CrossMark

ORIGINAL PAPER

Measuring Disorder: Observer Bias in Systematic Social Observations at Streets and Neighborhoods

E. M. Hoeben[1,2] · W. Steenbeek[2] · L. J. R. Pauwels[3]

Abstract

Objectives The present study focuses on Systematic Social Observation (SSO) as a method to investigate physical and social disorder at different units of analysis. The study contributes to the aggregation bias debate and to the 'social science of ecological assessment' in two ways: first, by presenting a new model that directly controls for observer bias in ecological constructs and second, by attempting to identify systematic sources of bias in SSO that affect the valid and reliable measurement of physical and social disorder at both street segments and neighborhoods.

Methods Data on physical disorder (e.g., litter, cigarette butts) and social disorder (e.g., loitering adults) from 1422 street segments in 253 different neighborhoods in a conurbation of the greater The Hague area (the Netherlands) are analyzed using cross-classified multilevel models.

Results Neighborhood differences in disorder are overestimated when scholars fail to recognize the cross-classified data structure of an SSO study that is due to allocation of street segments to observers and neighborhoods. Not correcting for observer bias and

---

✉ E. M. Hoeben
  ehoeben@nscr.nl

[1] Institutions, Inequalities, and Life Course Program (IIL), Amsterdam Institute for Social Science Research (AISSR), University of Amsterdam, P.O. Box 15508, Valckenierstraat 59, 1018 XE Amsterdam, The Netherlands

[2] Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Amsterdam, The Netherlands

[3] Department of Criminology, Penal Law and Social Law, Ghent University, Ghent, Belgium

observational conditions underestimates the disorder–crime association at street segment/grid cell level, but overestimates this association at the neighborhood level.

Conclusion Findings indicate that SSO can be used for measuring disorder at both street segment level and neighborhood level. Future studies should pay attention to observer bias prior to their data collection by selecting a minimum number of observers, offering extensive training, and collecting information on the urban background of the observers.

Keywords Physical and social disorder · Ecometrics · Neighborhoods · Systematic Social Observation · Crime

Introduction

Physical and social disorder has been related to mental and physical health problems (see overview given by Schaefer-McDaniel et al. 2010), community disruption (Steenbeek and Hipp 2011), fear of crime (Kelling and Coles 1996; Perkins and Taylor 1996), and crime itself (Skogan 1990). The broken window theory, which focuses on the associations between disorder and crime, has been influential in criminology and sociology and has also inspired a variety of policy programs (Braga et al. 2015). However, despite the societal and academic attention focused on disorder, a number of issues still hamper empirical studies of this phenomenon.

The current study takes up the call from Sampson and Raudenbush (1999, pp. 32) for the development of a 'science of ecological assessment'. One of the main tasks of the science of ecological assessment is dealing with the units of analysis by which phenomena and associations are measured and studied. This problem is inherent to ecological research because it lacks a natural unit such as a person. The current study adapts knowledge from psychometrics (e.g., concerning internal consistency or interrater reliability) to improve ecometric measures of disorder at the level of street segments and neighborhoods. Although many studies have implemented ecometrics, few have paid attention to observer bias in ecological constructs. The current study attempts to fill this gap by examining observer bias in Systematic Social Observation of physical and social disorder.

Systematic Social Observation (SSO) of disorder refers to systematically tallying all signs of disorder, such as cigarette butts, empty bottles, and litter in one location, for example, in a street segment or face block. The most important advantage of SSO over other methods for measuring disorder (census data, community surveys, and key informant interviews) is that it relies on the independent observation of locations by researchers, and not on conversations with respondents. It therefore does not have to deal with non-response, socially desirable answers or memory bias due to retrospective questioning. However, this is only the case if disorder observations obtained with SSO are not biased by the observers or other varying conditions; a disadvantage of the SSO method is that it is a snapshot in time. Some conditions under which observations are conducted may vary and bias the observation, such as the time of day, the day of the week, and the season in which the observation occurs (Jones et al. 2011; Raudenbush and Sampson 1999). Observers may bias the measures, because of their varying perceptions of disorder, or because of socialization or fatigue (Mastrofski et al. 2010; Spano 2005). A major shortcoming in most previous SSO studies is the lack of attention to sources of observer bias. Therefore, we built on previous research and present a refined model to directly control for observer bias

in ecological constructs. Data for the current study were collected in a conurbation in Europe: the areas surrounding The Hague in The Netherlands.

Theory

Disorder and Crime

Broken windows theory describes a process of urban decay in which signals of social disorder evoke fear of crime and fear of personal victimization. This causes a breakdown of community control as inhabitants turn away from what happens on the street (Wilson and Kelling 1982). The breakdown of community control provokes other forms of disorder as well as forms of crime, because such behavior is less likely to receive a response. Signs of disorder communicate to potential offenders that "no one cares". In the end, these processes of decreased control and increased disorder and crime result in a lack of confidence in police intervention and a more severe breakdown of community control (Skogan 1990; Wilson and Kelling 1982).

The empirical literature is still inconclusive about the direction of the relationship between disorder and crime. Although some studies suggest that disorder causes crime (e.g., Skogan 1990), and that reducing disorder helps to reduce crime rates (e.g., Braga et al. 1999), others have argued that the relationship is reciprocal (Boggess and Maskaly 2014); that disorder and crime are two ends of the same continuum caused by a third factor (Sampson and Raudenbush 1999), or that they are actually the same thing altogether (Gau and Pratt 2008). Even though the specifics of the disorder–crime relationship are still a subject to debate, the existence of a correlation between disorder and crime is well established (Skogan 2015), which emphasizes the importance of accurately measuring disorder.

Measuring Disorder Through Systematic Social Observation

In the 1980s and 1990s, Taylor, Perkins, and colleagues proposed Systematic Social Observation as a way to systematically observe physical and social disorder ('incivilities') at street block level (Perkins et al. 1992; Perkins and Taylor 1996; Taylor et al. 1984). Systematic Social Observation (SSO) refers to observation that is done systematically, in this case by filling in a checklist of disorder items. For example, 'Is litter present, yes or no?' Specific procedures dictate the unit of observation (e.g., streets, face blocks), the topic of observation (e.g., cigarette butts, dog feces), the duration of the observation (e.g., number of minutes) and the method of recording (e.g., on paper or by videotape; Reiss 1971). A typical SSO disorder study is organized as follows: in each neighborhood, a few locations are indicated as points of observation. The observers tally signs of disorder at these points of observation, for example, counting the number of empty bottles or abandoned bicycles. Points of observation can be houses, face blocks, or street segments. All points that have to be observed are allocated to a group of observers. To keep costs low, the number of observers is usually small, varying from a handful (e.g., Perkins and Taylor 1996) to a dozen (e.g., Clifton et al. 2007). This means that each observer visits tens to hundreds of locations, depending on the size of the research area. The small number of observers also means that observers visit multiple neighborhoods.

SSO differs from other methods for measuring disorder in several regards, and may be preferable to these methods in addressing specific topics. Census data are often not available for smaller areas. Key informant interviews[1] rely strongly on finding the appropriate respondents and, similarly to community surveys, run the risk of bias due to differing perceptions on the boundaries of the unit that is questioned (Coulton et al. 2001; see also the work on 'egohoods' of Hipp and Boessen 2013) or differing perceptions on types of disorder (e.g., What is graffiti? Do pieces count as well as tags? What defines 'a lot of' cigarette butts?). SSO relies on the observation of locations instead of on the interviewing of respondents. Therefore, by using SSO, we eliminate any issues with non-response, sampling decisions (e.g., whether researchers should interview adults versus adolescents, or new inhabitants versus individuals who lived in the area for a number of years), and the risk of socially desirable answers. Furthermore, SSO enables the precise recording of events prior to, during, and after the phenomena of interest and other conditions under which these phenomena are observed (Mastrofski et al. 2010; Reiss 1971). Measuring phenomena through interviewing residents or key informants is by definition retrospective and therefore filtered by "judgment and memory" (Carter et al. 1995, pp. 221). Thus, SSO may be a useful method if one wants to collect information about disorder at smaller levels of analysis (such as street segments), or that is unbiased by mental maps of the neighborhood, differing perceptions about disorder, retrospective questioning, or social desirability.

A disadvantage of the SSO method is that it gives information at one point in time, whereas an interview with a neighborhood resident may give an idea of the level of disorder over time. Replicability of SSO measurement is assumed because of its explicit procedures, disregarding the fact that some conditions under which observations are conducted may vary over time and thus bias the observation. This makes SSO more vulnerable to bias compared with methods that cover a longer period. Examples of biasing conditions are the time of day, day of the week, and the season in which the observation takes place (Jones et al. 2011; Raudenbush and Sampson 1999). Observers may also bias the observations. This will be elaborated on in the following section.

Sources of Observer Bias in Systematic Social Observation of Disorder

'Systematic' observations would be a lot less systematic if observers varied in their recordings of the topic of interest. Observer bias has even been referred to as the most serious challenge of field research (Spano 2005). Nevertheless, there has not been much attention paid to this problem in studies on SSO. This section summarizes three sources of observer bias: sources of *intra*-observer bias (socialization and fatigue), sources of *inter*-observer bias (based on individual characteristics and prior experiences), and reactivity. We concentrate on unintentional observer bias and do not take into account intentional bias caused by cheating.

