American Journal of Epidemiology

Vol. 180, No. 6
DOI: 10.1093/aje/kwu180
Advance Access publication:
August 13, 2014

Practice of Epidemiology

Validity of an Ecometric Neighborhood Physical Disorder Measure Constructed by Virtual Street Audit

Stephen J. Mooney, Michael D. M. Bader, Gina S. Lovasi, Kathryn M. Neckerman, Julien O. Teitler, and Andrew G. Rundle*

* Correspondence to Dr. Andrew G. Rundle, Department of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street, New York, NY 10032 (e-mail: agr3@columbia.edu).

Initially submitted January 15, 2014; accepted for publication June 9, 2014.

Neighborhood physical disorder is thought to affect mental and physical health, but it has been difficult to measure objectively and reliably across large geographical areas or multiple locales. Virtual street audits are a novel method for assessing neighborhood characteristics. We evaluated the ecometric properties of a neighborhood physical disorder measure constructed from virtual street audit data. Eleven trained auditors assessed 9 previously validated items developed to capture physical disorder (e.g., litter, graffiti, and abandoned buildings) on 1,826 block faces using Google Street View imagery (Google, Inc., Mountain View, California) dating from 2007–2011 in 4 US cities (San Jose, California; Detroit, Michigan; New York, New York; and Philadelphia, Pennsylvania). We constructed a 2-parameter item response theory scale to estimate latent levels of disorder on each block face and defined a function using kriging to estimate physical disorder levels, with confidence estimates, for any point in each city. The internal consistency reliability of the resulting scale was 0.93. The final measure of disorder was positively correlated with US Census data on unemployment and housing vacancy and negatively correlated with data on owner-occupied housing. These results suggest that neighborhood physical disorder can be measured reliably and validly using virtual audits, facilitating research on possible associations between physical disorder and health.

cities; data collection; epidemiologic methods; psychometrics; residence characteristics; social environment; spatial analysis; urban health

Abbreviations: CANVAS, Computer Assisted Visual Neighborhood Assessment System; IRT, item response theory.

In recent years, the epidemiology community has begun to assess the associations of physical disorder with health behaviors and outcomes, including associations with sexually transmitted infection incidence (1–3), obesity (4–6), and binge drinking (7, 8). Findings have been mixed; for example, one measure of disorder was unassociated with homicide rates in New York, New York (9), but a different measure was positively associated in Pittsburgh, Pennsylvania (10). As these results suggest, difficulties in measuring physical disorder objectively and reliably in multiple locales have led both to inconsistent findings and to difficulty comparing those findings between geographical locations (11).

Investigators use several methods to measure neighborhood physical disorder. First, some researchers have used data on neighborhood characteristics reported by study subjects (12, 13). While data can be efficiently collected in this manner, these data are subject to "same-source bias," which occurs when correlation exists between measurement error in self-reported individual health and behavior data and measurement error in self-reported neighborhood conditions (14–17). Furthermore, self-reported perception of neighborhood disorder may be influenced by stereotypes related to neighborhood racial composition (18).

Alternatively, measures of neighborhood physical disorder can be abstracted from governmental and commercial data sources and integrated into human health data sets using geographic information systems (4, 9, 19). However, such records are often collected for administrative purposes, may

not fully capture the construct of research interest, and may be collected at a spatial resolution that is not optimal for research purposes.

To avoid the limitations of self-reported and administrative data, researchers may employ trained observers to conduct in-person audits of neighborhood streets (7, 17, 20, 21). Because street audits assess only features visible at the time of the audit, measurement modeling techniques are typically employed to aggregate measurements of observable items to derive an estimate of a latent construct of interest and the amount of uncertainty in that estimate (14, 22). These techniques, derived from psychometric item response theory (IRT), are referred to as "ecometrics" in the context of neighborhood assessment (14, 22).

While the audit technique allows focus on specific constructs of researcher interest, systematic field audits can be expensive, especially if travel between cities is required; as a result, few studies have employed field audits, and those that have done so have focused on a single city. This prevents comparison of associations between neighborhood conditions and community health conditions across cities. However, recent studies have shown that it is possible to perform reliable street audits across multiple cities without requiring travel using visual imagery available online from sources such as Google Street View (Google, Inc., Mountain View, California) (23–29).

Combining ecometric measures with spatial interpolation offers more flexibility to define neighborhood boundaries. Spatial interpolation predicts the level of disorder on streets not sampled based on the spatial correlation of disorder ecometrically measured on a sample of streets, allowing researchers to estimate the level of physical disorder and confidence bounds around the estimate for all streets in a city (30). Estimating values for every street allows researchers to define neighborhoods according to their study populations and research questions, rather than rely on administrative boundaries (e.g., census tracts) (11, 30–34). The flexibility to define different boundaries also allows sensitivity analyses to assess the relationships between boundary choice and estimates of spatial associations (the "modifiable areal unit problem") (35).

In this paper, we assess the ecometric properties of a physical disorder measure constructed from virtual audits of 4 US cities, using an IRT model to combine items and kriging to spatially interpolate the resulting measure. We examine the internal consistency of the measure, spatial variation of the measure within each city, cross-validations of spatially interpolated results, and correlations between the physical disorder measure aggregated to census tracts within each city and 2010 US Census measures potentially associated with neighborhood physical disorder to assess the reliability and validity of the measure.

METHODS

Sample

The Computer Assisted Neighborhood Visual Assessment System (CANVAS) with Google Street View imagery, described in more detail elsewhere (M.D.M.B., unpublished manuscript, 2014), was used to virtually audit block faces within 4 US cities with varied spatial and economic profiles: New York, New York; Philadelphia, Pennsylvania; Detroit, Michigan; and San Jose, California. Sample points were selected in an approximately 2-km grid across each city, with a 1-km grid oversample in neighborhoods in the highest quartile of population density for the metropolitan area and a 0.5-km grid oversample in neighborhoods where subjects in the Fragile Families and Child Wellbeing Study resided (36). This sample resulted in 1,826 block faces with Google Street View imagery: 532 in New York, 503 in Philadelphia, 289 in San Jose, and 502 in Detroit. More details regarding sample selection are provided in Web Figure 1 (available at http://aje.oxfordjournals.org/). The variation in the number of block faces by city was a function of the geographical size of the city and the distribution of Fragile Families subjects within the city. Web Figure 2 provides a sample map of block faces selected for Philadelphia and the locations where no block face with Google Street View imagery was available.

