Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2013 January 1; 176(1): 227–249. doi:10.1111/j.1467-985X.2012.01065.x.

Assessing the measurement error properties of interviewer observations of neighbourhood characteristics

C. Casas-Cordero, **F. Kreuter**, **Y. Wang**, and **S. Babey**

Summary

Interviewer observations made during the process of data collection are currently used to inform responsive design decisions, to expand the set of covariates for nonresponse adjustments, to explain participation in surveys, and to assess nonresponse bias. However, little effort has been made to assess the quality of such interviewer observations. Using data from the Los Angeles Family and Neighbourhood Survey (L.A.FANS), this paper examines measurement error properties of interviewer observations of neighbourhood characteristics. Block level and interviewer covariates are used in multilevel models to explain interviewer variation in the observations of neighbourhood features.

1. Introduction

New insights from urban sociology and social epidemiology have revitalized interest in neighbourhood characteristics and their effects (Sampson et al., 2002), in particular the development of instruments measuring neighbourhood social processes and the physical environment. The relevance of these activities for survey research cannot be neglected, as neighbourhood characteristics have been found to be associated with survey outcomes (Sampson et al., 2002; Brooks-Gunn et al., 1997; Babey et al., 2008; Kawachi and Berkman, 2003) and with participation in household surveys (Couper and Groves, 1996; Campanelli et al., 1997; Groves and Couper, 1998; O’Muircheartaigh and Campanelli, 1999; Kennickell, 1999; Lynn et al., 2002; Kennickell, 2003; Johnson et al., 2006; Bates et al., 2008; Durrant and Steele, 2009; Casas-Cordero, 2010). For these reasons, neighbourhood characteristics are potentially ideal candidates to be used in responsive designs (Groves and Heeringa, 2006) and in nonresponse adjustments (Little, 1986; Kalton and Flores-Cervantes, 2003; Little and Vartivarian, 2003, 2005; Groves, 2006).

Most of the studies cited above used census demographic data rather than direct measures of the neighbourhood characteristics of interest (Groves and Couper, 1998; Kennickell, 2003; Johnson et al., 2006; Durrant and Steele, 2009). Other studies have used data available from commercial vendors. Schräpler et al. (2010), for example, linked commercial data on the immediate vicinity of the households from the MOSAIC data system to the German Socio-Economic Panel Study (SOEP). Unfortunately, not all neighbourhood constructs which are appropriate to explain the behaviour of households and interviewers during the survey recruitment process are available on census records or from commercial vendors. For example, in their survey participation model, Groves and Couper (1998) focused on the role of shared norms and values among potential respondents. The authors suggested that the lack of cohesion at the community level may have as its counterpart the isolation of individuals both from the local community and from society in general. This relative lack of participation or involvement in the community may reduce the willingness to engage in activities such as surveys (Groves and Couper, 1998, 177). Indicators that capture such behaviour can be collected through observation.

Urban sociologists are interested in similar theoretically-derived assessments of neighbourhoods (Diez-Roux, 2001; Sampson et al., 2002; Morenoff, 2003; Kawachi and Berkman, 2003), and have developed a series of reliable observational indicators (Taylor et al., 1995; Raudenbush and Sampson, 1999; Weich et al., 2001; Caughy et al., 2001; Craig et al., 2002; Brown et al., 2004; Andresen et al., 2006; Laraia et al., 2006; Zenk et al., 2007; Furr-Holden et al., 2008). But in these studies, observers are usually hired for the sole purpose of collecting neighbourhood data (Caughy et al., 2001; Weich et al., 2001; Dunstan et al., 2005; Zenk et al., 2007), or hired to code extensive videotapes collected of relevant neighbourhoods (Raudenbush and Sampson, 1999; Cohen et al., 2000). Thus these methods are time consuming and costly, and therefore less than ideal tools for regular survey production at the national level.

An alternative is the use of survey interviewers to collect neighbourhood observations. As part of face-to-face data collection efforts, many interviewers are already charged with making observations on *respondent* characteristics, attitudes and behaviours, and some surveys request interviewers to note *housing unit* or *neighbourhood* physical characteristics. Prominent surveys in the U.S. that collect such information are the Current Population Survey (CPS), the National Health Interview Survey (NHIS) and National Survey of Family Growth (NSFG). In Europe, countries participating in the European Social Survey (ESS) also collect similar observations, and in the U.K., examples are the British Crime Survey (BCS) and the Survey on British Attitudes (SBA).

Unfortunately, the predictive power of such interviewer observations of respondent, household, and neighbourhood characteristics for survey participation have not always been as high as theoretically expected or as would be needed for successful nonresponse adjustment (Campanelli et al., 1997; Lynn, 2003; Durrant and Steele, 2009; Maitland et al., 2009; Casas-Cordero, 2010; Kreuter et al., 2010). As Kreuter et al. (2010) discuss, this low correlation could in part be due to measurement error in interviewer observations.

While urban sociologists have looked at the measurement error properties of neighbourhood observations collected by specifically trained observers (Raudenbush and Sampson, 1999), the literature on the quality of neighbourhood observational data collected by interviewers is only now emerging. To our knowledge outside of L.A. FANS only one study has examined factors that influence interviewers' perceptions of neighbourhood characteristics (Eifler et al., 2009). The present paper attempts to fill this gap.

Data from the Los Angeles Family and Neighbourhood Survey (Sastry and Pebley, 2003b) allow such assessments. Two features make this study particularly relevant for survey research. First, the study trained a single group of interviewers to collect both survey data and neighbourhood data – which mirrors how regular surveys collect observational data. Second, the study collected multiple independent observations on each sampled neighbourhood – which allows the estimation of variance components associated with interviewers or geographical groupings. Furthermore, data on interviewers is available to test whether certain interviewer characteristics influence their perception of neighbourhood characteristics. That is, whether interviewers notice certain neighbourhood features, and how they rate such features. With these data we can address the following three research questions:

a. How good is the agreement between ratings of neighbourhood features on the same block face (the street on one side of a city block), and how large is the variability in agreement across these features?

b. How much interviewer variance is there in the neighbourhood ratings?

c. Do available interviewer or neighbourhood characteristics explain variability in the neighbourhood ratings?

2. Rationale and design of the L.A.FANS study

The Los Angeles Family and Neighbourhood Survey (L.A.FANS) is a study of families in Los Angeles County and the neighbourhoods in which they live, with a stratified multi-stage probability sample of census tracts, households, and individuals. In the United States, census tracts are the units most often used as geographical ‘sampling points’ (primary sampling units) for face-to-face household surveys. U.S. Census tracts are small statistical subdivisions of a county, with boundaries normally following visible features, but in some instances they may follow governmental unit boundaries and other non-visible features. They always nest within counties and average about 4,000 inhabitants. (For more details see https:ask.census.govappanswersdetaila_id245.)

A total of 65 U.S. census tracts were selected out of three sampling strata in the L.A.FANS design. The L.A. FANS strata correspond to census tracts that were very poor (those in the top 10% of the poverty distribution), poor (tracts in the 60–89th percentiles), and non-poor (tracts in the remaining 60% of the distribution). Tracts in the ‘poor’ and ‘very poor’ stratum were oversampled at this first phase. (See Sastry et al. (2003) for more details.) For the purpose of this paper, we defined neighbourhoods by their census tract boundaries, since this is the most ubiquitous definition in the urban sociology literature.

Within the selected tracts (neighbourhoods), a preliminary sample of approximately 9, 400 addresses was drawn to complete a screener interview, which consisted of answering a single question about the presence of children in the household. Among those successfully screened, a sub-sample of approximately 4, 100 households was selected for the L.A. FANS sample. Households with children were oversampled at this second phase. Household rosters were completed with 3, 083 households and individual interviews were completed with approximately 85% of the selected respondents.

L.A.FANS interviewers collected observations of the physical environment in all selected census tracts. A special data collection instrument and training protocol was designed for this purpose. The material provided to the interviewers had exact definitions of the items that they were asked to observe. This includes descriptions of terms “Graffiti: Spraypainting (or sometimes chalk) drawing or writing inscribed on rocks, walls, sidewalks, fences, etc. Does not include community murals.” as well as descriptions of the use of rating scales: “When estimating the amount of trash, your judgment is required. Below are some guidelines to help you get a feel for how the categories should work. None: No trash or junk is visible; Very little: If you look around carefully, you see trash in one or two places; Some: You notice trash or junk in three or four locations; A lot: You see trash or junk in several locations.” (Sastry and Pebley, 2003a).

Interviewers were trained to carry out their observations systematically, but fairly quickly, making the task similar to the one of interviewers in regular surveys. The observation protocol involved driving around the entire city block, and walking down each block face and recording the characteristics of that block face at the end of the walk. A block face is both sides of the street along one side of the block. If we think of a block as a rectangle, then a block face corresponds to the streets forming the sides of the rectangle. Many city blocks in the L.A.FANS data set are not proper rectangles but have more than four sides. Interviewers were instructed to complete these observations the first time they visited the sampled block. Neighbourhood observations were conducted between April 2000 and July 2001, with one third being done in April and May of 2000, and the remainder in April and May of 2001.

2.1. Analytic sample

One of the special features of the L.A.FANS data collection effort is that several interviewers completed the neighbourhood observations on the same areas independently of each other. In some neighbourhoods (census tracts) up to six independent assessments are available in the L.A.FANS data, though some with a time lag of more than six weeks. Having multiple observations of the same neighbourhoods is unusual in regular face-to-face surveys. Since the L.A.FANS had the study of neighbourhood effects as its primary objective, extra effort was made to reduce the measurement error of the neighbourhood observations by increasing the number of observations available for each item of the planned scales.

To minimize external sources of variation, such as changes in the neighbourhoods or differences in the interviewer training and the time since training, we restrict the analytic sample to the pairs of observations completed closest in time (usually completed within a two-week period (Peterson et al., 2007, p. 4)). Following this criterion, observations were dropped that corresponded to the 3rd, 4th, 5th or 6th rating on each block face. A total of seven of the 35 interviewers rated fewer than 30 block faces, compared to a median number of 108 block face ratings per interviewer. To ensure enough variability those seven interviewers, and a total of 30 unique block faces were excluded from the analysis.

