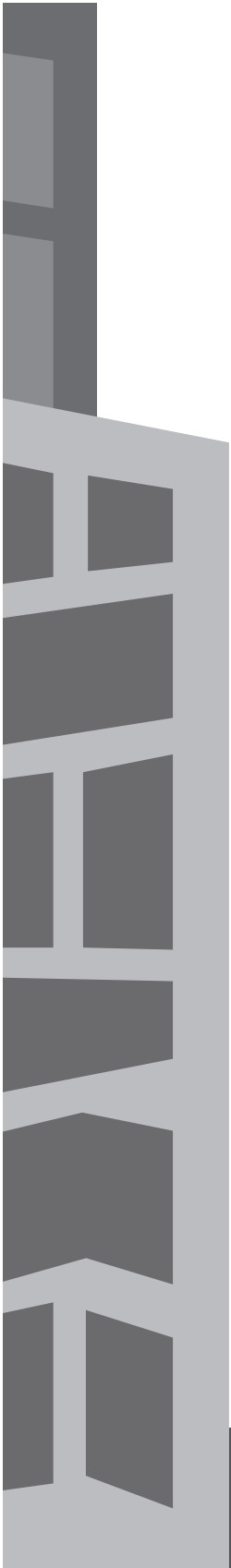


Chapter Four.

Systematic Social Observation

Hoeben, Evelien M., Wouter Steenbeek, and Lieven J.R. Pauwels. 2016. Measuring disorder: Observer bias in systematic social observations at streets and neighborhoods. *Conditionally accepted for publication in Journal of Quantitative Criminology.*



Systematic Social Observation

*Among the limitless attributes of a person's surroundings,
which ones are relevant to his behavior, and how does one identify
and measure them? – Roger Barker*

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, Welsh, and Schnell, 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: 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 inter-rater 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, Pebley, and Sastry, 2011; Raudenbush and Sampson, 1999). Observers may bias the measures, because of their varying perceptions of disorder, or because of fatigue or burnout cynicism (Mastrofski, Parks, and McCluskey, 2010; Spano, 2005). A major shortcoming in most previous SSO studies is the lack of attention to sources of observer bias. If characteristics of observers (such as their perceived vulnerability or urban background) affect their perceptions about disorder, not accounting for such characteristics would bias the obtained measures of disorder. Therefore we built on previous research and present a refined model to directly control for observer characteristics in ecological constructs. Data for the current study were collected in a conurbation in Europe: The areas surrounding The Hague, 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, Meeks, and Taylor, 1992; Perkins and Taylor, 1996; Taylor, Gottfredson, and Brower, 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, Smith, and Rodriguez, 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 interviews⁹ 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, Parks, and McCluskey, 2010; Reiss, 1971). Measuring phenomena through interviewing residents or key informants is by definition retrospective and therefore filtered by “judgment and memory” (Carter, Dougherty, and Grigorian, 1995: 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 the day, day of the week and the season in which the observation

9 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).

takes place (Jones, Pebley, and Sastry, 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 reactivity¹⁰.

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 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: 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, Parks, and McCluskey, 2010). Observer socialization and fatigue are both expected to result in less accurate and comprehensive data at later stages of the data

¹⁰ We concentrate on unintentional observer bias and do not take into account intentional bias caused by cheating.

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, Snipes, and Supina (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, social scientists and 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, 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, Louton, and Fornango, 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, Louton, and Fornango, 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, Louton, and Fornango, 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, Louton, and Fornango, 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. We do not expect this to bias the observation of physical disorder, as physical disorder only refers to objects, and not people. 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.

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. 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 4.1 illustrates such an allocation of observers to different neighborhoods. In Figure 4.1, 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).