Street auditing commenced in June 2012 and concluded in June 2013. Eleven auditors were trained and engaged in auditing, though no single auditor audited streets for the full year. We recorded Google Street View's report of the month and year in which imagery was captured at the starting point of each block face. Approximately 5% of the sampled block faces in each city were randomly selected to be part of a reliability subsample to be audited by all auditors auditing that city; this resulted in a subsample of 109 block faces, each audited by an average of 3.9 auditors. The remaining 95% of block faces were audited by 1 auditor each. Practical details regarding implementation of virtual street audits, including auditor training, protocol development, and the CANVAS system, are described more fully elsewhere (M.D.M.B., unpublished manuscript, 2014).

Measures

Nine virtual audit items designed to assess neighborhood physical disorder were developed from preexisting and validated scales: 7 items from the Project on Human Development in Chicago Neighborhoods (17) and 1 each from the Pedestrian Environment Data Scan tool (37) and the Irvine-Minnesota Inventory to Measure Built Environments (38). Items from the Project on Human Development in Chicago Neighborhoods were explicitly designed to assess physical disorder, while items from the Pedestrian Environment Data Scan and the Irvine-Minnesota Inventory (assessing vacant land and bars on windows, respectively) were added because of their potential relevance to neighborhood physical disorder. All items from the Pedestrian Environment Data Scan and the Irvine-Minnesota Inventory and some items from the Project on Human Development in Chicago Neighborhoods had previously been assessed for reliability in the virtual context (M.D.M.B., unpublished manuscript, 2014). Some items from the Project on Human Development in Chicago Neighborhoods were moderately reworded or condensed from previously tested items to enable virtual auditing within the CANVAS system while avoiding substantive change, similar to what has been done in other virtual audits (26). Training materials for our final audit protocol are available upon request.

To assess the potential for imagery limitations to undermine street audit reliability, we used 2 measures previously developed in the context of the CANVAS framework: an indication of camera quality problems (dark vs. bright, where 4 instances of "other" were recoded as bright after manual verification of the imagery by the data analyst) and the proportion of Street View vantage points along the block face that were obscured; obstructions were primarily parked cars.

Statistical analysis

Interrater reliability scores for each item were computed using average pairwise Cohen's κ values from the 109-block-face reliability subsample. Audit pairs wherein the imagery dates for the block face differed were excluded from the reliability analysis ($n = 3$; 0.5% of pairs). After computing interrater reliability scores, we recoded items in the reliability subsample for which there was not perfect agreement among raters ($n = 228$; 23% of items) for further analysis of the data. Where possible, the rating selected by the majority of raters was chosen to be the final rating for the block face; where raters were evenly split, we chose a response at random ($n = 44$; 19% of disagreements, 0.2% of observations overall). Because more raters contributed to the final reliability subsample ratings, ratings may have been somewhat more valid in this subsample. However, because reliability segments were chosen randomly and represented a small fraction of the overall sample, we do not expect this to have substantially influenced our findings. For interrater reliability analyses, "cannot tell" responses were considered to be missing responses and were excluded from analysis (range: from 0% of pairs for abandoned cars, building conditions, and bars on windows to 9.7% for beer or liquor bottles).

The internal consistency of the item set for the physical disorder items was assessed using Cronbach's α. To assess a latent level of disorder, all observations were modeled using an IRT model, in which the log odds of observing physical disorder for item i on block face j are modeled as a function of latent level of physical disorder θ_j, as follows:

$$\log\left(\frac{P(Y_{ij}=1|\theta_j)}{1-P(Y_{ij}=1|\theta_j)}\right)=\alpha_i(\theta_j-\beta_i).$$

When the model is fitted for all observed Y_{ij} values, α_i represents item i's discrimination, β_i represents item i's severity, and θ_j represents the latent level of physical disorder on block face j, which can be derived from the posterior distribution as described by Mislevy (39).

We selected an IRT model allowing for different discrimination levels between items due to an a priori expectation that items are differentially influenced by factors other than current conditions of physical disorder (e.g., the presence of protective bars on windows may reflect a history of neighborhood disorder in addition to current disorder, whereas graffiti is a more direct reflection of current disorder) and thus can be expected to have different levels of discrimination of physical disorder. We also tested a severity-only model, evaluating the fit of the 2 models using the Bayesian Information Criterion, which penalizes model complexity. We computed the internal consistency reliability of the IRT model as $1 - (1/I)$, where I represents the area under the total information curve (40).

To assess spatial autocorrelation of the latent construct measure estimate, we computed Moran's I within each city (41), considering the midpoint of the block to be the point observed. Next, we applied kriging methods to our data to estimate physical disorder scale scores across each city. Kriging is a geostatistical method that leverages the autocorrelation structure of spatially located observations to estimate values at unobserved locations (30, 42, 43). Following the procedure described by Bader and Ailshire (30), we first derived a semivariogram, or a plot measuring the degree to which observations covary as a function of separation distance, for distances ranging from 1,000 m to 100,000 m. From this semivariogram, we visually fitted an exponential curve with 3 parameters, as follows:

$$\gamma(h)=b+c\left(1-\exp\left\{\frac{-h}{\varphi}\right\}\right),$$
$$\text{where } b=\left\{\begin{array}{l} b_s, h>0 \\ b_m, h=0 \end{array}\right\}.$$

In this model, $\gamma(h)$ represents the variation in disorder as a function of the separation distance, h, between 2 points. The "nugget effect," b, represents the minimum variation as separation distance approaches zero. This nugget can be partitioned into b_s, the variation at a spatial scale too small to be captured in the spatial sample (e.g., the variation in disorder between 2 block faces on the same block), and b_m, variation due to measurement error. The sill—or maximum variation approached asymptotically at large separation distances—can be found by the sum of $b + c$. The range, φ, is the shape of the curve describing the increasing variation at larger separation distances, and it can be estimated visually by identifying where the variogram approaches the sill. Because the observations with the smallest separation distance carry the most weight in the kriging process, our judgment of visual fits emphasized a good fit at the distances under 2 km. To create maps, we estimated levels of disorder on a raster surface using 100-m^2 pixels across each city.