Table 1 displays the structure of the analytic data set. In total, there are 3, 998 records reflecting two independent assessments of 1, 999 block faces nested within 419 unique blocks, nested within 65 Census tracts. The models used in this paper combine the multilevel structure given by the geography of small areas, and the cross-classification with interviewers (n=28), which are given by the distribution of their work assignments. By cross-classification, we mean that interviewers are crossed with tracts and blocks, but interviewers are not nested within tracts or blocks. Each block and its block faces are nested in a single tract. However, blocks (and the block faces within) were observed by different interviewers, creating a partial interpenetration. Figure 1 illustrates this situation, where Interviewer 1 only collected data on Tract (neighbourhood) 1, whereas Interviewer 2, 3, and 4 collected observations in multiple tracts (neighbourhoods). Here ‘interviewers and blocks’ and ‘interviewers and tracts’ are cross-classified levels because none of them is completely nested within the other. In the analytic sample, each of the 65 sampled census tracts was observed by between 2 and 8 interviewers with an average of 5 interviewers per tract; and each of the 419 blocks was observed by two interviewers. Each interviewer observed between 2 and 31 tracts with an average of 12 tracts per interviewer and between 4 and 88 blocks with an average of 30 blocks per interviewer.

In our analysis of interviewer observations, we focus on 16 items that correspond to neighbourhood observations of ‘physical disorder’ (n=6), ‘residential decay’ (n=5) and ‘residential security’ (n=5). Signs of disorder and decay, also known as ‘incivilities’, are examined because they are theoretically linked to the mechanisms explaining cooperation in household surveys (Groves and Couper, 1998; Casas-Cordero, 2010). Items on residential security are included in the analyses because similar variables appear in many surveys and are therefore of wide interest beyond L.A.FANS. For the ‘social disorder’ indicators we will only report descriptive statistics. Those items were left out of our more complex analyses due to the very low prevalence in most of these categories (see Table 2). The items indicating abandoned cars and drugs were excluded for the same reason.

The physical disorder items and the items measuring residential security were collected using 4 and 5-category Likert-type questions. To facilitate comparisons and to address the skewness of the reported observations, all Likert-type items were dichotomized so that 1 means ‘presence’ and 0 means ‘absence’ of the characteristic being rated. (Note: The original scale for the physical disorder items was: 1=None; 2=A little; 3=Some; 4=A lot. The original scale for most of the residential decay and residential security items was: 1=None; 2=Very few; 3=Some; 4=Many; 5=All. The item ‘yards’ was reverse-coded before analysis. The item ‘buildings’ had a different scale (1=Very poor; 2=Poor; 3=Fair; 4=Very good; 5=Excellent) and also had to be reverse coded before analysis.) Exact wordings for each item are displayed in Tables 2 and 3. Within each panel in the table, the items are listed in the order that they appear in the observation forms. Unweighted estimates of the prevalence of each item, based on all records in the analytic dataset (n = 3, 998), are also provided.

Items in the physical disorder scale captured a wide spectrum of disorder phenomena. Consistent with the literature, items considered ‘less severe’ (e.g. litter) were reported much more frequently than ‘more severe’ items (e.g. drugs). Items in the social disorder scale showed lower prevalence, which is consistent with the higher severity of the types of observations covered (e.g. presence of gangs). Sample sizes are smaller for the residential decay and residential security items because some of them were not collected for block faces rated largely as nonresidential.

While urban sociologists might use the entire scale in their substantive analyses, this paper intentionally analyzes measurement error in individual items rather than scales, for two reasons: (a) to explore the properties of neighbourhood items that are currently being collected by large survey projects, and (b) to inform future studies that need to decide which items to pick to minimize the additional burden on the interviewer when making these observations during the normal data collection process. (Analyses conducted in sections 3 and 4 on the individual items also have been carried out on the full scales. The methodology and modeling results of the full scale analysis are discussed briefly in sections 3 and 4. Additional details on these results can be obtained from the authors.)

2.2. Explanatory variables

The knowledge we have about the factors driving perceptions of neighbourhood characteristics comes from studies investigating residents’ perceptions. Here Sampson and Raudenbush (2004) argue that increased past exposure to disorder increases the threshold at which disorder is perceived as a problem. Supporting this statement they and others found older residents less likely to perceive disorder than younger residents (Sampson and Raudenbush, 2004; Franzini et al., 2008), and residents involved in their communities less likely to perceive disorder than those that are not. The same mechanisms may not apply to non-residents, such as interviewers collecting data in unfamiliar neighbourhoods.

An interesting mechanism from the fear of crime literature links demographic characteristics to perception via the ‘vulnerability perspective’. This perspective emphasizes individual demographics to explain fear and is based on the assumption that fear is greatest when individuals perceive that they are at a physical disadvantage against potential assaults and/or when individuals believe that they are particularly vulnerable to being victims of crime (Wyant, 2008). Early research found that women (Clemente and Kleiman, 1977) and the elderly (Lee, 1983) were more fearful of crime – despite the fact that they were less likely to be victimized (Garofalo and Laub, 1978). No clear directionality was found across studies for the effect of socio-economic status variables on perception of disorder (Mujahid et al., 2007). Eifler et al. (2009) showed that perception of signs of incivility is increased through prior victimization experience, unfortunately those covariates are not available for this study.

Similar mechanisms might hold true for interviewers observing signs of disorder across different neighbourhoods. Interviewers who are more ‘involved in their communities’, for example, could perceive signs of disorder and decay more negatively when rating other communities. Similarly the fear of crime literature shows that the 'lack of familiarity' with a place is correlated with a heightened sense of insecurity and risk perception (Taylor et al., 1984). Interviewers working in unfamiliar places thus might also be more likely to perceive signs of disorder.

One result from the U.S. indicates that – compared to other residents in the same neighbourhood – black and minority residents are less likely to report signs of disorder (Sampson and Raudenbush, 2004; Mujahid et al., 2007; Franzini et al., 2008). Sampson and Raudenbush (2004) argue that increased past exposure to disorder increases the threshold at which disorder is perceived as a problem. Thus, given the history of racial segregation in the United States, it is possible that blacks have been exposed to more disorder than whites in the past and therefore it is possible that blacks and whites judge disorder differently.

Characteristics of the Interviewers—In the L.A. FANS study interviewer characteristics were recorded in the interviewer background questionnaire. In addition to demographic characteristics, the interviewer questionnaire captured information about the interviewers' own neighbourhoods (e.g. city of residence, satisfaction with own neighbourhood, how long lived there). Descriptive statistics for the variables used here are displayed in Table 4.

Race, the first variable in the table, was used here as an indicator of 'potential exposure to disorder'. Age represents a correlate of 'vulnerability'. The third set of variables was intended to capture interviewers' 'community involvement' and was derived by us from questions on marital status, presence of children, and community activities.

Table 4 shows that, while white interviewers represented 35% of the interviewer crew, they collected 51% of the observations. Interviewers who were married represented 42% of the interviewers but collected 64% of the observations. When looking at the analytic dataset (n = 3, 998) it is important to remember that this data reflects 'workloads' and not the distribution of the characteristics of the crew of interviewers.

The characteristics described here are considered 'fixed' for each interviewer, i.e. they do not change as the field work progresses. The next section describes a different set of variables that do not correspond to particular characteristics of interviewers or the blocks they are rating, but to the interaction of the two.

Characteristics of the Occasion of Measurement—In addition to neighbourhood and interviewer characteristics, there are some variables that arise as a combination of these characteristics. This is most prominently the case for interviewers' familiarity with the neighbourhood to be rated. The variable 'neighbourhood close' indicates if the distance between the interviewer's neighbourhood and the tract rated by him or her is less than or equal to 5 miles. We used longitude and latitude associated with the centroid of each census tract in the sample and the interviewer's place of residence to construct this indicator. The Stata 10 command 'geodist' was used to calculate ellipsoidal distances between two georeferenced points – tracts' location and interviewers' residence location. The procedure uses Vincenty's equations to approximate the distance between two points on the earth's surface (Vincenty, 1975). The variable 'experience with block' indicates whether the interviewer had any type of previous experience with the block (e.g. listing, interviewing) or not.

To capture the effect of 'temporal variability' (Raudenbush, 2003) on the ratings, indicators of time of day ('Rated After 5pm') and day of week ('Rated on Weekend') were constructed. Just like familiarity, the time of day at which a block was rated is not a property of the block itself nor is it a characteristic of the interviewer – the same block could have been rated at a different time if a different interviewer was assigned to it or if this interviewer had chosen a different time. Thus we summarize these variables as characteristics of the occasion of measurement. Table 5 displays the distribution of the variables based on the block-level data (*n* = 419).

Neighbourhood Structural Characteristics—The final set of correlates correspond to neighbourhood level attributes associated with neighbourhood socio-economic composition. Such measures are typically derived from Census records. Following Sampson and Raudenbush (1999) we use three measures of socio-economic composition: concentrated affluence, immigrant concentration, and concentrated disadvantage, plus an indicator of population density. The indicator of population density measures thousands of people per square mile and it is available at the census tract level. The measures of socio-economic composition are the result of a factor analysis based on variables from the 2000 Census (Casas-Cordero, 2010).