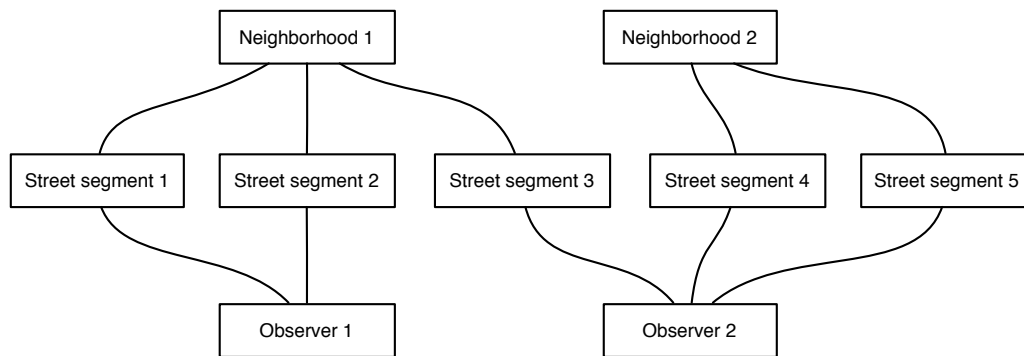


Figure 4.1. Street segments nested within neighborhoods and observers

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 4.1 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 observer characteristics (urban background, gender, perceived vulnerability to victimization), observational conditions (time, day, and weather), fatigue or socialization effects over time (number of observations

during the entire project and on a specific day), and by 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 States¹¹. 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 validation¹². To determine reliability and validity, we present a cross-classified model that takes into account observer bias.

11 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.

12 Numerous different types of validation have been suggested. We maintain the distinction of Adcock and Collier (2001: 539, 542), who distinguish 1) content validation, which is understood as "focusing on the relationship between an indicator and the concept of interest", 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". 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 current study, we test nomological/construct validation by studying the relationship between disorder and crime.

Data and methods

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 meters (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 centroids of every third grid cell were observed, resulting in a total of 1422 centroids, spanning 253 neighborhoods (neighborhood boundaries as defined by Statistics Netherlands). At every given address (based on the centroid of a grid cell), observers made photographs with a camera equipped with a GPS device. 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 meters of the centroid. The exact location of all observations was determined afterwards, based on the recorded GPS coordinates. Observers were instructed to walk 50 meters to the left and 50 meters to the right from the given address (or coordinates) and thereby observed street segments of 100 meters (both sides of the street). Each observation was carried out by one observer. 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 twenty to twenty four 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 workdays (Monday to Friday, except on holidays or during primary and secondary school vacations) between 10.00a.m. and 4.00p.m. Observations were not executed on days when garbage was collected by the municipality. Observation of one street segment took on average eight minutes and nine seconds. 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.

Prior to the data collection, all observers were trained to improve inter-rater reliability. The training took 1.5 days and also provided a short introduction to the theoretical background of the data collection. The observers had to practice with pictures taken during the pilot study and were confronted with common mistakes, also based on experiences from the pilot study. Then they were sent out to conduct field observations in groups of two or three observers. Their observations were analyzed by two researchers. Mistakes, problems and irregularities were discussed with the group, as suggested by Zenk et al. (2007). Observers that had scored items very differently were taken apart for additional instructions. Ten percent of the locations ($N = 147$) were coded double independently by two different observers to examine inter-rater reliability. Cohen's kappa was .731 for physical disorder and .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).

Figure 4.1 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)¹³.

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. The social disorder

¹³ 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 21, and the maximum was 120.

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 observed' and score 1 for 'observed'. For a complete overview and frequency distribution of items per scale, see Table 4.1.

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' (≥ 2500 addresses per km²), 2 was 'strongly urban' (1500-2500 addresses per km²), 3 was 'mixed rural and urban' (1000-1500 addresses per km²), 4 was 'moderately rural' (500-1000 addresses per km²), 5 was 'rural' (< 500 addresses per km²). *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. Descriptive statistics are given in Table 4B.1 in the Appendix of this chapter.

Six observational conditions were examined: *Time of day* referred to the hour in which the main part of the observation took place (between 10.00a.m. and 4.00p.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. To investigate fatigue and burnout cynicism of observers, we examined at the street segment level what number observation this was for the observer, being how many observations the observer had already conducted, both in total and that day. 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. Descriptive statistics are given in Table 4B.1 in the Appendix of this chapter.

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 geo-coded. The registered offenses had been reported by victims and bystanders or were noted by the police. All data were geo-coded 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 meters, and this information was matched to the disorder observations conducted at the grid cell centroids) and per neighborhood, with boundaries as defined by Statistics Netherlands. 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 chapter.

