 2 by 2 matrix with diagonal entries $1/(n_{pjk}w_{pjk})$ and $1/(n_{sjk}w_{sjk})$, with n_{pjk} being the number of physical disorder items assessed in face-block jk, n_{sjk} being the number of social disorder items in that face-block, w_{pjk} the average value of $\mu_{ijk}(1 - \mu_{ijk})$ for the physical disorder items in face-block jk and w_{sjk} the average value of $\mu_{ijk}(1 - \mu_{ijk})$ for social disorder items. Next, we reformulate equation (4) as

$$\hat{\pi}_{jk} = \beta_k + T_{jk}\theta + u_{jk} + (\hat{\pi}_{jk} - \pi_{jk}) \ , \tag{B3}$$

where $\beta_k = (\beta_{pk}, \beta_{sk})^T$, $u_{jk} = (u_{pjk}, u_{sjk})^T$, T_{jk} is a 2 by 10 matrix of time-of-day indicators, and θ is a 10 by 1 vector of time-of-day effects, assumed known. Here $u_{jk} \sim N(0, \tau)$. Generalized least squares estimation of β_k then yields the variance-covariance matrix

$$\mathrm{Var}(\hat{\beta}_k) = \left[\sum_{j=1}^{J_k} (\tau + V_{jk})^{-1} \right]^{-1} .$$

When $V_{jk} = V_k$ for all j, we have

$$\mathrm{Var}(\hat{\beta}_k) = J^{-1}(\tau + V_k)$$

with the first diagonal equal to the sum of the second and third terms of the denominator of equation (10). That equation thus represents an approximation that will be accurate when $V_{jk} = V_k$ for all j. Our results, however, are based on the estimated generalized least squares estimates for each NC, not on this approximation. The approximation is primarily useful in revealing the structure of measurement error in the three-level setting.

REFERENCES

Brennan, R., and M. Kane. 1979. “Generalizability Theory: A Review.” Pp. 33–51 in *New Directions for Testing and Measurement: Methodological Developments* edited by R. Traub. San Francisco, CA: Jossey-Bass.

Breslow, N., and D. G. Clayton. 1993. “Approximate Inference in Generalized Linear Mixed Models.” *Journal of the American Statistical Association* 88:9–25.

Breslow, N. E. and X. Lin. 1995. “Bias Correction in Generalized Linear Mixed Models with a Single Component of Dispersion.” *Biometrika* 82:81–91.

Brooks-Gunn, Jeanne, Greg Duncan, and Lawrence Aber, eds. 1997. *Consequences of Growing up Poor*, Vol. 1. New York: Russell Sage Foundation.

Bryk, A., S. Raudenbush, R. Congdon, and M. Seltzer. 1996. *HLM: Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs*. Chicago, IL: Scientific Software International.

Carter, Woody, Jody Dougherty, and Karen Grigorian. 1996. “Videotaping Neighborhoods.” National Opinion Research Center, University of Chicago, Working Paper.

Cook, T., S. Shagle, and Degirmencioglu. 1997. “Capturing Social Process for Testing Mediational Models of Neighborhood Effects.” Pp. 94–119 in *Neighborhood Poverty: Policy Implications in Studying Neighborhoods*, Vol. 2, edited by J. Brooks-Gunn, G. J. Duncan, and L. Aber. New York: Russell Sage Foundation.

Covington, J., and R. B. Taylor. 1991. “Fear of Crime in Urban Residential Neighborhoods: Implications of Between- and Within-Neighborhood Sources for Current Models.” *The Sociological Quarterly* 32:231–49.

Cronbach, L., G. Gleser, N. Harinder, and N. Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Duncan, Greg, and Stephen Raudenbush. 1997. “Getting Context Right in Quantitative Studies of Child Development.” Presented at the conference, “Research Ideas and Data Needs for Studying the Well-being of Children and Families,” October.

Elliott, Delbert, William Julius Wilson, David Huizinga, Robert J. Sampson, Amanda Elliott, and Bruce Rankin. 1996. “Effects of Neighborhood Disadvantage on Adolescent Development.” *Journal of Research in Crime and Delinquency* 33:389–426.

Gibbons, Robert D., and Donald Hedeker. 1997. “Random Effects Probit and Logistic Regression Models for Three-Level Data.” *Biometrics* 53:1527–37.

Goldstein, Harvey, Jon Rashbash, Ian Plewis, David Draper, William Browne, Min Yan, Geoff Woodhouse, and Michael Healy. 1998. *A User’s Guide to MLwiN*. London: University of London, Institute of Education.

Hedeker, D., and R. D. Gibbons. 1994. “A Random Effects Ordinal Regression Model for Multilevel Analysis.” *Biometrics* 50:933–44.

Joreskog, K., and D. Sorbom. 1988. *LISREL 7: A Guide to the Program and Applications*. Chicago, IL: Statistical Package for the Social Sciences.