The principal-factor method was used to analyze the correlation matrix. Under this method the factor loadings are computed using the squared multiple correlations as estimates of the communality. Factors were rotated using the Varimax (orthogonal) rotation method, all implemented as part of the 'factor' command in Stata 10. The three factors evolving from this analysis were similar factors to those used in contemporary neighbourhood research on child-developmental outcomes (Sampson et al., 1999). The factor 'concentrated affluence' had an Eigenvalue greater than 9 and had high loadings on Census variables such as Non-Spanish speakers, Non-Hispanic origin, higher education, high income, and executive/professional occupation. With an Eigenvalue greater than 2 and high positive loadings for percentage of foreign born and non-citizens, and negative loadings for owner-occupied, the second dimension in the factor analysis captured the degree of 'immigrant concentration'. The predominant interpretation for the third factor is concentrated disadvantage. This factor had an Eigenvalue larger than 1.5 and loaded primarily on four variables: percentage in poverty, on public assistance, female head-of-household, and black residents. The percentage of variance explained by the first 3 factors was 0.9269 (original solution) and 0.7657 (rotated solution). By construction, all factor variables have a mean of zero and a standard deviation of one. For the original items included in each factor see the Appendix. The variable population density was also standardized to facilitate comparability of the results.

3. Methodology

First, agreement was assessed using Kappa statistics (Cohen, 1960), which are typically used to examine the agreement between two observers in categorical rating tasks. Cohen's Kappa is a measure of interrater reliability and ranges generally from 0 to 1.0. Large numbers mean better reliability, values near zero would suggest that agreement between the observers is due to chance. Next, the measurement error properties of the neighbourhood observations were examined using cross-classified multilevel logistic regression models (Goldstein, 2010). These models account for the clustering of housing units within interviewers and geographic areas, and they allow us to analyze how much of the variation in perception is due to the interviewers or areas they work in. We examine which interviewer characteristics contribute to any variation across interviewers, and how the interviewer variance component compares to those estimated for geographic units such as tracts. The availability of repeated observations for interviewers and neighbourhoods allows for the estimation of such variance components. In a second step, covariates will be included in the model to see if their fixed effects explain the variance in the random effects that were initially observed.

3.1. Analyses of Random Influences

The standard measurement error model used for neighbourhood constructs (Raudenbush and Sampson, 1999) has been conceptualized as using a three-level multilevel model, where neighbourhood observations are at the lowest level, blockfaces are at the second level, and tracts are at the highest level. Looking back at Figure 1, the model can be viewed as an item response model at level 1, embedded within a multilevel structure in which the secondary units of measurement (here the blocks) at level 2 are nested within the units of primary interest, the neighbourhoods (here tracts which form level 3). The model has been extended by allowing for multiple characteristics (factors) to be measured simultaneously (Raudenbush and Sampson, 1999). Here, we propose to modify the standard model by allowing the estimation of interviewer random effects alongside the area level effects.

The multilevel structure of the data used in this paper is illustrated in Figure 1, and is represented in model (1), where the dependent variable is a binary indicator taking on a value of 1 if a neighbourhood characteristic, e.g. graffiti, is observed at a given occasion of measurement *i* in block *j* in tract *k*, rated by interviewer *r*, and has a value of 0 if it was not observed. The model includes three random effects to take into account the dependency among the observations within blocks (b_j), tracts (t_k), and interviewers (o_r). We also allow for a random interaction effect ($l_{(kr)}$) of the cross-classified factors tracts and interviewers.

$$log(\frac{\pi_{ij(kr)}}{1-\pi_{ij(kr)}})=\beta+b_j+t_k+o_r+l_{(kr)} \quad (1)$$

The model in (1) corresponds to an ‘unconditional’ model because it fits the probability of observing a neighbourhood characteristic, e.g. graffiti, as a function of an overall mean (β) without covariates. The random effects are assumed to be normally distributed with variances $\gamma_b^2, \gamma_t^2, \gamma_o^2$, and γ_l^2 respectively. The models were fit using SAS®version 9.2 with the GLIMMIX procedure. PROC GLIMMIX fits the specified model by maximizing an approximation to the likelihood integrated over the random effects using adaptive Gaussian quadrature; here Laplacian approximation was used (Wolfinger, 1993). The residual follows a logistic distribution, and the scale parameter is set to one in PROC GLIMMIX, so the residual variance is (π^2)/3 by definition of the probability density function (Long, 1997, p. 47–48). Following Rabe-Hesketh and Skrondal (2012, p. 536), we used likelihoodratio tests for the null hypothesis that each respective variance component is zero. For each neighbourhood item, five models are fit starting with an empty model and adding one random effect at a time for interviewers (o_r), tracts (t_k), blocks (b_j), and tracts by interviewers ($l_{(kr)}$).

The model expressed with equation (1) can also be used to derive estimates of the unique interviewer (ρ_{int}) and tract (ρ_{sp}) contribution to the variation in the observed item prevalence (Snijders and Bosker, 1999, p. 224). These estimates represent, respectively, the intraclass correlation between two observations made by the *same* interviewer in *different* blocks (eq. 2), and the correlation between two observations made on the *same* tract by *different* interviewers on different blocks (eq. 3). Given that neighbourhoods vary in their social composition and many other aspects, one would not be surprised to find natural variation associated with the neighbourhoods (Schnell and Kreuter, 2005). Estimates of ρ_{int} and ρ_{sp} will be used to compare the extent of interviewer effects across different neighbourhood items and to gauge their magnitude. Items with high interviewer effects will be examined further.

$$\rho_{int}=\frac{\gamma_o^2}{\gamma_t^2+\gamma_b^2+\gamma_o^2+\gamma_l^2+\pi^2/3} \quad (2)$$

$$\rho_{sp}=\frac{\gamma_t^2}{\gamma_t^2+\gamma_b^2+\gamma_o^2+\gamma_l^2+\pi^2/3} \quad (3)$$

3.2. Analyses of Systematic Influences

Analyses described above are used to assess different measurement properties conceptualized as random variation. In addition we employed models to assess systematic sources of variation associated with the interviewers, the occasion of measurement, and the neighbourhoods (blocks and tracts). Here too, the multilevel structure of the data is taken into account. Thus model (1) is extended to include covariates hypothesized to influence perceptions of disorder, as well as additional controls found to be influential on perceptions of disorder in the past. The extended model is given by the equation below,

$$log\left(\frac{\pi_{ij(kr)}}{1-\pi_{ij(kr)}}\right)=\beta+\sum_{p=1}^{5}\alpha_p\mathbf{I}_{pr}+\sum_{q=1}^{6}\theta_q\mathbf{B}_{qjr}+\sum_{s=1}^{4}\tau_s\mathbf{T}_{sk}+t_k+b_j+o_r+l_{(kr)} \quad (4)$$

where I_{pr} is a vector containing five interviewer variables (p = 5), B_{qj} is a vector containing four block level covariates (q = 4), and T_{sk} contains four tract-level variables (s = 4). The regression coefficients associated with these covariates are α_p, θ_q and τ_s. The same procedures used to estimate the unconditional model in equation (1) were used to estimate the conditional model in equation (4). The model includes random effects for interviewers (o_r), tracts (t_k), blocks (b_j) and an interaction term between interviewers and tracts ($l_{(kr)}$). The random effects are assumed to be normally distributed when conditioned on the covariates. The between-within method (DDFM=BETWITHIN option) was used in computing the denominator degrees of freedom for the significance tests of fixed effects. This method partitions the residual degrees of freedom into between-subject and within-subject parts (Schluchter and Elashoff, 1990). All analyses were done with unweighted data. Because the current analyses focus on interviewers' observations of blocks rather than on households or individual respondents, the L.A. FANS sample weights are not appropriate here.

Cross-classified linear multilevel models were estimated to evaluate the effects on three summary scales; physical disorder, residential decay, and residential security. Here too an unconditional model was estimated with random effects for interviewers, tracts, and blocks, and a residual term that captures remaining variation including the interaction of interviewers and tracts. The same set of covariates was used.

4. Results

4.1. Agreement, reliability and measurement error in interviewer observations

At the outset of this paper, we posed the following research questions: “How good is the agreement between ratings?” and “How large is the variability in agreement across these features?” A typical estimate of the extent of agreement between two binary outcomes is the percentage agreement between two observers. Agreement in the L.A. FANS study is very high for many items, in part because of the low prevalence of the disorder items. Most of the time the two independent observers agreed on ‘not having seen’ a certain sign of disorder. The problem with this estimator is that it does not correct for agreement due to chance.

Figure 2 displays both the percentage agreement for all 25 items (in grey), as well as Cohen's Kappa (in black), which corrects for chance agreement (Cohen, 1960; Hintze, 2005). Percentage agreement was relatively high across all items ($min = 0.66$, $max = 0.99$). The social disorder items, however, achieved the highest scores ($min = 0.87$, $max = 0.99$). This result is not surprising since, given the 'severity' of the disorder items, most of the time the two independent observers agreed on 'not having seen' signs of social disorder. Once agreement due to chance is taken into account by the Kappa statistic, the performance of the disorder items decreased ($min = 0.00$, $max = 0.17$). The Kappa statistics for the remaining items varied considerably ($min = 0.12$, $max = 0.62$).

Overall, the values for Cohen's Kappa are moderate to low, and lower than most reliability estimates reported in the urban sociology literature – where specialized observers are typically used. One important limitation of the L.A.FANS data, for the purpose of estimating reliability, is that in most cases the observations available for each block face were made at two different time points. Thus, we cannot completely disentangle interobserver variability from temporal variability. Among the 1,999 pairs of observations included in this analysis, only 23% were made on the same day and – among those observed on the same day – only 64% were made at the same time of day. To test the post-hoc hypothesis that temporal effects impact observations, we replicated the analysis on the full sample ($n = 3, 998$) with a subset of observations collected on the 'same day' ($n = 908$), and again for those observations collected on the 'same day and time' ($n = 586$). Overall the results did not vary much between the 'full' sample, the 'same day' sample, and the 'same day and time' sample. Results suggest that most neighbourhood features under observation are not very sensitive to time difference. An exception was trash, whose estimate of agreement did not change much when going from the 'full' sample to the 'same day' sample, but increased when using the 'same day and time' sample. Other items showed unexpected patterns, such as security gates for which agreement decreased when going from the full sample to the 'same day' and 'same day and time' sample. These results suggest that interviewer factors may be contributing to some of the relatively low agreement rates found here.