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 were also estimated (1999). We refer to those models as ‘traditional econometrics method’ throughout this chapter. Results of the traditional econometrics models are given in Table 4C.1 in the Appendix of this chapter, and in Table 4E.1 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

Table 4.1 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 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).

Table 4.1. Occurrence of physical disorder and social disorder ($N = 1422$ street segments)

Scales and items	Frequency	Percentage
Physical disorder		
Litter or broken glass	967.0	68.0
Cigarette butts	849.0	59.7
Dog feces	350.0	24.6
Empty bottles or cans	348.0	24.5
Graffiti tags (small surface, tag)	211.0	14.8
Abandoned bicycles	44.0	3.1
Graffiti pieces (big surface, piece)	36.0	2.5
Social disorder		
Adults loitering	110.0	7.7
Loud music	39.0	2.7
Teenagers loitering	30.0	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

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

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 4.2. 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 percent of the total variance in physical disorder reflects differences between neighborhoods (6.340, Table 4.2.) and 12.4 percent (12.412) reflects differences between street segments. For social disorder, 17.2 percent (17.228) of the total variance reflects differences between neighborhoods and 24.2 percent (24.243) reflects differences between street segments. One can compare these findings with those derived from a traditional econometrics model, which does not take into account the allocation of street segments to observers. Variance components in Table 4C.1 in the Appendix of this chapter indicate that in a traditional econometrics model, it appears that approximately 10.7 percent of the total variance in physical disorder reflects differences between neighborhoods, compared with 6.3 percent in the cross-classified model, and 27.7 percent reflects differences between street segments, compared with 12.4 percent in the cross-classified model. Differences for social disorder are less substantial. A traditional econometrics model (as presented in Table 4C.1 in the Appendix) indicates that about 17.7 percent of the variance in social disorder is at neighborhood level, compared with 17.2 percent in the cross-classified model, and that about 29.9 percent is at street segment level, compared with 24.2 percent in the cross-classified model.

The intra-class correlation coefficients (ICCs), presented in Table 4.2, express the extent to which street segments within one neighborhood that are *observed by different observers* are alike (physical disorder: 6.3 percent, social disorder: 17.2 percent). 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 percent for physical and social disorder (.248 and .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 percent; social disorder: 24.9 percent) compared to when they are observed by *different* observers (physical disorder: 6.3 percent; social disorder: 17.2 percent). This indicates the presence of observer bias.

Table 4.2. 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	(.077)	-6.620	(.160)
Variance between street segments	.650	(.009)	1.569	(.230)
Variance between neighborhoods	.332	(.005)	1.115	(.106)
Variance between observers	.965	(.243)	.498	(.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 ^a	.063	-	.172	-
ICC neighborhood*observers ^b	.248	-	.249	-
Lambda street segment	.806	-	.886	-
Lambda neighborhood	.389	-	.683	-
Cronbach's alpha	.570	-	.304	-
DIC	7456.170	-	1450.890	-

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

ABBREVIATIONS: SE = standard error; ICC = intra-class correlation; DIC = Deviance Information Criterion.

^aICC neighborhood: variance neighborhood/total variance.

^bICC neighborhood*observers: (variance observers + variance neighborhood)/ total variance.

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 4A at the end of this chapter. The ecological reliability measures (lambdas) for the disorder constructs are presented in Table 4.2. These lambdas are based on empty cross-classified models¹⁴. The values of lambda parameters vary, similar to the Cronbach's

14 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).

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 4.2 presents the mean lambdas. The street segment lambda differs with the variance on street segment level and the number of items per scale: Equation 2 in Appendix 4A. On average, the street segment lambdas are acceptably high for both physical and social disorder (respectively .806 and .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 the number of observers that observed the street segments in that neighborhood: Equation 1 in Appendix 4A. The neighborhood lambdas for physical disorder vary between .137 and .676. For social disorder, they vary between .310 and .911. The average neighborhood lambda is acceptable for social disorder (.683), but rather low for physical disorder (.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 Figure 4.2. Figure 4.2 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 .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 .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 .549. To obtain neighborhood lambdas of .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.