We used a jackknife resampling cross-validation strategy, comparing the measured value at each sampled point with estimates kriged using all other measured values to obtain an empirical estimate of kriging error. We assessed the cross-validation error's sample characteristics to ensure that mean error was approximately zero and that the distribution of error was approximately normal (30).

To assess the construct validity of our interpolated disorder measure, we assessed the correlation between the average interpolated disorder for all raster surface points falling in each US Census tract and Census characteristics we expected to be associated with disorder. We obtained the following characteristics from the 2010 Census and from 2006–2011 5-year American Community Survey estimates: unemployment rate (strong positive correlation expected (14)), housing vacancy (strong positive correlation expected (10)), population density (weakly positive correlation expected (17)), and owner-occupied housing rates (weak negative correlation expected (14)). We excluded tracts without housing units and

Table 1. Selected Characteristics of 1,826 Block Faces in 4 US Cities Virtually Audited for Neighborhood Physical Disorder Using Google Street View Imagery Dating From 2007–2011

Characteristic	Location									
	New York, NY ($n = 532$)		Philadelphia, PA ($n = 503$)		Detroit, MI ($n = 502$)		San Jose, CA ($n = 289$)		Overall ($n = 1,826$)	
	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)	%	Mean (SD)
High-resolution camera	66		42		67		88		64	
Imagery recorded in 2009 or later	67		42		98		93		73	
Time elapsed between image capture and block face audit, years		2.5 (1.7)		3.7 (0.8)		3.6 (0.4)		1.8 (0.7)		3.0 (1.3)
No. of Street View vantage points on block face		17.8 (11.6)		15.3 (12.2)		18.5 (9.8)		20.6 (12.9)		17.7 (11.6)
Proportion of vantage points obstructed	31		17		10		15		19	
Distance from start of block face to end of block face, m		190 (119)		171 (136)		204 (105)		207 (129)		192 (123)
Prevalence of autosampled segments requiring manual adjustment	11		4		9		8		8	

Abbreviations: CA, California; MI, Michigan; NY, New York; PA, Pennsylvania; SD, standard deviation.

Philadelphia Census Tract 364, which was 84.2% vacant due to redevelopment of the Philadelphia State Hospital at Byberry at the time of the 2010 Census.

Kriged interpolations cannot measure effects of small-scale variation and measurement error, and like all regression-based estimates, they do not represent the full variance of the estimated characteristic. Estimates that fail to account for this underrepresentation create an overly smooth estimation of disorder (30, 44). To account for oversmoothing, we created 10 potential values or "conditional realizations" based on the observed conditions at sampled locations (specific realizations are based on both the distance of the interpolated point from the sampled block face and the distance between sampled block faces). The disorder estimates based on these 10 conditional realizations represent multiple imputations of the disorder surface (30). We used these values to estimate the correlation between disorder and Census characteristics using a multiple-imputation framework (45).

All analyses were performed using R, version 2.15.3 (R Foundation for Statistical Computing, Vienna, Austria). The "ltm" package (version 0.9-9) was used for construction of IRT models (46), the "geoR" package (version 1.7-4) was used for spatial analysis and kriging (47), and the "mi" package (version 0.09-18) was used for estimation across imputations (48).

RESULTS

Block faces included in the sample

Our sampling process searched 2,060 locations across the 4 cities and selected 1,826 (88.6%) block faces for systematic audit. Visual inspection of maps generated by the sampling process to assess selection trends revealed that most grid sample locations where no auditable block face could be found were within large parks or industrial areas. Imagery capture dates for the block faces sampled ranged from July 2007 to October 2011; elapsed time between image capture and virtual audit ranged from 10 months to nearly 6 years, with a median of 3.2 years (Table 1). Imagery was generally older in Philadelphia and more recent in Detroit and San Jose. Blocks were shortest both in distance and in number of distinct images in Philadelphia, while view of the block face was obstructed (usually by parked cars) more often in New York than in the other cities, which may have decreased sensitivity to small-scale items in New York.

Ecometrics and spatial variation of the measure

The final interrater reliability analytical subsample included about 600 pairs of observations for each item. Individual-item average pairwise Cohen's κ values computed from the reliability subsample ranged from 0.34 ("fair agreement") to 0.80 ("substantial agreement") (49). Table 2 shows κ values for each item and the number of pairings used to compute each. In a sensitivity analysis, κ varied between block faces with high- and low-resolution imagery (Web Table 1); interrater reliability for the presence of bottles, graffiti, abandoned cars, burned-out buildings, and bars on windows was lower on block faces with high-resolution imagery.

The IRT scale's internal consistency scores were consistently in the 0.8–0.9 range, suggesting that the scale measures a coherent construct, and were higher than Cronbach's α scores, as expected given varying severities of assessed items. Bayesian Information Criterion comparison of the IRT models confirmed that a model in which discrimination was allowed to vary between items (Bayesian Information Criterion: 12,712.89) fitted the data better than a model with fixed discrimination (Bayesian Information Criterion: 12,881.77). These results indicate varying influence of forces other than physical disorder (including measurement error) on some items. Item severities ranged from −2.02 (very common) for the presence of litter to 3.80 for abandoned cars (very rare), and discrimination ranged from 0.36 (weak) for bars on windows to 2.45 (strong) for abandoned buildings.