First, sources of *intra*-observer bias—if observers change their observation over time—include observer socialization and fatigue. *Observer socialization*, also referred to as 'going native', occurs if observers change their attitude toward the topic of interest during

---

[1] The key informant interview is a method for data collection that collects information by interviewing one or more persons that have a central role in the community, and who are therefore able to provide specific information on the neighborhood as a whole or aspects of the neighborhood that are of specific interest. Key informants are not approached because they represent residents, but because of their professionally acquired knowledge on the topic of interest (Pauwels and Hardyns 2009; Tremblay 1957).

the project (Spano 2005). Over the course of a project, observers can become more sympathetic toward the topic under investigation. This may translate into increasing involvement with their research subjects, or even participation in activities under study (Adler and Adler 1987). *Fatigue*, or 'burnout cynicism', occurs if observers become bored or tired, and therefore less accurate in their recordings. Fieldwork can be mentally and physically demanding, because observers have to maintain focus, "be polite at all costs", "play the fool" in interaction with research subjects (Spano 2005, pp. 586), and, in the case of the current study, spend long periods of time walking outside on the streets and traveling from one research location to the other. Exhaustion may undermine observers' accuracy or memory, but may also trigger 'shirking', which occurs when observers unintentionally or intentionally reduce their workload by avoiding the recording of events that require additional coding (Mastrofski et al. 2010). Observer socialization and fatigue are both expected to result in less accurate and comprehensive data at later stages of the data collection. Therefore, we hypothesize that observers will report fewer signs of disorder as the number of observations increases over the course of the *project*, and over the course of the *day* (Hypothesis 1).

Second, sources of *inter*-observer bias—differences in observations between observers—can be found in observers' personal characteristics and prior experiences. Individual characteristics and prior experiences shape the feelings, images, and memories that observers bring to the field. These unconscious perspectives and thoughts may shape observers' judgments and understanding, and thereby bias observations (Hunt 1989). Empirical research on this form of observer bias in SSO is fairly limited. Mastrofski et al. (1996) investigated whether observers' personal views on community policing implementation biased their observation of police officers' community policing orientation and the officers' success in achieving compliance from citizens, but did not find evidence for such bias. On the other hand, Reiss (1971) found that observers' professional expertise (i.e., police training, a background in social science, or a background in law) affected their observation of police behavior. Additionally, in an experimental study, Yang and Pao (2015) investigated whether police officers perceived disorder (in photos) differently than students. This indeed appeared to be the case, and more so for social disorder than physical disorder. Studies on individual perceptions of neighborhood disorder generally derive information from community surveys, to examine whether some respondents are more likely to report disorder in their own neighborhood than others. Findings of these studies indicate that several demographics are indeed predictive of reports of disorder: females report more disorder than males (Hipp 2010; Sampson and Raudenbush 2004; Wallace et al. 2015; null-findings by Franzini et al. 2008; Latkin et al. 2009; reversed effect reported by Hinkle and Yang 2014), younger individuals report more disorder than older individuals (Hinkle and Yang 2014; Hipp 2010; Latkin et al. 2009; Sampson and Raudenbush 2004; Wallace et al. 2015; null-finding by Franzini et al. 2008), and individuals from ethnic minority backgrounds report less disorder (Franzini et al. 2008; Hipp 2010; Sampson and Raudenbush 2004; Wallace et al. 2015; null-finding by Hinkle and Yang 2014). The studies also point at other characteristics that may be relevant in determining individuals' perceptions of disorder, such as having a history of depression (Latkin et al. 2009) and marital status (Franzini et al. 2008; Latkin et al. 2009; Sampson and Raudenbush 2004; Wallace et al. 2015). It is unclear to what extent these findings are generalizable to the situation of SSO, as residents may perceive disorder very differently compared with observers who are conducting systematic observations (Hinkle and Yang 2014).

Building on the theoretical and empirical context outlined by these studies, we theorize that two features of observers may be relevant in explaining their perceptions of physical

and social disorder: their urban background, and their perceived vulnerability for victimization. Individuals may be cognitively adjusted to disorder in their own neighborhood and their experiences with disorder in prior neighborhoods (Taylor and Shumaker 1990). This may affect their assessment of disorder in other neighborhoods (Hipp 2010; Sampson and Raudenbush 2004) and potentially make them more aware of their surroundings in areas where the level of disorder is very different from what they perceive as normal. Experimental research suggests that individuals' urban background (the urbanicity of their own neighborhood) is indeed a relevant factor in making accurate inferences about communities based on signs of physical disorder (O'Brien et al. 2014). As the data collection for the current study took place in a highly urbanized area, we hypothesize that observers from similarly urbanized neighborhoods will report less disorder than observers from rural backgrounds (Hypothesis 2).

Observers' perceived vulnerability to victimization may also affect their observations of disorder. The literature on fear of crime suggests that some individuals are more aware of their surroundings than others because of personal safety reasons (Hale 1996; LaGrange and Ferraro 1989). Individuals who perceive themselves to be more vulnerable will be more aware of their surroundings, and potentially also of signs of physical and social disorder (Hipp 2010), as disorder gives a signal of potential threats in that area (Innes 2004). This idea is supported by empirical findings that females perceive more disorder than males (Hipp 2010; Sampson and Raudenbush 2004) and that residents who reported to 'feel unsafe' perceived more social disorder in their street segments than others (Hinkle and Yang 2014). In the current study, we test the hypothesis that females, as well as observers who perceive themselves to be more vulnerable toward victimization are likely to report more disorder (Hypothesis 3).

A third source of bias is not linked to characteristics or experiences of observers (such as the previously discussed sources of intra-observer and inter-observer bias), but to the reaction of the observed to the presence of the observer. This source of bias is referred to as 'reactivity', or the 'Hawthorne effect' (Spano 2003, 2005; Sykes 1978). Reactivity occurs if subjects change their behavior in reaction to the presence of the observer. For physical disorder, this would mean that the presence of observers alters, for example, people's littering behavior. We do not expect that the observations, as conducted in the present study, biased physical disorder through reactivity, because the observations were brief and not repeated over time. As for the observation of social disorder, we deem it possible, yet improbable, that beggars, prostitutes, or people under the influence of alcohol will leave the area if they see an observer taking notes. Furthermore, because the Systematic Social Observation of disorder usually takes place within a few minutes and does not require interaction with subjects, we expect that reactivity is not a major source of observer bias for the observation of disorder.[2]

Observer Bias in Ecological Constructs

Disorder measures derived from Systematic Social Observations, as well as from other sources such as community surveys, are generally used to construct aggregated measures:

---

[2] The observers were asked to note 'how people in the street acted toward them'. They could quote more than one category. For most of the observations, observers noted that people barely paid attention to them (quoted in 59% of the observations), only made eye contact (26%) or greeted them (14%). On few occasions, people asked the observers what they were doing (5.6% of the observations) or responded suspiciously to the presence of the observer (0.7% of the observations).
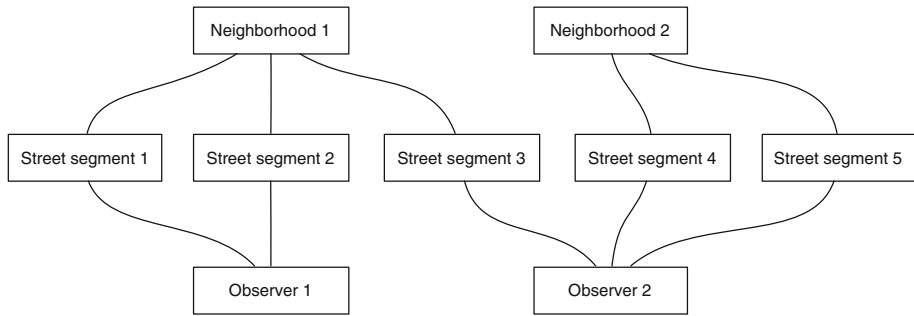
Fig. 1 Street segments nested within neighborhoods and observers (items are clustered in street segments)

for example, at the level of street segments, neighborhoods, census tracts or city districts. In assessing observer bias for these units of analysis, we are challenged with an additional issue: the allocation of observers to different areas. We previously noted that SSO disorder studies are often organized by letting a handful of observers observe tens to hundreds of locations. These observers conduct observations across several neighborhoods or census tracts. Figure 1 illustrates such an allocation of observers to different neighborhoods. In Fig. 1, neighborhood 1 consists of three street segments, two of which are observed by observer 1 and one that is observed by observer 2. Neighborhood 2 consists of two street segments that are both observed by observer 2. Assessment of observer bias in aggregated constructs of SSO disorder measures requires taking into account this cross-classified data structure (Fielding and Goldstein 2006).

Current Study

The current study contributes to existing research in three ways. First, the study presents an innovative ecological model to directly control for observer bias in observations of physical and social disorder. The model builds on the 'model for uncertainty in SSO', also called ecometrics, of Raudenbush and Sampson (1999), and refines it by taking into account the allocation of street segments to observers across neighborhoods. The model as proposed by Raudenbush and Sampson (1999) allows the studying of item inconsistency within a street segment, and street segment variation within neighborhoods. These goals were achieved by measuring a three-level hierarchical model with items at level 1, face blocks (in our case street segments) at level 2, and neighborhoods at level 3, with the control variable 'time' at the level of face blocks. Our refinement of the model of Raudenbush and Sampson (1999) is the addition of the 'observers' level, crossed with neighborhoods at level 3. This new crossed three-level model enables us to gauge the extent of observer bias (street segment variation within observers) and to explain this bias using several observer characteristics. Figure 1 is a schematic representation of the cross-classified data structure.

A second contribution to existing research is that the study thoroughly examines the extent to which Systematic Social Observations of disorder are biased by observational conditions (time, day, and weather), sources of intra-observer bias (fatigue or socialization effects over time; investigated by the effect of what number of observation it was during the entire project and on a specific day), and sources of inter-observer bias (urban background, gender, perceived vulnerability to victimization, and observers' feelings of safety at the observed locations).

Third, the study uses data collected in a European city, which extends the scope of earlier SSO studies on ecological disorder assessment that were mostly conducted in the United States.[3] To our knowledge, there have been no studies that have applied SSO to the measurement of crime and disorder in a European city.

In summary, the current study investigates whether Systematic Social Observation enables reliable and valid measurement of physical and social disorder at both the street segment level and the neighborhood level. With 'reliability', we refer to internal consistency of the measure and to ecological reliability, which is the extent to which the observed characteristics can be interpreted as characteristics of neighborhoods, as opposed to characteristics of the smaller units on which they are observed (in this case street segments). 'Validity' refers to whether a measure captures the idea contained in the intended concept. In the current study, validity specifically refers to the absence of systematic bias by observer characteristics or observational conditions and is also studied as nomological validity, with crime as a variable for validation.[4] To determine reliability and validity, we present a cross-classified model that takes into account observer bias.