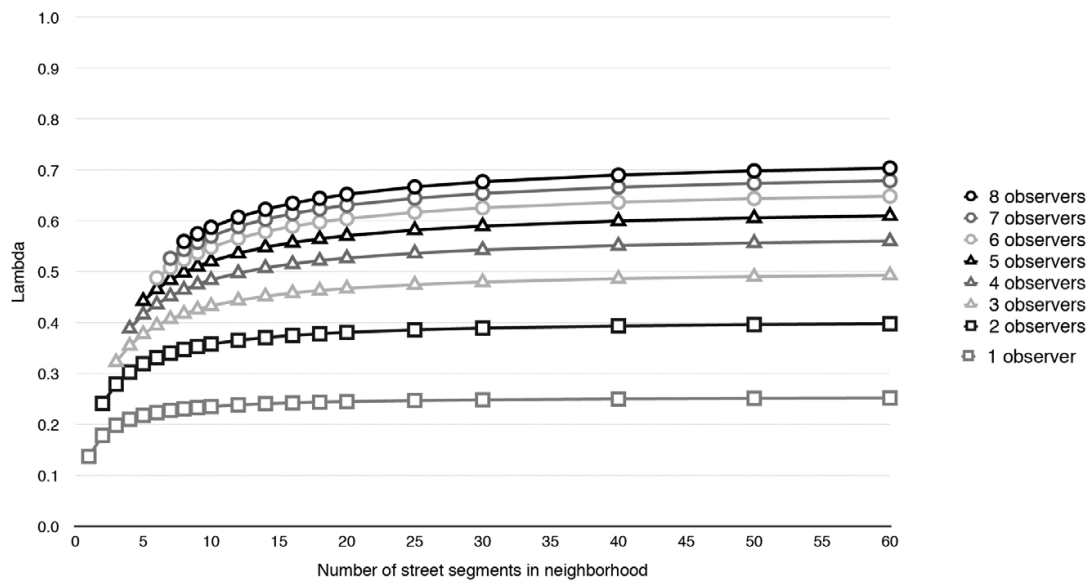


Figure 4.2. Neighborhood reliability (lambda) for physical disorder as a function of sampled street segments and number of observers where the number of items is constant

Observer characteristics

The findings of the variance components analyses that we discussed previously suggested 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 4.3, and in Tables 4F.1 (physical disorder) and 4F.2 (social disorder) in the supplementary material. Results of additional Wald tests are presented in Table 4D.1 in the Appendix of this chapter.

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 4.3). The other characteristics did not appear to be relevant (the results for these characteristics are presented in Tables 4F.1 and 4F.2 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 4.3 indicate that observers who grew up in moderately rural areas observed less physical disorder than observers from very strongly urban backgrounds ($\beta_{\text{moderately urban}} = -1.456, p < .05$). Also, we found that observers from strongly urban backgrounds observed less social disorder than observers from very strongly urban backgrounds ($\beta_{\text{strongly urban}} = -1.237, p < .05$). These findings contradict our Hypothesis 2, that observers from very strongly urban areas would observe fewer signs of physical and social disorder.

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 4.3¹⁵ and Table 4G.1 in the supplementary material. The latter presents models that do not include observers' urban backgrounds. Results of additional Wald tests are presented in Table 4D.1 in the Appendix of this chapter.

15 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).

Table 4.3. Bias by observer characteristics and observational conditions ($N = 1422$ street segments)

