

**Econometrics in the Age of Big Data:  
Measuring and Assessing “Broken Windows” Using Large-scale Administrative Records**

Daniel O’Brien, Robert J. Sampson and Christopher Winship

Harvard University

We thank the National Science Foundation (Grant # SMA 1338446), the John D. and Catherine T. MacArthur Foundation (Grant # 13-105766-000-USP), and the Radcliffe Institute for Advanced Study for funding assistance. We also thank the City of Boston’s “Office of New Urban Mechanics” and “Department of Innovation and Technology” for supporting our examination of government data, and the Editor and reviewers of *Sociological Methodology* for helpful comments on earlier drafts. A version of this paper was presented at the annual meeting of the American Association for the Advancement of Science in Chicago, February 15<sup>th</sup>, 2014.

## **Econometrics in the Age of Big Data:**

### **Measuring and Assessing “Broken Windows” Using Large-scale Administrative Records**

#### **Abstract:**

The collection of large-scale administrative records in electronic form by many cities provides a new opportunity for the measurement and longitudinal tracking of neighborhood characteristics, but one that will require novel methodologies that convert such data into research-relevant measures. The current paper illustrates these challenges by developing measures of “broken windows” from Boston’s “Constituent Relationship Management” (CRM) system (aka 311 hotline). A sixteen-month archive of the CRM database contains more than 300,000 address-based requests for city services, many of which reference physical incivilities (e.g., graffiti removal). We carry out three econometric analyses, each building off the previous. Analysis 1 examines the *content* of the measure, identifying 28 items that constitute two independent constructs, *private neglect* and *public denigration*. Analysis 2 assesses the *validity* of the measure by examining the “civic response rate” across neighborhoods using investigator-initiated neighborhood audits. Indicators of civic response were then extracted from the CRM database, so that measurement adjustments could be automated. These adjustments were calibrated against measures of litter from the objective audits. Analysis 3 examines the *reliability* of the composite measure of physical disorder at different spatio-temporal windows, finding that census tracts can be measured at two-month intervals, and census block groups at six-month intervals. The final measures are highly detailed, can be tracked longitudinally, and are virtually costless. Our framework thus provides an example of how new forms of large-scale administrative data can yield econometric measurement for urban science while at the same time illustrating the methodological challenges that must be addressed.

**Keywords:** Econometrics; urban sociology; “big data”; computational social science; physical disorder; broken windows; 311 hotlines

The global move towards digital technology has instigated a marked shift in the practice of science over the last two decades. Surveys and experiments are now often conducted through internet platforms; GPS devices and other sensors allow us to track patterns of movement and behavior; and computer processing technology has supported the development of new forms of statistical analysis. A recent consequence of this “digital revolution” is the availability of large-scale administrative data that might prove useful in research. Many public agencies and private companies systematically collect information on services and clients and compile it in digital databases. Some of these are more detailed versions of familiar data, like crime reports, while others, like cell phone records or citizen requests for governmental services, are novel. These “big” or *next-generation data* offer the opportunity to paint a comprehensive picture of cities, which has the potential to transform theoretical models of urban governance and social behavior (Lazer et al. 2009).

Despite considerable excitement at this prospect, big data have not yet become commonplace in contemporary social science research, in part, it seems, because researchers do not entirely know what to make of them. Without a clear understanding of how these new data sources contribute to our ongoing debates and the questions facing our fields, some might reasonably consider their promise as being overblown. There is thus a need for methodologies that can connect big data with the current practice of social science.

We offer one such “proof of concept” in the present paper, utilizing a database of over 300,000 citizen-generated requests for public services in Boston, MA to measure the conditions of urban neighborhoods across space and time. Building on the methodology of econometrics (Raudenbush and Sampson 1999a), we construct and assess a measure of physical disorder, one of the most widely used and popular concepts in urban sociology, criminology, and public

policy. Although the idea of disorder has a long history in sociology, it has received increased attention in recent decades owing to the influential “broken windows” theory of crime and urban decline (Raudenbush and Sampson 1999b; Ross et al. 2001; Skogan 1992; Wilson and Kelling 1982), making it an ideal test case for assessing the potential for econometrics based on large-scale administrative data.

## **1. AN ECONOMETRIC APPROACH TO DISORDER**

Almost fifteen years ago, Raudenbush and Sampson (1999) proposed a systematic approach to the measurement of neighborhood social ecology, what they termed “econometrics.” They encouraged researchers to borrow three tools developed by psychometricians for the measurement of behavior: 1) item-response models, which call for the use of scales whose multiple items vary in their difficulty, allowing for greater precision in measurement across neighborhoods; 2) factor analysis, in order to address the interrelation between items, and to identify one or a few latent constructs that the items reflect; and 3) generalizability theory, which requires criteria for ensuring that a given measurement of a neighborhood is reflective of the “true” score on the characteristic of interest, and not overly influenced by either stochastic or confounding processes. These guidelines, along with the illustrative examples that accompanied them, provided researchers with a step-by-step methodology for developing survey and observational protocols that could measure econometrics, one that has been implemented by hundreds of researchers in dozens of cities.

The advent of large administrative data represents a new opportunity for econometric study. The giga- and terabytes of data being collected by both public and private sector entities are a rich, low-cost resource for measuring the characteristics of neighborhoods, but using them in this

manner poses clear methodological and substantive challenges. Administrative data are not collected according to any research question or plan, and thus, in their raw state, lack some of the characteristics expected of researcher-collected data. These challenges are well suited to the techniques common to econometric study, which can act as a guide to both what is missing or occluded in such data sets, as well as how a researcher might address such issues.

We focus here on one of the most influential concepts in the urban sciences—that of *physical disorder*, including the iconic “broken window,” the accumulation of litter, the presence of graffiti, or other indications that a neighborhood is poorly maintained and monitored. Such incivilities are often associated with elevated crime rates (Raudenbush and Sampson 1999b; Wilson and Kelling 1982) and lower mental, physical and behavioral health among residents (Burdette and Hill 2008; Caughy et al. 2008; Furr-Holden et al. 2012; Mujahid et al. 2008; O'Brien and Kauffman 2013; Wen et al. 2006), attracting attention from a variety of disciplines (Caughy et al. 2001; Cohen et al. 2000; O'Brien and Wilson 2011; Taylor 2001). Physical disorder's importance as a neighborhood characteristic is such that it was also one of the two test cases that Raudenbush and Sampson (1999) used to illustrate their methodological approach to econometrics.

Physical disorder is traditionally measured either through surveys or detailed neighborhood audits (e.g., Taylor, 2001; Sampson and Raudenbush 1999), but the effort and cost associated with such protocols have made whole-city assessments challenging and precise longitudinal tracking nearly impossible. Modern technology used by city agencies, however, is now recording similar information in real time. These databases have the potential to supplement traditional econometric protocols. One such database is a result of a recent policy innovation called the constituent relationship management (CRM) system. Colloquially known as a Mayor's

Hotline or 311 line, these systems provide constituents with a variety of channels for directly requesting services from the city government, using phone, internet or smart phone applications that communicate requests to the appropriate department. The resultant database is a detailed documentation of constituent needs, leading some of its initial implementers to refer to it as “the eyes and ears of the city”—Jane Jacobs (1961) meets “big data” as it were. Many of the requests refer to individual instances of physical disorder, like graffiti or abandoned housing, giving the database the ability to reflect their prevalence across neighborhoods.

Although the potential impacts of big data on science have been over-hyped (Pigliucci 2009) and there have been highly visible failures of prediction based on large-scale data (Lazer et al. 2014), the CRM database offers a number of possibilities as an alternative or supplement to expensive new data collection—especially in a time of declining research support. For one, the system receives hundreds of cases every day, each attributed to a particular address or intersection, giving researchers considerable flexibility in how they might geographically divide the city. It also lends itself to the longitudinal tracking of physical disorder, a major advance considering that no whole-city protocol to date has been conducted more than once in a five-year period. Further, the database differentiates between dozens of case types, allowing greater precision in defining the events that comprise disorder than has previously been possible.

The CRM database was not created for the purposes of disorder research, however, and has three weaknesses that any methodology must address. First, the substantive content of the database is noisy, and it is not immediately apparent what it can measure nor how it can do so. Some cases, like requests for graffiti removal, are clear examples of physical disorder, but others, like scheduling a bulk item pickup, are not. Second, there may be some aspect of data collection that creates systematic biases in measurement. For instance and quite importantly, the CRM

system may suffer from skewed reporting in the incidence of disorder across neighborhoods. Last, there is no information about what scale of geographical analysis the database can support—for example, census block groups or tracts—nor over what time spans.

Whereas Raudenbush and Sampson (1999) forwarded criteria for survey- and observation-based measurement across geographical units, the CRM database makes clear the need for a new set of guidelines for the utilization of administrative data in the creation of ecometrics. The current manuscript uses the CRM database from Boston to illustrate the multiple analytic steps in the formulation of original ecometrics. This process is reported in three parts, each requiring its own distinct logic, data sources, and analytical approach. Analysis 1 examines the *content* within CRM database that reflects physical disorder, and uses correlational analyses to identify an underlying factor structure. Analysis 2 then addresses the *validity* of any measure extracted from the CRM database by assessing biases in reporting through original data collection involving neighborhood audits. A method is then developed for using auxiliary measures from within the CRM database to estimate these biases, and to help account for over- or underreporting. Analysis 3 then examines the *reliability* of these composite measures by identifying the spatial and temporal ranges at which their measurement is consistent. In each case we spell out the necessary assumptions in our analysis and that are inherent to big data.

## **2. ANALYSIS 1: OPERATIONALIZING PHYSICAL DISORDER**

When Raudenbush and Sampson (1999a) developed their methodology for ecometrics, they emphasized the development of item-response models, and their examination through factor analysis, an approach that had been in common use in the field of psychometrics for decades. When conducting a neighborhood audit, for example, a protocol might measure a variety of

items that collectively capture an overall pattern. A factor analysis based on their intercorrelations would then help determine which of these items in fact measured the desired construct, while also testing whether they reflected one or multiple constructs regarding the neighborhood's ecology. The challenge with next-generation data, however, is that it is not immediately apparent what they *can* measure. Traditionally, research measures are derived from protocols written by the researchers themselves, and their items are based on an underlying theoretical construct. Administrative data are not endowed with an *a priori* theoretical organization of this sort. The CRM database, for example, is a byproduct of a system intended to transmit the needs of constituents to the appropriate government agencies, and its organization reflects this function, rather than a deliberate intent to measure neighborhood characteristics. Nonetheless, with thousands of requests spanning over 150 case types, the CRM offers a rich store of information for measuring neighborhood characteristics. But before factor analysis can be considered, it falls to researchers to use existing theory to identify those specific items that are likely to be relevant.

Physical disorder is typically defined as any aspect of a neighborhood's visual cues that reflect a "breakdown of the local social order" (pg. 2, Skogan 1992), though this has come to mean two different things in practice. Raudenbush and Sampson's (1999) measure focused specifically on the artifacts of physical incivilities that were publicly visible and that denigrated the public space according to broken windows theory, such as graffiti and various forms of litter indicating illegal or typically problematic behavior (e.g., used condoms, empty beer bottles, hypodermic needles). A variety of other researchers have expanded this definition to include any item that might be evidence that "spaces are not being kept or used properly" (pg. 5, Taylor, 2001). This had led to a variety of protocols that also include items that, while not the result of



flagrant incivilities, reflect an overall pattern of neglect, including deteriorating or abandoned housing, unkempt lawns or vegetation, and litters of all kinds (Caughy et al. 2001; Cohen et al. 2000; Furr-Holden et al. 2008; O'Brien and Wilson 2011; Ross and Mirowsky 1999; Rundle et al. 2011; Skogan 1992; Taylor 2001). One important consequence of this approach is that it extends measurement to elements of the neighborhood that are technically private, but whose appearance and use are a visible part of the local scenery, like front porches, lawns, and the facades of houses. Despite this distinction, factor analyses on such protocols often identify a single factor, though Ross and Mirowsky (1999) found evidence for two latent constructs they referred to as disorder and decay, approximating the dichotomy described here.

In order to make the greatest use of the CRM database, this study will identify case types that reflect either private neglect or public denigration. Some will correspond directly to items in previous methodologies, like a report of an abandoned house, or a request for graffiti removal. But others will be novel, either because they are too uncommon to be measured through one-time neighborhood audits (e.g., cars illegally parked on a lawn), or because they are more likely to be experienced in private spaces, like a rodent infestation. This latter opportunity to “look” at the conditions inside houses could potentially add a new dimension to the measurement of disorder, one that has been hinted at in previous protocols that examine visible deterioration, but has not been completely accessible. Altogether, it is possible to construct a battery of “items” that offers a greater breadth and depth than any previous measure of physical disorder. The second stage of the analysis will then use factor analysis to explore the dimensionality of these items. Given their large number, it seems feasible that they will not describe a unitary construct, but one with multiple aspects that are related but distinct.

## 2.1. *The CRM Database*

Boston's CRM system received 365,729 requests for service via its three channels (hotline calls, internet self-service portal and smartphone application) between March 1, 2010 and June 29, 2012, 334,874 of which had a geographic reference. March 1<sup>st</sup> was chosen as the start date because that is when a standardized data entry form was implemented.