Given this result, one could also argue that much of what is observed could have no relationship to the spatial characteristics of the areas but be due to idiosyncratic characteristics of the interviewers. It is therefore interesting to evaluate unique interviewer effects ρ_{int} in the perception of neighbourhood characteristics relative to the geographic effects ρ_{sp}, where we would expect to find a source of variation. Figure 3 displays this result. Estimates of ρ_{int} correspond to the unique correlation between the observations collected on different block faces by the same interviewers. The higher this correlation, the stronger the effect of the interviewer on the observations he or she collects. Estimates of ρ_{sp} correspond to the correlation between the observations collected on different block faces that belong to the same neighbourhood (census tract). The higher this correlation, the stronger the evidence for the effect of the phenomenon we are measuring.

Two results are worth highlighting here: (1) the relative size of interviewer effects across different items, and (2) the relative size of interviewer effects to sampling point effects $\left(\frac{\rho_{sp}}{\rho_{int}}\right)$. Items such as indicators of bars on windows, boarded up housing, and vacant lots showed very low interviewer effects compared to tract level effects, suggesting that observers have a relatively clear – and common – understanding about how to rate these features. In the context of this study, this 'common ground' most likely comes from the special training they received prior to conducting the observations, and because these items are more salient, and easier to understand. For other items, the influence of interviewer 'idiosyncratic' judgments was probably stronger. Examples of items with large interviewer effects include the observations of the presence of cigarettes, litter and security gates. When examining factors that influence the perception of neighbourhood characteristics, we will focus on those that show higher interviewer effects than tract effects.

4.2. Factors influencing interviewer perceptions

In light of the significant contributions of interviewers and geographic areas to the variation in perception the question arises: do available interviewer or neighbourhood characteristics explain any of those variance components? We will answer this question in two parts. For two selected variables, we show the results for each modelling step in detail. For a large set of indicators, the final model is displayed directly.

Table 6 displays the estimated coefficients for the conditional multilevel models in equation (4) for the two selected items: trash, and graffiti. The panel at the bottom displays estimates of the variance components associated with interviewers, tracts, blocks, and the variance associated with the interaction of the cross-classified terms.

The unconditional model (Model 1) is used as a reference to compare the conditional models that incorporate the fixed effects of interviewers (Model 2), the occasions of measurement (Model 3) and the neighbourhoods (Model 4). To allow comparison across models, the four models associated with each item were estimated on the same estimation sample – which corresponds to the sample used in the model with all covariates (Model 4).

The interviewer variables in Model 2 did not reach statistical significance for the perception of trash. Model 3 added the covariates associated with the occasion of measurement. Again, none of the covariates reached statistical significance. Model 4 incorporated the last set of covariates, which were derived from census records and represented features of the socio-economic composition of the neighbourhoods being rated. Consistent with prior research, indicators of concentrated disadvantage, concentrated affluence and immigrant concentration were significant predictors of perceiving disorder. As expected, the indicator of population density was not significant.

When modeling the perception of graffiti, two interviewer characteristics stood out. White interviewers are $e^{1.53} = 4.62$ times more likely than non-white interviewers to record graffiti, and older interviewers are almost twelve times less likely ($1/e^{-2.47} = 11.82$) to record graffiti than their younger colleagues. Including characteristics of the occasion of measurement did not change those coefficients (Model 3), however including neighbourhood characteristics did. The probability of perceiving signs of graffiti was higher in neighbourhoods with higher values on the variables disadvantage and immigrant concentration. As expected, the probability of observing graffiti in affluent neighbourhoods was lower. The same pattern was observed for the perceptions of trash.

Two other findings are worth noting. First, the effect sizes of the coefficients of neighbourhood covariates were larger for graffiti than for trash. The second finding is the dramatic reduction in the coefficient associated with the tract-level random effect when neighbourhood characteristics were included, which occurred for both trash and graffiti. This finding suggests that the neighbourhood covariates were successful in explaining the variability associated with tracts (neighbourhoods).

The results discussed for trash and graffiti were replicated, to a large extent, across the other items of physical disorder, residential decay and residential security. These results are displayed in Table 7. The lack of predictive power of some of the covariates in the model is evident in the panel with the fixed effects. None of the variables involved in our hypotheses showed a consistent pattern – neither in terms of effect sizes nor statistical significance of the results. Interviewers with children, for example, were not more likely to see signs of litter on the block face. And interviewers living close to the area they were rating were just as likely to see signs of deteriorated buildings as those living further away. The strongest influence on the ratings was, by far, neighbourhood socio-economic composition. Neighbourhoods with higher levels of immigrant concentration and concentrated disadvantage are positively associated with perceptions of all signs of disorder and decay, however, they were negatively associated with perceptions of security signs (secsign) and neighbourhood watch signs (ngwatch). These results make sense, since these latter signs are found in neighbourhoods with lower prevalence of disorder and decay. Accordingly, neighborhoods with higher levels of concentrated affluence were negatively associated with perceptions of all signs of disorder and decay, but positively associated with perceptions of security signs (secsign) and neighbourhood watch signs (ngwatch).

The results for the entire scales (not shown but available from the authors upon request) match the patterns seen here, with interviewer age having some influence on the perception of residential decay, but not on physical disorder or residential security. No other interviewer characteristics influenced the scale perception. The tract level covariates showed effects on physical disorder, and residential decay, but not on residential security.

5. Discussion

Survey researchers are beginning to assess the potential use of observational data for methodological and practical purposes. Many surveys routinely require survey interviewers to collect observations on survey respondents. In this setting, adding neighbourhood observations to current call record forms may seem relatively easy. However, such requests create additional burden for the interviewers, and these data are not without measurement error. As a result, a revised version of the Raudenbush and Sampson (1999) model of measurement error for neighbourhood data was developed here and hypotheses about the influence of interviewer characteristics on the ratings were examined.

The revised model had three important features. First, it set up the analysis based on individual items rather than on a group of items (scales). Second, it incorporated interviewers as an additional level of clustering, which enabled the derivation of estimators for interviewer clustering effects. Finally the model incorporated different sets of covariates which were used to test hypotheses about the systematic influence of interviewers, occasions of measurement, and neighbourhood characteristics on the perceptions of signs of disorder and decay.

We report three main findings, corresponding to the research questions that we posed initially. First, we note that there is only moderate reliability of perceptions among interviewers rating the same block faces. However, the reliability in perception varies considerably across items, and is particularly high for the assessments of building conditions and appearances, as well as security related measures. These items are also collected by survey interviewers more often than the specialized items composing the social disorder measure.

Second, interviewer effects on the probability of perceiving neighbourhood features are generally rather small and tend to be smaller than or similar to the sampling point effects. Among the few exceptions are the perception of neighbourhood watch signs in the neighbourhoods, and the presence of security gates. Third, there was no evidence of the systematic influence of measured interviewer characteristics on the neighbourhood ratings.

The results suggest that idiosyncratic characteristics of interviewers do not influence their perceptions of disorder to a large extent. Despite the presence of random variation in the interviewer observations, the absence of systematic variation in those observations is good news to data collectors who plan to use interviewers to collect additional data used in nonresponse adjustment or responsive design decisions. One could argue though, that interviewer characteristics that explain the variation in interviewer observations have yet to be identified.

These results can also be use to inform the selection of neighbourhood characteristics for interviewers to observe. For example, in a study that involves the observation of a wide range of neighbourhoods, similar to L.A. FANS, even moderate amounts of variability due to interviewers may be tolerable as long as the variability due to the neighbourhoods themselves is still larger. In this case, it may be acceptable to include observations of items such as neighbourhood watch signs. However, in a survey that occurs in an area with less variation between neighbourhoods, interviewer influences on observations are a greater concern. In this case, survey methodologists may want to select items that are less influenced by interviewer effects relative to neighbourhood effects (such as presence of vacant lots or boarded up buildings) to capitalize on the ability to detect useful variation. The results from this analysis can help to identify which types of observations may be appropriate in each of these situations.

Of course there are some limitations in the present paper that are important to address: (1) lack of random assignment of interviewers to areas; (2) focus on the analysis of individual items rather than scales; and (3) limits on the generalizability of the results from this study. We will briefly discuss each issue here. The lack of randomization in investigations of interviewer variability could lead to overestimation of the interviewer effect (ρ_{int}) (Kish, 1962). As Kish points out, the overestimation of ρ_{int} could be great in those sampling operations where the interviewer has wide latitude in choosing his workload, but it might be small in surveys carried out at one limited site, where an approximation to randomization occurs automatically. The L.A. FANS study is much closer to the latter case, because all observations were completed in a single county.

In this paper we analyzed individual items rather than scales. As a result of using this approach, estimates of measurement error derived from the current analyses are most likely larger (i.e. provide an upper bound) than those derived from analysis of multiple items or composite scores. Researchers should consider these implications when evaluating whether to use measures of disorder derived from single items or composites scores for their substantive analysis.

The analyses presented here focused on variability across time and items. We did not examine estimates of reliability varying across places. As one of the reviewers to this paper suggested, it is conceivable that reliability decreases in areas where observations might be more difficult. Such differential measurement errors would be important to examine in future research.

One of the key features of the L.A. FANS study is that it combined (a) a state of the art questionnaire and training protocols for the collection of observational data, and (b) a regular crew of survey interviewers to collect those observations. This study, however, was conducted in a single city in the Unites States, thus results presented here might not be generalizable to a broader setting. It is an empirical question, however, whether these results can be replicated in a multi-site study or studies where training and other conditions in the field would vary greatly.

These findings are particularly relevant for the National Children’s Study (NCS), which will collect data on neighbourhood environments throughout the U.S. using the same items as were used in the L.A. FANS questionnaire. The NCS is designed to be a long-running, observational panel study of a nationally representative probability sample of 100,000 births to be followed from before birth to age 21. The study will examine the effects of the environment, broadly defined to include factors such as air, water, diet, sound, family dynamics, community and cultural influences, and genetics on the growth, development, and health of children across the United States (see http:www.nationalchildrensstudy.gov). Unlike the L.A. FANS study, however, it will use different contractors across the U.S. to collect the survey data and (most likely) the neighbourhood observational data. Different contractors might use different data collection protocols to collect the observational data. The L.A. FANS instruments, the interviewer manual and the training protocol, are already state of the art methods and the results presented here most likely provide an upper bound to the sources of error for contractors aiming to collect this type of data.