Table 2. Items Measured on 1,826 Block Faces in 4 US Cities as Part of a Virtual Audit of Neighborhood Physical Disorder Using Google Street View Imagery Dating From 2007–2011, Including Audit Source and Interrater Reliability Score

Question Identification, Including Source	Full Question	Categorization	Frequency of Response Indicating Disorder in Overall Sample, %	Average Pairwise κ[a] Score in Reliability Subsample	No. of Pairs Used to Compute Pairwise κ[b]
PHDCN 1	Is there garbage, litter, or broken glass in the street or on the sidewalks?	Yes (1) vs. no (0)	89.1	0.35	642
PHDCN 2	Are there empty beer or liquor bottles visible in streets, yards, or alleys?	Yes (1) vs. no (0)	13.1	0.34	588
PHDCN 3	Is there graffiti, or evidence of graffiti that has been painted over, on buildings, signs, or walls?	Yes (1) vs. no (0)	41.1	0.55	626
PHDCN 4	Are there abandoned cars?	Yes (1) vs. no (0)	2.3	0.63	651
PHDCN 5	How would you rate the condition of most of the buildings on the block face?	Fair condition or poor/badly deteriorated condition (1) vs. very well kept/good condition or moderately well kept condition (0)	50.7	0.48	636
PHDCN 6	Do you see burned-out buildings in the block face?	Yes (1) vs. no (0)	1.3	0.69	635
PHDCN 7	Do you see boarded-up or abandoned buildings in the block face?	Yes (1) vs. no (0)	17.4	0.80	635
PEDS 1.7	Is there vacant or undeveloped land?	Yes (1) vs. no (0)	22.8	0.55	621
IMI 130	Do any buildings have windows with bars?	Yes (1) vs. no or not applicable (0)	27.7	0.53	651

Abbreviations: IMI, Irvine-Minnesota Inventory (31); PEDS, Pedestrian Environment Data Scan (13); PHDCN, Project on Human Development in Chicago Neighborhoods (10).

[a] Cohen's κ value.

[b] Items with more "cannot tell" responses resulted in fewer pairings.

Internal consistency reliability for the scale was 0.93 when pooling data from all cities, suggesting that a consistent construct underlies the scale. Item characteristic curves for the 2-parameter model, which estimate the probability of finding each item conditional on a latent level of physical disorder, are provided in Web Figure 3.

The measure displayed spatial autocorrelation in all cities; all Moran's I values were significantly positive ($P < 0.001$) in a 2-sided test. Moran's I values for the measure ranged from 0.030 for San Jose to 0.091 for Philadelphia, indicating that disorder was more strongly spatially patterned in Philadelphia than in San Jose. Detroit and Philadelphia showed the highest levels of spatial variation overall. Figure 1 presents the relationship of variance in the disorder measure with distances between sampled pairs (semivariograms) for each city. The "nugget effect" (i.e., the distance at which small-scale variance cannot be detected) was largest in Philadelphia, possibly reflecting the compact scale of Philadelphia's urban grid, and smallest in San Jose, a comparatively sprawling city. Sills were higher in Philadelphia and Detroit, where more extreme disorder was present, than in San Jose or New York. Overall, the median level of disorder was highest in Detroit and lowest in San Jose (Table 3).

Maps of estimated disorder indicated a spatial variation consistent with our knowledge regarding neighborhood disinvestment in all 4 cities. For example, Figure 2 maps the measure of physical disorder in Philadelphia, wherein the largest cluster of disorder is in the "Badlands," a neighborhood that suffers from abandonment (50). Jackknife cross-validation showed error to be approximately normally distributed for all 4 cities (Web Figure 4), with a mean of −0.002 and variance of 0.42 across all cities (Table 3). Median error was lowest in San Jose (median absolute deviation, 0.35) and highest in Detroit (median absolute deviation, 0.53).

Construct validity of the measure

Census measures were generally correlated as expected with the average estimated level of neighborhood physical disorder within a census tract across 10 conditional realizations (Table 4). For example, a scatterplot of the relationship between neighborhood housing vacancy and average physical disorder level by census tract in Philadelphia in one conditional realization is displayed in Figure 3 (see Web Figure 5 for scatterplots for the other 9 realizations). Correlations were closest to our a priori expectations in Philadelphia and most