The requests for service included 178 different case types. A subset of types reflected examples of physical disorder arising from either human negligence or denigration of the neighborhood (e.g., illegal dumping, abandoned bicycle). Other case types either did not indicate physical disorder (e.g., general request, bulk item pickup) or indicated deterioration that was not the fault of local residents (e.g., street light outage).

Each case record included the date of the request, the address or intersection where services were to be rendered, as well as the case type. These locations came from a master geographical database of the addresses and intersections of Boston based on the City's tax assessor and roads data, with each address keyed to the appropriate census geographies (from the 2005-2009 ACS, the most recent census with socioeconomic data when the database was built). The main measures for this analysis were counts of events that occurred in a neighborhood, which we operationalize as the census block group (CBG). CBGs are smaller than the more typically used census tract (avg. population  $\approx$  1,000 vs. 4,000), but the volume of CRM calls enables measurement and analysis at this finer scale. Boston contains 543 CBGs with a substantial population. All CRM reports document an event at a parcel or intersection, and these are each attributed to a CBG. From this, neighborhood-level counts for all case types can be calculated.

## 2.2. Defining Physical Disorder from Case Types

An initial examination of the 178 case types produced a list of 33 that might be evidence of human neglect or denigration in public spaces (see Table 1). Counts were tabulated for each of these 33 case types for each CBG over the period covered by the database. As a first step to identifying an underlying factor structure, an exploratory factor analysis was run on the 33 count variables (Tabachnick and Fidell 2006). The final solution produced 5 factors with an eigenvalue > 1. These factors, whose constituent types and loadings are in Table 1, might be described as:

- *Housing issues*, including eleven items referring to poor maintenance by landlords (e.g., poor heating, chronic dampness) and the presence of pests (e.g., bedbugs).
- “*Uncivil*” *use of space*, including seven items that reflect how private actions can negatively impact the public sphere (e.g., illegal rooming house, poor conditions of property, and abandoned building).
- *Big buildings complaints*, including three different case types regarding problems with the upkeep of big buildings, like condos.
- *Graffiti*, including two different case types regarding graffiti, one generated by constituents, the other by the Public Works Department.
- *Trash*, including five items related to incivilities regarding trash disposal: illegal dumping, improper storage of trash barrels, empty litter basket, abandoned bicycle, and rodent activity. This last is not itself an incivility, but is a consequence of poor trash storage.

Five items did not load on any factor and were discarded before the foregoing analyses. Four other items that loaded at <.4, though, were maintained based on conceptual similarity: abandoned buildings loaded at .36 on the factor of *uncivil use*; requests to empty a litter basket

loaded  $>.3$  on both trash and graffiti, and was maintained on the former factor based on its substantive content; two items were added to the housing factor as they were conceptually identical to the definition of the factor and likely did not load in the factor analysis because of their low frequency.

### 2.3 Exploring the Dimensions of Physical Disorder

New measures were created from these five factors in order to evaluate their higher-order factor structure. We accomplished this by summing counts for each of the constituent case types for each CBG over the period covered by the database. These measures had substantial outliers and were all log-transformed before analysis. Correlations between them were all significant (except uncivil use and graffiti), although modest if they are considered to be manifestations of a single super-construct (see Table 2); only two bivariate correlations were above  $r = .4$  (housing and uncivil use; graffiti and trash), and two others were above  $r = .3$  (housing and big buildings). Given both content and the pattern of correlations, the five factors appear to suggest two main groupings: *denigration of the public space*, comprised of trash and graffiti; and *poor care or negligence for private space*, comprised of big buildings, housing, and uncivil use.

Confirmatory factor analysis, via structural equation modeling, was used to compare this two-factor structure to a one-factor structure in which all five measures loaded together on an overarching measure of physical disorder. The two-factor model was superior by all measures. It had better fit ( $CFI = .82$  vs.  $.61$ ;  $SRMR = .07$  vs.  $.10$ ;  $\Delta\chi^2_{df=1} = 89.27, p < .001$ ), and accounted for 42% of the variation across factors, as opposed to 26%. The model estimated the correlation between the two factors at  $r = .38$  ( $p < .001$ ). One will note that although the two-factor model was stronger, it still had a poor fit. Because the hypothesis in question was the efficacy of a one- or two-factor model, there were no assumptions that the components of each were completely

independent. Thus, we took the exploratory step of examining modification indices, leading to the addition of a covariance between uncivil use and trash was added to the model, greatly improving fit (CFI = .95, SRMR = .05,  $\chi^2_{df=5} = 24.26, p < .001$ ). The final parameter estimates for this model are presented in Figure 1.

Analysis 1 thus suggests that the CRM database is at least in principle capable of measuring two distinct but related aspects of physical disorder: *private neglect* and *public denigration*. This result provides a more nuanced measurement than existing scales of physical disorder, particularly with the ability to go beyond elements visible from the street, and to access conditions inside of buildings. Many previous protocols for measuring disorder have combined items from each of these categories (for example, abandoned or deteriorating housing with graffiti), and thus it is not surprising that the two constructs are correlated. It may also explain why previous longitudinal work has found that such items become uncoupled across time (Taylor, 2001). It is important to note that correlational constructs of this sort reflect a shared process, but it is not clear what this process actually is. It is possible, for example, that housing issues and uncivil use of private space are generated by the same behavioral tendencies, but it is equally feasible that one of these causes the other, or even that they are mutually reinforcing. These are questions that go beyond the scope of our paper and thus are ripe for future study. For present purposes, the reliable co-occurrence of these elements across neighborhoods provides two different sets of measures we subject to an econometric analysis: two of a generalized sort, *private neglect* and *public denigration*; and five lower level categories that are more specific, *housing*, *uncivil use of space*, *big buildings*, *graffiti*, and *trash*.

### 3. ANALYSIS 2: VALIDITY AND BIAS IN ADMINISTRATIVE DATA

Although it is tempting to treat the CRM database as the “eyes and ears of the city,” and thereby a direct reflection of neighborhood conditions across space and time, its accuracy in these regards cannot be assumed because each case in the database is in fact the coincidence of two events: the issue itself; and the decision of a resident or passer-by to report it. This fact suggests that assumptions must be imposed to analyze the data. In guiding this process we invoke a simple behavioral model for the distribution of calls defined not only by the probability of an issue in a given space ( $P_1$ ), but also the probability that it will be reported ( $P_2$ ).<sup>1</sup> If  $P_2$  varies across neighborhoods, it could in turn create systematic biases in any measure based on the CRM system. For example, in regions where residents are not inclined to make such calls, an issue might sit unnoted for a lengthy period, or even indefinitely, creating a gap or false negative in the database. Conversely, the residents of some neighborhoods might be notable in their vigilance, generating multiple reports for a single issue, leading to false positives that exaggerate the actual prevalence of disorder. This variation in  $P_2$  might be referred to as the *civic response rate*, which we thus account for in order to establish validity for the measures identified in Analysis 1.

In pursuing this goal, we look to develop a methodology that accounts for the local civic response rate, producing final measures that more accurately reflect neighborhood conditions. We focus particularly on issues in the public domain, like street light outages, as these are likely to be the most vulnerable to such biases being that the responsibility for reporting them belongs to no specific individual, but to the neighborhood as a whole. Developing this methodology entails three steps, utilizing data from the CRM system and a series of neighborhood audits. First, there must be an independent or “objective” measure of response rate that captures the

---

<sup>1</sup> We thank an anonymous reviewer for spurring our thoughts on this issue.

propensity of a neighborhood's residents or visitors to report a given issue. We use two such measures, one identifying street light outages, and the other evaluating sidewalk quality. Second, it is necessary to create a measure of civic response rate that is based on measures from within the CRM system. This is critical because such a measure would allow the continual estimation of response rate, and in turn the production of valid measures of disorder, in lieu of regular neighborhood audits. In the next subsection we develop the theoretical basis for how particular patterns in the CRM database might be reflective of civic response rate. By examining the multivariate relationships between these internal measures and the objective measures of response rate, it is possible to construct a new measure from within the CRM system that can be used as an adjustment factor.

Third and last, we develop an equation that combines counts of cases with the adjustment factor to calculate final measures of physical disorder. This requires a measure of objective physical disorder, against which it is possible to calibrate the adjustment factor, determining how heavy its influence should be. This is done through an additional neighborhood audit that assessed loose litter on streets and sidewalks, an item that has been central to measures of physical disorder. In sum, this process produces a complete methodology for translating a raw database of CRM calls into a measure of physical disorder across a city. To conclude, we examine the construct validity of the measures produced by this methodology, comparing it to a series of other demographic, economic, and social indicators traditionally associated with disorder.

### *3.1. Sources of the Civic Response Rate*

Reporting rates in the CRM database for public issues can be seen as having two distinct elements. The first entails knowledge of the CRM system and willingness to use it. The second is

a decision to take action or responsibility for the public space. To the former, a large part of the battle for any public service agency is informing residents of available services, and making them comfortable with utilizing them. The CRM system also requires direct interaction between constituents and the government, something that those from disadvantaged or minority groups are sometimes less inclined towards, either because they distrust the government in general, or because they do not expect the requested services to actually be delivered (Putnam 1993; Verba et al. 1995). The sum of these effects might be described as *engagement*, or the likelihood that a person would use the CRM system in any case. Given the evidence that such patterns cluster demographically, it is likely to vary across neighborhoods, potentially contributing to measurement bias.

Knowing of and being willing to use the CRM system is not sufficient for using it to report a public issue, however. When calling in a report about something like graffiti or illegal dumping, one is taking responsibility for the public space, something that might have a different set of motivations than a call addressing one's personal needs (e.g., a request for a bulk item pick-up). There are a number of mechanisms that might cause this *concern for public space* to vary systematically across neighborhoods. First, such variation could be owed to differences in the cognitive perception of disorder. One striking finding of city-wide neighborhood surveys is that resident ratings of local disorder vary within the same neighborhood and only moderately correlate with observational (e.g., video or research rating) measures (Franzini et al. 2008; Sampson and Raudenbush 2004; Taylor 2001). This would indicate that individuals and communities vary in their definition of "disorder," something that might play an important role in how likely they are to feel compelled to report such issues. At the same time, it by definition reveals that surveys reports are not "objective" measures of disorder either.



Another factor could be the variation across individuals in the level of personal responsibility they feel for the public space. For example, homeowners tend to be more engaged with public maintenance (O'Brien 2012), likely because of the long-term investment they have made by purchasing a house (Fischel 2005). Consistent with this, our preliminary analysis of the CRM data indicated that homeowners are four times more likely to report public issues than renters (O'Brien 2013). A third possibility is that the accumulation of physical disorder might incline residents to see the act of reporting new issues as useless, as it will be unable to overcome the consistent generation of such problems (Ross et al. 2001). The truth may involve any one of these mechanisms, or some combination thereof, but the point stands that concern for public space could contribute to cross-neighborhood variations in the rate of reporting actual instances of physical disorder.

There are two features of the CRM database that will prove useful in the development of measures that reflect *engagement* and *concern for public space*. As noted, CRM case records indicate the type of services requested. From these, there is a subset that indicates issues in the public space. This subset overlaps with, but is not equivalent to, the subset of case types regarding physical disorder. Second, users of the CRM system are able to register, creating an account for tracking their reports.<sup>2</sup> Reports made by a registered user are then attributed to the individual's account using an anonymous code, making it possible to determine how often an individual uses the system, and to approximate the individual's home location. Though this ignores those who have used the system but not established an account, this information still provides insights on an individual's calling patterns that we would not otherwise have.

---

<sup>2</sup> This procedure is encouraged by the directors of the CRM system, who see following-up with constituents as central to their goal of establishing open communication between citizens and the government.

The most direct way to measure engagement would be to tabulate the number of individuals who do and do not know about the CRM system. This can be approximated as the proportion of neighborhood residents who have an account with the CRM system. A less direct approach would be to identify case types whose need might be even across the city, that is, for which  $P_1$  would be constant across neighborhoods. In these cases, measuring their geographic distribution would then provide access to  $P_2$ , the likelihood of utilizing the system. For example, one might expect the need for general requests, which entail questions about city services and other government-related items, to be driven solely by interest and engagement with government. Another example would be requests for sanitation services to pick up bulk items. It is reasonable to assume that residents of all neighborhoods have a similar need for this service as it is not determined by external, neighborhood processes. A third example of an evenly distributed issue is the need for snow plows during a snowstorm. During a snowstorm, all neighborhoods should have a roughly equal need for snow plows, controlling for certain infrastructural characteristics (e.g., the total road length, dead ends). We then have four candidate measures of a neighborhood's engagement: *total registered users*, *general requests*, *bulk item pick-ups*, and *snow plow requests*.

Measuring concern for public space requires a focus on reports that document a case of public deterioration, and, in turn, a constituent's decision to take action regarding it. This requires a list of case types that indicate a public issue. It is not possible to use any one of these types as a benchmark, as done with general requests, bulk item pick-ups and snow plow requests, because the very issue at hand is whether public issues are uniformly distributed across the city. Instead, we focus on the other two techniques described for engagement. First, it is possible to identify a subset of users who have made one or more reports of a public issue. This could be

used to tabulate the number of individuals in each neighborhood that have used the CRM system for such a purpose. Additionally, some of these “public reporters” make a disproportionate number of reports. Given their zeal for neighborhood maintenance, these individuals might be referred to as “exemplars.” Public issues in a neighborhood with either a greater number of average or exemplar public reporters would be expected to instigate reports to the CRM system more often and more quickly. Second, it is possible to measure the proportion of reports of public issues that were made by registered users. This would indicate how consistently such calls are part of a sustained relationship between a resident and government services. This amounts to three measures of concern for public space: *public reporters*, *exemplars*, and *proportion of calls made by registered users*. Importantly, none of these measures is fully independent of engagement itself. For example, regardless of one’s inclination to report a street light outage, he or she must first know that the CRM system exists. Consequently, the following analysis allows these measures to load on one or both of these constructs.