Results from this paper assessed the magnitude of the measurement error associated with neighbourhood observations collected by survey interviewers. Surveys that use less developed material than L.A. FANS might need to examine the reliability of interviewer observations. But even the detailed training manual was not enough in L.A. FANS to remove interviewer variability in perceptions. Visually enhanced material could be helpful. From our analyses we can not infer how such material should look. However, in the material used here there seems to be a disconnect between the fairly precise explanations given in the training material (i.e. for cigarettes: None – No cigarette or cigar butts or packages are visible. Little – You may see one or two items; Some – You notice more than 3 items or you see items in more than two locations. More than what you would encounter from a single careless passerby; A lot – You see 4 or more items or an item in several locations) and the rather coarse answer categories that the interviewers see when doing the rating ('none', 'very little', 'some', and 'a lot'). The reduction in scale labels could be due to reduced space on the paper-and-pencil form. As interviewers move towards hand-held devices for listing and screening, explanations of categories could be embedded into the material. The next step is to develop instruments or training protocols aimed at reducing those sources of errors. Future research along these lines might be inspired by more recent studies on interviewer observations of other elements aside from neighbourhood characteristics. For example, West (2010) and McCulloch et al. (2010) are currently examining measurement error in interviewer observations of respondent characteristics, and Sinibaldi et al. (2011) are exploring measurement error in interviewer observations of housing unit characteristics. Alternatively neighbourhood and housing unit observations could be collected by a separate set of observers, an approach more common among neighbourhood researchers. Observers can then be well trained, and asked to do all ratings in pairs. The day and time that observations are taken can also be better controlled with this approach. However, there is a trade-off between quality and cost. Asking interviewers to make those observations in addition to their primary data collection does not add additional travel cost but increases the risk of higher measurement error. Training special observers, and having them make neighbourhood and housing unit observations will increase costs but might help to reduce measurement error. Future research will need to determine the value-added of those neighbourhood observations for example when used in nonresponse adjustment.

Acknowledgments

We thank Gabriele Durrant, Stephanie Eckman, and two anonymous reviewers for their very helpful comments on earlier versions of the manuscript. We thank the L.A. FANS team for providing interviewer characteristics. Those data are not part of the regular public or the restricted L.A. FANS datasets. We gratefully acknowledge administrative support from the Eunice Kennedy Shriver National Center for Child Health and Human Development grant R24-HD041041 Maryland Population Research Center. This research was funded in part by grant R01-HL104213 from the National Heart, Lung, and Blood Institute and by a grant funded by the Economic and Social Research Council (ESRC), UK, titled 'The Use of Paradata in Cross-Sectional and Longitudinal Surveys', grant number: RES-062-23-2997. Any views expressed in this paper are those of the authors.

References

Andresen E, Malmstrom T, Miller D, Wolinsky F. Reliability and validity of observer ratings of neighborhoods. Journal of Aging and Health. 2006; 18(1):28–36. [PubMed: 16468180]

Babey S, Hastert T, Yu H, Brown E. Physical activity among adolescents: When do parks matter? American Journal of Preventive Medicine. 2008; 34:345–348. [PubMed: 18374249]

Bates N, Dahlhamer J, Singer E. Privacy concerns, too busy, or just not interested: Using doorstep concerns to predict survey nonresponse. Journal of Official Statistics. 2008; 24:591–612.

Brooks-Gunn, J.; Duncan, G.; Aber, J. Neighborhood Poverty: Vol. I: Context and Consequences for Children. New York: Russell Sage Found; 1997.

Brown B, Perkins D, Brown G. Incivilities, place attachment and crime: Block and individual effects. Journal of Environmental Policy. 2004; 24(3):359–371.

Campanelli, P.; Sturgis, P.; Purdon, S. Can You Hear me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates. London: National Centre for Social Research; 1997.

Casas-Cordero, C. Neighborhood Characteristics and Participation in Household Surveys. College Park: University of Maryland; 2010. Ph. D. thesis

Caughy M, O'Campo P, Patterson J. A brief observational measure for urban neighborhoods. Health and Place. 2001; 7(3):225–236. [PubMed: 11439257]

Clemente F, Kleiman MB. Fear of crime in the United States: A multivariate analysis. Social Forces. 1977; 56:519–531.

Cohen D, Spear S, Scribner R, Kissinger P, Mason K, Wildgen J. 'Broken Windows' and the risk of gonorrhea. American Journal of Public Health. 2000; 90(2):230–236. [PubMed: 10667184]

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20:37–46.

Couper MP, Groves RM. Social environmental impacts on survey cooperation. Quality and Quantity. 1996; 30:173–188.

Craig C, Brownson R, Cragg S, Dunn A. Exploring the effect of the environment on physical activity - A study examining walking to work. American Journal of Preventive Medicine. 2002; 23(2):36–43. [PubMed: 12133736]

Diez-Roux AV. Investigating neighborhood and area effects on health. American Journal of Public Health. 2001; 91(11):1783–1789. [PubMed: 11684601]

Dunstan F, Weaver N, Araya R, Bell T, Lannon S, Lewis G, Patterson J, Thomas H, Jones P, Palmer S. An observation tool to assist with the assessment of urban residential environments. Journal of Environmental Psychology. 2005; 25(3):293–305.

Durrant GB, Steele F. Multilevel modelling of refusal and non-contact in household surveys: evidence from six uk government surveys. Journal Of The Royal Statistical Society Series A. 2009; 172(2): 361–381.

Eifler, S.; Thume, D.; Schnell, R. Unterschiede zwischen subjektiven und objektiven messungen von zeichen öffentlicher unordnung ("signs of incivility"). In: Weich-bold, M.; Bacher, J.; Wolf, C., editors. Umfrageforschung: Herausforderungen und Grenzen (Österreichische Zeitschrift fr Soziologie Sonderhelft 9). Vol. 9. Wiesbaden: VS Verlag für Sozialwissenschaften; 2009. p. 415-441.

Franzini L, Caughy MO, Nettles SM, O'Campo P. Perceptions of disorder: Contributions of neighborhood characteristics to subjective perceptions of disorder. Journal of Environmental Psychology. 2008; 28(1):83–93.

Furr-Holden MJ, Smart JL, Pokorni NSI, Leaf PJ, Holder HD, Anthony JC. The nifety method for environmental assessment of neighborhood-level indicators of violence, alcohol, and other drug exposure. Prevention Science. 2008; 9:245–255. [PubMed: 18931911]

Garofalo J, Laub J. The fear of crime: Broadening our perspective. Victimology. 1978; 3:242–253.

Goldstein, H. Multilevel Statistical Models. 4th Edition. London: Arnold; 2010.

Groves, R.; Couper, M. Nonresponse in Household Interview Surveys. New York: Wiley; 1998.

Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly. 2006; 70(5):646–675.

Groves RM, Heeringa SG. Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society, Series A. 2006; 3:439–457.

Hintze JM. Psychometrics of direct observation. School Psychology Review. 2005; 34(4):507–519.

Johnson T, Cho Y, Campbell R, Holbrook A. Using community-level correlates to evaluate nonresponse effects in a telephone survey. Public Opinion Quarterly. 2006; 70(5):704–719.

Kalton G, Flores-Cervantes I. Weighting methods. Journal of Official Statistics. 2003; 19:81–97.

Kawachi, I.; Berkman, L. Neighborhoods and Health. New York: Oxford University Press Inc; 2003.

Kennickell AB. Analysis of nonresponse effects in the 1995 survey of consumer finances. Journal of Official Statistics. 1999; 15(2):283–303.

Kennickell AB. Reordering the Darkness: Application of Effort and Unit Nonresponse in the Survey of Consumer Finances. Proceedings of the American Statistical Association, Section on Survey Research Methods. 2003

Kish L. Studies of interviewer variance for attitudinal variables. Journal of the American Statistical Association. 1962; 57(297):92–115.

Kreuter F, Olson K, Wagner J, Yan T, Ezzati-Rice TM, Casas-Cordero C, Lemay M, Peytchev A, Groves RM, Raghunathan TE. Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. Journal Of The Royal Statistical Society Series A. 2010; 173(2):389–407.

Laraia BA, Messer L, Kaufman JS, Dole N, Caughy M, O'Campo P, Savitz DA. Direct observation of neighborhood attributes in an urban area of the U.S. south: characterizing the social context of pregnancy. International Journal of Health Geographics. 2006; 5:11. [PubMed: 16545132]

Lee G. Social integration and fear of crime among older persons. Journal of Gerontology. 1983; 6:745–750. [PubMed: 6630912]

Little R. Survey nonresponse adjustments for estimates of means. International Statistical Review. 1986; 54:139–157.

Little R, Vartivarian S. Does weighting for nonresponse increase the variance of survey means? Survey Methodology. 2005; 31:161–168.

Little RJ, Vartivarian S. On weighting the rates in non-response weights. Statistics in Medicine. 2003; 22(9):1589–1599. [PubMed: 12704617]

Long, LS. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publication; 1997.

Lynn P. PEDAKSI: Methodology for collecting data about survey non-respondents. Quality and Quantity. 2003; 37:239–261.

Lynn, P.; Clarke, P.; Martin, J.; Sturgis, P. The effects of extended interviewer efforts on nonresponse bias. In: Groves, R.; Dillman, D.; Eltinge, J.; Little, R., editors. Survey Nonresponse. New York: Wiley; 2002.

Maitland, A.; Casas-Cordero, C.; Kreuter, F. An evaluation of nonresponse bias using paradata from a health survey. Washington, D.C: Proceedings of the Section on Government Statistics, Joint Statistical Meetings; 2009.