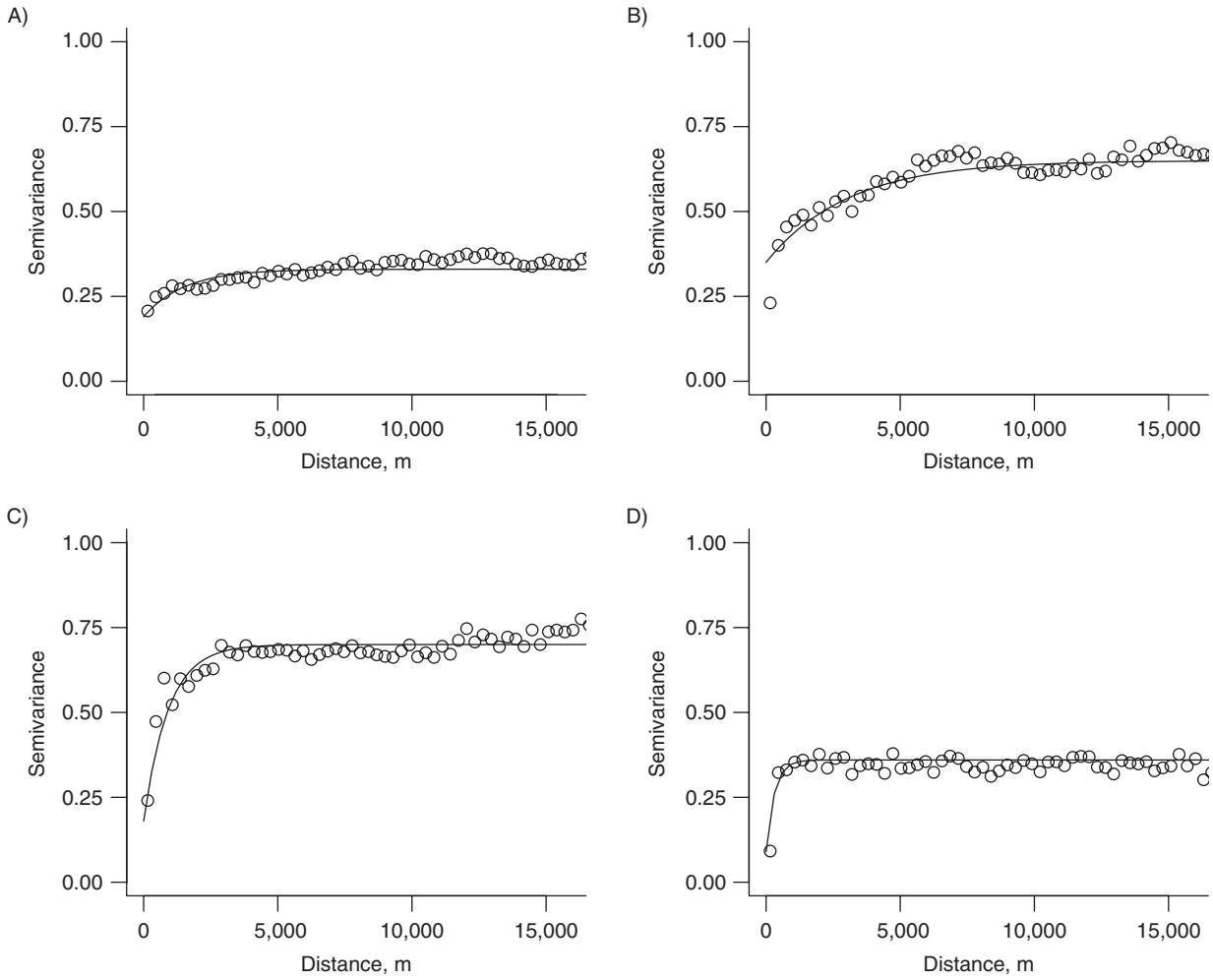

Figure 1. Variation in a measure of neighborhood physical disorder as related to distance between points (semivariograms) across 4 US cities: A) New York, New York; B) Philadelphia, Pennsylvania; C) Detroit, Michigan; and D) San Jose, California. The curve on each plot represents the exponential function visually fitted to that semivariogram. The measure was constructed using Google Street View imagery that was initially captured between 2007 and 2011.

different in Detroit and San Jose. Linear regression and correlation coefficients for each measure in each city are displayed in Table 4.

DISCUSSION

In this study, we explored the measurement properties of a scale measuring physical disorder in 4 US cities with data collected by virtual street audit using Google Street View and the CANVAS system. Items were generally reliable and consistent across all cities (per-item κ scores ranged from 0.34 to 0.80, while internal consistency reliability scores ranged from 0.86 to 0.89 in the 4 cities), and the interpolated measure covaried as expected when compared with both convergent and divergent census measures.

The interrater reliability of items included in our physical disorder scale ranged from κ values considered to represent "fair agreement" (for garbage and empty bottles) to levels at the top end of "substantial agreement" (for abandoned buildings) (49). This range of κ scores was comparable to scores Franzini et al. (6) observed for an in-person audit in 3 cities using similar measures of neighborhood physical disorder. Jones et al. (51) also reported on an in-person survey of physical disorder using similar measures; while the intraclass correlation coefficients they reported are not directly comparable to κ scores, we note that they did observe higher levels of agreement for the presence of abandoned buildings and lower levels for trash and empty bottles (51), generally following the interrater reliability trend we observed. Interrater reliability was worse on block faces with high-resolution imagery for most indicators, particularly the presence of bottles, graffiti, and bars on windows. This may be because high-resolution imagery presents more opportunities for some raters to detect small-scale items that would be missed by all raters in lower-resolution imagery. While it is frequently noted that κ scores trend lower for low-prevalence items (52), we observed

Table 3. Selected Ecometric and Spatial Properties of a Measure of Neighborhood Physical Disorder Computed From 1,826 Block Faces in 4 US Cities Using Google Street View Imagery Dating From 2007–2011

Location	No. of Block Faces Audited	Cronbach's α for Raw Measures	Internal Consistency of IRT Model	Median Level of Latent Disorder, *z* Score[a] (IQR[b])	Moran's *I*	Kriging Parameter			Cross-Validation Result	
						b (Nugget)	Φ (Range)	*c* (Sill – Nugget)	Residual Mean (SD)	Median Absolute Deviation
New York, NY	532	0.48	0.86	0.07 (−0.55 to 0.42)	0.041	0.19	5,000	0.14	−0.002 (0.54)	0.40
Philadelphia, PA	503	0.69	0.89	0.07 (−0.69 to 0.70)	0.092	0.35	10,000	0.30	−0.003 (0.66)	0.46
Detroit, MI	502	0.63	0.89	0.47 (−0.21 to 1.10)	0.063	0.18	3,000	0.52	0.001 (0.77)	0.53
San Jose, CA	289	0.48	0.87	−0.55 (−0.69 to 0.00)	0.030	0.09	1,000	0.27	−0.002 (0.57)	0.35
Overall	1,826	0.62	0.93	0.07 (−0.69 to 0.51)					−0.002 (0.65)	0.45

Abbreviations: CA, California; IQR, interquartile range; IRT, item response theory; MI, Michigan; NY, New York; PA, Pennsylvania; SD, standard deviation.

[a] *z* scores were calculated from the posterior distribution of the disorder IRT model.

[b] 25th–75th percentiles.