### 3.2. *Estimating Civic Response Rate from the CRM Database*

In order to be concurrent with the neighborhood audits (described below), the current analysis uses only CRM reports from 2011, amounting to 161,703 cases with geographic reference across 154 case types. This analysis incorporates two new ways of utilizing the CRM database. First, similar to the identification of case types reflecting physical disorder in Analysis 1, 59 case types were identified as reflecting issues in the public space (e.g., street light outage, pothole, graffiti removal; complete list in Appendix A). Such a report indicates a concern for the maintenance of the public space on the part of the reporter. Other case types reflected personal needs rather than public concerns (e.g., general request, bulk item pickup). Second, all individuals who have registered with the CRM system have an anonymous ID code that is

appended to each of their reports. In 2011, there were 29,439 constituent users, accounting for 38% of all requests for service.<sup>3</sup> The ID code makes it possible to construct a database of users with variables describing each individual's pattern of reporting across time and space. This two-part database of calls and users was used to calculate the measures hypothesized to reflect a CBG's civic response rate.<sup>4</sup>

The call database was used to measure four of the seven proposed measures. Bulk item pick-ups (*bulk items*) and *general requests* were measured as the number of such requests occurring within a CBG. *Proportion of public issues reported by registered users* was measured as the number of public issues reported in a CBG attributed to a registered user divided by the total number of public issues reported in the CBG. *Snow plow requests* were first tabulated as a count for each CBG, but were then adjusted for the total population, road length, and the length of dead end roads.<sup>5</sup>

The other three measures were calculated from the database of registered users which included three main pieces of information: 1) the total number of calls a user had made; 2) the total number of calls a user had made regarding a public issue; 3) an estimate of the user's home location, based on the locations at which they requested services.<sup>6</sup> In 2011, 46% of registered

---

<sup>3</sup> Users who made one or more reports as a department member at any time (including 2010 or 2012), were removed because city employees differ from other constituents in their motivation for making reports. This excluded five individuals, a number that is low because for many employee-specific case types, user ID's were stripped before data sharing.

<sup>4</sup> Two CBGs were excluded from analysis as there were concerns that calls from there might not reflect usage of the CRM system by actual residents: 1) the CBG that contains City Hall, because many reports without an address are attributed to that location; and 2) the CBG that contains a large park, zoo and golf course, but includes the houses that ring the park.

<sup>5</sup> The number of snow plow requests (log-transformed to adjust for a skewed distribution) was regressed upon these three measures, accounting for 14% of the variation across CBGs.

<sup>6</sup> The home location was estimated in two ways, depending on the geographic range of an individual's requests for service. If the individual reported cases over a range with diameter smaller than .5 miles (90% of users), location was defined as the centroid of all reports made,

users were public reporters. Of these, 87% reported two or fewer public issues, though there were those who were considerably more active (18 made over 100 reports). Given this distribution, *total users* were measured as the number of registered users whose estimated location fell within the CBG; *average public reporters* were measured as the number of a CBG's total users who had reported one or two public issues during the year of 2011; and *exemplars* were measured as those who had reported 3 or more public issues.

### 3.3. *Objective Measures from Neighborhood Audits*

Objective neighborhood conditions were assessed through two separate audits. One identified street light outages and the level of street garbage in 72 of Boston's 156 census tracts (46%) between June 1 and August 31, 2011. In total, 4,239 street segments were assessed, and 244 street light outages were identified. Street light outages were attributed to the nearest address. Garbage was rated for each street block on a 5-point scale, with higher scores indicating more and larger piles of garbage. More detail on this protocol is provided in Appendix B.

In the second audit a consulting group hired by the City of Boston's Public Works Dept. assessed the quality of all of the city's sidewalks between November, 2009 and April, 2012. The unit of analysis was each continuous stretch of sidewalk that ran from intersection to intersection ( $N = 27,388$ ). For each sidewalk, the assessors noted the proportion of panels that required

---

which was then attributed to the appropriate CBG. Because of the small range, this estimate can be assumed to be reasonably precise. For those whose range had a diameter greater than .5 miles, this precision was weaker. These individuals were attributed not to a centroid, but to the census block group from which they made the most calls. This was done using the entire period of the database (March 2010 – June 2012) in order to make the greatest use of available information. This estimation technique was validated against a sample of 7,433 users for whom home locations were known. Of these, 78% were attributed to the correct CBG. More importantly, the counts generated by this process correlated with actual counts at  $r = .93$ . There is reason to believe that this correlation is underestimated. The sample used in the validation had an above average number of calls per person, a subsample for which the estimates had greater error.

replacement (i.e., cracked, broken), and subtracted this from the total. This generated a 0-100 measure of *sidewalk quality* (100 being a sidewalk with no panels requiring replacement).

Street light outages and sidewalks were each cross-referenced with the CRM database to identify reports regarding them. For street light outages, we sought to identify the date on which each was reported. This was defined as the earliest case of an outage reported on the street segment in question that was fixed by the city after the date an auditor noted the outage.<sup>7</sup> This was then used to create a series of dichotomous measures indicating whether the outage had been reported by a constituent within a certain time window (e.g., one month).<sup>8</sup> For sidewalks, all requests for sidewalk repair were joined to the nearest sidewalk polygon from the same road. We were able to exclude those created by City employees as an additional code was included with such cases. The count of constituent reports for every sidewalk was then tabulated. Of the 27,388 sidewalk polygons, 1,168 generated requests for repair (4%; *min* = 1, *max* = 19).

Because the three audits described events or conditions on a single street segment within a neighborhood, multilevel models were run to create CBG-level measures (Raudenbush et al. 2004). These models controlled for micro-spatial characteristics of the street (e.g., zoning), and the second-level residuals were then used as CBG-level measures. The outcome measures for these models were: the likelihood of a sidewalk generating one or more requests; the likelihood of a street light outage being reported within one month; and the continuous 1-to-5 measure of

---

<sup>7</sup> Note that this means a street light might have been reported before the audit, as long as the city had not completed the job until after.

<sup>8</sup> It was possible to distinguish whether a report was made by a constituent or a City employee. Thus a continuous measure of the time before reporting would not necessarily reflect the strength of constituent response. Instead, the dichotomous measures were created so that employee-reported outages could be considered not-reported until the date the employee report appeared. Thereafter they were omitted from the data, as it is not possible to know whether a constituent would have reported up to that point. For example, a street light outage reported 16 days later by a City employee takes the value “0” for the measure of being reported within two weeks, but would take no value (omitted) for the measure one month.

garbage. See Appendix C for more detail on these models and the specification of outcome measures.

Two deviations from this approach are important to note. First, the number of outages per CBG was low for a multilevel model (244 outages in 127 CBGs), so the models were run instead with tracts as the second-level ( $N = 56$  tracts with outages). Each CBG then took the measure for its containing tract. Second, because sampling for the garbage audit occurred at the tract level, CBGs varied in the number of street segments that were rated. In order to be certain that neighborhood-level measures were reliable, the ensuing analysis was limited to the 196 CBGs with 10 or more street segment measures (see also Sampson and Raudenbush, 1999).

### *3.4. Evaluating the Proposed Model of Civic Response Rate*

Descriptive statistics for both objective measures of response rate and CRM-based measures proposed to estimate response rate are reported in Table 3, as well as the correlations among them. All tabular variables had a skewed distribution with a long tail of CBGs that used the system extensively, leading us to log-transform them before correlational and regression analyses. As hypothesized, all variables indicating use of the system (general requests, bulk item pick-ups, all users, users reporting public issues, exemplary reporters) were strongly correlated ( $r$ 's = .36 - .93, all  $p$ -values < .001). Because of the very high correlation between all users and average users reporting public issues ( $r = .93$  in the full sample and  $r = .95$  in the subsample with values for all measures), the two were deemed to be the same measure. The “all users” measure was thus dropped from all proceeding analyses to avoid issues of multicollinearity.

Requests for sidewalk repairs and propensity to report street lights were modestly correlated ( $r = .18$ ,  $p < .05$ ). Each also shared stronger correlations with those measures from the CRM database intended to measure concern for the public space (public reporters, exemplars,

percentage of public issues reported by registered users) than those intended to measure engagement (general requests, bulk item pick-ups, all users). The reverse was true for requests for snow plows, as predicted. They were significantly positively correlated with the sidewalk measure ( $r = .14, p < .05$ ) but not the street light outage measure ( $r = -.12, p = ns$ ) and had a stronger correlation with measures of engagement than with concern for public space.

Structural equation modeling was used to determine how well the proposed constructs fit the data. The model analyzed those 195 CBGs with a measure for propensity to report street lights. The best-fitting model, depicted in Figure 2, had good fit ( $CFI = .95$ ,  $SRMR = .06$ ,  $\chi^2_{df=9} = 25.44, p < .01$ ), and was quite similar to the model proposed in the introduction to this analysis. The measures derived from the CRM system did indeed separate into the two proposed latent constructs, engagement and concern for public disorder. It is notable, however, that the two objective measures of civic response rate loaded on the latent construct of concern for public disorder (sidewalks:  $\beta = .34, p < .001$ ; street light outages:  $\beta = .18, p < .05$ ), but not on engagement.

As with the models in Analysis 1, the novelty of the various measures required that we take a partially exploratory approach, tweaking the theoretically-based model to specify the best fit. Consequently there were a few alterations that bear mentioning. 1) With the removal of the measure of all users, it was necessary to have average reporters of public issues load on both latent constructs (engagement:  $\beta = .50, p < .001$ ; concern for the public space:  $\beta = .52, p < .001$ ). 2) The measure of general requests was also removed as its strong correlation with other variables made the factor structure unstable. 3) The percentage of public calls from registered users was discarded as doing so strengthened the model's fit. 4) Total population was used as a control variable predicting average reporters of public issues ( $\beta = .13, p < .05$ ) and exemplars ( $\beta$



= .16,  $p < .01$ ). Modification indices for the final model suggested that no significant bivariate relationships had been omitted.

### 3.5. *Evaluating the Adjustment Factor*

The results of the previous model suggest that the estimate of the civic response rate, and therefore the desired adjustment factor, is based on measures of concern for the public space. A composite measure for each CBG was created using the parameter estimates from Figure 2. We then established its efficacy as an adjustment factor by examining how well it improved the relationship between the raw measures from Analysis 1 and objective measures of physical disorder, as indicated by street garbage. The analysis was performed in two parts. First, the raw counts of case types in each category (log-transformed to better approximate normality) were entered into five separate regressions predicting the level of street garbage in a CBG. Second, an adjustment factor was created for each count as an interaction with the civic response rate, which was then added to the corresponding regression.<sup>9</sup> This analysis was limited to residential neighborhoods (excluding regions dominated by institutions, parks, or downtown areas), as the predictive relationship between local behavior and loose litter would be most clear in these areas; in others, litter would be subject to dynamics that would not necessarily influence the other components of physical disorder (e.g., graffiti) in the same way ( $N = 135$  residential CBGs).

The first set of regressions found that all but one of the raw measures (graffiti) significantly predicted levels of street garbage (complete details in Table 4). The strongest relationships were with housing ( $B = .63$ ,  $p < .001$ ) and uncivil use of space ( $B = .38$ ,  $p < .001$ ). Big buildings ( $B = .21$ ,  $p < .05$ ) and trash ( $B = .18$ ,  $p < .05$ ) had more moderate relationships.

---

<sup>9</sup> The response rate was standardized and centered before the interaction was calculated. The physical disorder measures were left uncentered, with a minimum of zero, so that the response rate would adjust up or down proportional to the total number of actual reports.

The fit of all five regressions increased significantly with the introduction of the adjustment factor (see Table 4), with the strongest improvement occurring for trash ( $\Delta R^2 = .06, p < .01$ ) and uncivil use ( $\Delta R^2 = .05, p < .05$ ). Notably, the variance explained more than doubled for both trash and graffiti, which had the weakest initial relationships with street garbage.

### *3.5. Construct Validity for the Composite Measures*

As a last step, we evaluated the construct validity of these final measures by examining their relationship with other popular indicators of neighborhood conditions, drawn from three different data sources: median income, homeownership, and measures of ethnic composition from the census' American Community Survey (ACS; 2005-2009); survey measures of perceived physical disorder and collective efficacy (i.e., social cohesion and social control between neighbors) from the Boston Neighborhood Survey (BNS; 2008-2010 estimates,  $N = 3,428$ );<sup>10</sup> and reports of gun-related incidents from Boston's 911 call record (2011). Because the time points of these data sources vary, we analyze their relationship to the CRM-based measure for the most concurrent year: 2010 for the ACS and BNS, and 2011 for 911. As before, we focus the analysis on residential neighborhoods, but in this case we analyze at the broader spatial scale of census tracts rather than block groups ( $N = 121$  residential census tracts). We do so because the interpretation of the analysis depends in important ways on comparison with findings from previous studies, particularly Raudenbush and Sampson (1999) and Sampson and Raudenbush

---

<sup>10</sup> The BNS was a telephone survey based on the methodology from Raudenbush and Sampson (1999) with 3,428 participants in two waves (2008:  $N = 1710$ ; 2010:  $N = 1718$ ) recruited by random-digit dial. The two waves were combined to provide a reasonable number of respondents to create measurements at the scale of CBGs. Scales measuring physical disorder and collective efficacy were calculated first for each individual respondent. Neighborhood-level measures were then calculated by fitting multilevel models that nested individuals within their CBG and controlled for individual-level demographic characteristics (gender, age, ethnicity, and parental status). The Bayes residuals for the neighborhood-level model were then extracted as neighborhood measures adjusted for measurement error.