McCulloch, S.; Kreuter, F.; Calvano, S. In Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research. Chicago, IL: 2010. Interviewer observed vs. reported respondent gender: Implications on measurement error.

Morenoff J. Neighborhood mechanisms and the spatial dynamics of birth weight. American Journal of Sociology. 2003; 108(5):976–1017.

Mujahid MS, Roux AVD, Morenoff JD, Raghunathan T. Assessing the measurement properties of neighborhood scales: From psychometrics to ecometrics. American Journal of Epidemiology. 2007; 165(8):858–867. [PubMed: 17329713]

O'Muircheartaigh C, Campanelli P. A multilevel exploration of the role of interviewers in survey non-response. Journal of the Royal Statistical Society, Series A. 1999; 162(Part 3):437–446.

Peterson CE, Sastry N, Pebley AR. The Los Angeles Family and Neighborhood Survey: Neighborhood Observations Codebook. Technical Report WR-240/13-LAFANS, Labor and Population Program, RAND Corporation. 2007

Rabe-Hesketh, S.; Skrondal, A. Multilevel and Longitudinal Modeling using Stata, Third Edition, Volume II: Categorical Responses, Counts, and Survival. College Station, TX: Stata Press; 2012.

Raudenbush, S. The quantitative assessment of neighborhood social environment. In: Kawachi, I.; Berkman, L., editors. Neighborhoods and Health. New York, NY: Oxford University Press; 2003. p. 112-131.

Raudenbush S, Sampson RJ. "Ecometrics": Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. Sociological Methodology. 1999; 29:1–41.

Sampson R, Morenoff J, Earls F. Beyond social capital: Spatial dynamics of collective efficacy for children. American Sociological Review. 1999; 64(5):633–660.

Sampson R, Morenoff J, Gannon-Rowley T. Assessing "neighborhood effects": Social processes and new directions in research. Annual Review of Sociology. 2002; 28:443–478.

Sampson R, Raudenbush S. Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. American Journal of Sociology. 1999; 105(3):603–651.

Sampson R, Raudenbush S. Seeing disorder: Neighborhood stigma and the social construction of "Broken windows". Social Psychology Quarterly. 2004; 67(4):319–342.

Sastry N, Ghosh-Dastidar B, Adams J, Pebley A. The design of a multilevel longitudinal study of children, families and communities: the Los Angeles Family and Neighborhood Study. Labor and Population Program, RAND Corporation (DRU-2400/1-LAFANS). 2003

Sastry N, Pebley A. The Los Angeles Family and Neighborhood Survey: Neighborhood observation forms and interviewer manual. Technical Report DRU-2400/6-LAFANS, Labor and Population Program, RAND Corporation. 2003a

Sastry N, Pebley A. Non-response in the Los Angeles Family and Neighborhood Study. Technical Report DRU-2400/7-LAFANS, Labor and Population Program, RAND Corporation. 2003b

Schluchter MD, Elashoff JT. Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structure. Journal of Statistical Computation and Simulation. 1990; 37:69–87.

Schnell R, Kreuter F. Separating interviewer and sampling-point effects. Journal of Official Statistics. 2005; 21(3):389–410.

Schräpler J-P, Schupp J, Wagner GG. Individual and neighborhood determinants of survey nonresponse. Technical report, SOEP papers on Multidisciplinary Panel Data Research, Deutsches Institut for Wirtschaftsforschung (DIW Berlin). 2010

Sinibaldi J, Durrant G, Kreuter F. Evaluating the measurement error of interviewer observed paradata: Evidence from six uk surveys. Working Paper. 2011

Snijders, T.; Bosker, R. Multilevel Analysis. An introduction to basic and advanced multilevel modeling. Thousand Oaks, CA: Sage; 1999.

Taylor RB, Gottfredson SD, Brower S. Neighborhood naming as an index of attachment to place. Population and Environment. 1984; 7:101–111.

Taylor RB, Koons BA, Kurtz EM, Greene JR, Perkins DD. Streetblocks with more nonresidential land use have more physical deterioration: Evidence from baltimore and philadelphia. Urban Affairs Review. 1995; 31:120–136.

Vincenty T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review. 1975; 22(176):88–93.

Weich S, Burton E, Blanchard M, Prince M, Sproston K, Erens B. Measuring the built environment: Validity of a site survey instrument for use in urban settings. Health and Place. 2001; 7:283–292. [PubMed: 11682328]

West, B. An examination of the quality and utility of interviewer estimates of household characteristics in the national survey of family growth; Chicago, IL. In Paper presented at the 2010 Annual Meeting of the American Association for Public Opinion Research; 2010.

Wolfinger R. Laplace's approximation for nonlinear mixed models. Biometrika. 1993; 80:791–795.

Wyant BR. Multilevel impacts of perceived incivilities and perceptions of crime risk on fear of crime: Isolating endogenous impacts. Journal of Research in Crime and Delinquency. 2008; 45:39–64.

Zenk SN, Schulz AJ, Mentz G, House JS, Gravlee CC, Miranda PY, Miller P, Kannan S. Inter-rater and test-retest reliability: Methods and results for the neighborhood observational checklist. Health and Place. 2007; 13(2):452–465. [PubMed: 16809060]

6. Appendix

Appendix Table 1

Descriptive Statistics for Tract Characteristics Used to Construct Neighbourhood Structural Characteristics Factors (unweighted estimates)

Variable	Tracts	Mean	Std Dev	Minimum	Maximum
Percent foreign-born	65	40.22	15.35	11.51	74.89
Percent non-citizens	65	27.23	14.83	3.72	61.5
Percent hispanic-latino	65	55.10	29.57	1.84	96.12
Percent adults spanish speakers	65	47.20	27.65	2.84	88.71
Percent of individuals in poverty	65	22.95	13.88	2.72	51.03
Percent HH receiving public assistance	65	9.26	7.41	0.00	31.10
Percent unemployed in civilian labor force	65	5.59	3.39	0.41	24.06
Percent HH headed by females with children	65	10.52	5.96	1.48	24.96
Percent non-hispanic black	65	8.30	9.98	0.00	48.06
Percent of families with income $>= 75k$	65	22.20	20.19	1.85	78.66
Percent adults 25+ yrs with 13+ yrs school	65	35.41	23.76	5.97	87.00
Percent workers in exec/prof occupations	65	26.70	18.36	7.12	73.40
People per square mile	65	14,836.44	10,461.92	64.78	44,790.46

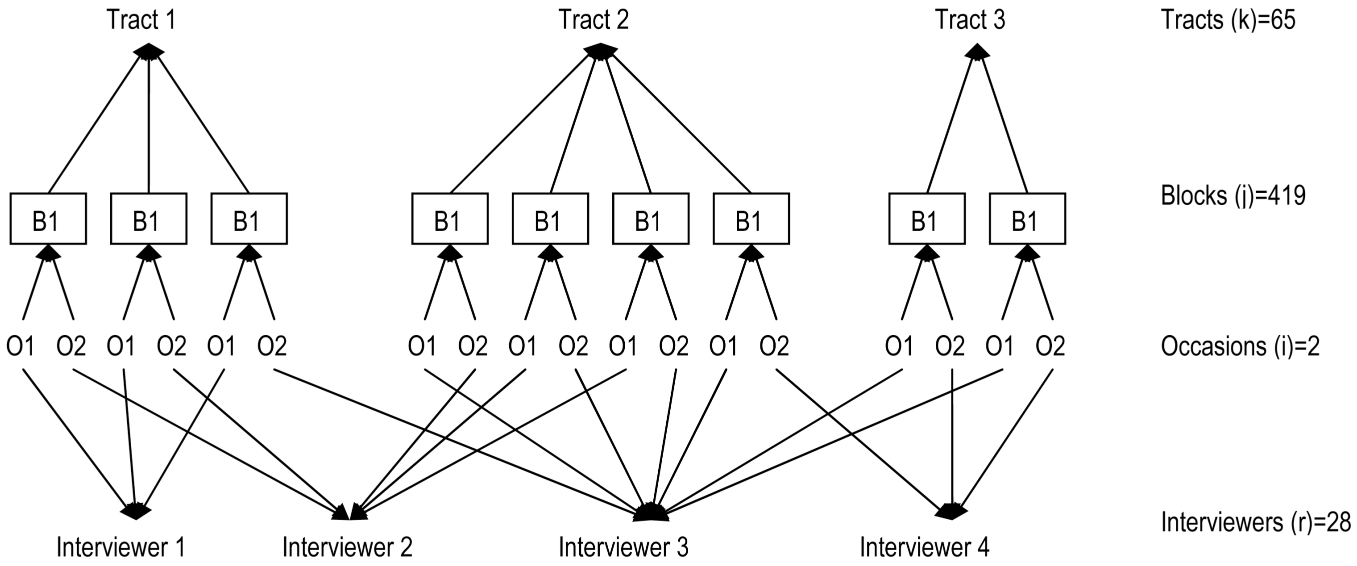

Fig. 1.