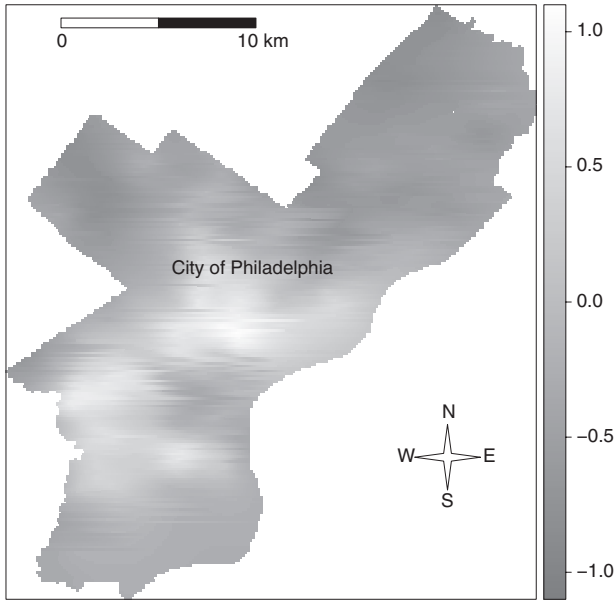

Figure 2. Interpolated levels of physical disorder in Philadelphia, Pennsylvania, constructed using Google Street View imagery that was initially captured between 2007 and 2011. Lighter areas have more physical disorder; the large, central cluster of physical disorder corresponds roughly with North Philadelphia, and the smaller clusters south and west of it correspond with West Philadelphia and the Gray's Ferry and Point Breeze areas of South Philadelphia.

κ values in the "substantial agreement" range (κ = 0.61–0.80) for our lowest-prevalence items: presence of abandoned cars (2.3%) and burned-out buildings (1.3%).

Measurement properties of the scale were encouraging. The internal consistency of the raw measures across all cities (α = 0.62) was lower than Franzini et al. (6) observed (α = 0.78). However, our IRT scale's consistency estimate was high (reliability = 0.93), consistent with prior ecometric measures of physical disorder using similar items (17, 30). The low α score of the raw items but high internal consistency reliability of the IRT scale may reflect the fact that these items serve as indicators of a consistent latent construct of physical disorder across a wide range of latent disorder levels. Indeed, item severities in the IRT model ranged from −2.0 to 3.8, suggesting that the scale was able to pick up on variation at a wide range of latent levels of physical disorder. Previous IRT scales of neighborhood physical disorder have also found a wide distribution of item severities (14, 51). The capacity of physical disorder scales to discern variation across the continuum of physical disorder is encouraging for the use of physical disorder in scientific research.

Our measure varied spatially in accordance with our expectations. Keyes et al. (7) reported a somewhat higher (Moran's $I = 0.14$) level of autocorrelation for a neighborhood disorder measure in Detroit constructed from street audit measures of abandonment. While Keyes et al. used a random sample of census blocks, we used a systematic sample, which can diminish observed autocorrelation (53).

Table 4. Correlation Coefficients and Slope Estimates[a] Comparing Mean Physical Disorder Levels Computed From 1,826 Block Faces in 4 US Cities (Using Google Street View Imagery Dating From 2007–2011) With Selected 2010 US Census Estimates for Each Tract[b]

Characteristic	Expected Correlation	Location									
		New York, NY ($n = 532$)		Philadelphia, PA ($n = 503$)		Detroit, MI ($n = 502$)		San Jose, CA ($n = 289$)		Overall ($n = 1,826$)	
		r	β	*r*	β	*r*	β	*r*	β	*r*	β
Unemployment rate	Strongly positive	0.20[c]	0.69	0.44[c]	0.97[c]	0.23[c]	0.50[c]	0.15	1.00	0.43[c]	1.15[c]
Housing vacancy rate	Strongly positive	−0.08[c]	−0.31[c]	0.71[c]	1.52[c]	0.44[c]	0.96[c]	0.04	0.12	0.38[c]	0.99[c]
Population density	Weakly positive	0.19[c]	0.67[c]	0.31[c]	0.69[c]	−0.15[c]	−0.29[c]	0.22[c]	1.73	0.09[c]	0.26[c]
Owner-occupied housing	Weakly negative	−0.40[c]	−1.43[c]	−0.30[c]	−0.67[c]	−0.03	−0.06	−0.20[c]	−1.15	−0.22[c]	−0.60[c]

Abbreviations: CA, California; MI, Michigan; NY, New York; PA, Pennsylvania.
[a] Slope estimates denote the estimated increase in census characteristic *z* score associated with a 1-unit change in physical disorder score.
[b] All estimates were computed from 10 conditional realizations using a multiple-imputation framework.
[c] $P < 0.05$.

Cross-validation results showed that errors were unbiased and normally distributed.

The correlations between our interpolated estimates of physical disorder, computed using conditional realizations to minimize bias due to oversmoothing, and US Census measures of tract-level unemployment, housing vacancy, population density, and owner-occupied housing were mostly in the expected direction. The direction and strength of association fitted our expectations better in Philadelphia and New York than in Detroit and San Jose. This was due in part to the fact that Philadelphia and New York had more variation in disorder than Detroit and San Jose and in part to the large-scale abandonment of some neighborhoods in Detroit, which creates an inverse association between disorder and population density. These differences highlight the different social processes at work across cities and the value of comparative research.

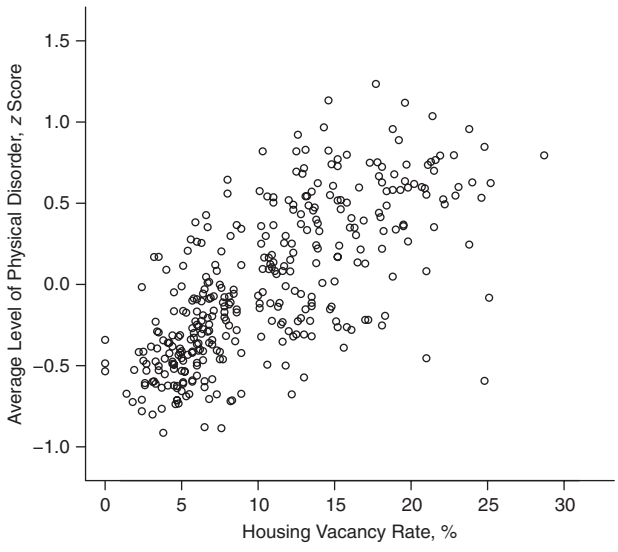

Figure 3. Average level of physical disorder as related to 2010 US Census estimates of housing vacancy rates by census tract in Philadelphia, Pennsylvania, estimated using Google Street View imagery dating from 2007–2011.