Data and Methods

Sample

Data were collected as part of a larger NSCR research project: the Study of Peers, Activities and Neighborhoods (SPAN). The SPAN project used observation units (grid cells) of 200 by 200 m (656 by 656 feet), which were determined independently of the neighborhood boundaries as defined by the local government. The research area concerns the municipality of The Hague, the third largest city in the Netherlands, but also includes parts of the surrounding municipalities of Westland, Leidschendam-Voorburg, Delft, Wassenaar, Pijnacker-Nootdorp, and Rijswijk. The entire research area incorporated 4561 grid cells. The street segments at centroids of every third grid cell were observed, resulting in a total of 1422 street segments, spanning 253 neighborhoods (neighborhood boundaries as defined by Statistics Netherlands). A visual overview of the units of measurement is provided in Fig. 2.

Observers were instructed to walk 50 m to the left and 50 m to the right from a given address or location[5] (based on the centroid of a grid cell) and thereby observed and immediately coded street segments of 100 m, both sides of the street. Additionally, at

---

[3] A few exceptions are studies in Canada (Craig et al. 2002), Australia (Jago et al. 2005), the United Kingdom (Weich et al. 2001), and one study in the Netherlands (De Vries et al. 2007). See Schaefer-McDaniel et al. (2010) for an extensive overview of previous SSO studies. Of the studies that were *not* conducted in the United States, a majority concern health issues related to neighborhood characteristics.

[4] Numerous different types of validation have been suggested. We maintain the distinction of Adcock and Collier (2001), who distinguish (1) content validation, which is understood as 'focusing on the relationship between an indicator and the concept of interest' (pp. 539), and is concerned with whether all relevant elements are covered; (2) convergent/discriminant validation, which exists if indicators of a similar concept are empirically associated with each other (they are convergent) and more weakly associated to indicators of a second, different concept (they discriminate among concepts); (3) nomological/construct validation can be used 'in a domain of research in which a given causal hypothesis is reasonably well established' (pp. 542). Evidence for validity is found if the given hypothesis is again confirmed with the 'new', to be validated, indicator of one of the variables in the hypothesis. In the present study, we test nomological/construct validity by studying the relationship between disorder and crime.

[5] For 268 points of observation, observers were asked to observe a location based on the coordinates of the nearest street point, as there was no address within 200 m of the centroid.

Fig. 2 Units of measurement: neighborhoods, grid cells, and street segments. *Notes* The disorder data were collected at street segments located at the centroids of grid cells. For the comparison with crime rates, we make the assumption that the disorder measures at those street segments reflect the level of disorder in the entire grid cell. The disorder observations at street segments were aggregated to obtain measures of disorder at the neighborhood level. Crime rates reflect the registered number of offenses in public places that have been committed in the grid cells and neighborhoods between 2007 and 2009

every observation site, observers made four photographs (still images) with a camera equipped with a GPS device. The exact location of all observations was determined afterwards based on the recorded GPS coordinates. The photographs were not used to code disorder from. Each observation was carried out by one observer. Allocation of observers to neighborhoods and street segments occurred at random, while making sure the locations in one neighborhood were allocated to as many different observers as possible. In total, thirteen observers participated in the data collection, all of which were undergraduate or graduate students in the social sciences. The students were aged between 20 and 24 years, and twelve of them were of native Dutch descent. Six of the thirteen observers were female.

The data collection took place between March and June 2012. Observations were restricted to weekdays (Monday to Friday, except on holidays or during primary and secondary school vacations) between 10.00 a.m. and 4.00 p.m. Observations were not executed on days on which garbage was collected by the municipality. Observation of one street segment took on average 8 min and 9 s. The observation form included 61 items concerning land use, physical disorder, social disorder, physical condition of buildings, territoriality, traffic, formal and informal control, and guardianship. The instrument contained both dichotomous items (yes or no) and items with an ordinal scale (none, one, and more than one). A first version of the instrument was tested in a pilot study in September and October 2011.

Figure 1 gives a simplified representation of the cross-classified data structure. In the actual data, there were 253 neighborhoods, 1422 street segments, and thirteen observers. In

each neighborhood, on average 5.62 street segments were observed. Each of the 13 observers observed on average 109 street segments. The median number of different observers in one neighborhood was three; 64 of the 253 neighborhoods were observed by three different observers. An observer observed on average 58 different neighborhoods (median is 37).[6]

Training of Observers

Prior to the data collection, all observers were trained to improve inter-rater reliability. The training took 1.5 days. First, the observers were provided with an introduction into the theoretical background of the data collection (e.g., regarding broken windows theory). Second, we explained how they had to navigate to the centroids of the grid cells using address lists and GPS, provided them with safety information, and informed them on how to enter their observations into data collection software. Third, the observation form and corresponding protocols were discussed plenary. After this more theoretical introduction, the observers had to practice with coding disorder from pictures taken during the pilot study. They were confronted with common mistakes, also based on experiences from the pilot study, and were trained to apply operational definitions for specific forms of disorder. Subsequently, the observers were sent out to conduct field observations in groups of two or three observers. They separately coded disorder at the same time on the same location. Their observations were analyzed afterwards by two researchers, to investigate on what occasions observers differed in their coding. The next day, we held a group discussion on these irregularities as well as on questions about operational definitions brought forward by the observers. We also provided individual feedback on observers' performances and, when necessary, gave them additional instructions. This practice was suggested by Zenk et al. (2007) to promote inter-rater reliability.

During the data collection, 10% of the locations ($N = 147$) were observed twice, on different occasions by two different observers, to examine inter-rater reliability. Cohen's kappa was 0.731 for physical disorder and 0.957 for social disorder; percentages of agreement varied between 61.9 and 100.0 across the items. For more details about the data collection, see metadata at DANS (PID urn:nbn:nl:ui:13-wngr-5q) or at www.spanproject. nl.

Measures

The dependent variables were the items of a physical disorder construct and a social disorder construct. The physical disorder construct consisted of 7 items (e.g., dog feces, abandoned bicycles, and graffiti) and the social disorder construct consisted of 8 items (e.g., teenagers loitering and loud music playing). The items of physical disorder were initially measured on an ordinal scale (none; 1; more than 1), but were dichotomized because most items behaved as dichotomous items and because it was more consistent with the analyses for social disorder. All indicators of disorder were recoded to score 0 for 'not

---

[6] More specifically: In 32 neighborhoods we observed only one street segment (=minimum), whereas in one neighborhood we observed 51 street segments (=maximum). One observer observed 28 street segments (=minimum), the maximum amount of street segments observed by one observer is 226. Thirty-eight neighborhoods were observed by only one observer, whereas one neighborhood was observed by seven different observers. The minimum amount of neighborhoods that were observed by one observer is 20, and the maximum was 125.

observed' and score 1 for 'observed'. For a complete overview and frequency distribution of items per scale, see Table 1.

Independent variables were *observer characteristics* and *observational conditions* that potentially biased the disorder observations. Five observer characteristics were investigated. *Urban background* referred to the population density of the area where the observers grew up ('where did you live most of the years between birth and your 18th birthday? Please note down the address'), based on census data of Statistics Netherlands. Urbanicity of the area was expressed in five categories: 1 was 'very strongly urban' ($\geq$2500 addresses per km^2), 2 was 'strongly urban' (1500–2500 addresses per km^2), 3 was 'mixed rural and urban' (1000–1500 addresses per km^2), 4 was 'moderately rural' (500–1000 addresses per km^2), 5 was 'rural' (<500 addresses per km^2). *Gender* was a dichotomous variable that expressed whether the observer was male (1) or female (0). *Perceived chance of victimization* consisted of three items that each concerned a different type of victimization: victimization of threat, abuse, and burglary (e.g., 'how do you estimate your risk of becoming a victim of threat in the coming year?'). Each item originally had seven answer categories, varying from 'very big chance' (1) to 'very small chance' (7). As none of the observers scored 1, 2, or 3, the scale consisted of four categories, coded such that a higher score indicated a bigger perceived chance of victimization. *Perceived response to threat* consisted of one item: 'In the event of an assault on the street by a young, unarmed man, which of the following categories applies?' (1) I'm sure I'd be able to escape or to defend myself, (2) I'd probably be able to escape or to defend myself, (3) it depends, (4) I'd probably give in and do what he says, (5) I'm sure I'd give in and do what he says. This construct was derived from Killias and Clerici (2000) and translated to Dutch. The observers only scored in the categories 1, 2, and 3. The scale therefore consisted of three categories, coded such that a higher score indicated higher perceived vulnerability. *Feelings of safety at observation locations* expressed to what extent observers reported feeling safe in a street segment, varying from 'unsafe, not at all at ease' (1) to 'safe, completely at ease' (5). This was asked for every observed street segment, and we included observers' mean scores across all the sites they observed. Descriptives are given in Table A1 in the supplementary material.

Six observational conditions were examined: *time of day* referred to the hour in which the main part of the observation took place (between 10 a.m. and 4 p.m.); *day of week* referred to the weekday on which the observation took place; *weather condition* expressed the weather on the moment of observation, categorized with five different conditions: 'sun, clear blue sky'; 'sun with an incidental white cloud'; 'mainly cloudy, with sun shining through'; 'drizzle rain, sun shines through the clouds'; 'sky is completely clouded, clouds are grey, no sun shining through'. Observers were instructed not to perform observations in the case of snow, pouring rain or hail, or a thunderstorm. *Fatigue* and *socialization* were investigated by examining, at the street segment level, the effect of the number of observations the observers had already conducted, respectively, on that day and during the entire project. We also examined observers' *feeling of safety at that observation location*, as a deviation of the observers' overall reported feeling of safety across all of their observations. A positive deviation indicated that the observer felt safer at that location than average, and a negative deviation indicated that they felt less safe than usual. The last three 'observational conditions' are of course observational conditions as well as observer characteristics. As they were investigated at the street segment level, we discuss them as part of the observational conditions. Descriptives are given in Table A1 in the supplementary material.