Schematic diagram of the L.A. FANS data structure

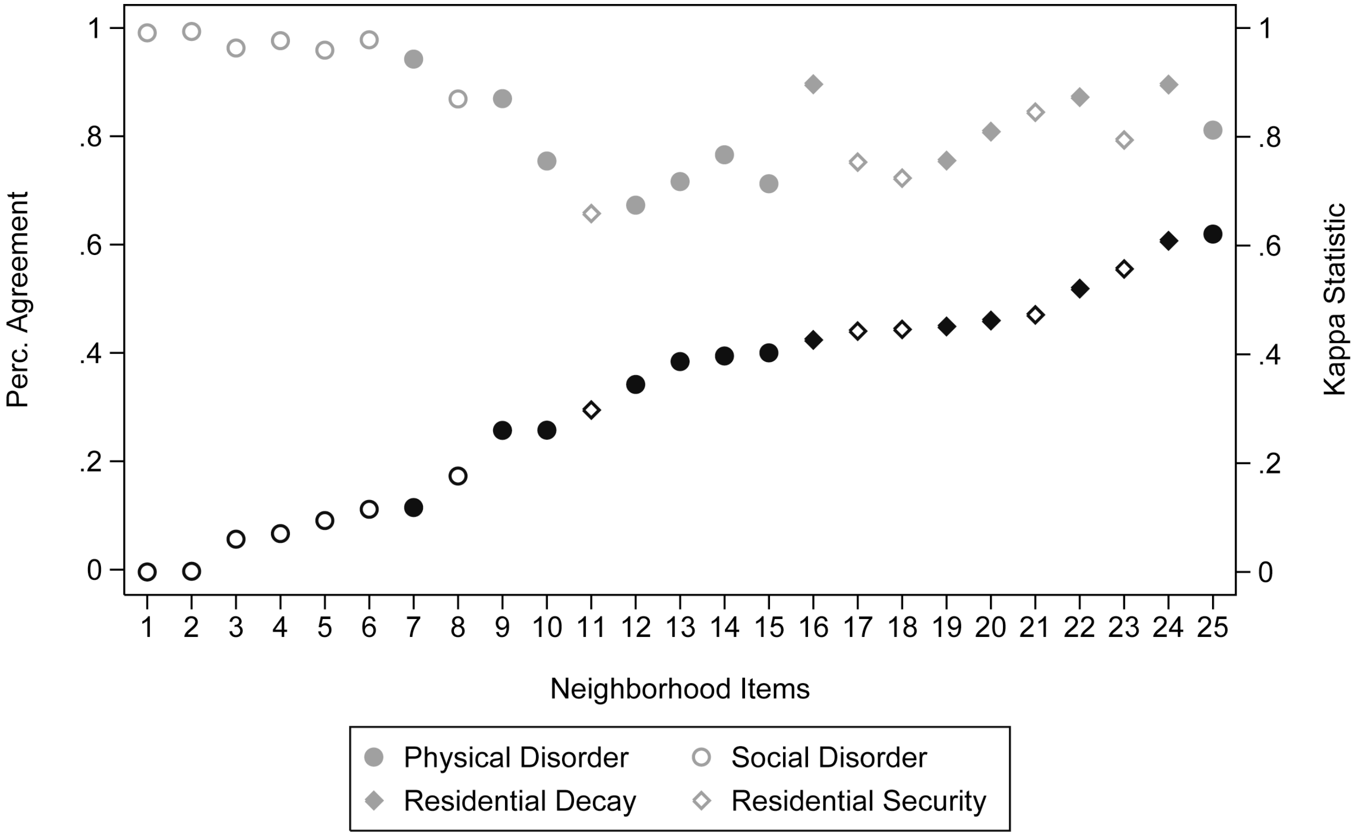

Fig. 2.

Estimates of Interobserver Agreement

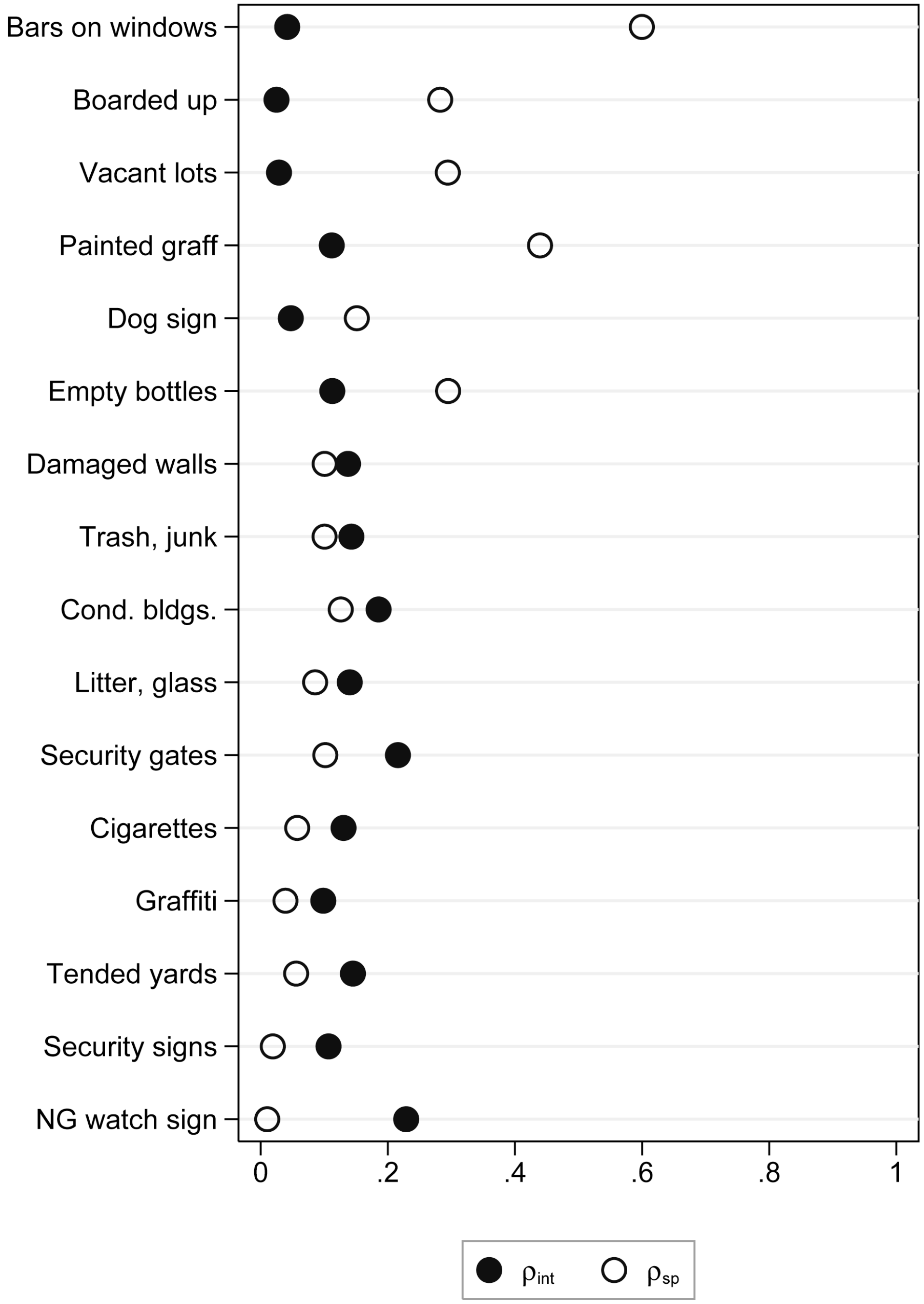

Fig. 3.
Estimates of Intraclass Correlation from the Unconditional Logistic Models (unweighted) Intraclass Correlations

Table 1

Structure and frequencies of L.A.FANS observation data. Unweighted estimates.

Level	Analytic Cases	Original Data
Number of data records	3,998	5,966
Number of repeated observations per item	2	2–6
Number of unique Interviewers	28	35
Number of unique Tracts	65	65
Number of unique Blocks	419	422
Number of unique Block Faces	1,999	2,029

Table 2

Percentage distribution of physical and social disorder items (unweighted)

Neighbourhood Items	Label	Perc.	n
Physical Disorder Items (n=8)		%	
Are there abandoned cars on the street or in alleys or lots?	cars	9.8	3,998
Is there trash or junk on the street or sidewalks, in yards/lots?	trash	52.7	3,998
Is there garbage, litter, or broken glass on the street or sidewalk, in yards, or vacant lots?	litter	73.6	3,998
Are there needles, syringes, condoms, or drug-related paraphernalia on the street or sidewalk, in yards/lots?	drugs	3.4	3,998
Are there empty beer containers or liquor bottles on the street or sidewalks, in yards, or vacant lots?	bottles	21.0	3,998
Are there cigarettes or cigar butts or discarded cigarette packages on the street or sidewalks, in yards/lots or gutters?	cigars	59.7	3,998
Is there graffiti on buildings, sidewalks, walls, or signs?	graffiti	53.5	3,998
Is there painted-over graffiti on buildings, sidewalks, walls, or signs?	pograff	36.3	3,998
Social Disorder Items (n=7)			
Did any of the groups of teens you saw appear to be a gang?	gang	1.2	3,962
Did you see any adults on the block face loitering, congregating or hanging out?	loitering	8.6	3,982
Did you see any prostitutes on the block face?	prostit	0.3	3,986
Did you see any homeless people or people begging on the block face?	homeless	2.0	3,996
Did you see people who were selling illegal drugs on the block face?	selling	0.5	3,996
Did you see any people drinking alcohol openly on the block face?	drinking	2.3	3,996
Did you see any drunken or otherwise intoxicated people on the block face?	intox	1.3	3,996

Table 3

Percentage distribution of residential decay and residential security items (unweighted)

Neighbourhood Items	Label	Perc.	n
Residential Decay Items (n=5)		%	
What is the overall condition of the residential buildings?	bldgs	84.6	3,627
How many houses/apartments are burned out, boarded	boarded	10.0	3,627
How many vacant lots are there on this block?	vacant	16.0	3,627
How many houses/apartments have peeling paint or damaged exterior walls?	walls	66.7	3,627
How many houses/apartments have well-tended yards or gardens?	yards	77.1	3,627
Residential Security Items (n=5)			
How many houses/apartments have window bars or gratings on doors or windows?	barswin	63.8	3,627
How many houses/apartments have signs indicating they are protected by private security services?	secsign	52.2	3,627
How many houses/apartments have signs indicating they are protected by dogs?	dogsign	32.4	3,627
How many houses/apartments have security gates or security fences?	gates	59.2	3,627
Are there signs indicating there is a neighbourhood watch on this block?	ngwatch	17.7	3,618

Table 4

Percentage distribution of interviewer characteristics (unweighted)

Indicators of		Interviewer Data (n=28)	Block Face Data (n=3,998)
Potential Exposure		%	%
White	No	65.4	48.6
	Yes	34.6	51.4
Vulnerability			
55+ yrs	No	92.0	84.4
	Yes	8.0	15.6
Community Involvement			
Ever Married	No	57.7	36.0
	Yes	42.3	64.0
Has Kids	No	65.4	53.4
	Yes	34.6	46.6
Com. Activ.	No	46.2	54.9
	Yes	53.8	45.1

Table 5

Characteristics of the Measurement Occasion (unweighted)

Indicators of		Block Data (n=419)
Familiarity with Area		%
Neighbourhood Close	No	87.3
	Yes	12.7
Experience with Block	No	22.9
	Yes	77.1
Temporal Variability		
Rated after 5pm	No	88.7
	Yes	11.3
Rated on Weekend	No	74.2
	Yes	25.8

Table 6

Cross-classified multilevel logistic regression models for observing trash and grafitti (unweighted).