Littel, R. C., G. A. Millilen, W. W. Strong, and R. D. Wolfinger. 1996. *SAS System for Mixed Models*. Cary, NC: SAS Institute.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Mayer, Susan E., and Christopher Jencks. 1989. “Growing up in Poor Neighborhoods: How Much Does It Matter? *Science* 243:1441–45.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*, 2d ed. London: Chapman and Hall.

Muthen, B. 1991. “Multilevel Factor Analysis of Class and Student Achievement Components. *Journal of Educational Measurement* 28:338–54.

———. 1997. “Longitudinal and Multilevel Modeling: Latent Variable Modeling of Longitudinal and Multilevel Data.” Pp. 453–80 in *Sociological Methodology 1997*, edited by Adrian Raftery. Cambridge, MA: Blackwell Publishers.

National Opinion Research Center (NORC). 1995. PHDCN Project 4709. Systematic Social Observation Coding Manual, June 1995. NORC/University of Chicago.

Pallas, A. 1988. “School Climate in American High Schools.” *Teachers College Record* 89:541–53.

Rasch, G. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago.

Raudenbush, S. W. 1995. “Hierarchical Linear Models to Study the Effects of Social Context on Development.” Pp. 165–201 in *The Analysis of Change*, edited by J. Gottman. Hillsdale, NJ: Lawrence Erlbaum.

Raudenbush, S., and R. J. Sampson. Forthcoming. “Assessing Direct and Indirect Associations in Multilevel Designs with Latent Variables.” *Sociological Methods and Research*.

Raudenbush, S., B. Rowan, and S. Kang. 1991. “A Multilevel, Multivariate Model for Studying School Climate in Secondary Schools with Estimation via the EM Algorithm. *Journal of Educational Statistics* 16:295–330.

Raudenbush, S. W., and Meng-Li Yang. 1998. “Maximum Likelihood for Hierarchical Models via High-Order, Multivariate LaPlace Approximation.” Paper submitted to the *Journal of Computational and Graphical Statistics*.

Reiss, A. J., Jr., 1971. “Systematic Observations of Natural Social Phenomena.” Pp. 3–33 in *Sociological Methodology 1971*, edited by H. Costner. San Francisco: Jossey-Bass.

———. 1975. “Systematic Observation Surveys of Natural Social Phenomena.” Pp. 132–50 in *Perspectives on Attitude Assessment: Surveys and their Alternatives*. NTIS.

Rodriguez, G., and N. Goldman. 1995. “An Assessment of Estimation Procedures for Multilevel Models with Binary Responses.” *Journal of The Royal Statistical Society, Series A* 56:73–89.

Sampson, R., and W. Groves. 1989. “Community Structure and Crime: Testing Social-Disorganization Theory.” *American Journal of Sociology* 94:774–802.

Sampson, R., and S. Raudenbush. (Forthcoming). “Systematic Social Observations of Public Spaces: A New Look at Neighborhood Disorder.” To appear in *American Journal of Sociology*.

Sampson, R., S. Raudenbush, and F. Earls. 1997. “Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy.” *Science* 277:918–24.

Sirotnik, K. 1980. “Psychometric Implications of the Unit-of-Analysis Problem (with Examples from the Measurement of Organizational Climate).” *Journal of Educational Measurement* 17:245–82.

Skogan, W. 1990. *Disorder and Decline: Crime and the Spiral of Decay in American Neighborhoods*. Berkeley: University of California Press.

Taub, Richard, D. Garth Taylor, and Jan Dunham. 1984. *Paths of Neighborhood Change: Race and Crime in Urban America*. Chicago: University of Chicago Press.

Taylor, R. B., S. D. Gottfredson, and S. Brower. 1984. “Block Crime and Fear: Defensible Space, Local Social Ties, and Territorial Functioning.” *Journal of Research in Crime and Delinquency* 21:303–31.

Taylor, R. B., S. Shumaker, and S. D. Gottfredson. 1985. “Neighborhood-Level Links Between Physical Features and Local Sentiments: Deterioration, Fear of Crime, and Confidence.” *Journal of Architectural Planning and Research* 21:261–75.

Wright, B., and G. Masters. 1982. *Rating Scale Analysis: Rasch Measurement*. Chicago, IL: MESA Press.

Wright, B., and M. Stone. 1979. *Best Test Design: Rasch Measurement*. Chicago, IL: MESA Press.

Yang, Meng-Li. 1998. “Increasing the Efficiency in Estimating Multilevel Bernoulli Models.” Ph. D. dissertation, Michigan State University, East Lansing.

Zeger, S. L., and M. R. Karim. 1991. “Generalized Linear Models with Random Effects: A Gibbs Sampling Approach.” *Journal of the American Statistical Association* 86:79–86.