Areal crime rates were used to investigate nomological validity of the disorder constructs. The crime rates were operationalized with police registered offenses in public places, committed between 2007 and 2009. These were the most recent available data; police data are not usually geocoded. The registered offenses had been reported by victims and bystanders, or were noted by the police. All data were geocoded with the exact location of where the crime had occurred. For the current study, we aggregated that information to count the number of crimes per grid cell—grid cells were 200 by 200 m, and this information was matched to the disorder observations conducted in the street segments at the grid cell centroids—and per neighborhood, with boundaries as defined by Statistics Netherlands (see also Fig. 2). The distinction in 'private', 'semi-public', and 'public' places was made by the police. 'Public places' are, for example, a market, parking lot, or train station. We specifically studied crime in public places based on the assumption that behavior in public spaces is more strongly related to the presence of disorder than behavior elsewhere. Additional analyses with 'general' crime showed substantially similar results as the ones presented in this paper.

Analytical Strategy

Three models were estimated for both physical and social disorder: (1) an empty three-level model with items at level 1, street segments at level 2, and both neighborhoods and observers at level 3 (cross-classified model); (2) the cross-classified model extended with one control variable at the observer level; (3) the cross-classified model extended with one control variable at the observer level and variables on observational conditions at the street segment level. To every model, item dummies were entered as independent variables, centered on their grand mean (following Raudenbush and Sampson 1999). Centering occurred separately for physical and social disorder. The hierarchical models as proposed by Raudenbush and Sampson (1999) were also estimated. We refer to those models as 'traditional ecometrics method' throughout the paper. Results are given in Tables B1 and B2 in the supplementary material.

Random intercept models were estimated with Markov Chain Monte Carlo (MCMC) procedures in MLwiN 2.20 (Browne 2012), using IGLS estimates as starting values. Logit functions were used because of the dichotomous nature of the physical and social disorder items; variance at level 1 was fixed (Snijders and Bosker 2012, Section 17.3). The posterior means of the Bayesian estimation are considered to be the best unbiased measures of disorder (Snijders and Bosker 2012, Section 4.8). These estimates are thus our adjusted measures of disorder, used to study disorder–crime correlations. The measures at street segment level are the sum of the posterior mean at observer level, the posterior mean at the neighborhood level, and the posterior mean at street segment level. The measures at the neighborhood level represent the posterior mean at the neighborhood level.

Findings

Descriptives

Table 1 shows the frequency distribution of the individual disorder items. The frequencies and percentages express in how many street segments these items were observed at least once. Signals of social disorder were far less frequently observed than signals of physical

Scales and items	Frequency	Percentage
Table 1 Occurrence of physical disorder and social disorder (*N* = 1422 street segments)		
Physical disorder		
Litter or broken glass	967	68.0
Cigarette butts	849	59.7
Dog feces	350	24.6
Empty bottles or cans	348	24.5
Graffiti—tags (small surface, tag)	211	14.8
Abandoned bicycles	44	3.1
Graffiti—pieces (big surface, piece)	36	2.5
Social disorder		
Adults loitering	110	7.7
Loud music	39	2.7
Teenagers loitering	30	2.1
People fighting or arguing	9	0.6
Smell of marihuana	3	0.2
People drinking alcohol	2	0.1
People using drugs	2	0.1
Beggars	1	0.1

Frequencies and percentages express in how many street segments these items were observed at least once

disorder. The frequency of the items is in line with findings of Raudenbush and Sampson (1999): more serious signals of disorder (e.g., abandoned bicycles, people using drugs) are reported less often than less serious signals of disorder (e.g., cigarette butts, adults loitering).

Variance Components: Street Segments, Neighborhoods, and Observers

One way to establish the presence of inter-observer bias is to investigate the variation in observed disorder between and within observers. But of course, disorder varies also between and within neighborhoods. As a first step in building our model, we therefore investigated the variance components of the disorder items. In other words, we investigated to what extent the total variance in disorder was attributed to (a) variance between observers, (b) variance between neighborhoods, and (c) variance between street segments.

Table 2 shows the variance components per disorder construct of an empty three-level model with items at level 1, street segments at level 2, and neighborhoods crossed with observers at level 3. Approximately 6.3% of the total variance in physical disorder reflects differences between neighborhoods (6.340, Table 2), and 12.4% (12.412) reflects differences between street segments. For social disorder, 17.2% (17.228) of the total variance reflects differences between neighborhoods and 24.2% (24.243) reflects differences between street segments. One can compare these findings with those derived from a traditional ecometrics model, which does not take into account the allocation of street segments to observers. Variance components in Table B1 in the supplementary material indicate that in a traditional ecometrics model, it appears that approximately 10.7% of the total variance in physical disorder reflects differences between neighborhoods, compared with 6.3% in the cross-classified model, and 27.7% reflects differences between street segments, compared with 12.4% in the cross-classified model. Differences for social disorder are less substantial. A traditional ecometrics model (as presented in Table B1 in the

Table 2 Variance components, internal consistency, and lambdas for physical disorder and social disorder ($N = 1422$ street segments)

	Physical disorder		Social disorder	
	Mean	SE	Mean	SE
Intercept	−1.732	0.077	−6.620	0.160
Variance between street segments	0.650	0.009	1.569	0.230
Variance between neighborhoods	0.332	0.005	1.115	0.106
Variance between observers	0.965	0.243	0.498	0.153
Total variance	5.237		6.472	
% of variance street segment	12.412		24.243	
% of variance neighborhood	6.340		17.228	
% of variance observer	18.427		7.695	
ICC neighborhood	0.063		0.172	
ICC neighborhood × observers	0.248		0.249	
Lambda street segment	0.806		0.886	
Lambda neighborhood	0.389		0.683	
Cronbach's Alpha	0.570		0.304	
DIC	7456.170		1450.890	

Variance at level 1 (items) is fixed at $\pi^2/3$. Results obtained from empty cross-classified multilevel logistic regression models

Calculation of the ICCs:

ICC neighborhood: variance neighborhood/total variance

ICC neighborhood × observers: (variance observers + variance neighborhood)/total variance

supplementary material) indicates that about 17.7% of the variance in social disorder is at neighborhood level, compared with 17.2% in the cross-classified model, and that about 29.9% is at street segment level, compared with 24.2% in the cross-classified model.

The intra-class correlation coefficients (ICCs), presented in Table 2, express the extent to which street segments within one neighborhood that are *observed by different observers* are alike (physical disorder: 6.3%, social disorder: 17.2%). These ICCs can be compared to the ICC neighborhood × observers values, which express the likeness of street segments within one neighborhood *observed by the same observer*. These ICCs are 24.8 and 24.9% for physical and social disorder (0.248 and 0.249), respectively. Thus, street segments within one neighborhood appear to be more similar when they are observed by the *same* observer (physical disorder: 24.8%; social disorder: 24.9%) compared to when they are observed by *different* observers (physical disorder: 6.3%; social disorder: 17.2%). This indicates the presence of observer bias.[7]

---

[7] As an anonymous reviewer pointed out, the variation among observers *might* be a function of true differences across street segments (e.g., one observer visited the only three disordered street segments in an otherwise tranquil neighborhood, and four other observers visited other street segments in that neighborhood that did not have disorder). However, in order for such a scenario to affect our analyses, this should occur for *many* observers and street segments. This is unlikely, because allocation of observers to neighborhoods and street segments occurred at random, while making sure the locations in one neighborhood were allocated to as many different observers as possible. Nevertheless, to examine this possibility, we investigated the median disorder rates and crime rates across the street segments allocated to different observers. We found that, despite the fact that some observers had been allocated to more 'criminal' grid cells than others, the observers sent to the most criminal areas were not the ones who recorded most disorder, and vice versa.

Ecological Reliability

We then assessed the ecological reliability of the disorder constructs at street segment level and neighborhood level, while taking into account the allocation of street segments to different observers. To do so, formulas were developed to calculate reliability measures—lambdas—that incorporate the cross-classified nesting of street segments in neighborhoods and observers. These formulas are presented in "Appendix". The ecological reliability measures (lambdas) for the disorder constructs are presented in Table 2. These lambdas are based on empty cross-classified models.[8] The values of lambda parameters vary, similar to the Cronbach's Alpha, between 0 and 1, where 0 is 'not reliable' and 1 is 'highly reliable'. Each street segment and neighborhood gets its own lambda; Table 2 presents the mean lambdas. The street segment lambda differs with the variance on street segment level and the number of items per scale: Eq. (2) in "Appendix". On average, the street segment lambdas are acceptably high for both physical and social disorder (respectively 0.806 and 0.886). The neighborhood lambdas differ with the unexplained variance in a neighborhood, with the amount of street segments that were observed in the neighborhood, and with the number of observers that observed the street segments in that neighborhood: Eq. (1) in "Appendix". The neighborhood lambdas for physical disorder vary between 0.137 and 0.676. For social disorder, they vary between 0.310 and 0.911. The average neighborhood lambda is acceptable for social disorder (0.683), but rather low for physical disorder (0.389). This indicates that the number of street segments or observers per neighborhood was low in the current study.

The key findings with regard to the proposed lambdas, which take into account the allocation of street segments to observers, are summarized in Fig. 3. Figure 3 shows how the neighborhood reliability scale for physical disorder behaves with varying numbers of street segments and observers per neighborhood. The graph clearly illustrates how important it is to include a sufficient number of observers in the data collection. Although the addition of street segments per neighborhood results in some improvement of the ecological reliability, the inclusion of extra observers is far more relevant. To give an example, imagine a data collection where four observers observe ten street segments on average per neighborhood. The average neighborhood lambda for physical disorder, given the variance components found in the current study, would then be 0.484. To improve reliability, one could either *increase the number of street segments per neighborhood*, where an increase of twenty street segments, resulting in an average of 30 street segments per neighborhood, would result in a lambda of 0.544. Alternatively, one could *increase the number of observers per neighborhood*, where the inclusion of two extra observers per neighborhood—for a total of six—would result in a lambda of 0.549. To obtain neighborhood lambdas of 0.6 or higher, a study would need at least 14 different street segments per neighborhood, allocated to at least 7 different observers, or 20 different street segments per neighborhood allocated to at least 6 different observers, given the variance components found in the current study.