	Trash				Graffiti			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Fixed Effects								
Interviewer Characteristics								
Race White		0.69 (0.99)	0.70 (0.99)	0.77 (0.91)		1.53 * (0.66)	1.47 * (0.66)	1.23 * (0.62)
Age 55+ yrs		−1.06 (1.28)	−1.07 (1.29)	−0.91 (1.18)		−2.47 * (0.80)	−2.39 * (0.80)	−1.89 * (0.76)
Has Kids		1.15 (0.88)	1.18 (0.88)	0.92 (0.81)		0.93 (0.61)	0.87 (0.61)	0.44 (0.58)
Comm. Involvement		−0.32 (0.65)	−0.31 (0.66)	−0.39 (0.60)		0.18 (0.41)	0.16 (0.41)	0.27 (0.40)
Ever Married		−1.37 (1.04)	−1.35 (1.05)	−1.05 (0.96)		−0.91 (0.71)	−0.89 (0.71)	−0.50 (0.67)
Occasion of Measurement								
Prior Experience			−0.04 (0.21)	−0.20 (0.21)			−0.23 (0.26)	−0.42 (0.26)
Neighbourhood close			0.10 (0.36)	0.22 (0.31)			−0.39 (0.41)	−0.08 (0.36)
Rated after 5pm			0.12 (0.23)	0.16 (0.23)			0.02 (0.31)	0.17 (0.31)
Rated on Weekend			0.14 (0.20)	0.00 (0.19)			−0.09 (0.25)	−0.16 (0.24)
Neighbourhood Characteristics								
Immigrant				0.51 * (0.13)				0.95 * (0.17)
Disadvantage				0.63 * (0.11)				1.19 * (0.16)
Affluence				−0.88 * (0.13)				−1.85 * (0.17)
Pop. Density				−0.05 (0.15)				0.38 (0.20)
Intercept	0.67 (0.37)	0.88 (0.54)	0.82 (0.58)	0.65 (0.52)	0.76 (0.44)	0.42 (0.50)	0.73 (0.56)	0.28 (0.42)
Variance Components								
Interviewers γ_o^2	2.25 * (0.89)	1.90 * (0.77)	1.92 * (0.79)	1.58 * (0.65)	1.03 * (0.44)	0.53 * (0.28)	0.52 * (0.28)	0.49 * (0.26)
Tracts γ_t^2	1.48 * (0.38)	1.47 * (0.38)	1.47 * (0.39)	0.10 * (0.13)	8.14 * (1.80)	8.27 * (1.82)	8.28 * (1.84)	0.32 * (0.21)
Blocks γ_b^2	0.35 * (0.14)	0.35 * (0.14)	0.35 * (0.14)	0.36 * (0.15)	0.82 * (0.25)	0.83 * (0.26)	0.80 * (0.25)	0.98 * (0.29)
Int & Tracts γ_I^2	0.82 * (0.23)	0.83 * (0.23)	0.82 * (0.23)	0.86 * (0.24)	0.54 * (0.23)	0.54 * (0.24)	0.56 * (0.24)	0.67 * (0.27)

	Trash				Graffiti			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Summary Statistics								
n	2892	2892	2892	2892	2892	2892	2892	2892
Log-likelihood	−1437.71	−1436.31	−1435.86	−1400.68	−1092.59	−1087.62	−1086.74	−1023.62

* $p < 0.05$

Table 7

Cross-classified multilevel logistic regression models for selected neighbourhood items (unweighted).

	Physical Disorder Items				Res. Decay Items			Res. Security Items		
	trash	litter	cigars	graffiti	bldgs	walls	yards	secsign	gates	ngwatch
Fixed Effects										
Interviewer Characteristics										
Race White	0.77 (0.91)	0.06 (0.85)	0.25 (1.30)	1.23* (0.62)	−0.03 (1.54)	0.27 (0.43)	0.02 (0.91)	−1.02* (0.45)	0.17 (0.96)	−0.24 (0.39)
Age 55+ yrs	−0.91 (1.18)	−0.84 (1.08)	1.25 (1.84)	−1.89* (0.76)	0.87 (1.87)	−2.26* (0.46)	−1.60 (1.06)	−0.45 (0.56)	2.15 (1.23)	−0.49 (0.43)
Has Kids	0.92 (0.81)	−0.48 (0.77)	0.61 (1.20)	0.44 (0.58)	0.98 (1.56)	1.10* (0.39)	1.39 (0.88)	−0.02 (0.41)	−1.54 (0.86)	0.50 (0.36)
Comm. Involvement	−0.39 (0.60)	0.73 (0.57)	0.19 (0.84)	0.27 (0.40)	1.45 (1.06)	0.25 (0.29)	0.04 (0.58)	−0.05 (0.30)	1.45* (0.63)	0.13 (0.24)
Ever Married	−1.05 (0.96)	1.48 (0.91)	0.82 (1.38)	−0.50 (0.67)	0.02 (1.77)	0.59 (0.47)	−0.63 (1.03)	0.99* (0.48)	−0.04 (1.01)	−0.06 (0.41)
Occasion of Measurement										
Prior Experience	−0.20 (0.21)	0.38 (0.25)	0.13 (0.24)	−0.42 (0.26)	0.04 (0.46)	−0.28 (0.23)	0.14 (0.31)	0.11 (0.19)	−0.39 (0.25)	−0.25 (0.21)
Neighbourhood close	0.22 (0.31)	−0.26 (0.37)	−0.39 (0.38)	−0.08 (0.36)	0.06 (0.73)	−0.45 (0.35)	0.87 (0.47)	0.31 (0.28)	−0.12 (0.41)	−0.12 (0.31)
Rated after 5pm	0.16 (0.23)	0.56 (0.34)	0.68* (0.30)	0.17 (0.31)	0.57 (0.55)	0.22 (0.27)	−0.34 (0.35)	−0.10 (0.22)	0.12 (0.30)	−0.06 (0.27)
Rated on Weekend	0.00 (0.19)	0.19 (0.26)	0.23 (0.23)	−0.16 (0.24)	−0.99* (0.46)	0.06 (0.22)	0.20 (0.30)	0.11 (0.17)	0.23 (0.26)	−0.12 (0.20)
Neighbourhood Composition										
Immigrant Conc.	0.51* (0.13)	0.87* (0.17)	0.76* (0.14)	0.95* (0.17)	0.93* (0.42)	0.49* (0.18)	1.15* (0.25)	−0.37* (0.14)	0.55* (0.20)	−0.43* (0.17)
Cone. Disadvantage	0.63* (0.11)	1.00* (0.18)	0.91* (0.15)	1.19* (0.16)	3.10* (0.64)	0.78* (0.16)	1.44* (0.25)	−0.42* (0.12)	0.55* (0.17)	−0.27 (0.14)
Conc. Affluence	−0.88* (0.13)	−1.23* (0.16)	−1.05* (0.00)	−1.85* (0.17)	−3.11* (0.42)	−1.31* (0.17)	−1.51* (0.23)	1.11* (0.14)	−0.88* (0.19)	0.11 (0.16)
Pop. Density	−0.05 (0.15)	0.23 (0.22)	0.42* (0.18)	0.38 (0.20)	−0.56 (0.54)	0.05 (0.21)	−0.33 (0.30)	−0.03 (0.16)	0.16 (0.23)	0.01 (0.19)
Intercept	0.65 (0.52)	0.64 (0.51)	−0.45 (0.74)	0.28 (0.42)	5.27* (1.04)	0.64 (0.32)	2.65* (0.57)	−0.18 (0.31)	1.04 (0.57)	−1.74* (0.31)
Variance Components										
Interviewers γ_{o}^{2}	1.58* (0.65)	1.24* (0.50)	3.36* (1.28)	0.49* (0.26)	3.11* (1.47)	0.06* (0.11)	0.99* (0.49)	0.28 (0.15)	1.62* (0.65)	0.11 (0.11)
Tracts γ_{t}^{2}	0.10* (0.13)	0.06* (0.20)	0.00* (0.24)	0.32* (0.21)	1.28* (0.73)	0.58* (0.24)	0.83* (0.45)	0.44* (0.16)	0.68* (0.33)	0.49* (0.21)
Blocks γ_{b}^{2}	0.36* (0.15)	0.65* (0.22)	1.00* (0.26)	0.98* (0.29)	1.82* (0.70)	0.32* (0.17)	1.33* (0.39)	0.57* (0.15)	1.25* (0.31)	1.20* (0.27)
Int & Tracts γ_{I}^{2}	0.86* (0.24)	1.13* (0.32)	0.87* (0.24)	0.67* (0.27)	1.54* (0.82)	0.95* (0.27)	0.79* (0.36)	0.16* (0.11)	1.08* (0.34)	0.09* (0.13)

	Physical Disorder Items				Res. Decay Items			Res. Security Items		
	trash	**litter**	**cigars**	**graffiti**	**bldgs**	**walls**	**yards**	**secsign**	**gates**	**ngwatch**
Summary Statistics										
n	2892	2892	2892	2892	2631	2631	2631	2631	2631	2623
Log-likelihood	−1400.68	−1092.72	−1222.77	−1023.62	−515.92	−1082.85	−870.89	−1443.03	−1198.17	−1137.04

* $p < 0.05$