Our conclusions are strengthened by the consistency of our results across 4 cities with diverse spatial and socioeconomic profiles. Further, our use of geospatial techniques to construct convergent and divergent validity tests increases our confidence in the measure's validity not just at the sampled points but across each of the 4 cities.

Our study also had several important limitations. First, Google Street View imagery represents a view of a street at a particular time; some measures (e.g., the presence of litter or bottles) may be affected by time of day, which we were unable to assess with this method. Second, we were unable to assess physically small indicators of physical disorder that could have increased the precision of the measure (e.g., the presence of discarded hypodermic needles or condoms on the sidewalk). Third, the spatial interpolation procedure we used assumed that distance was the only driver of covariance and did not account for barriers (e.g., a river or highway) separating sampled points. We note, however, that any error due to diurnal variation, imagery limitations, or failure to account for barriers was incorporated into the measures used to assess divergent and convergent construct validity; the relatively strong correlations we observed somewhat mitigate our concern about these limitations. Fourth, a limitation inherent in street audits is that auditor or imagery characteristics may affect data integrity: An auditor familiar with a neighborhood may interpret imagery differently than an auditor to whom the neighborhood is unknown (54). Most of our raters were familiar with New York but not the other 3 cities; our consistent results across 4 cities with varied racial and socioeconomic composition mitigate this concern somewhat. However, in future research, investigators may consider whether residents of urban areas differ from suburban- or rural-area residents in their ratings of urban environments.

In conclusion, neighborhood physical disorder is of considerable interest to epidemiologists and other social scientists but has been expensive and difficult to assess reliably. Our virtual audit approach has yielded a measure of neighborhood physical disorder with desirable ecometric properties across multiple cities. Reliable and valid measurement

is a necessary precursor to investigations of neighborhood physical disorder's association with health outcomes.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York (Stephen J. Mooney, Gina S. Lovasi, Andrew G. Rundle); Department of Sociology and Center on Health, Risk and Society, American University, Washington, DC (Michael D. M. Bader); Columbia Population Research Center, Columbia University, New York, New York (Kathryn M. Neckerman); and School of Social Work, Columbia University, New York, New York (Julien O. Teitler).

Funding for this project was provided by the Eunice Kennedy Shriver National Institute of Child Health and Human Development under grants R21HD062965 and K01HD067390 and development grant R24HD058486, awarded to the Columbia Population Research Center; and by the National Cancer Institute under grant T32-CA09529. We also acknowledge the support of the Robert Wood Johnson Health & Society Scholars Program.

We thank the virtual street auditors: Dominic Abordo, Jonathon Costa, Payal Desai, Silvano DiMonte, Emerald Gearing, Marcelo Gelormini, Erin Huie, Lizzy Iuppa, Kelsey Kirkwood, Shiva Kooragalaya, and Daniel Wasserman.

Portions of this work were previously presented at the SERdigital Student Novel Methods Web Conference on November 6, 2013.

Conflict of interest: none declared.

REFERENCES

1. Cohen D, Spear S, Scribner R, et al. "Broken windows" and the risk of gonorrhea. *Am J Public Health*. 2000;90(2):230–236.
2. Bobashev GV, Zule WA, Osilla KC, et al. Transactional sex among men and women in the South at high risk for HIV and other STIs. *J Urban Health*. 2009;86(suppl 1):32–47.
3. Latkin CA, Curry AD, Hua W, et al. Direct and indirect associations of neighborhood disorder with drug use and high-risk sexual partners. *Am J Prev Med*. 2007;32(6 suppl): S234–S241.
4. Lovasi GS, Bader MD, Quinn J, et al. Body mass index, safety hazards, and neighborhood attractiveness. *Am J Prev Med*. 2012;43(4):378–384.
5. Burdette AM, Hill TD. An examination of processes linking perceived neighborhood disorder and obesity. *Soc Sci Med*. 2008;67(1):38–46.
6. Franzini L, Elliott MN, Cuccaro P, et al. Influences of physical and social neighborhood environments on children's physical activity and obesity. *Am J Public Health*. 2009;99(2):271–278.
7. Keyes KM, McLaughlin KA, Koenen KC, et al. Child maltreatment increases sensitivity to adverse social contexts: neighborhood physical disorder and incident binge drinking in Detroit. *Drug Alcohol Depend*. 2012;122(1–2):77–85.
8. Reboussin BA, Preisser JS, Song EY, et al. Geographic clustering of underage drinking and the influence of community characteristics. *Drug Alcohol Depend*. 2010;106(1):38–47.
9. Cerdá M, Tracy M, Messner SF, et al. Misdemeanor policing, physical disorder, and gun-related homicide: a spatial analytic test of "broken-windows" theory. *Epidemiology*. 2009;20(4): 533–541.
10. Wei E, Hipwell A, Pardini D, et al. Block observations of neighbourhood physical disorder are associated with neighbourhood crime, firearm injuries and deaths, and teen births. *J Epidemiol Community Health*. 2005;59(10):904–908.
11. Entwisle B. Putting people into place. *Demography*. 2007; 44(4):687–703.
12. Weir LA, Etelson D, Brand DA. Parents' perceptions of neighborhood safety and children's physical activity. *Prev Med*. 2006;43(3):212–217.
13. Gómez JE, Johnson BA, Selva M, et al. Violent crime and outdoor physical activity among inner-city youth. *Prev Med*. 2004;39(5):876–881.
14. Raudenbush SW, Sampson RJ. Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociol Methodol*. 1999; 29(1):1–41.
15. Shenassa ED, Liebhaber A, Ezeamama A. Perceived safety of area of residence and exercise: a pan-European study. *Am J Epidemiol*. 2006;163(11):1012–1017.
16. Blacksher E, Lovasi GS. Place-focused physical activity research, human agency, and social justice in public health: taking agency seriously in studies of the built environment. *Health Place*. 2012;18(2):172–179.
17. Sampson RJ, Raudenbush SW. Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. *Am J Sociol*. 1999;105(3):603–651.
18. Sampson RJ, Raudenbush SW. Seeing disorder: neighborhood stigma and the social construction of "broken windows." *Soc Psychol Q*. 2004;67(4):319–342.
19. Lovasi GS, Schwartz-Soicher O, Neckerman KM, et al. Aesthetic amenities and safety hazards associated with walking and bicycling for transportation in New York City. *Ann Behav Med*. 2013;45(suppl 1):S76–S85.
20. Pikora TJ, Bull FC, Jamrozik K, et al. Developing a reliable audit instrument to measure the physical environment for physical activity. *Am J Prev Med*. 2002;23(3):187–194.
21. Reiss AJ. Systematic observation of natural social phenomena. *Sociol Methodol*. 1971;3(1):3–33.
22. Mujahid MS, Diez Roux AV, Morenoff JD, et al. Assessing the measurement properties of neighborhood scales: from psychometrics to ecometrics. *Am J Epidemiol*. 2007;165(8): 858–867.
23. Rundle AG, Bader MD, Richards CA, et al. Using Google Street View to audit neighborhood environments. *Am J Prev Med*. 2011;40(1):94–100.
24. Badland HM, Opit S, Witten K, et al. Can virtual streetscape audits reliably replace physical streetscape audits? *J Urban Health*. 2010;87(6):1007–1016.
25. Kelly CM, Wilson JS, Baker EA, et al. Using Google Street View to audit the built environment: inter-rater reliability results. *Ann Behav Med*. 2013;45(suppl 1):S108–S112.
26. Odgers CL, Caspi A, Bates CJ, et al. Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method. *J Child Psychol Psychiatry*. 2012;53(10):1009–1017.
27. Wilson JS, Kelly CM, Schootman M, et al. Assessing the built environment using omnidirectional imagery. *Am J Prev Med*. 2012;42(2):193–199.
28. Clarke P, Ailshire J, Melendez R, et al. Using Google Earth to conduct a neighborhood audit: reliability of a virtual audit instrument. *Health Place*. 2010;16(6):1224–1229.