	Physical disorder				Social disorder							
	β	(SE)	OR	β	(SE)	OR	β	(SE)	OR			
Observer level												
<i>Urban background</i>												
Very strongly urban (ref)												
Strongly urban	-.501	(.426)	.606	-.570	(.415)	.556	-1.237*	(.332)	.290*	-.948	(.360)	.388
Mixed rural and urban	.658	(.413)	1.931	.851	(.403)	2.342	-.057	(.268)	.945	.440	(.305)	1.553
Moderately rural	-1.456*	(.373)	.233*	-1.370*	(.336)	.254*	-1.019	(.288)	.361	-.595	(.329)	.552
Rural	-.893	(.434)	.409	-.932	(.384)	.394	-1.014	(.324)	.363	-.760	(.387)	.468
Street segment level												
<i>Time of day</i>												
10.00-11.00				-.024	(.029)	.976				-.454	(.217)	.635
11.00-12.00				-.134	(.022)	.875				-.368	(.166)	.692
12.00-13.00 (ref)												
13.00-14.00				-.077	(.022)	.926				.663	(.138)	1.941
14.00-15.00				.188	(.027)	1.207				1.150*	(.163)	3.158*
15.00-16.00				.400	(.042)	1.492				.970*	(.222)	2.638*
<i>Day of week</i>												
Monday				-.077	(.026)	.926				-.353	(.162)	.688
Tuesday				.361*	(.022)	1.435*				.014	(.144)	1.014
Wednesday (ref)												
Thursday				.274	(.023)	1.315				.386	(.127)	1.471
Friday				.004	(.025)	1.004				.218	(.144)	1.244
<i>Weather</i>												
Sunny				-.083	(.021)	.920				.747*	(.118)	2.111*
Sunny with clouds				-.168	(.019)	.845				-.263	(.136)	.769
Cloudy with sun (ref)												

Continuation of Table 4.3

	Physical disorder				Social disorder				
	β	(SE)	OR	β	(SE)	OR	β	(SE)	OR
Drizzle rain				-.485	(.159)	.616	-1.344	(2.152)	.261
Grey and cloudy				-.211	(.018)	.810	-.367	(.136)	.693
No. of observation total				-.003*	(.000)	.997*	-.006*	(.000)	.994*
No. of observation that day				-.035*	(.000)	.966*	-.043	(.001)	.958
Feeling at location ^a				-.276*	(.008)	.759*	-.589*	(.040)	.555*
Intercept	-1.122*	(.217)	.326*	-.771	(.236)	.463	-5.907	(.240)	.003
DIC	6313.32			6309.29			1268.67		1252.26
Variance observer level	.441			.380			.141		
Variance neighborhood level	.409			.316			1.010		.460
Variance street segment level	.803			.819			1.817		2.066

NOTES: 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 4F.1 and 4F.2).

ABBREVIATIONS: SE = standard error; OR = odds ratio; ref = reference category; DIC = Deviance Information Criterion.

^a Expressed as deviation of observers' general feeling of safety across all observed locations.

* $p < .05$ (two-tailed).

With regard to physical disorder (Table 4.3), 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 4.00p.m. than earlier in the day (more than between 11.00a.m. and noon: $\chi^2 = 5.592$; and more than between 1.00 and 2.00p.m.: $\chi^2 = 6.223$, Table 4D.1 in the Appendix of this chapter). 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 4D.1 in the Appendix), more disorder was observed on Thursdays than on Mondays ($\chi^2 = 4.452$, Table 4D.1 in the Appendix), more disorder was observed on Tuesdays than on Fridays ($\chi^2 = 4.776$, Table 4D.1 in the Appendix), and more disorder was observed on Tuesdays than on Wednesdays ($\beta_{\text{Tuesday}} = .361, p < .05$, Table 4.3). Furthermore, we found that observers reported less physical disorder as their number of conducted observations increased (Table 4.3). Although these effects were modest, they were visible both for the number of observations across the entire project ($\beta_{\text{Nr of observation total}} = -.003, p < .05$), and for the number of observations that had been conducted that day ($\beta_{\text{Nr of observation that day}} = -.035, p < .05$). These findings support Hypothesis 1 and imply that fatigue or observer 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 ($\beta_{\text{Feeling at location}} = -.276, p < .05$). 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 4H.1 in the supplementary material.

With regard to social disorder (Table 4.3), 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 2.00p.m. than around noon: $\beta_{2-3pm} = 1.150, p < .05$; $\beta_{3-4pm} = .970, p < .05$, significantly less often before noon than after 1.00p.m., as illustrated by the findings of the Wald tests in Table 4D.1 in the Appendix), and more often when it was sunny compared with when it was sunny with clouds ($\chi^2 = 7.166$, Table 4D.1 in the Appendix), cloudy ($\beta_{\text{sunny}} = .747, p < .05$, Table 4.3) or grey ($\chi^2 = 9.911$, Table 4D.1 in