Footnote 7 continued

Therefore, we have no reason to assume that observer differences in recordings of disorder are due to selective allocation of observers to street segments.

[8] These lambdas are not compared with lambdas from other models (e.g., models that include explanatory variables), because the fixed residual variance at level one complicates the comparison of sequential models (Snijders and Bosker 2012, Section 17.3.5).
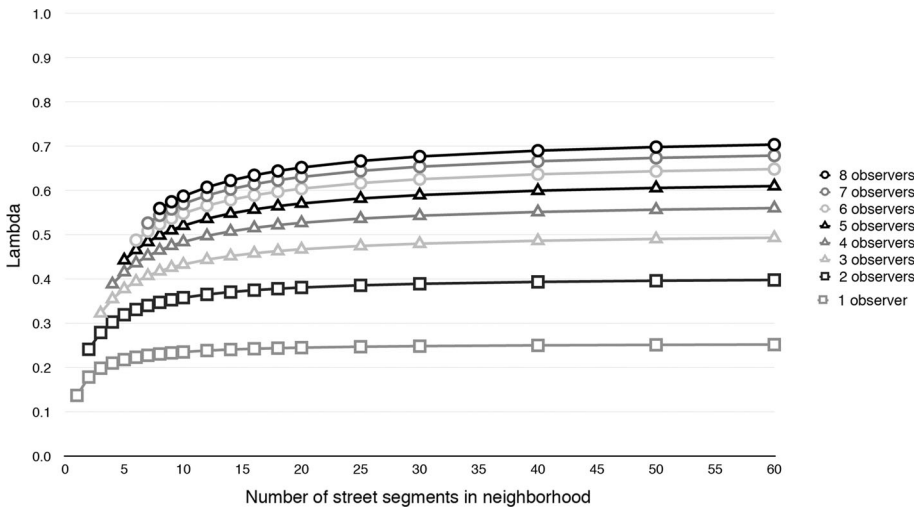
Fig. 3 Neighborhood reliability—lambda—for physical disorder as a function of sampled street segments and number of observers where the number of items is constant

Observer Characteristics

The findings of the variance components analyses suggest that the disorder observations differed between observers. We now turn to attempting to explain these differences by examining the influence of five observer characteristics: urban background, gender, perceived chance of victimization, perceived response to threat, and general feeling of safety at observation locations. As the number of observers was relatively low ($N = 13$), only one explanatory variable at a time was added to the observer level of the model. Results are presented in Table 3, and in Tables C1 (physical disorder) and C2 (social disorder) in the supplementary material. Results of additional Wald tests are presented in Table F1 in the supplementary material.

Of the investigated observer characteristics, only one appeared to affect the disorder observations, namely the urbanicity of the area where the observers grew up; their urban or rural background (results are presented in Table 3). The other characteristics did not appear to be relevant (the results for these characteristics are presented in Tables C1 and C2 in the supplementary material). It is possible that, due to the small number of observers, we have overlooked the effects of these characteristics. Nevertheless, based on these findings, we have to reject Hypothesis 3 regarding the effect of perceived vulnerability on the observation of disorder.

Regarding urban background, the results presented in Table 3 indicate that observers who grew up in moderately rural areas observed less physical disorder than observers from very strongly urban backgrounds ($\beta_{moderately\ urban}$: -1.456, $p < 0.05$). Also, we found that observers from strongly urban backgrounds observed less social disorder than observers from very strongly urban backgrounds ($\beta_{strongly\ urban}$: -1.237, $p < 0.05$). These findings contradict our Hypothesis 2, that observers from very strongly urban areas would observe fewer signs of physical and social disorder.

Table 3 Bias by observer characteristics and observational conditions (N = 1422 street segments)

	Physical disorder			Social disorder		
	β	SE	Odds ratio	β	SE	Odds ratio
Observer level						
Urban background						
Very strongly urban (ref)						
Strongly urban	−0.501	0.426	0.606	−1.237*	0.332	0.290*
Mixed rural and urban	0.658	0.413	1.931	−0.057	0.268	0.945
Moderately rural	−1.456*	0.373	0.233*	−1.019	0.288	0.361
Rural	−0.893	0.434	0.409	−1.014	0.324	0.363
Street segment level						
Time of day						
10.00–10.59	−0.024	0.029	0.976	−0.454	0.217	0.635
11.00–11.59	−0.134	0.022	0.875	−0.368	0.166	0.692
12.00–12.59 (ref)						
13.00–13.59	−0.077	0.022	0.926	0.663	0.138	1.941
14.00–14.59	0.188	0.027	1.207	1.150*	0.163	3.158*
15.00–16.00	0.400	0.042	1.492	0.970*	0.222	2.638*
Day of week						
Monday	−0.077	0.026	0.926	−0.353	0.162	0.688
Tuesday	0.361*	0.022	1.435*	0.014	0.144	1.014
Wednesday (ref)						
Thursday	0.274	0.023	1.315	0.386	0.127	1.471
Friday	0.004	0.025	1.004	0.218	0.144	1.244
Weather						
Sunny	−0.083	0.021	0.920	0.747*	0.118	2.111*
Sunny with clouds	−0.168	0.019	0.845	−0.263	0.136	0.769

Table 3 continued

	Physical disorder						Social disorder					
	β	SE	Odds ratio	β	SE	Odds ratio	β	SE	Odds ratio	β	SE	Odds ratio
Cloudy with sun (ref)												
Drizzle rain				−0.485	0.159	0.616				−1.344	2.152	0.261
Grey and cloudy				−0.211	0.018	0.810				−0.367	0.136	0.693
No. of observation total				−0.003*	0.000	0.997*				−0.006*	0.000	0.994*
No. of observation that day				−0.035*	0.000	0.966*				−0.043	0.001	0.958
Feeling at location[a]				−0.276*	0.008	0.759*				−0.589*	0.040	0.555*
Intercept	−1.122*	0.217	0.326*	−0.771	0.236	0.463	−5.907	0.240	0.003	−6.131*	0.551	0.002*
DIC	6313.32			6309.29			1268.67			1252.26		
Var. observer level	0.441			0.380			0.145			0.141		
Var. neighborhood level	0.409			0.316			1.010			0.460		
Var. street segment level	0.803			0.819			1.817			2.066		

Results obtained from cross-classified multilevel logistic regression models, β is expressed in log odds. Observations of both physical and social disorder are only affected by 'urban background'. Models with the other observer variables are presented in the supplementary material (Tables C1 and C2)

* $p < 0.05$ (two-tailed tests)

[a] Expressed as deviation of observers' general feeling of safety across all observed locations

Observational Conditions

We further investigated whether the conditions under which the observations took place would affect the disorder observations. Six observational conditions were examined: the time of day, day of the week, weather, the number of observations the observer had conducted prior to the observation across the entire project and during that day (as an expression of socialization or fatigue experienced by the observer at the moment of observation) and how safe the observer felt at that location. Results are presented in Table 3[9] and Table D1 in the supplementary material. The latter presents models that do not include observers' urban backgrounds. Results of additional Wald tests are presented in Table F1 in the supplementary material.

With regard to physical disorder (Table 3), we found that weather conditions did not affect systematic observations. Time of day appeared to be relevant, in that more disorder was observed between 3.00 and 4.00 p.m. than earlier in the day (more than between 11.00 a.m. and noon: $\chi^2 = 5.592$; and more than between 1.00 and 2.00 p.m.: $\chi^2 = 6.223$, Table F1 in the supplementary material). Day of the week appeared to be relevant, in that there was more physical disorder observed in the middle of the week than on Mondays or Fridays. More physical disorder was observed on Tuesdays than on Mondays ($\chi^2 = 7.211$, Table F1 in the supplementary material), more disorder was observed on Thursdays than on Mondays ($\chi^2 = 4.452$, Table F1), more disorder was observed on Tuesdays than on Fridays ($\chi^2 = 4.776$, Table F1), and more disorder was observed on Tuesdays than on Wednesdays ($\beta_{Tuesday}$: 0.361, $p < 0.05$, Table 3). Furthermore, we found that observers reported less physical disorder as their number of conducted observations increased. Although these effects were modest, they were visible both for the number of observations across the entire project ($\beta_{No.\ of\ observation\ total}$: -0.003, $p < 0.05$), and for the number of observations that had been conducted that day ($\beta_{No.\ of\ observation\ that\ day}$: -0.035, $p < 0.05$). These findings support Hypothesis 1 because they imply that fatigue or socialization may indeed be relevant factors in Systematic Social Observations of physical disorder. A larger effect was found for the effect of observers' feeling of safety at the observation location; the safer observers felt at a location, the less physical disorder they reported ($\beta_{Feeling\ at\ location}$: -0.276, $p < 0.05$). Note, however, that observers may feel safer in locations with fewer signs of physical disorder. We replicated all models without that variable, but its inclusion did not seem to affect the coefficients of the other predictors. Results of these additional analyses are presented in Table E1 in the supplementary material.

With regard to social disorder (Table 3), we found that the time of day and weather conditions affected the observations, but not day of the week. As we would expect, social disorder was more often observed in the afternoon than in the morning (significantly more often after 2 p.m. than around noon: $\beta_{2-3p.m.}$: 1.150, $p < 0.05$; $\beta_{3-4p.m.}$: 0.970, $p < 0.05$, significantly less often before noon than after 1 p.m., as illustrated by the findings of the Wald tests in Table F1 in the supplementary material), and more often when it was sunny compared with when it was sunny with clouds ($\chi^2 = 7.166$, Table F1 in the supplementary material), cloudy (β_{sunny}: 0.747, $p < 0.05$, Table 3) or grey ($\chi^2 = 9.911$, Table F1). We expect that these conditions are more predictive of the occurrence of social disorder than

---

[9] The results regarding observers' urban backgrounds should only be interpreted with findings from models that do not include the observational conditions. The coefficients for the variables at observer level, here urban background, are biased by the addition of the variables at street segment level (e.g., time, day, weather). This is due to our dichotomous dependent variables; the level one variance is fixed at $\pi^2/3$, therefore macro-level parameters change after adding micro-level variables (Snijders and Bosker 2012, Section 17.3.5).

that they explain bias in observations: people are more likely to be outside in the afternoon (after work or school) and when the weather is nice. Furthermore, we found that observers reported less social disorder as their number of conducted observations increased across the project ($\beta_{No.\ of\ observation\ total}$: -0.006, $p < 0.05$), which supports the notions of observer socialization or fatigue. We also found that observers' feeling of safety was negatively associated with the observed social disorder; the safer the observer felt at a location, the less social disorder was reported ($\beta_{Feeling\ at\ location}$: -0.589, $p < 0.05$). But, as we stated previously, the presence of disorder may affect observers' feeling of safety. See also the models without this variable in Table E1 in the supplementary material, which show similar findings.