29. Griew P, Hillsdon M, Foster C, et al. Developing and testing a street audit tool using Google Street View to measure environmental supportiveness for physical activity. *Int J Behav Nutr Phys Act*. 2013;10:103.
30. Bader MDM, Ailshire JA. Creating measures of theoretically relevant neighborhood attributes at multiple spatial scales [published online ahead of print February 7, 2014]. *Sociol Methodol*. (doi:10.1177/0081175013516749).
31. Sampson RJ, Morenoff JD, Gannon-Rowley T. Assessing "neighborhood effects": social processes and new directions in research. *Annu Rev Sociol*. 2002;28:443–478.
32. O'Campo P. Invited commentary: advancing theory and methods for multilevel models of residential neighborhoods and health. *Am J Epidemiol*. 2003;157(1):9–13.
33. Diez Roux AV. Estimating neighborhood health effects: the challenges of causal inference in a complex world. *Soc Sci Med*. 2004;58(10):1953–1960.
34. Lovasi GS, Grady S, Rundle A. Steps forward: review and recommendations for research on walkability, physical activity and cardiovascular health. *Public Health Rev*. 2012;33(2): 484–506.
35. Fotheringham AS, Wong DWS. The modifiable areal unit problem in multivariate statistical analysis. *Environ Plan A*. 1991;23(7):1025–1044.
36. Reichman NE, Teitler JO, Garfinkel I, et al. Fragile families: sample and design. *Child Youth Serv Rev*. 2001;23(4–5): 303–326.
37. Clifton KJ, Smith ADL, Rodriguez D. The development and testing of an audit for the pedestrian environment. *Landscape Urban Plan*. 2007;80(1–2):95–110.
38. Day K, Boarnet M, Alfonzo M, et al. The Irvine-Minnesota Inventory to Measure Built Environments: development. *Am J Prev Med*. 2006;30(2):144–152.
39. Mislevy RJ. Bayes modal estimation in item response models. *Psychometrika*. 1986;51(2):177–195.
40. Wainer H, Dorans NJ, Green BF, et al. *Computerized Adaptive Testing: A Primer*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 1990:166.
41. Moran PA. Notes on continuous stochastic phenomena. *Biometrika*. 1950;37(1–2):17–23.
42. Jerrett M, Burnett RT, Ma R, et al. Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology*. 2005; 16(6):727–736.
43. Auchincloss AH, Diez Roux AV, Brown DG, et al. Filling the gaps: spatial interpolation of residential survey data in the estimation of neighborhood characteristics. *Epidemiology*. 2007;18(4):469–478.
44. Lantuéjoul C. *Geostatistical Simulation: Models and Algorithms*. Berlin, Germany: Springer Verlag; 2002.
45. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 2nd ed. Boca Raton, FL: CRC Press; 2003:532–536.
46. Rizopoulos D. ltm: an R package for latent variable modeling and item response theory analyses. *J Stat Softw*. 2006;17(5): 1–25.
47. Ribeiro PJ Jr, Diggle PJ. geoR: a package for geostatistical analysis. *R News*. 2001;1(2):14–18.
48. Su Y-S, Yajima M, Gelman AE, et al. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *J Stat Softw*. 2011;45(2):1–31.
49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
50. Neuman SB, Celano DC. Worlds apart: one city, two libraries, and ten years of watching inequality grow. *Am Educator*. 2012; 36(3):13–19.
51. Jones M, Pebley AR, Sastry N. Eyes on the block: measuring urban physical disorder through in-person observation. *Soc Sci Res*. 2011;40(2):523–537.
52. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46(5):423–429.
53. Fortin M-J, Drapeau P, Legendre P. Spatial autocorrelation and sampling design in plant ecology. *Vegetatio*. 1989;83(1–2): 209–222.
54. Neckerman KM, Lovasi GS, Davies S, et al. Disparities in urban neighborhood conditions: evidence from GIS measures and field observation in New York City. *J Public Health Policy*. 2009;30(suppl 1):S264–S285.