the Appendix). We expect that these conditions are more predictive of the occurrence of social disorder than that they explain bias in observations: People are more likely to be outside after work or school, in the afternoon, 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 (Table 4.3: $\beta_{Nr\ of\ observation\ total} = -.006, p < .05$), 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 (Table 4.3: $\beta_{Feeling\ at\ location} = -.589, p < .05$). But, as we stated previously, the presence of disorder may affect observers' feeling of safety. See also the models without this variable in Table 4H.1 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 econometrics 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 4.3 (including observers' urban background and the observational conditions). The measures obtained through traditional econometrics, to which we will refer as 'econometrics measures', were created by taking the posterior means of the models presented in Table 4E.1 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 chapter (in Figure 4.3), where we compared the cross-classified measures with the

ecometrics measures for neighborhood constructs of physical disorder. The left side of Figure 4.3 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 Figure 4.3 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.

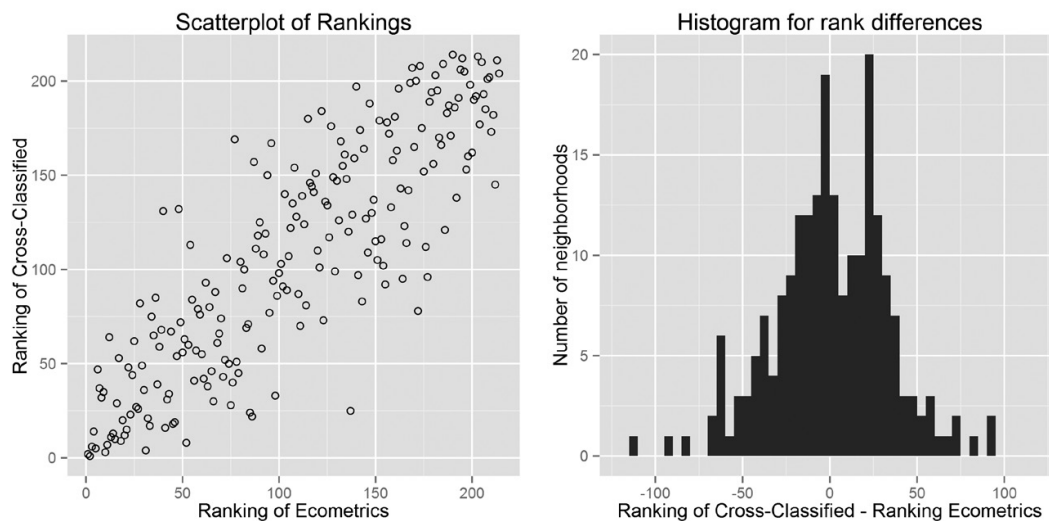


Figure 4.3. Comparing rank order of neighborhoods based on measures for physical disorder as derived from a traditional ecometric model with those derived from the proposed cross-classified model

Second, we inspected the *correlations between the different disorder measures*. As shown in Table 4.4, the measures correlated relatively highly at neighborhood level. For the neighborhood constructs of physical disorder, we found correlations of .795 and .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 .721 and .895 between the cross-classified measure and the respective simple mean and econometrics measures. Correlations were lower at the street segment level. The cross-classified measure for physical disorder at street segment level showed a correlation of .819 with the simple mean measure, and of .296 with the econometrics measure. For social disorder at street segment level, we found correlations of .581 and .542 between the cross-classified measure and the simple mean measure and econometrics 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 1.000, 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 *crime rates* differed across the three disorder measures. The correlations are presented in Table 4.4. At the street segment 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 econometrics measures. For example, the physical disorder-crime correlation at street segment level based on the cross-classified model was .287, whereas that correlation based on the econometrics model was .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 econometrics measures. For example, the physical disorder-crime correlation at neighborhood level based on the cross-classified model was .299, whereas that correlation based on the econometrics model was .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 street segment 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 street segment level. For the cross-classified measures for physical disorder, these correlations were respectively .299 and .287 for neighborhood and street segment level, and for social disorder the correlations were respectively .340 and .224.