(1999), which were conducted on census tracts and clusters of tracts. In addition, because of the smaller sample size of the BNS compared to the Chicago study, the BNS has greater between-neighborhood reliability for tracts than for block groups. For the sake of brevity, we conducted this analysis on the higher-order measures of private neglect and public denigration. Results for the five lower-order measures as well as block groups are available upon request.

The measure of private neglect was lower where there was higher median income ( $r = -.59, p < .001$ ), homeownership ( $r = -.36, p < .001$ ), and collective efficacy ( $r = -.38, p < .001$ ), and greater where there were greater black ( $r = .61, p < .001$ ) and Hispanic populations ( $r = .27, p < .001$ ). It also co-occurred with gun-related incidents ( $r = .68, p < .001$ ). Further, it was higher where residents perceived more disorder ( $r = .44, p < .001$ ). The measure of public denigration had largely similar relationships with these measures—it was lower in areas with more homeowners ( $r = -.49, p < .001$ ) collective efficacy ( $r = -.48, p < .001$ ), and a higher median income ( $r = -.39, p = .001$ ). Public denigration was higher where there was a greater Hispanic population ( $r = .41, p < .001$ ) and more gun-related incidents ( $r = .27, p < .01$ ). It was also higher where residents perceived more disorder ( $r = .48, p < .001$ ). The one unexpected correlation was that it held no correlation with the proportion of black residents ( $r = -.05, p = ns$ ).

These validation correlations are lower than those reported in Raudenbush and Sampson (1999: 31) for survey-reported disorder. For example, public denigration correlates with perceived disorder at .48 in Boston but .71 in Chicago. However, at least four factors differ between studies beside the method (observation vs. CRM for non-survey based indicators of disorder)—the items in the measure, city, reliability of the surveys, and time period—making direct comparability difficult. We would note though, that the correlations for structural characteristics are similar—for example, the correlation between physical neglect and income in

Boston is -.59 and in Chicago the correlation of observed disorder with poverty is .64. And the correlations for residential stability are -.36 in Boston and -.25 in Chicago. Moreover, the CRM correlations are on par with previous comparisons between perceived and objective disorder in other studies (Brown et al. 2004; Franzini et al. 2008; Sampson and Raudenbush 2004; Taylor 2001).

Overall the results suggest that it is possible to construct a measure from within the CRM database that adjusts counts of case types to better reflect neighborhood conditions, though there are differences between the two classes of physical disorder that should be noted. In particular, private neglect had a stronger relationship with street garbage, with two of its constituent metrics (housing and uncivil use) surpassing the threshold of ~15% shared variance typically seen between domains of physical disorder (Taylor 2001). The relationships between the indicators of public denigration and street garbage were a bit weaker, but the correlations with other indicators of disorder were of similar magnitude, even stronger in cases. This could be owed to one of two possible explanations. The first is that issues of trash storage and graffiti are in fact less linked to patterns of litter than expected. The second is that these issues are more susceptible to reporter bias and potentially in ways that audits of natural patterns in deterioration, like street light outages and sidewalk cracks, might not fully capture. The same norms that lead to garbage-laden streets might also be responsible for diminished motivation to report graffiti or other issues in the public space. If so then the assumption that a neighborhood's civic response rate,  $P_2$ , is consistent for a given neighborhood across all case types is called into question. Future validation efforts should carefully evaluate the most effective measures both for objective comparison and internal adjustment, as these might differ depending on the particular set of conditions that are intended to be the focus, a theme we return to below.

#### **4. ANALYSIS 3: ASSESSING RELIABILITY ACROSS SPACE AND TIME**

Analyses 1 and 2 have provided a methodology for measuring physical disorder using the CRM system, but without a guideline for how such measures should be bounded in space and time. Thus far, measures have been developed for CBGs over the entire available time course (two years and four months). It is desirable, however, to assess measures for smaller time windows, allowing researchers to examine local conditions at more precise intervals, and facilitating longitudinal analysis. In addition to CBGs, it would be appropriate to determine the optimal time window for census tracts, the unit at which most urban research is conducted.

Determining an “optimal” time window for measurement requires a balance of two contrasting dimensions: smaller time windows are more precise, but are more sensitive to random events. To do this, we must examine how consistent the multiple measures of a single neighborhood are for different time intervals (using the intraclass correlation, or ICC), and the ability to statistically distinguish between neighborhoods (using the reliability coefficient, or  $\lambda$ ); these characteristics can be assessed using multilevel models. The goal is to identify the smallest time interval for which measures within a neighborhood are sufficiently consistent, and not overly sensitive to error or stochastic processes. Because the measures of interest are in fact composites that combine counts of cases with the measures of concern for the public space, the establishment of reliability requires two steps. First, we must identify a time interval for which all of the constituent measures (e.g., instances of housing issues) attain a desired threshold for reliability and ICC. Once the appropriate time interval for the constituent measures is determined, it must be confirmed that the same time interval is appropriate for the composite measure. One will note, however, that step one is not possible for the measure of exemplar reporters, as they are defined by their behavior over the course of a complete year. For this

reason, exemplars will always be calculated as the number of public reporters in a region obtaining exemplar status over the previous 365 days.

The last question we seek to answer is that of longitudinal tracking. If the final time intervals are small enough, it would be possible to examine patterns of change across time. The multilevel models can assess the slope for a measure at both the global and neighborhood levels. If the reliability for the slope is high enough, the model is capable of discerning varying trajectories across neighborhoods, which could then be used in subsequent analyses.

#### *4.1. Creating Measures for Spatio-Temporal Windows*

The complete CRM database, including all requests for service received between March 1, 2010 and June 29, 2012, is the basis for the temporal analysis. All requests are categorized by case type and include the date of the request and the address or intersection where services were to be rendered, allowing all requests to be geocoded to the appropriate census geographies.

The focal variables are those that constitute the composite measures of physical disorder, including both the raw counts of cases that reflect the five categories of physical disorder, and the measures of response rate. Drawing from Analysis 1, the five categories of physical disorder were housing, uncivil use of space, big buildings, graffiti and trash. Based on Analysis 2, response rate was calculated as the number of individuals reporting public issues, divided into two counts: those who made two or fewer calls in a year's time, and those who made three or more calls in a year's time (i.e., exemplars).

Measures for each variable, excepting exemplars, were created for all CBGs and tracts for eight temporal windows—1, 2, and 3 weeks, and 1, 2, 3, 4, and 6 months. For each, the original database was split into intervals of the given size, starting with March 1, 2010 and

ending with the last complete interval. A count was then produced for each interval for each element in the given level of analysis (i.e., block group or tract).<sup>11</sup>

#### 4.2. Multilevel Models

Hierarchical Linear Modeling (Raudenbush et al. 2004) was used to compare the consistency of counts within a CBG over time. A natural-log link was used to account for the Poisson distributions of all outcome variables. The first level equation predicted the outcome for a given time point relative to other measures for that region, and included: the number of time intervals elapsed since the start of the database, in order to estimate the rate and direction of change over time; dummy variables controlling for seasonal effects, based on the month of the midpoint of the given time-interval:

$$Y_{jk} = \beta_{0k} + \beta_{1k} * time_{jk} + \beta_{2k} * season_{jk} + r_{jk}$$

$$r_{jk} \sim N(0, \sigma^2)$$

The second level equation was an intercepts-only model, estimating the average level of a measure for a neighborhood across time. In addition, the parameter relating time to changes in a measure,  $\beta_1$ , was allowed to vary across CBGs, permitting the model to estimate different trajectories of change for different CBGs:

$$\beta_{0k} = \gamma_{00} + \mu_{0k}$$

$$\beta_{1k} = \gamma_{10} + \mu_{1k}$$

$$\mu_{0k} \sim N(0, \tau_0)$$

$$\mu_{1k} \sim N(0, \tau_1),$$

---

<sup>11</sup> For example, there were 28 1-month time windows between 3/10 and 6/12, generating that many counts for each CBG. The resultant dataset then contained 15,176 counts (28\*542), each attributed to a CBG and a time window.

where  $\tau_0$  is the measure of variation in the outcome measure between CBGs and  $\tau_1$  is a measure of variation between CBGs in the linear relationship between time and the outcome variable. Furthermore,  $\sigma^2$  is a measure of the variation in the outcome measure within CBGs (i.e., differences within a CBG across time).

The intraclass correlation coefficient (ICC) is then calculated as the proportion of variation that lies between groups:

$$ICC = \frac{\tau_0}{\sigma^2 + \tau_0}$$

Reliability is calculated as:

$$\lambda_0 = \frac{\tau_0}{\tau_0 + \sigma^2/n}$$

where  $n$  is the number of observations per CBG. As one can see, this measure grows both with a greater ICC, but also with more observations.

Variation across CBGs in the linear relationship between time and the outcome measure is assessed in two ways. First, the significance of the magnitude of  $\tau_1$  is assessed using a  $\chi^2$  test. Second, its reliability is measured as:

$$\lambda_1 = \frac{\tau_1}{\tau_1 + \sigma^2/SS_{Time}}$$

where  $SS_{Time}$  is the sums of squares for the measure of time.

#### 4.3. Comparing Spatio-Temporal Windows

The reliabilities and ICCs from the multilevel models described above are reported in Tables 5 (CBGs) and 6 (tracts). As expected, the proportion of variation attributable to differences between both CBGs and tracts (measured by the ICC) increased monotonically as



time windows became larger, owing both to greater consistency and fewer measures per neighborhood. As would also be expected, ICCs were higher when comparing tracts than CBGs.

The six measures varied in their consistency, with housing, graffiti and trash holding the strongest reliabilities and ICC's among the different measures of physical disorder. These differences seem largely attributable to the frequency of these categories. For example, there were three times as many events reflecting housing issues than uncivil use of space. With a lower frequency, counts of the latter will be more stochastic and therefore less consistent at smaller time intervals. Interestingly, counts of public reporters, though far fewer in number than actual calls, featured greater consistency within a region than any of the measures of physical disorder.

All ICCs in Tables 5 and 6 were significant at  $p < .001$ . The intent here, however, is not to find significant between-region variation, but to identify spatio-temporal windows at which a single measure is indicative of a region's "actual" value on that measure. The ICC, in that case, is used as an evaluation of how strongly a single measure of a neighborhood correlates with all other measures of that neighborhood. If we elect .7 as a threshold for a reliable neighborhood-level measure, then there are acceptable spatio-temporal windows available for all of the measures apart from big buildings. For those measures with greater consistency, the options are many: housing, for example, could be measured at two-month intervals for tracts or four-month intervals for CBGs. For others, like uncivil use, there is a need for six-month intervals at the tract level, and no time interval satisfies this criterion for CBGs.

The slope reliabilities in Tables 5 and 6 indicate the ability of the model to distinguish between the trajectories of different regions over time. Variation in slopes across CBGs and tracts were significant at  $p < .05$  (or some lower threshold) in nearly all models, excepting all

those for big buildings, and those for public reporters of intervals longer than four months long. This variation was somewhat more discernible in tract-level models.

We then examined whether these cross-time consistencies hold for the composite measures. For the sake of simplicity, this was done for all variables using six-month windows for tracts. One will note that the generation of the composite measures requires the incorporation of the number of exemplars, measured for the full year preceding the last day of the given time window. Consequently, the first time window analyzed must be that which ends at or after the end of the 12<sup>th</sup> month of the available database, diminishing the number of measurements per tract. For this reason, this analysis does not examine change over time.

Similar to above, multilevel models were run to examine the consistency of the composite measures across space and time. The reliabilities and ICC's from these are reported in the bottom row of Table 6. Across the board, reliabilities and ICC's were lower for the composite measures, but not alarmingly so. All measures (other than big buildings) maintained ICC's around .6 or higher, and housing had an ICC greater than .7. Reliabilities were typically around .8.

Last, we replicated the analysis for the two higher-order constructs, private neglect and public denigration. The statistical advantage is that the combination of multiple measures amplifies the number of cases in the average time interval, thereby enabling higher reliabilities and ICC's at smaller time windows. This is particularly important when considering a measure like big buildings. Though it has low reliability when measured on its own, it might be incorporated into a more comprehensive description of the neighborhood in this manner.

Reliabilities and ICC's for these higher-order counts were higher than their constituent categories. For each, the criterion of ICC = .7 was attained at six-month intervals for CBGs and

two-month intervals for tracts (complete results available on request). This remained largely consistent when they were combined with measures of concern for the public space to create composite measures, though the consistency in public denigration for small time windows was somewhat diminished. For tracts with two-month intervals, public denigration had an ICC = .44, and a reliability coefficient of .88. Private neglect had an ICC = .65 and a reliability coefficient of .94. For CBGs with six-month intervals, public denigration had an ICC = .51, and a reliability coefficient of .76. Private neglect had an ICC = .68 and a reliability coefficient of .87.