Comparing Measures of Disorder

In the current study, we proposed a new cross-classified model to account for observer bias in disorder constructs at the street segment and neighborhood levels. In this section, we will illustrate how these new disorder constructs differ from constructs that were created by the traditional ecometrics method (Raudenbush and Sampson 1999), and the often applied method of simply taking means. Our new measures, to which we will refer as 'cross-classified measures', were created by taking the posterior means of the models presented in Table 3 (including observers' urban background and the observational conditions). The measures obtained through traditional ecometrics, to which we will refer as 'ecometrics measures', were created by taking the posterior means of the models presented in Table B2 in the supplementary material (including observational conditions). The measures obtained by simply taking mean scores, to which we will refer as 'simple mean measures', were created by taking the means of all items to construct street segment measures, and by taking the means of these street segment measures per neighborhood to construct neighborhood measures. We compared the different disorder measures in three ways.

First, we compared the *rank order of neighborhoods* based on the different disorder measures. We present one example in this paper (in Fig. 4), where we compared the cross-classified measures with the ecometrics measures for neighborhood constructs of physical
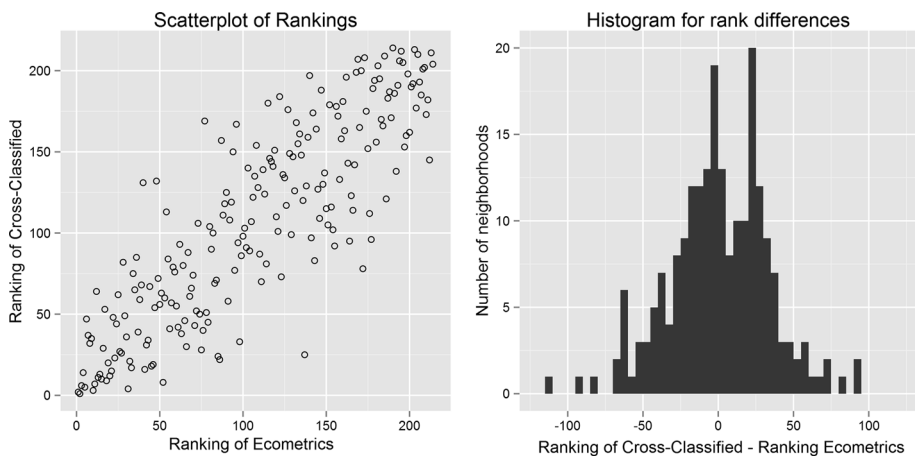
Fig. 4 Comparing rank order of neighborhoods based on measures for physical disorder as derived from a traditional ecometric model with those derived from the proposed cross-classified model

disorder. The left side of Fig. 4 shows the scatterplot of all neighborhoods, ranked from the lowest to the highest score on the ecometrics measure (x-axis) and the cross-classified measure (y-axis). The right side of Fig. 4 shows a more precise comparison of the rank order differences. If observer bias did not affect a neighborhood's ranking of physical disorder, all points would lie on the diagonal of the scatterplot, and thus most ranking differences would be close to zero. However, the scatterplot and histogram show that taking observer bias into account changes the neighborhood's score dramatically. When accounting for observer bias, about half of all neighborhoods receive a higher ranking. These neighborhoods actually have more physical disorder than a standard ecometrics approach would have led us to believe. On the other hand, about half of all neighborhoods are also perceived to experience more physical disorder than they actually do.

Second, we inspected the *correlations between the different disorder measures*. As shown in Table 4, the measures correlated relatively highly at neighborhood level. For the neighborhood constructs of physical disorder, we found correlations of 0.795 and 0.857 between the cross-classified measure and the simple mean and ecometrics measures, respectively. For the neighborhood constructs of social disorder, we found correlations of 0.721 and 0.895 between the cross-classified measure and the respective simple mean and ecometrics measures. Correlations were lower at the street segment level. The cross-classified measure for physical disorder at street segment level showed a correlation of 0.819 with the simple mean measure, and of 0.296 with the ecometrics measure. For social disorder at street segment level, we found correlations of 0.581 and 0.542 between the cross-classified measure and the simple mean measure and ecometrics measure, respectively. Disregarding observer bias thus affects disorder measures more strongly at street segment level than at neighborhood level. Nevertheless, at both levels, we find correlations that differ from 1.00, suggesting that it matters for disorder estimates whether or not one takes into account the allocation of street segments to observers.

Third and finally, we examined how correlations with *areal crime rates* differed across the three disorder measures. The correlations are presented in Table 4. For the comparison with areal crime rates, we make the assumption that the disorder measures, observed in the street segments at the centroids of grid cells (see also Fig. 2) reflect the level of disorder in the entire grid cell. At the grid cell level, we found that disorder–crime correlations, as based on the cross-classified measures for disorder, were overall slightly higher than those based on the simple mean and ecometrics measures. For example, the physical disorder–crime correlation at grid cell level based on the cross-classified model was 0.287, whereas that correlation based on the ecometrics model was 0.240. On the other hand, at the neighborhood level, we found that disorder–crime correlations as based on the cross-classified measures for disorder were generally lower than those based on the simple mean and ecometrics measures. For example, the physical disorder–crime correlation at neighborhood level based on the cross-classified model was 0.299, whereas that correlation based on the ecometrics model was 0.403. These patterns were visible for physical disorder as well as for social disorder. These correlations therefore indicate that inadequate consideration of observer bias leads to slightly underestimating the disorder–crime association at grid cell level, and to overestimating that association at neighborhood level. The differences between the disorder–crime associations across levels of aggregation may therefore be less substantial than appeared to be the case in previous studies. Overall, we found that the disorder–crime correlations were slightly stronger at the neighborhood level than at the grid cell level. For the cross-classified measures for physical disorder, these correlations were respectively 0.299 and 0.287 for neighborhood and grid cell level, and for social disorder the correlations were respectively 0.340 and 0.224.

Table 4 Spearman's rho correlations between different measures of disorder and crime at street segment/grid cell level and neighborhood level

	Simple mean		Ecometrics		Cross-classified	
	Phys. disorder	Soc. disorder	Phys. disorder	Soc. disorder	Phys. disorder	Soc. disorder
Street segment/grid cell						
Phys. disorder (simple mean)	1.000**	0.146**	0.286**		0.819**	
Soc. disorder (simple mean)	0.146**	1.000**		0.299**		0.581**
Phys. disorder (ecometrics)	0.286**		1.000**	0.165**	0.296**	
Soc. disorder (ecometrics)		0.299**	0.165**	1.000**		0.542**
Phys. disorder (cr.-classified)	0.819**		0.296**		1.000**	0.120**
Soc. disorder (cr.-classified)		0.581**		0.542**	0.120**	1.000**
Crime	0.290**	0.222**	0.240**	0.268**	0.287**	0.224**
Neighborhood						
Phys. disorder (simple mean)	1.000**	0.235**	0.930**		0.795**	
Soc. disorder (simple mean)	0.235**	1.000**		0.811**		0.721**
Phys. disorder (ecometrics)	0.930**		1.000**	0.286**	0.857**	
Soc. disorder (ecometrics)		0.811**	0.286**	1.000**		0.895**
Phys. disorder (cr.-classified)	0.795**		0.857**		1.000**	0.224**
Soc. disorder (cr.-classified)		0.721**		0.895**	0.224**	1.000**
Crime	0.382**	0.506**	0.403**	0.388**	0.299**	0.340**

The *simple mean measures* for disorder are the average scores of items at street segment level and averages of street segments at the neighborhood level. The *ecometrics measures* for disorder are the posterior means from the hierarchical multilevel models that control for observational conditions (as presented in Table B2 in the supplementary material). The *cross-classified measures* for disorder are the posterior means from the cross-classified models that control for observers' urban background and observational conditions (as presented in Table 3). *Crime* is measured as the registered number of offenses in public places that have been committed between 2007 and 2009

* $p < 0.05$; ** $p < 0.01$

Discussion and Conclusion

The correct measurement of disorder is an important endeavor in criminology. Disorder may cause fear of crime (Kelling and Coles 1996; Perkins and Taylor 1996) and has been related to crime as cause and as consequence (Boggess and Maskaly 2014; Skogan 1990). However, empirical studies regarding disorder are plagued by several issues, including

aggregation bias and disagreement about the best method of measurement. The aim of the current study was to examine the extent to which SSO is a reliable and valid measurement method for disorder on different units of analysis. In accounting for observer bias, the study elaborated on the ecometrics model of Raudenbush and Sampson (1999), and proposed a cross-classified model that accounts for the allocation of street segments to observers. The study thereby builds on a long tradition of discovering and solving methodological and statistical problems within ecological crime research (for overviews see Sampson and Lauritsen 1994; Weisburd et al. 2009b) and connects to state-of-the-art methods by applying cross-classified modeling and focusing on smaller units of analysis—street segments—alongside the more traditional unit of the neighborhood (Weisburd et al. 2012).