Table 4.4. Spearman's rho correlations between different measures of disorder and crime at street segment level and neighborhood level ($N_{\max} = 1422$ street segments, 253 neighborhoods)

	Simple mean		Ecometrics		Cross-classified	
	Physical disorder	Social disorder	Physical disorder	Social disorder	Physical disorder	Social disorder
Street segment						
Physical disorder (sm)	1.000**	.146**	.286**		.819**	
Social disorder (sm)	.146**	1.000**		.299**		.581**
Physical disorder (eco)	.286**		1.000**	.165**	.296**	
Social disorder (eco)		.299**	.165**	1.000**		.542**
Physical disorder (cc)	.819**		.296**		1.000**	.120**
Social disorder (cc)		.581**		.542**	.120**	1.000**
Crime	.290**	.222**	.240**	.268**	.287**	.224**
Neighborhood						
Physical disorder (sm)	1.000**	.235**	.930**		.795**	
Social disorder (sm)	.235**	1.000**		.811**		.721**
Physical disorder (eco)	.930**		1.000**	.286**	.857**	
Social disorder (eco)		.811**	.286**	1.000**		.895**
Physical disorder (cc)	.795**		.857**		1.000**	.224**
Social disorder (cc)		.721**		.895**	.224**	1.000**
Crime	.382**	.506**	.403**	.388**	.299**	.340**

NOTES: 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 4E.1 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 4.3). *Crime* is measured as the registered number of offenses in public places that have been committed between 2007 and 2009.

ABBREVIATIONS: sm = simple mean; eco = traditional ecometrics; cc = cross-classified.

* $p < .05$; ** $p < .01$ (two-tailed).

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 (Bogges 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, Bernasco, and Bruinsma, 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 traditionally used unit of the neighborhood (Weisburd, Groff, and Yang, 2012).

Methodological implications

An important implication of the current 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 street segment level and overestimate the association at the neighborhood level.

These findings 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 indicate, however, 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. Our findings thus imply that more observers are not strictly necessary, providing 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.

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, Snipes, and Supina, 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 current 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, Bernasco, and Bruinsma, 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 (street segment 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 systematic sources of bias such as observational conditions and observer characteristics 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 an exemplar of 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, Louton, and Fornango, 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. 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 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 workdays between 10.00a.m. and 4.00p.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 to affect individual behavior: “Individuals’ actions and development are only influenced by the environments they can access with their senses” (Oberwittler and Wikström, 2009: 57), and those environments are generally smaller than the neighborhood or city where people reside. Street segments may serve as behavior settings (Barker, 1968; Taylor, 1987), because people who live or otherwise spend time within street segments become aware of each other’s routines and standing patterns of behavior (Taylor, 1987). 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.

Appendices Chapter Four

- Appendix 4A. Formulas
- Appendix 4B. Descriptive statistics
- Appendix 4C. Variance components traditional econometrics
- Appendix 4D. Wald tests

Supplementary material

(enclosed in a separate document available from the author):

- Appendix 4E. Traditional econometrics
- Appendix 4F. Observer characteristics
- Appendix 4G. Observational conditions
- Appendix 4H. Models without variable ‘feel’

Appendix 4A

The below formulas express construct reliability of measures at neighborhood level in Equation 1 and street segment level in Equation 2:

$$\lambda_k = \frac{\omega}{\omega + \frac{\nu}{O_k} + \frac{\tau}{J_k} + \frac{\sigma}{nJ_k}} \quad [1]$$

$$\lambda_{jk} = \frac{\omega + \nu + \tau}{\omega + \nu + \tau + \frac{\sigma}{n}} \quad [2]$$

With:

λ_k = reliability of neighborhood construct

λ_{jk} = reliability of street segment construct

ω = variance at the neighborhood level

ν = 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

n = number of items per street segment

Appendix 4B

Table 4B.1. Descriptive statistics of observer characteristics and observational conditions