## **5. SUMMARY AND IMPLICATIONS**

The current study sought to demonstrate how a citizen-initiated administrative database might act as “the eyes and ears of the city” in the spirit of Jane Jacobs (1961) while at the same time providing a low-cost, real-time measure of physical disorder. To accomplish this goal we needed to address three major issues: the lack of interpretable constructs; the potential that the database might not objectively or accurately reflect real-world conditions; and the need for criteria for reliability when bounding measures in space and time. Creating a set of theoretically-guided factors first required an item-response model, in this case 28 case types that reflected deterioration or incivilities within a neighborhood. The subsequent factor analysis revealed five separate categories of physical disorder. It is worth noting that these constructs were extant in the data, but that it was necessary to distinguish them from the noise surrounding them. Skipping forward to Analysis 3, once these measures were fully developed, criteria for reliability were established both for one-time measures and cross-time trajectories using multilevel modeling.

In between these two steps, Analysis 2 addressed the question of validity, which is a perhaps underappreciated concern for econometric study. Neighborhood audit protocols are

developed and administered to measure specific things as accurately as possible, meaning they have an inherent validity for those items that they assess. In contrast, administrative data are the byproduct of processes whose idiosyncrasies might bias their reflection of ground-truth. The CRM database is the product of constituent reports, and therefore is vulnerable to inconsistencies in reporting across neighborhoods. Because the nature of the bias was known, however, it was possible to account for it. The final methodology used indicators of civic response rate, derived from the CRM database itself, to systematically adjust raw measures to better reflect objective conditions. Reaching this point entailed considerable work, including two independent data collections and a lengthy set of analyses. Nonetheless, that investment of cost and effort would be necessary for any traditional protocol for measuring disorder, and in our case laid the groundwork for a methodology that can be reproduced at little cost both within Boston across time, and in other cities with their own CRM systems. It is also worth noting that cities frequently conduct audit studies, so it is reasonable to assume that there will be an ongoing stream of potential sources of data from which to derive validation measures.

The final product was a multidimensional measure of physical disorder that is not only nearly costless to the researcher, but also more comprehensive and precise than other measures currently available. Further, the programming code published along with this manuscript facilitates reproduction of the measure wherever similar databases exist. Given these apparent upsides to the use of administrative data, it seems appropriate to forward a new, three-step process for carrying econometrics into the age of big data:

- 1) *Extract constructs* by identifying item-specific models that are reflective of the theoretical concept of interest, and then examining their underlying factor structure

2) *Validate the measure* by identifying and adjusting for any bias that the information source might impart to the data, and examining in conjunction with external data.

3) *Establish reliability* in the measure's ability to track information across space and time.

With this methodology in hand, the opportunity before econometric urban science is considerable, as there is a veritable trove of information on cities that sits largely untapped, of which the CRM database is but one example. Cities collect and now make available many other data points—such as tax assessments, building permits, zoning decisions, restaurant inspections, environmental assessments, housing code violations, pedestrian flows, and bicycle collisions, to name a few—each providing their own insights on the social and physical ecology of neighborhoods. Going further, there are private databases, such as Twitter, cell phone records, and Flickr photo collections that are also geo-coded and might be equally informative in building innovative measures of urban social processes. . These various resources could be used to develop new versions of traditionally popular measures, like we have done here, or to explore new ones that have not been previously accessible. An illustration of the latter comes from our own analysis, where a byproduct of validation has provided two unanticipated behavioral measures—one related to civic engagement and the other capturing attitudes towards disorder in the public space. The potential of new forms of large-scale data underscores the central inspiration of this manuscript: as the volume of data on urban areas continues to grow and diversify, they provide new and distinctive ways to measure neighborhood characteristics, often in ways not previously foreseen. Such advances can be appropriated to shed light on some of the most salient themes in urban science, from the structure and function of the social organization,

to the role of cognition and culture in generating local patterns, to the nascent examination of relationships between neighborhoods and the higher-order social structure of the city.

Apart from its implications for econometric science more broadly, the current methodology represents an advance for the direct measurement of physical disorder in urban neighborhoods. It incorporates a broad range of phenomena and is the first physical disorder measure to divide these items into independent subcategories, suggesting new avenues for research. For example, do the five subcategories relate differently to a neighborhood's other social and demographic characteristics? If so, do they each reflect a different set of processes occurring within the neighborhood? Further, what is the source of the higher-order constructs suggested by Analysis 1, private neglect and public denigration? Is it that their constituent types are all manifestations of the same social and behavioral patterns, or do they share other causal relationships that reinforce their correlation? It is crucial that we not over interpret this single case and inappropriately reify these particular constructs. It will be necessary to confirm their consistency with data from other time points and cities, something that is likely to be possible with the continued proliferation of CRM systems throughout North America and Western Europe.

In addition, the measures enable a variety of analytical approaches that could prove useful in the extension of research surrounding "broken windows" and other theories of neighborhood well-being. All of the measures describe neighborhood conditions at the level of census tracts, and some can be used for CBGs. Future work could likely find ways to measure and interpret patterns of disorder for streets or even individual buildings. The measures can also be tracked across time, allowing for analyses that evaluate not only what a neighborhood's current level of physical disorder is, but whether it is on an upwards or downward trajectory.

Finally, the CRM data are continuously generated as part of administrative operations. A new study with up-to-date data requires only a download and some data manipulation. In an effort to assist others in initiating such work, we will be publishing the computer code for constructing the measures developed in the current manuscript (along with the data). As CRM systems become more numerous around the world, typically in the form of 311 Hotlines, this sort of measurement is becoming possible in a variety of cities. Some of these cities have established common standards for publishing CRM data, meaning that the data are not only being made readily available, but are compatible in ways that would support cross-city comparisons.

## **6. BALANCING LIMITATIONS AND OPPORTUNITIES**

We have thus far focused predominantly on the opportunity presented by “big data,” but we must also take stock of the limitations that they carry and the challenges to be addressed. Indeed, the methodology presented here is only a first step—an illustration of what is possible—and future work will need to refine it further, particularly in terms of the validation process. We would likewise stress that traditional, well-established methods of urban data collection, such as community surveys and social observation (Sampson 2012), will continue to play an essential role in any future analysis. Claims to the contrary are merely “big data hubris” as aptly put by Lazer et al. (2014). Each approach has its pros and cons, the balancing of which will depend on the research question. Surveys and observation are expensive and cannot realistically be carried out in real time, for example, but they can be calibrated to be representative of the population. By contrast, the CRM data analyzed here are cheap and in principle can be measured at very fine-grained geographic scales and almost in real time, but issues of reliability and what the data are really measuring remain.

For example, complaints about big buildings were not particularly common in our database, making it difficult to measure that construct reliably. Techniques that aggregate cases at higher levels, by increasing the geographical range, the temporal window, or, as we did here, combining multiple related constructs, will be critical. Future research is needed to examine these issues, especially in a context that can directly compare and contrast different methodologies of data collection, such as systematic social observation. Perhaps more important, although our validation process is promising we still cannot be entirely certain that we have directly accessed the intended information, especially for those things that occur within private spaces or out of the public observation. In particular, our measures of public denigration were not as closely correlated with other indicators of urban social structure as might be expected from past research. It may be, as we noted earlier, that the techniques for measuring and accounting for bias in this case were not sufficient to fully calibrate public denigration.

Another potential weakness is our working assumption that reporting bias is consistent across case types. Our data here seem to suggest that the situation is more nuanced, as reporting rates for street light outages and broken sidewalks were only moderately correlated. This finding points to important improvements for future versions of the measure, while also highlighting the need to tailor the validation process to the specific measure of interest. In some cases, like that presented here, there is a need to adjust for biases inherent in the data, and the objective measures necessary for doing so will need to be carefully constructed and measured.

In other cases, however, such a process of construct validation or bias adjustment may be less necessary or not applicable even though reliability assessment by temporal and geographic scale remains at issue. For example, building permits and zoning approvals or variances are legal requirements for major building renovations and additions, meaning they should be largely



objective in the information they provide. Allocation of city resources (e.g., beautification efforts or economic development) or distribution of the city budgets by amount and location are also largely “bias free” in their measurement and now widely available electronically. The availability of such measures could provide new insight into processes such as gentrification and inequality in the delivery of city services (see e.g., Hwang and Sampson 2014).

Furthermore, in certain cases, the raw contents of the data are exactly what a researcher wants, and their face validity is sufficient to offer concrete interpretation. In such situations, researchers do not actually want to adjust for any biases. For example, a recent paper used noise complaints from New York City’s 311 hotline as a direct reflection of social conflict between neighbors (Legewie 2014). Regardless of actual noise levels or norms of reactivity, each call in this analysis reflects an objective case of one neighbor asking the government to regulate the behavior of another. More generally, the electronic availability in many cities of citizen reporting systems offers a wide variety of domains (in our Boston data, 178 unique types of service calls) to test Black’s (1976) theory of the behavior of law and citizen initiation of government control.

Of course, the fundamental issues of measurement error and validity bear down at some level on all methodologies: survey reports can be skewed by other perceptual factors, including implicit judgments of race and class (Sampson and Raudenbush 2004), and observational work is dependent on inter-rater reliability that is always less than unity. The dominant investigator-driven research method is to conduct surveys or interviews, but even here there is continuing controversy over the idea that researcher control leads to validity. For example, a recent critique argues that interviews are a weak basis for studying culture or inferring the motives for an

individual's behaviors (Jerolmack and Khan 2014, and responses). While we would not go that far, our point is that assumptions must always be invoked in the analysis of social science data.

It remains significant, however, that administrative data are outside of a researcher's direct control and that assumptions may sometimes be required that are uncheckable because of unobserved processes related to reporting or administrative filtering. It follows that validation is imperfect and should be viewed as a continuing process, one that will need to be undertaken for new administrative data sets that become available. Some might argue that these caveats obviate the usefulness of administrative data and other forms of "naturally occurring" digital information. Our position is to recognize both strengths and weaknesses of the new data being made available, using the most rigorous methods possible to address limitations. In defense of our approach, we would also note the significant advantages of 311 data we analyze relative to big data more generally. The CRM data are characterized by their richness and geographic precision, and their longitudinal nature permits long-term tracking. In addition, some of the content that is most difficult to validate stems from the fact that it cannot be measured by direct means and was thus previously unavailable, making it novel. If used properly, such data can broaden the range of questions that we can examine, and the manner in which we do so. Our hope is that future efforts will capitalize on the advantages of large-scale administrative records and to combine them in meaningful ways with survey and observational protocols.

In sum, instead of an either/or approach, the debate between those who believe that only data generated directly as part of the research process is valid and those that believe that administrative and other types of naturally occurring data can be of use pushes both sides to improve the quality of their research, which certainly can only lead to better science. In the meanwhile, owing to their increasing availability at little or no cost and at unprecedented

temporal and geographic scales, big data remain a resource to be tapped, and it is incumbent upon researchers to develop methodologies that do so in ways that fulfill the expectations of rigorous science.

## **7. METHODOLOGY, THEORY, AND THE FUTURE OF BIG DATA**

Although this manuscript was tailored to the specifics of econometrics, the rise of big data illustrates the challenges facing computational social science writ large. There is a clear need to demonstrate what these novel data sources can measure and how constructed metrics are theoretically relevant. Further, they must be demonstrated to be both reliable and valid in their measurement before modeling can begin, which unfortunately seems to be the default approach in many current approaches which emphasize “econometrics” over “ecometrics” or simply the power to predict. However powerful predictive analytics may be, it does not answer the substantive questions about social processes and mechanisms that motivate most social scientists.

This paper therefore set out to accomplish a linked set of measurement goals that was rooted in substantive concerns. We grounded our study in a measure that is influential in urban research and theory, and we closely examined validity in a manner that goes beyond previous work. Though others have used supplementary data to give context to the patterns in Facebook or cell phone calls (Eagle et al. 2009; Kosinski et al. 2013), ours is the only study that we know of that has gone a step further, using multiple internal measures to reduce measurement error and then validating this technique with external sources and substantive theory—our approach was not simply data driven. Given the size and novelty of aptly termed “big” data, there is the temptation to allow them to guide analysis and, in turn, dictate theory. Indeed, some have claimed that the era of big data will eliminate the need for theory, as it will be derived from the

massive size of the data available. The former Editor of *Wired* magazine was perhaps the most bold, claiming “the end of theory” and that “the data deluge makes the scientific method obsolete” (quoted in Pigliucci, 2009). We strongly disagree. Purely data driven approaches run the risk of producing models and algorithms that are over fit to the idiosyncrasies of a particular data set, leading to new theoretical models that are artifactual or just plain wrong—something that has been partially blamed for the failure to predict the crash of the housing bubble in 2008. Big data hubris is indeed a problem (Lazer et al. 2014). Accordingly, we have had theory take the lead throughout, determining the case types reflecting physical disorder and the measures of civic response rate.

A balance must nonetheless be maintained. As Lazer et al. (2009) rightfully point out, our current theories are not well-suited to the complexity of information contained in these sorts of data, and consequently are often unequipped to offer conjectures about them. In the current case, there was no model of disorder that was sufficiently articulated to predict *a priori* categories for the 28 indicators that we identified. Thus, there is something to be learned from these data about the causal dynamics that underpin disorder and its various manifestations, though such insights will of course be subject to the same rigorous evaluation required of any new theory. This “checks-and-balances” relationship between theory and empirics is instructive, and will probably characterize the continued efforts of scientists to incorporate big data into their work, as well as the emergence of a fully mature field of computational social science.