Methodological Implications

An important implication of the present study is its contribution to the 'social science of ecological assessment', by gauging the effect of observer bias in Systematic Social Observations. With the introduction of a method to reliably aggregate variables to a higher level, Raudenbush and Sampson (1999) made a vital contribution to ecological crime research. Our study aims to refine their method by extending the traditional ecometrics model with a level for observers, crossed with neighborhoods at the highest level of the model. The proposed model thereby extends the existing measures in taking into account the allocation of street segments to observers. Our findings of variance components analyses indicate that neighborhoods may be more alike than we would think based on traditional methods: part of the variance in disorder is actually explained by differences between observers. Furthermore, application of the proposed cross-classified model showed to be definitive for the correlations between disorder and police recorded crime. Traditional methods of aggregation, such as the simple means method or traditional ecometrics, appeared to underestimate the disorder–crime association at grid cell level and overestimate the association at the neighborhood level.

With multilevel study designs, a balance must be found between sampling extra macro-units and sampling extra micro-units. In this case, the balance is between a few individual observers per neighborhood observing many street segments within one neighborhood (scenario a), or many individual observers per neighborhood observing a few street segments within one neighborhood (scenario b)—and this balance must be found given other design and budget constraints on the data collection (e.g., wanting a minimum number of different observers to be able to tease out observer bias in the first place). An important disadvantage of *scenario a*, say with two observers per neighborhood, is that these two observers can strongly affect neighborhood-level inferences in the presence of observer bias, while the extra street segment-level observations per observer does not provide more data to investigate such observer bias. The ideal situation would be to follow *a combination of* scenario a and b, by sampling many street segments within one neighborhood, as well as to let different observers observe all street segments twice, but time and budget constraints did not allow for such a design. One important contribution of the present study is to show how cross-classified multilevel models can model exactly the nested relationships between neighborhood, street segment, and observer, with which observer bias can be partialled out (if not always explained).

Our findings indicate that the reliability of the collected data strongly improves with the use of more observers; future SSO studies can acquire neighborhood reliability scores for physical disorder of 0.6 and higher if they select 14 street segments per neighborhood and

allocate these to 7 different observers per neighborhood, or if they select 20 street segments per neighborhood and allocate these to 6 different observers per neighborhood. Thus, the more different observers conduct observations within one neighborhood, the more likely it is that their potential biases level-out, resulting in a disorder measure that is more likely to be replicated had another set of observers coded the street segments within that neighborhood. This might bring about practical implications for future data collections. To keep down costs, most studies prefer to let observers observe additional street segments, rather than training more observers. Our findings imply that more observers are not strictly necessary, provided that thought is given to the allocation of these observers to observation locations.

Reliability issues that arise from inadequate allocation of locations to observers may also be relevant to other data collection methods, although perhaps to a lesser extent. Fatigue and socialization are likely to plague any data collection method that makes use of face-to-face interview or observer techniques, and other issues may be relevant depending on the method and topic of interest. For example, when collecting data based on community surveys that are conducted in face-to-face interviews or based on key informants' interviews, it may be relevant whether interviewers vary in their emotional approach and interactions with subjects. These and other differences between interviewers potentially translate into less ecologically reliable measures if no thought is given to the allocation of interviewers over areas. On the other hand, it is possible that observations based on photo or video material may be less biased than observations conducted in the field. These issues should be further investigated in future studies.

Similarly, our findings may have implications for studies into phenomena other than physical and social disorder. Observer bias is a problem that extends to almost any phenomenon that is studied through observation. Previous studies in criminology have applied (Systematic) Social Observation to examine shoplifting (Buckle and Farrington 1984), police behavior (Mastrofski et al. 1996), and aggression in barrooms (Graham et al. 2006) among other examples. Our findings suggested that, for the observation of physical and social disorder, the urban background of observers was a factor of influence. We further found indications of intra-observer bias, such as fatigue. We expect that observer socialization, fatigue, and observers' prior experiences and unconscious perspectives may also shape observations of other topics. To avoid observer bias as much as possible, we suggest that future SSO studies, regardless of their topic, select a minimum number of observers, pay attention to allocation of observers over locations, subjects, or events, offer extensive training prior to observations and organize feedback meetings during the period of data collection. For some topics, it may be necessary or fruitful to select or reject observers based on their pre-existing attitudes or expertise; previous studies have shown that police officers and college students differed significantly in their observations of police behavior (Reiss 1971) and social disorder (Yang and Pao 2015). Fatigue and observer socialization may be avoided or reduced by ensuring short observation sessions, restricting the maximum number of observations per observer, and by organizing discussions among the observers about changed perceptions and feelings.

Finally, the present study contributes to the growing body of literature about aggregation bias and the prediction of crime. It is now widely recognized that correlations depend on the level of (geographical) aggregation (Openshaw 1984; Robinson 1950). However, despite this knowledge, it is still unclear what the appropriate level of analysis should be for relations between crime and important predictors of crime, if an appropriate level of analysis can even be said to exist (Hipp 2007; Weisburd et al. 2009a). In the current study, we found that if we applied the proposed cross-classified model to aggregate

disorder—thus accounting for the allocation of street segments to observers—differences in disorder–crime correlations across levels of aggregation (grid cell level and neighborhood level) were less substantial, compared with our results if we applied more traditional methods to aggregate disorder. Although this finding needs further investigation, it suggests that correction for observer bias and bias introduced by varying observational conditions in SSO may help to reduce the problem of aggregation.

Limitations and Future Research

One limitation of the current study is the small number of observers; thirteen observers conducted all observations in the greater The Hague area. This small number decreases the statistical power of the estimation at the observer level and may therefore only acknowledge extremely large effects of observer characteristics. The current study should therefore be viewed as a first step in explaining observer bias in Systematic Social Observations of disorder. Nonetheless, the study is exemplary for how SSO studies of disorder are generally organized: a handful of observers, often students, are asked to observe hundreds of locations (Schaefer-McDaniel et al. 2010). Our findings entail a cautionary note for such studies regarding the allocation of observers over locations. Further research is necessary to investigate a wider range of observer characteristics among a larger group of observers. Findings of previous studies on perceptions of disorder, based on data from community surveys, suggest that it might be interesting to investigate observers' age, ethnic background, and marital status (Franzini et al. 2008; Hinkle and Yang 2014; Hipp 2010; Latkin et al. 2009; Sampson and Raudenbush 2004; Wallace et al. 2015). These characteristics were not incorporated in the current study, because there was insufficient variation among the observers on these features. It would also be interesting to further gauge whether observer socialization effects are dependent on observer characteristics. For example, it is possible that observers from rural backgrounds show stronger effects over time as they grow accustomed to urban settings, compared with their colleagues from urban backgrounds who were already familiar with the urban setting.

A second, potential, limitation is that our training process may have been insufficient to prepare the observers for their fieldwork. We took great care in designing the observer training and incorporated suggestions based on previous data collections (e.g., Zenk et al. 2007). Nevertheless, it is difficult to isolate the effects of the training on observer bias afterwards, unless a subgroup of observers participates in a different training. Future research, perhaps in experimental setup, is necessary to specify what educational factors contribute most to observers' performance in the field.

A third limitation of this study is that crime rates from 2007 to 2009 were compared with disorder measures collected in 2012. As explained previously, these crime data were the most recent available data, as police data are not usually geocoded. Nevertheless, we would like to point out that crime rates at grid cells and neighborhoods were quite stable over time: Spearman's rho correlations at both levels of aggregation varied between 0.89 and 0.98 for all consecutive years between 2001 and 2009 (see also Steenbeek and Weisburd 2016). Also, as a way of dealing with the variability of crime rates over time that is inherent to studying small spatial units, we took the crime rate over the years 2007–2009 instead of the crime rate of only 1 year.

A fourth limitation was the scarcity of observations of social disorder, which restricted the variance in these observations and made the analyses more prone to inaccuracy. Systematic Social Observation might be less suitable for measuring social disorder, due to its instantaneous character. Whereas physical disorder remains mostly unchanged during

the course of the day, the observation of social disorder depends largely on the time, the day (weekday or weekend), and the duration of the observation. We decided to restrict the observations to weekdays between 10.00 a.m. and 4.00 p.m., because small signs of physical disorder would be better visible in daylight, and because we did not want rush hours to affect the coding of traffic. This decision has undoubtedly affected the chance to observe social disorder. Our findings regarding the effects of 'time of day' and 'day of week' on social disorder observations should thus be interpreted in light of this restriction. Future study of disorder may require separate data collections for physical and social disorder, where physical disorder is observed in daylight and social disorder in the evenings.

Concluding Remarks

Despite these limitations, the study makes some important contributions to the literature on Systematic Social Observations, to studies on disorder and crime and, more generally, to the 'social science of ecological assessment'. As a final remark, we want to emphasize the need for further attention to the measurement of disorder in micro settings such as street segments. Street segments may be more appropriate units to understand environmental influences on individual behavior, because they are small enough: "Individuals' actions and development are only influenced by the environments they can access with their senses" (Oberwittler and Wikström 2009, pp. 57), and those environments are generally smaller than the neighborhood or city where people reside. Measurement of phenomena at smaller units also enables the aggregation to a wider variety of higher units (Brantingham et al. 2009; Oberwittler and Wikström 2009), which is necessary to truly grasp the crime–disorder relationship (Hipp 2007). Further research is necessary to determine the minimal amount of Systematic Social Observations, key informant interviews, and resident surveys needed to construct reliable measures at these smaller units. The need for assessment at micro settings in the crime–disorder debate makes measurement of disorder, more than ever, a daunting task.

Appendix: Formulas

The below formulas express construct reliability of measures at neighborhood level in Eq. (1) and street segment level in Eq. (2):

$$\lambda_k = \frac{\omega}{\omega + \frac{v}{O_k} + \frac{\tau}{J_k} + \frac{\sigma}{nJ_k}} \tag{1}$$

$$\lambda_{jk} = \frac{\omega + v + \tau}{\omega + v + \tau + \frac{\sigma}{n}} \tag{2}$$

with λ_k reliability of neighborhood construct, λ_{jk} reliability of street segment construct, ω variance at the neighborhood level, v variance at observer level, τ variance at street segment level, σ variance at item level (fixed at $\pi^2/3$), O_k number of observers per neighborhood, J_k number of street segments per neighborhood, and n number of items per street segment.