	N	Mean	(SD)	Min	Max
Observational conditions					
Time of day					
10.00-11.00	1422	.149	(.356)	.000	1.000
11.00-12.00	1422	.210	(.408)	.000	1.000
12.00-13.00	1422	.196	(.397)	.000	1.000
13.00-14.00	1422	.190	(.392)	.000	1.000
14.00-15.00	1422	.158	(.364)	.000	1.000
15.00-16.00	1422	.098	(.297)	.000	1.000
Day of week					
Monday	1422	.213	(.410)	.000	1.000
Tuesday	1422	.166	(.372)	.000	1.000
Wednesday	1422	.279	(.448)	.000	1.000
Thursday	1422	.170	(.376)	.000	1.000
Friday	1422	.173	(.378)	.000	1.000
Weather conditions					
Sunny	1422	.277	(.448)	.000	1.000
Sunny with clouds	1422	.271	(.445)	.000	1.000
Cloudy with sun	1422	.215	(.4110)	.000	1.000
Drizzle rain	1422	.013	(.112)	.000	1.000
Grey and cloudy	1422	.224	(.417)	.000	1.000
Number of observation of observer total	1422	78.459	(59.458)	1.000	226.000
Number of observation that day	1422	7.930	(5.819)	1.000	35.000
Feeling of safety at location ^a	1422	.000	(.498)	-2.980	1.370
Observer characteristics					
Gender (male = 1)	13	.539	(.519)	.000	1.000
Urban background	11	3.091	(1.446)	1.000	5.000
Perceived chance of victimization	12	5.778	(.656)	4.670	7.000
Perceived response to threat	12	2.167	(.718)	1.000	3.000
General feeling of safety	13	4.293	(.476)	3.630	4.980

ABBREVIATIONS: SD = standard deviation; Min = minimum; Max = maximum.

^a Expressed as deviation of observers' general feeling of safety across all observed locations.

Appendix 4C

Table 4C.1. Variance components, internal consistency and lambdas for physical and social disorder ($N = 1422$ street segments)

	Physical disorder		Social disorder	
	Mean	(SE)	Mean	(SE)
Intercept	-1.613	(.007)	-6.519	(.117)
Variance between street segments	1.481	(.021)	1.874	(.288)
Variance between neighborhoods	.573	(.012)	1.114	(.110)
Total variance	5.344	-	6.278	-
% of variance street	27.714	-	29.851	-
% of variance neighborhood	10.723	-	17.745	-
ICC neighborhood ^a	.279	-	.373	-
Lambda street segment	.814	-	.879	-
Lambda neighborhood	.623	-	.733	-
Cronbach's alpha	.570	-	.304	-
DIC	7701.71	-	1461.54	-

NOTES: Variance at level 1 (items) is fixed at $\pi^2/3$. Results obtained from empty hierarchical three-level logistic models.

ABBREVIATIONS: SE = standard error; ICC = intra-class correlation; DIC = Deviance Information Criterion.

^aICC neighborhood: variance neighborhood/total variance.

Appendix 4D

Table 4D.1. Results of Wald tests (χ^2) comparing categories of observer characteristics and observational conditions

	Physical disorder	Social disorder
Urban background		
Strongly urban-mixed rural and urban	5.053*	6.469*
Strongly urban-moderately rural	1.852	.359
Strongly urban-rural	.320	.097
Mixed rural and urban-moderately rural	16.054**	3.980*
Mixed rural and urban-rural	8.916**	4.630*
Moderately rural-rural	.589	.080
Time of day		
10.00-11.00	.480	.035
10.00-13.00	.077	5.297*
10.00-14.00	1.003	9.064**
10.00-15.00	2.774	5.661*
11.00-13.00	.126	6.275*
11.00-14.00	3.094	11.137**
11.00-15.00	5.592*	6.484*
13.00-14.00	2.868	1.898
13.00-15.00	6.223*	.528
14.00-15.00	1.349	.208
Day of week		
Monday-Tuesday	7.211**	.714
Monday-Thursday	4.452*	3.153
Monday-Friday	.261	1.924
Tuesday-Thursday	.300	.863
Tuesday-Friday	4.776*	.248
Thursday-Friday	2.583	.186
Weather conditions		
Sunny-sunny with clouds	.335	7.166**
Sunny-drizzle rain	.987	2.019
Sunny-grey	.795	9.911**
Sunny with clouds-drizzle rain	.621	.536
Sunny with clouds-grey	.092	.067
Drizzle rain-grey	.465	.438

NOTES: Comparing coefficients of Models as presented in Table 4.3.

** $p < .01$; * $p < .05$ (two-tailed).