## REFERENCES

- Black, Donald. 1976. *The Behavior of Law*. New York: Academic.
- Brown, Barbara B., Douglas D. Perkins and Graham Brown. 2004. "Incivilities, Place Attachment and Crime: Block and Individual Effects." *Journal of Environmental Psychology* 24:359-371.
- Burdette, A. M. and T. D. Hill. 2008. "An Examination of Processes Linking Perceived Neighborhood Disorder and Obesity." *Social Science and Medicine* 67:38-46.
- Caughy, M. O., S. M. Nettles and P. J. O'Campo. 2008. "The Effect of Residential Neighborhood on Child Behavior Problems in First Grade." *American Journal of Community Psychology* 42:39-50.
- Caughy, Margaret O., Patricia J. O'Campo and Jacqueline Patterson. 2001. "A Brief Observational Measure for Urban Neighborhoods." *Health and Place* 7:225-236.
- Cohen, Deborah, Suzanne Spear, Richard Scribner, Patty Kissinger, Karen Mason and John Widgen. 2000. "'Broken Windows' and the Risk of Gonorrhea." *American Journal of Public Health* 90:230-236.
- Eagle, Nathan, Alex Pentland and David Lazer. 2009. "Inferring Friendship Network Structure by Using Mobile Phone Data." *Proceedings of the National Academy of Sciences* 106:15274-15278.
- Fischel, William A. 2005. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance, and Land-Use Policies*. Cambridge, MA: Harvard University Press.
- Franzini, Luisa, Margaret O'Brien Caughy, Sandra Murray Nettles and Patricia O'Campo. 2008. "Perceptions of Disorder: Contributions of Neighborhood Characteristics to Subjective Perceptions of Disorder." *Journal of Environmental Psychology* 28:83-93.
- Furr-Holden, C. D. M., M. J. Smart, J. L. Pokorni, N. S. Ialongo, P. J. Leaf, H. D. Holder and J. C. Anthony. 2008. "The Nifety Method for Environmental Assessment of Neighborhood-Level Indicators of Violence, Alcohol, and Other Drug Exposure." *Prevention Science* 9:245-255.
- Furr-Holden, C. Debra M., Adam J. Milam, Elizabeth K. Reynolds, Laura MacPherson and Carl W. Lejuez. 2012. "Disordered Neighborhood Environments and Risk-Taking Propensity in Late Childhood through Adolescence." *Journal of Adolescent Health* 50:100-102.
- Hwang, Jackelyn and Robert J. Sampson. 2014. "Divergent Pathways of Gentrification: Racial Inequality and the Social Order of Renewal in Chicago Neighborhoods." *American Sociological Review* 79: DOI: 10.1177/0003122414535774.
- Jacobs, Jane. 1961. *The Death and Life of Great American Cities*. New York: Random House.
- Kosinski, Michal, David Stillwell and Thore Graepel. 2013. "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior." *Proceedings of the National Academy of Sciences* 110:5802-5805.
- Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343:1203-1205.
- Lazer, David, Alex Pentland, Lada Adamic, Siana Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas A. Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne. 2009. "Computational Social Science." *Science* 323:721-723.
- Legewie, Joscha. 2014. "Contested Boundaries: Explaining Where Ethno-Racial Diversity Provokes Neighborhood Conflict." New York University: Working Paper.

- Mujahid, M. S., A. V. D. Roux, M. W. Shen, D. Gowda, B. Sanchez, S. Shea, D. R. Jacobs and S. A. Jackson. 2008. "Relation between Neighborhood Environments and Obesity in the Multi-Ethnic Study of Atherosclerosis." *American Journal of Epidemiology* 167:1349-1357.
- O'Brien, Daniel Tumminelli. 2012. "Managing the Urban Commons: The Relative Influence of Individual and Social Inventives on the Treatment of Public Space." *Human Nature* 23:467-489.
- O'Brien, Daniel Tumminelli. 2013. "Custodians and Custodianship in Urban Neighborhoods: A Methodology Using Reports of Public Issues Received by a City's 311 Hotline." *Environment and Behavior*.
- O'Brien, Daniel Tumminelli and David S. Wilson. 2011. "Community Perception: The Ability to Assess the Safety of Unfamiliar Neighborhoods and Respond Adaptively." *Journal of Personality and Social Psychology* 100:606-620.
- O'Brien, Daniel Tumminelli and Richard A. Kauffman. 2013. "Broken Windows and Low Adolescent Prosociality: Not Cause and Consequence but Co-Symptoms of Low Collective Efficacy." *American Journal of Community Psychology* 51:359-369.
- Pigliucci, Massimo 2009. "The End of Theory in Science?" *European Molecular Biology Organization Reports* 10:534.
- Putnam, Robert. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, NJ: Princeton University Press.
- Raudenbush, Stephen W. and Robert J. Sampson. 1999a. "Ecometrics: Toward a Science of Assessing Ecological Settings, with Application to the Systematic Social Observation of Neighborhoods." *Sociological Methodology* 29:1-41.
- Raudenbush, Stephen W. and Robert J. Sampson. 1999b. "'Ecometrics': Toward a Science of Assessing Ecological Settings, with Application to the Systematic Social Observation of Neighborhoods." *Sociological Methodology* 29:1-41.
- Raudenbush, Stephen W., Anthony Bryk, Yuk Fai Cheong, Richard Congdon and Mathilda du Toit. 2004. *Hlm 6: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.
- Ross, Catherine E. and John Mirowsky. 1999. "Disorder and Decay: The Concept and Measurement of Perceived Neighborhood Disorder." *Urban Affairs Review* 34:412-432.
- Ross, Catherine E., J. Mirowsky and S. Pribesh. 2001. "Powerlessness and the Amplification of Threat: Neighborhood Disadvantage, Disorder, and Mistrust." *American Sociological Review* 66:568-591.
- Rundle, A. G., M. D. M. Bader, C. A. Richards, K. M. Neckerman and J. O. Teitler. 2011. "Using Google Street View to Audit Neighborhood Environments." *American Journal of Preventive Medicine* 40:94-100.
- Sampson, Robert J. 2012. *Great American City: Chicago and the Enduring Neighborhood Effect*. Chicago: University of Chicago Press.
- Sampson, Robert J. and Stephen W. Raudenbush. 2004. "Seeing Disorder: Neighborhood Stigma and the Social Construction of 'Broken Windows'." *Social Psychology Quarterly* 67:317-342.
- Skogan, Wesley G. 1992. *Disorder and Decline*. Berkeley, CA: University of California Press.
- Tabachnick, Barbara G. and Linda S. Fidell. 2006. *Using Multivariate Statistics*. New York: Allyn and Bacon.

- Taylor, Ralph B. 2001. *Breaking Away from Broken Windows: Baltimore Neighborhoods and the Nationwide Fight against Crime, Grime, Fear, and Decline*. Boulder, CO: Westview.
- Verba, Sidney, Kay Lehman Schlozman and Henry E. Brady. 1995. *Voice and Equality: Civic Volunteerism in American Politics*. Cambridge, MA: Harvard University Press.
- Wen, Ming, Lousie C. Hawkley and John T. Cacioppo. 2006. "Objective and Perceived Neighborhood Environment, Individual Ses and Psychosocial Factors, and Self-Rated Health: An Analysis of Older Adults in Cook County, Illinois." *Social Science and Medicine* 63:2575-2590.
- Wilson, James Q. and George Kelling, L. 1982. "The Police and Neighborhood Safety: Broken Windows." *Atlantic Monthly* 127:29-38.

**Table 1.** Counts of case types that reflect human neglect or denigration of the neighborhood, including the factors and loadings from an exploratory factor analysis.

Case Type	Count	Factor Loading	Case Type	Count	Factor Loading
<i>Housing Issues</i>			<i>Big Buildings</i>		
Bed Bugs	871	.49	Big Buildings Enforcement	236	.68
Breathe Easy	590	.53	Big Buildings Online Request	274	.72
Chronic Dampness/Mold	442	.44	Big Buildings Resident Complaint	209	.60
Heat - Excessive, Insufficient Maintenance	2175	.62	<i>Graffiti</i>		
Complaint – Residential	687	.54	Graffiti Removal	8826	.83
Mice Infestation – Residential	796	.59	PWD Graffiti	847	.50
Pest Infestation – Residential	330	.52	<i>Trash</i>		
Poor Ventilation <sup>a</sup>	26	—	Abandoned Bicycle	144	.45
Squalid Living Conditions <sup>a</sup>	128	—	Empty Litter Basket <sup>b</sup>	802	.30
Unsatisfactory Living Conditions	8948	.85	Illegal Dumping	2292	.87
Unsatisfactory Utilities – Electrical, Plumbing	174	.41	Improper Storage of Trash (Barrels)	4756	.91
<i>Uncivil Use of Space</i>			Rodent Activity	3287	.40
Abandoned Building	238	.36	<i>No Factor (Discarded)</i>		
Illegal Occupancy	642	.42	Illegal Auto Body Shop	105	—
Illegal Rooming House	471	.47	Illegal Posting of Signs	236	—
Maintenance – Homeowner	180	.41	Illegal Use	137	—
Parking on Front/Back Yards (Illegal Parking)	336	.42	Overflowing or Unkept Dumpster <sup>a</sup>	526	—
Poor Conditions of Property	2438	.80	Pigeon Infestation	82	—
Trash on Vacant Lot	432	.57			

<sup>a</sup> – Items did not load on initial factor analysis, but were added based on content similar to factor or one or more of its constituent items.



<sup>b</sup> – Item loaded at  $>.3$  on both the trash and graffiti factors. It was maintained on the trash factor for reasons of content.

Note: For factor analysis,  $N = 544$  census block groups. An iterated principal factors estimation was used with a promax rotation.

**Table 2.**

Descriptive Statistics for and Correlations between Five Sub-Measures of Physical Disorder.

	1.	2.	3.	4.	5.
1. <i>Housing</i>	—	.47***	.34***	.18***	.20***
2. <i>Uncivil Use</i>	—	—	.10*	.04	.33***
3. <i>Big Buildings</i>	—	—	—	.14**	.21***
4. <i>Graffiti</i>	—	—	—	—	.45***
5. <i>Trash</i>	—	—	—	—	—
<b>Median</b>	17.5	5	0	7	10
<b>(Range)</b>	(0 – 183)	(0 – 49)	(0 – 18)	(0 – 216)	(0 – 279)

Note:  $N = 544$  census block groups. All variables were log-transformed before correlations.\*\* -  $p < .01$ , \*\*\* -  $p < .001$

**Table 3.**

Descriptive statistics for and correlations between proposed indicators of response rate.

	<i>General Requests</i>	<i>Bulk Item Pick-Ups</i>	<i>Registered Users</i>	<i>Average Public Reporters</i>	<i>Exemplars</i>	<i>% Pub Issues by Users</i>	<i>Snow Plowing</i>	<i>Side-walk Repairs</i>	<i>Street Light Outages</i>	<i>Total Pop</i>
<i>General Requests<sup>a</sup></i>	—	.37***	.67***	.61***	.53***	.02	.26***	.18***	.01	.30***
<i>Bulk Item Pick-Ups<sup>a</sup></i>	.37***	—	.78***	.68***	.36***	-.24***	.43***	.19***	-.09	.07
<i>All Registered Users<sup>a</sup></i>	.69***	.78***	—	.93***	.62***	-.07 <sup>+</sup>	.44***	.28***	.04	.27***
<i>Registered Users Reporting Public Issues<sup>a</sup></i>	.68***	.68***	.95***	—	.60***	.03	.42***	.30***	.13 <sup>+</sup>	.26***
<i>Exemplary Reporters of Public Issues<sup>a</sup></i>	.51***	.40***	.66***	.64***	—	.16***	.29***	.28***	.09	.25***
<i>Percentage of Public Issues Reported by Registered Users</i>	.04	-.25***	-.08	.05	.22**	—	-.36***	.07 <sup>+</sup>	.18*	.09*
<i>Requests for Snow Plowing<sup>b</sup></i>	.31***	.49***	.51***	.48***	.35***	-.36***	—	.14***	-.12	-.00

<i>Propensity to Request Sidewalk Repairs<sup>b</sup></i>	.13 <sup>+</sup>	.08	.18*	.26***	.26***	.15*	.14 <sup>+</sup>	—	.18*	.09*
<i>Propensity to Report Street Light Outages within One Month<sup>b</sup></i>	.01	-.09	.04	.13 <sup>+</sup>	.09	.18*	-.12 <sup>+</sup>	.18*	—	-.01
<i>Total Population</i>	.25***	.00	.17*	.14 <sup>+</sup>	.16*	.04	.04	.00	-.01	—
<b>Mean (SD)</b>	47.13 (34.07)	56.28 (36.28)	53.49 (30.68)	21.37 (13.33)	3.06 (2.99)	0.44 (0.09)	0.00 (1.00)	0.00 (0.37)	0.00 (0.60)	1153 (565)
<b>Range</b>	1 – 461	0 – 199	2 – 232	0 – 104	0 – 29	0.17 – 0.76	-3.29 – 2.64	-.89 – 1.18	-1.01 – 2.22	246 – 4719

<sup>+</sup> -  $p < .10$ , \* -  $p < .05$ , \*\* -  $p < .01$ , \*\*\* -  $p < .001$

Note:  $N = 541$  for all measures except propensity to report street light outages within one month ( $N = 195$ ). Descriptive statistics reported for all census block groups; correlations including all CBGs with measures on both variables reported above the diagonal, correlations for those CBGs with values for all measures ( $N = 195$ ) reported below the diagonal. See text for more details on the derivation of each measure.