References

Adcock R, Collier D (2001) Measurement validity: a shared standard for qualitative and quantitative research. Am Polit Sci Rev 95:529–546

Adler PA, Adler P (1987) Membership roles in field research. Sage, Newbury Park

Boggess LN, Maskaly J (2014) The spatial context of the disorder–crime relationship in a study of Reno neighborhoods. Soc Sci Res 43:168–183

Braga AA, Weisburd D, Waring EJ, Green Mazerolle L, Spelman W, Gajewski F (1999) Problem-oriented policing in violent crime places: a randomized controlled experiment. Criminology 37:541–580

Braga AA, Welsh BC, Schnell C (2015) Can policing disorder reduce crime? A systematic review and meta-analysis. J Res Crime Delinq 52:567–588

Brantingham PL, Brantingham P, Vajihollahi M, Wuschke K (2009) Crime analysis at multiple scales of aggregation: a topological approach. In: Weisburd D, Bernasco W, Bruinsma GJN (eds) Putting crime in its place: units of analysis in geographic criminology. Springer, New York, pp 87–107

Browne WJ (2012) MCMC Estimation in MLwiN: Version 2.26. Centre for Multilevel Modelling, University of Bristol, Bristol

Buckle A, Farrington DP (1984) An observational study of shoplifting. Br J Criminol 24:63–73

Carter W, Dougherty J, Grigorian K (1995) Videotaping neighborhoods. In: Earls FJ, Raudenbush SW, Reiss AJJ, Sampson RJ (eds) Project on Human Development in Chicago Neighborhoods (PHDCN): systematic social observation. National Opinion Research Center (NORC), Inter-University Consortium for Political and Social Research (ICPSR), Ann Arbor

Clifton KJ, Smith ADL, Rodriguez D (2007) The development and testing of an audit for the pedestrian environment. Landsc Urban Plan 80:95–110

Coulton CJ, Korbin J, Chan T, Su M (2001) Mapping residents' perceptions of neighborhood boundaries: a methodological note. Am J Community Psychol 29:371–383

Craig CL, Brownson RC, Cragg SE, Dunn AL (2002) Exploring the effect of the environment on physical activity: a study examining walking to work. Am J Prev Med 23:36–43

De Vries SI, Bakker I, Van Mechelen W, Hopman-Rock M (2007) Determinants of activity-friendly neighborhoods for children: results from the SPACE study. Am J Health Promot 21:312–316

Fielding A, Goldstein H (2006) Cross-classified and multiple membership structures in multilevel models: an introduction and review. Department for Education and Skills, Birmingham

Franzini L, O'Brien Caughy M, Murray Nettles S, O'Campo P (2008) Perceptions of disorder: contributions of neighborhood characteristics to subjective perceptions of disorder. J Environ Psychol 28:83–93

Gau JM, Pratt TC (2008) Broken windows or window dressing? Citizens' (in)ability to tell the difference between disorder and crime. Criminol Public Policy 7:163–194

Graham K, Tremblay P, Wells S, Pernanen K, Purcell J, Jelley J (2006) Harm, intent, and the nature of aggressive behavior: measuring naturally occurring aggression in barroom settings. Assessment 13:280–296

Hale C (1996) Fear of crime: a review of the literature. Int Rev Victimol 4:79–150

Hinkle J, Yang S-M (2014) A new look into Broken Windows: what shapes individuals' perceptions of social disorder? J Crim Justice 42:26–35

Hipp JR (2007) Block, tract, and levels of aggregation: neighborhood structure and crime and disorder as a case in point. Am Sociol Rev 72:659–680

Hipp JR (2010) Resident perceptions of crime and disorder: how much is 'bias' and how much is social environment differences? Criminology 48:475–508

Hipp JR, Boessen A (2013) Egohoods as waves washing across the city: a new measure of "neighborhoods". Criminology 51:287–327

Hunt JC (1989) Psychoanalytic aspects of fieldwork. Sage, Beverly Hills

Innes M (2004) Signal crimes and signal disorders: notes on deviance as communicative action. Br J Sociol 55:335–355

Jago R, Baranowski T, Zakeri I, Harris M (2005) Observed environmental features and the physical activity of adolescent males. Am J Prev Med 29:98–104

Jones M, Pebley AR, Sastry N (2011) Eyes on the block: measuring urban physical disorder through in-person observation. Soc Sci Res 40:523–537

Kelling GL, Coles CM (1996) Fixing broken windows: restoring order and reducing crime in our communities. Touchstone, New York

Killias M, Clerici C (2000) Different measures of vulnerability in their relation to different dimensions of fear of crime. Br J Criminol 40:437–450

LaGrange RL, Ferraro KF (1989) Assessing age and gender differences in perceived risk and fear of crime. Criminology 27:697–719

Latkin CA, German D, Hua W, Curry AD (2009) Individual-level influences on perceptions of neighborhood disorder: a multilevel analysis. J Commun Psychol 37:122–133

Mastrofski SD, Snipes JB, Supina AE (1996) Compliance on demand: the public's response to specific police requests. J Res Crime Delinq 33:269–305

Mastrofski SD, Parks RB, McCluskey JD (2010) Systematic social observation in criminology. In: Piquero AR, Weisburd D (eds) Handbook of quantitative criminology. Springer, Berlin, pp 225–247

Oberwittler D, Wikström P-OH (2009) Why small is better: Advancing the study of the role of behavioral contexts in crime causation. In: Weisburd D, Bernasco W, Bruinsma GJN (eds) Putting crime in its place: units of analysis in geographic criminology. Springer, New York, pp 35–59

O'Brien DT, Norton CC, Cohen J, Wilson S (2014) Local adaptation in community perception: how background impacts judgments of neighborhood safety. Environ Behav 46:213–240

Openshaw S (1984) The modifiable areal unit problem. Concepts and techniques in modern geography. Geo Books, Norwich

Pauwels LJR, Hardyns W (2009) Measuring community (dis)organizational processes through key informant analysis. Eur J Criminol 6:401–417

Perkins DD, Taylor RB (1996) Ecological assessments of community disorder: their relationship to fear of crime and theoretical implications. Am J Community Psychol 24:63–107

Perkins DD, Meeks JW, Taylor RB (1992) The physical environment of street blocks and resident perceptions of crime and disorder: implications for theory and measurement. J Environ Psychol 12:21–34

Raudenbush SW, Sampson RJ (1999) Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. Sociol Methodol 29:1–41

Reiss AJJ (1971) Systematic observation of natural social phenomena. Sociol Methodol 3–33

Robinson WS (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15:351–357

Sampson RJ, Lauritsen JL (1994) Violent victimization and offending: Individual-, situational-, and community-level risk factors. In: Reiss AJJ, Roth J (eds) Understanding and preventing violence: social influences, vol 3. National Academy Press, Washington DC, pp 1–114

Sampson RJ, Raudenbush SW (1999) Systematic social observation of public spaces: a new look at disorder in urban neighborhoods. Am J Sociol 105:603–651

Sampson RJ, Raudenbush SW (2004) Seeing disorder: neighborhood stigma and the social construction of "Broken Windows". Soc Psychol Q 67:319–342

Schaefer-McDaniel N, O'Brien Caughy M, O'Campo P, Gearey W (2010) Examining methodological details of neighborhood observations and the relationship to health: a literature review. Soc Sci Med 70:277–292

Skogan WG (1990) Disorder and decline: crime and the spiral of decay in American neighborhoods. The Free Press, New York

Skogan WG (2015) Disorder and decline: the state of research. J Res Crime Delinq 52:464–485

Snijders TAB, Bosker RJ (2012) Multilevel analysis: an introduction to basic and advanced multilevel modeling, 2nd edn. Sage, Los Angeles

Spano R (2003) Concerns about safety, observer sex, and the decision to arrest: evidence of reactivity in a large-scale observational study of police. Criminology 41:909–932

Spano R (2005) Potential sources of observer bias in police observational data. Soc Sci Res 34:591–617

Steenbeek W, Hipp JR (2011) A longitudinal test of social disorganization theory: feedback effects among cohesion, social control, and disorder. Criminology 49:833–871

Steenbeek W, Weisburd D (2016) Where the action is in crime? An examination of variability of crime across different spatial units in The Hague, 2001–2009. J Quant Criminol 32:449–469

Sykes RE (1978) Toward a theory of observer effect in systematic field observation. Hum Organ 37:148–156

Taylor RB, Shumaker SA (1990) Local crime as a natural hazard: implications for understanding the relationship between disorder and fear of crime. Am J Community Psychol 18:619–641

Taylor RB, Gottfredson SD, Brower S (1984) Block crime and fear: defensible space, local social ties, and territorial functioning. J Res Crime Delinq 21:303–331

Tremblay MA (1957) The key informant technique: a nonethnographic application. Am Anthropol 59:688–701

Wallace D, Louton B, Fornango R (2015) Do you see what I see? Perceptual variation in reporting the presence of disorder cues. Soc Sci Res 51:247–261

Weich S, Burton E, Blanchard M, Prince M, Sproston K, Erens B (2001) Measuring the built environment: validity of a site survey instrument for use in urban settings. Health Place 7:283–292

Weisburd D, Bernasco W, Bruinsma GJN (2009a) Putting crime in its place: units of analysis in geographic criminology. Springer, New York

Weisburd D, Bernasco W, Bruinsma GJN (2009b) Units of analysis in geographic criminology: historical development, critical issues, and open questions. In: Weisburd D, Bernasco W, Bruinsma GJN (eds) Putting crime in its place: units of analysis in geographic criminology. Springer, New York, pp 3–31

Weisburd D, Groff ER, Yang S-M (2012) The criminology of place: street segments and our understanding of the crime problem. Oxford University Press, Oxford

Wilson JQ, Kelling GL (1982) Broken windows: the police and neighborhood safety. Atl Mon 29–38

Yang S-M, Pao C-C (2015) Do we "see" the same thing? An experimental look into the black box of disorder perception. J Res Crime Delinq 52:534–566

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