<sup>a</sup> – Log-transformed before correlations to account for skewed distribution.

<sup>b</sup> – Deviation from regression equation controlling for key variables; see text for more detail.

**Table 4.**

Comparison of results from regressions using the five categories of physical disorder derived from the CRM database to predict objective measures of garbage, with and without the CRM-based adjustment factor.

	<b>Housing</b>	<b>Uncivil Use</b>	<b>Big Building s</b>	<b>Graffiti</b>	<b>Trash</b>
	<b>B</b>	<b>B</b>	<b>B</b>	<b>B</b>	<b>B</b>
<i>Raw Measure</i>	.63***	.38***	.21*	.13	.18*
<b>R<sup>2</sup></b>	<b>.40***</b>	<b>.14***</b>	<b>.05*</b>	<b>.02</b>	<b>.03*</b>
<i>Raw Measure</i>	.62***	.39***	.27**	.29**	.33**
<i>Adjustment Factor</i>	-.13 <sup>+</sup>	-.19*	-.20*	-.27*	-.29**
<b>Total R<sup>2</sup></b>	<b>.42***</b>	<b>.18***</b>	<b>.08**</b>	<b>.07*</b>	<b>.09**</b>
<b>Δ R<sup>2</sup></b>	<b>.02*</b>	<b>.04*</b>	<b>.03*</b>	<b>.05*</b>	<b>.06**</b>

Note:  $N = 135$  census block groups classified as residential and with measures of garbage for ten or more street segments. All CRM-based variables were log-transformed before regressions.

\* -  $p < .05$ , \*\* -  $p < .01$ , \*\*\* -  $p < .001$

**Table 5.**

Intraclass correlations (ICC) and reliabilities ( $\lambda$ ) for level (intercept) and cross-time change (slope) in measures of public denigration and private neglect across census block groups for various time windows.

	<b>Housing</b>			<b>Uncivil Use</b>			<b>Big Buildings</b>		
	<b>Intercept</b>		<b>Slope</b>	<b>Intercept</b>		<b>Slope</b>	<b>Intercept</b>		<b>Slope</b>
	<b>ICC</b>	<b><math>\lambda</math></b>	<b><math>\lambda</math></b>	<b>ICC</b>	<b><math>\lambda</math></b>	<b><math>\lambda</math></b>	<b>ICC</b>	<b><math>\lambda</math></b>	<b><math>\lambda</math></b>
<i>1 week</i>	.13	.91	.36	.02	.78	.31	.01	.46	.10
<i>2 weeks</i>	.23	.91	.36	.03	.78	.30	.02	.46	.10
<i>3 weeks</i>	.31	.91	.35	.05	.78	.30	.03	.46	.10
<i>1 month</i>	.39	.91	.36	.07	.78	.30	.03	.46	.11
<i>2 month</i>	.56	.91	.37	.13	.78	.29	.04	.46	.10
<i>3 month</i>	.65	.91	.37	.19	.77	.30	.10	.46	.09
<i>4 month</i>	.72	.91	.37	.25	.78	.29	.19	.46	.12
<i>6 month</i>	.77	.90	.33	.39	.74	.25	.31	.48	.01

**Table 5 (cont).**

	<b>Graffiti</b>			<b>Trash</b>			<b>Public Reporters</b>		
	<b>Intercept</b>		<b>Slope</b>	<b>Intercept</b>		<b>Slope</b>	<b>Intercept</b>		<b>Slope</b>
	<b>ICC</b>	$\lambda$	$\lambda$	<b>IC C</b>	$\lambda$	$\lambda$	<b>ICC</b>	$\lambda$	$\lambda$
<i>1 week</i>	.07	.87	.51	.05	.88	.48	.20	.96	.47
<i>2 weeks</i>	.13	.87	.51	.09	.88	.48	.32	.96	.40
<i>3 weeks</i>	.18	.87	.51	.13	.88	.48	.39	.96	.35
<i>1 month</i>	.24	.87	.51	.17	.88	.48	.47	.96	.30
<i>2 month</i>	.38	.87	.51	.30	.88	.48	.63	.95	.21
<i>3 month</i>	.47	.86	.51	.41	.88	.45	.68	.95	.03
<i>4 month</i>	.56	.87	.51	.47	.88	.47	.75	.95	.04
<i>6 month</i>	.60	.84	.56	.63	.86	.39	.80	.93	.01

Note: *N*'s vary based on the number of time intervals possible for the 28 month period in the database, nested in 541 census block groups. All ICCs significant at  $p < .001$ .

**Table 6.**

Intraclass correlations (ICC) and reliabilities ( $\lambda$ ) for level (intercept) and cross-time change (slope) in measures of public denigration and private neglect across census tracts for various time windows.

	<u>Housing</u>			<u>Uncivil Use</u>			<u>Big Buildings</u>		
	<u>Intercept</u>		<u>Slope</u>	<u>Intercept</u>		<u>Slope</u>	<u>Intercept</u>		<u>Slope</u>
	<b>IC C</b>	$\lambda$	$\Lambda$	<b>IC C</b>	$\lambda$	$\lambda$	<b>ICC</b>	$\lambda$	$\lambda$
<i>1 week</i>	.30	.97	.44	.06	.92	.38	.02	.68	.26
<i>2 weeks</i>	.46	.97	.45	.12	.92	.37	.04	.68	.27
<i>3 weeks</i>	.56	.97	.45	.17	.92	.37	.07	.68	.26
<i>1 month</i>	.64	.97	.46	.22	.92	.37	.09	.68	.27
<i>2 month</i>	.78	.97	.46	.35	.92	.36	.12	.68	.27
<i>3 month</i>	.84	.97	.44	.46	.91	.40	.24	.68	.24
<i>4 month</i>	.88	.97	.45	.55	.92	.37	.40	.68	.29
<i>6 month</i>	.90	.96	.36	.75	.90	.31	.41	.70	.01
<i>Composite</i> <sub>a</sub>	.78	.92	—	.64	.84	—	.20	.43	—



**Table 6 (cont).**

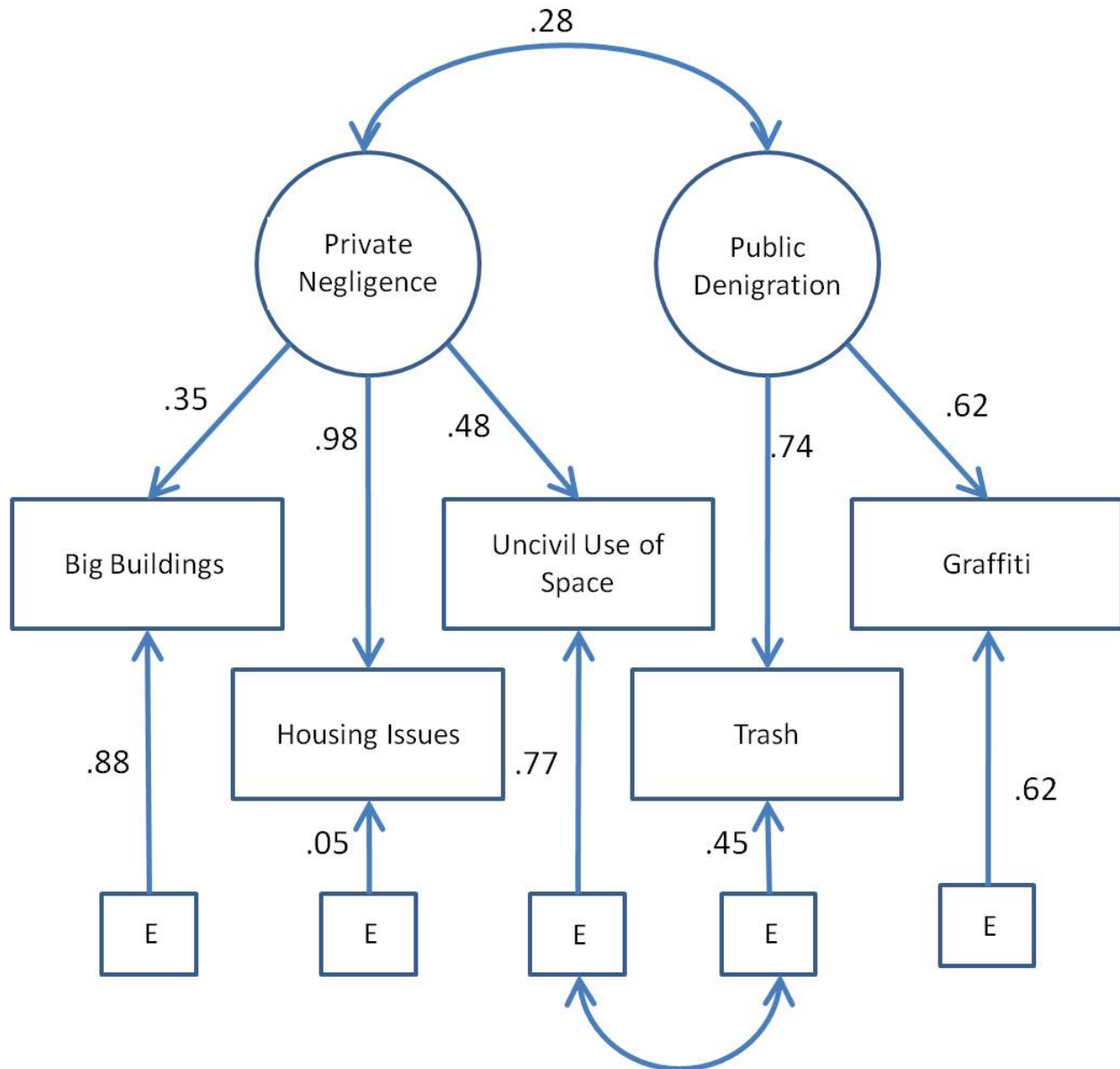
	Graffiti			Trash			Public Reporters		
	Intercept		Slope	Intercept		Slope	Intercept		Slope
	ICC	$\lambda$	$\lambda$	IC C	$\lambda$	$\lambda$	ICC	$\lambda$	$\lambda$
<i>1 week</i>	.18	.95	.71	.14	.96	.68	.43	.99	.48
<i>2 weeks</i>	.31	.95	.71	.25	.96	.67	.59	.99	.39
<i>3 weeks</i>	.39	.95	.71	.32	.96	.67	.67	.99	.34
<i>1 month</i>	.49	.95	.71	.40	.96	.67	.73	.98	.28
<i>2 month</i>	.64	.95	.72	.59	.96	.67	.84	.98	.06
<i>3 month</i>	.73	.95	.72	.67	.96	.62	.88	.98	.07
<i>4 month</i>	.79	.95	.71	.74	.96	.66	.90	.98	.05
<i>6 month</i>	.86	.94	.77	.86	.95	.55	.93	.98	.00
<i>Composite</i> <sup>a</sup>	.53	.77	—	.59	.82	—	—	—	—

<sup>a</sup> – A combination of the raw count and the measures of concern for the public space, calculated for six-month windows only. See text for more details on construction.

Note: *N*'s vary based on the number of time intervals possible for the 28 month period in the database, nested in 156 census tracts. All ICCs significant at  $p < .001$ .

**Figure 1.**

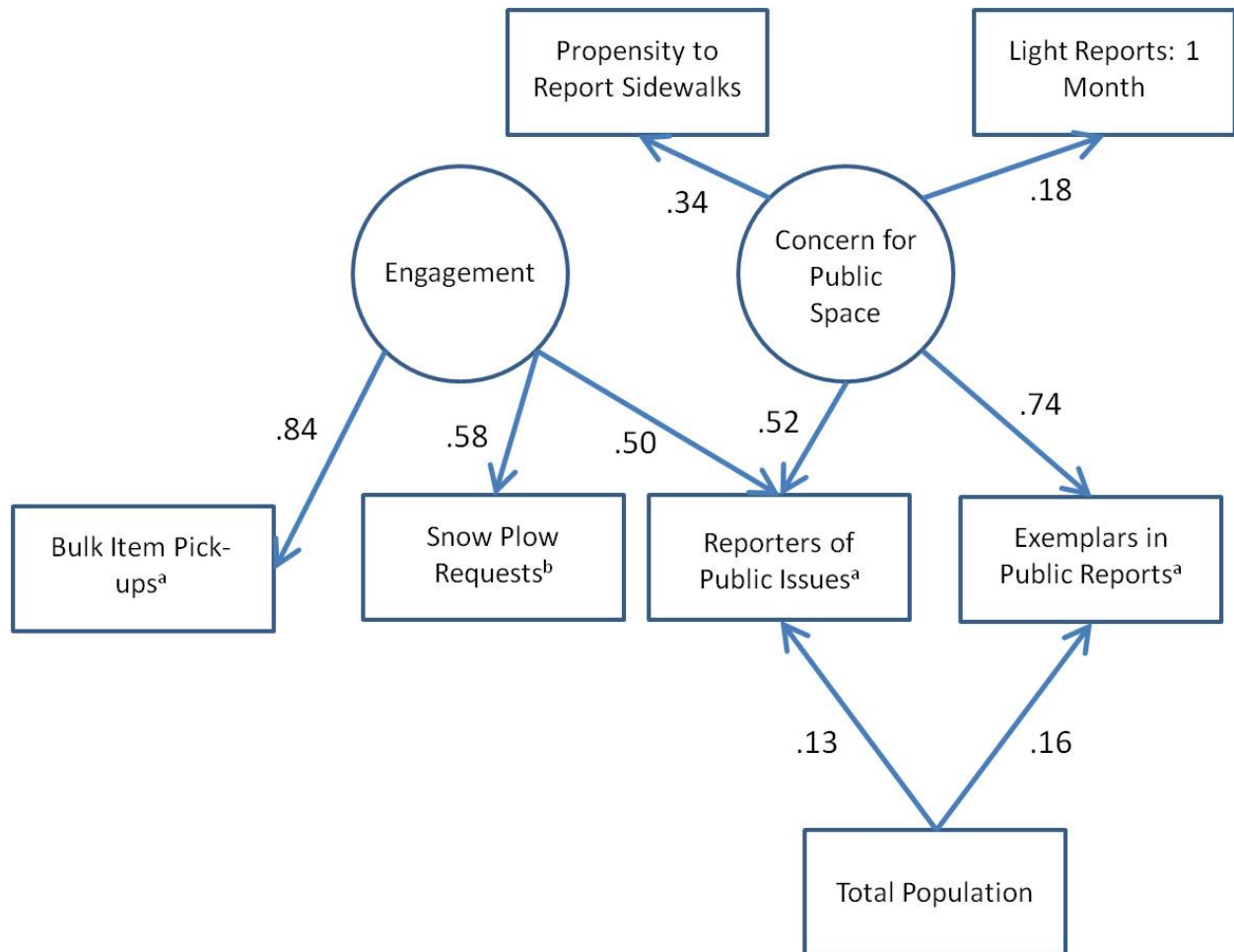
Estimated relationships between categories of physical disorder with standardized parameters from best-fitting confirmatory factor analysis.



Note: CFI = .95; SRMR = .05;  $N = 543$  census block groups. All parameters significant at  $p < .001$ .

**Figure 2.**

Relationships between objective and CRM-derived measures of response rate with standardized parameters from the best-fitting structural equation model.



Note: CFI = .95, SRMR = .06;  $N = 195$  census block groups with measures on all variables. All parameters significant at  $p < .05$ .

<sup>a</sup> – Log-transformed before analysis

<sup>b</sup> – Controlled for total population, total street length, and \_\_\_dead end length before analysis.

## APPENDIX A

**Table A1. Case Types that Reflect an Issue in the Public Space**

Case Type	Count	Case Type	Count
Abandoned Bicycle	71	Park Safety Notifications	2
Abandoned Building	103	Parking Enforcement	685
Abandoned Vehicles	2233	Parking Meter Repairs	139
Bridge Maintenance	29	Parking on Front/Back Yards (Illegal Parking)	132
Building Inspection Request	822	Parks General Request	106
Catchbasin	13	Parks Lighting Issues	5
Construction Debris	101	Pavement Marking Maintenance	272
Empty Litter Basket	292	Pick up Dead Animal	1374
Exceeding Terms of Permit	68	Pigeon Infestation	29
Fire Hydrant	8	PWD Graffiti	160
General Lighting Request	460	Request for Litter Basket Installation	80
Graffiti Removal	3893	Request for Pothole Repair	4603
Highway Maintenance	3297	Request for Snow Plowing	7270
Illegal Auto Body Shop	46	Requests for Street Cleaning	953
Illegal Dumping	831	Requests for Traffic Signal Studies or Reviews	96
Illegal Occupancy	263	Roadway Repair	306
Illegal Posting of Signs	116	Rodent Activity	1241
Illegal Rooming House	177	Sidewalk Cover / Manhole	3
Illegal Use	62	Sidewalk Repair	1294
Illegal Vending	32	Sidewalk Repair (Make Safe)	2119
Improper Storage of Trash (Barrels)	1745	Sign Repair	1172
Install New Lighting	25	Snow Removal	2103
Misc. Snow Complaint	1407	Street Light Knock Downs	476
Missed Trash/Recycling/Yard Waste/Bulk Item	6211	Street Light Outages	8127
Missing Sign	671	Traffic Signal Repair	2585
New Sign, Crosswalk or Pavement Marking	976	Trash on Vacant Lot	121
New Tree Requests	831	Tree Emergencies	3446
Overflowing or Un-kept Dumpster	149	Tree Maintenance Requests	3336
Park Improvement Requests	3	Upgrade Existing Lighting	15
Park Maintenance Requests	87		

## APPENDIX B

Neighborhood audits identifying street light outages and assessing levels of street garbage were conducted in 72 of Boston's 156 census tracts (46%) between June 1 and August 31, 2011 as part of an undergraduate seminar. The sample was constructed in a multistep process, intended to cover about half of the city, while capturing the full range of demographic, socioeconomic, and geographic diversity.

First, tracts were attributed to one of Boston's 16 planning districts, contiguous regions with characteristic demographic and socioeconomic profiles.<sup>12</sup> The population-weighted mean for tract-level median income was calculated for each planning district. A stratified sample of three or four tracts was then created for each planning district (depending on the size), including one tract more than a standard deviation above the local weighted mean for median income, one more than a standard deviation below, and either one or two within a standard deviation of the weighted mean. Because planning districts vary in the number of tracts they contain ( $min = 1$ ,  $max = 24$ ), the sample was completed by random selection from planning districts with a high number of tracts. The final sample was representative of the diversity across all Boston tracts, both in terms of its central tendency and range (See Table B1).

During each audit teams of two walked the streets of a particular tract. Highways, service roads, and other roads rarely used by pedestrians were omitted. One person walked each side of the street. The goal was to cover all other roads, though this was sometimes not possible given time constraints on audits. On each street segment (intersection-to-intersection or intersection–

---

<sup>12</sup> A distinction created by the Boston Redevelopment Authority for administrative purposes, but based on historically salient regions, many of which are once-independent municipalities that were annexed. Using an ANOVA, the planning districts account for about 50% of the variation in ethnic composition and median income across census tracts.

to-dead end), each person recorded the level of garbage and the presence of any street light outages on his or her side of the street. In total, 4,239 street segments were assessed. Garbage was rated on a 5-point scale, with higher scores indicating larger piles of garbage and more of them, for both the street and the sidewalk (if present). These measures were then adjusted for street sweeping.

The date of data collection was used in conjunction with the City's street sweeping schedule to fit a linear model that used the number of days since that side of the street was swept to predict level of garbage. The linear model indicated that streets swept within the past three days had lower-than-expected litter at the rate of .06/day on our scale. After three days had passed, there was no substantial difference in garbage ratings. Sidewalks were not adjusted in this fashion as they are not swept. Following this, an average of the adjusted street measure and the sidewalk measure on each side of the street was calculated as the total garbage rating for the street segment. Before data collection, inter-rater reliability was established through training PowerPoint slides and neighborhood walks.

**Table B1. Comparison of Demographic Characteristics between Tracts Sampled in Street Light Outage and Garbage Audits and All Tracts**

	<b>All Tracts</b>		<b>Sampled Tracts</b>	
	<b>Mean (SD)</b>	<b>Range</b>	<b>Mean (SD)</b>	<b>Range</b>
<i>Median Income</i>	\$52,572 (\$23,607)	\$10,250 - \$143,819	\$55,256 (\$27,436)	\$10,250 - \$143,819
<i>Population Density<sup>a</sup></i>	22.83 (16.24)	1.36 – 93.07	23.96 (17.55)	3.18 – 93.07
<i>% Homeowners</i>	.36 (.19)	.00 - .88	.38 (.21)	.00 - .88
<i>% White</i>	.51 (.31)	.00 - .99	.52 (.33)	.00 - .98
<i>% Black</i>	.21 (.25)	.00 - .92	.22 (.26)	.00 - .92
<i>% Hispanic</i>	.17 (.16)	.00 - .84	.15 (.1f)	.00 - .62

<sup>a</sup> – Thousands per sq. mil

## APPENDIX C

The street light outages, sidewalk reports, and garbage assessments all describe events or conditions on a single street segment within a CBG. To create CBG-level measures that controlled for the microspatial effects of street characteristics, multilevel models (Raudenbush et al. 2004) were developed in which two simultaneous equations were estimated, one at the level of streets (first-level), the second at the level of CBGs (second-level). The street-level equation was defined as:

$$Y_{jk} = \beta_{0k} + \sum_i \beta_i X_{ijk} + r_{jk}$$

$$r_{jk} \sim N(0, \sigma^2)$$

where  $Y_{jk}$  represents the  $j$ th street in CBG  $k$ , and  $\beta_{0k}$  is the estimated mean for neighborhood  $k$ . Each  $X_i$  is a first-level predictor, and each  $\beta_i$  is the corresponding regression parameter, explaining differences between streets within the same CBG. The errors of measurement  $r_{ij}$  for street  $j$  in neighborhood  $k$  are assumed to be normally distributed with variance  $\sigma^2$ . The estimated mean for neighborhood  $k$  is modeled as:

$$\beta_{0k} = \gamma_{00} + \mu_{0k}$$

$$\mu_{0k} \sim N(0, \tau)$$

where  $\gamma_{00}$  is the estimated mean value for the neighborhood-level measure across neighborhoods, and  $\mu_{0k}$  is the random neighborhood effect for neighborhood  $k$ . The latter can also be described as the deviation of the average value in neighborhood  $k$  from the cross-neighborhood mean. These random neighborhood effects are assumed to be normally distributed with variance  $\tau$ , and are the values extracted for the desired CBG-level measure. For example, in the case of garbage,  $\mu_{0k}$  indicates the extent to which the average street in CBG  $k$  has more or less loose garbage than the average street in the average neighborhood. In addition, the magnitude of  $\tau$  in relation to  $\sigma^2$  is



valuable in determining how well  $Y$  captures differences between neighborhoods. This is evaluated with a  $\chi^2$  test.

There were slight variations between the models for street light outages, sidewalks and garbage, including in first-level predictors and the link function used. For sidewalks, the sidewalk care index was the lone first-level predictor. The binary outcome (whether a sidewalk generated any reports) used a logit link, and the continuous outcome (how many reports a sidewalk generated) used a zero-inflated Poisson link. The models for both garbage and street light outages incorporated a dichotomous variables distinguishing between main and non-main streets, and between streets with different types of zoning. For garbage, dichotomous variables for all non-residential zonings were included (i.e., commercial, industrial, exempt, and un-zoned). To conserve degrees of freedom for the analysis of street light outages, this was simplified to a single dichotomous variable distinguishing between residential and non-residential zonings. The garbage model used a standard regression as garbage was a continuous, normal variable. The street light outage outcomes were dichotomous, necessitating a logit link.

In determining the proper event-level outcome to use as the basis for the CBG-level measure, there were multiple options for the street light outages and sidewalk reports. Multilevel models were run using each option and their results were compared.

For sidewalks, there were two candidate measures: whether a sidewalk polygon generated one or more reports (binary model); and if a polygon had generated any reports, how many it had generated (continuous model). Since not all CBGs contained a sidewalk that generated a request for repair, the continuous model only analyzed 416 CBGs. In each model, the sidewalk care index was entered as the sole first-level predictor, in order to control for the objective need for repair.

Both models indicated significant CBG-level variation, with the binary measure appearing to do so more effectively (binary:  $\chi^2_{df=541} = 940.47, p < .001$ ; continuous:  $\chi^2_{df=415} = 505.11, p < .01$ ). In addition, the binary measure was predicted by the sidewalk care index in the expected direction (i.e., a higher index predicts a lower likelihood of requests for repair), while the continuous measure was not (binary:  $\beta = -0.003, p < .01$ ; continuous:  $\beta = 0.001, p = ns$ ), suggesting it to be the superior measure for the subsequent analyses. Neighborhood-level residuals for this measure were extracted, with higher values indicating a CBG with a greater likelihood of requesting a sidewalk repair, controlling for quality of the sidewalk.

For street light outages, it was necessary to run the models at the tract level, owing to the low number of street light outages per CBG (244 in 127 CBGs and 56 tracts with outages<sup>13</sup>). The model was used to predict the likelihood of an outage being reported by a constituent at six time points after being identified: one week, two weeks, one month, two months, three months, and four months. Of these, the one-month ( $\chi^2_{df=54} = 78.39, p < .05$ ), two-month ( $\chi^2_{df=53} = 80.80, p < .01$ ), three-month ( $\chi^2_{df=53} = 73.95, p < .05$ ), and four-month ( $\chi^2_{df=53} = 73.30, p < .05$ ) models identified significant differences between tracts. Given the strong statistical similarity between the one-month and two-month measures, the former was selected for subsequent analyses because it indicates relatively quicker action on the part of constituents.

The continuous measure of garbage was also assessed for neighborhood-level variation. The model indicated significant CBG-level variation ( $\chi^2_{df=350} = 4,765.56, p < .001$ ). The neighborhood-level residual was extracted as the measurement of garbage.

---

<sup>13</sup> This number diminishes with some measures that allow greater time between identification of an outage and reporting, being that those that were reported by City employees in that time span were removed.