Journal of Criminal Justice 51 (2017) 74–88

Contents lists available at ScienceDirect

Journal of Criminal Justice

journal homepage: www.elsevier.com/locate/jcrimjus

New windows into a broken construct: A multilevel factor analysis and DIF assessment of perceived incivilities

Jeffrey T. Ward[a,*], Nathan W. Link[b], Ralph B. Taylor[a]

[a] Temple University, Department of Criminal Justice, United States
[b] Rutgers University-Camden, Department of Sociology, Anthropology & Criminal Justice, United States

ARTICLE INFO

Keywords:
Incivilities
Disorder
Neighborhood
Broken windows
Differential item functioning
Multilevel CFA
Multilevel MIMIC

ABSTRACT

Objectives: (1) To determine whether perceived physical and social incivilities are distinguishable at the individual and/or neighborhood levels and, if so, whether there are differences in effects on fear of crime. (2) To identify, characterize, and account for differential item functioning (DIF) to understand differences in subjective perceptions of incivilities across demographic groups.
Methods: This study uses data from a probability sample of 1622 residents nested within 66 ecologically valid neighborhoods and employs multilevel SEM to identify factor structure, assess DIF, and examine structural relations at individual and neighborhood levels.
Results: Physical and social incivilities are distinguishable at the individual level but not at the neighborhood level. Three physical incivilities items exhibit DIF for race and three social incivilities items exhibit DIF for age. Residents in neighborhoods with higher concentrations of African Americans report greater levels of combined incivilities, but, within neighborhoods, African Americans perceive lower levels of physical and social incivility. Within neighborhoods, social incivilities link to fear of crime.
Conclusions: Demographic factors affect how individuals use response categories for gauging perceived incivilities in their locale. Discriminability of underlying separate physical and social components only at the individual level points to needed areas of theoretical elaboration in incivilities models.

1. Introduction

Over forty years of scholarship has found that resident perceptions of incivilities (disorder) link with a number of adverse outcomes, including heightened crime risk perceptions, fear, health issues, and withdrawal from active outdoor activity (Gallagher et al., 2010; Perkins, Brown, & Taylor, 1996; Wyant, 2008). At the community level, incivilities are argued to facilitate the decline of entire neighborhoods (Skogan, 1992). The fundamental idea that signs of incivility that remain unchecked reflect (Hunter, 1978) or pave the way for more serious crime (see Skogan, 1992; Wilson & Kelling, 1982) serves as the theoretical underpinning of several policing initiatives (Kelling, 2015).[1] Public officials often attribute crime reduction—such as that seen in New York City in the 1990s—to the eradication of uncivil behavior. Scholars have contributed by demonstrating that community and problem-solving strategies focusing on disorder reduction may account for some of these declines (Braga, Welsh, & Schnell, 2015). Other research shows, however, that the incivilities-crime connection—if the two constructs are fundamentally distinct (Gau & Pratt, 2008)—is relatively weak (Sampson & Raudenbush, 2004; Taylor, 2001). Given the importance of investigating the validity of the incivilities thesis to inform cost-effective community enhancement and crime reduction strategies, recent calls have been made to reexamine the broken windows thesis; Kubrin (2008), p. 204) has argued incisively that "the most important step in this process is to reevaluate the central concept of disorder itself".

There are core, unresolved conceptual and measurement issues hampering research on incivilities, and therefore hindering clarity on possible implications for crime and for individual and community quality of life (Gau & Pratt, 2008; Kubrin, 2008; see also Taylor, 1999). Specifically, Kubrin (2008), extending a list of concerns previously discussed (Taylor, 1999, 2001), notes widespread variation across studies and a lack of consensus on terminology (e.g., disorder vs. incivilities), conceptualization (e.g., distinction between physical and

* Corresponding author at: 527 Gladfelter Hall, 1115 Polett Walk, Philadelphia, PA 19122, United States
E-mail addresses: jeffrey.ward@temple.edu (J.T. Ward), nathan.link@temple.edu (N.W. Link), rbrecken@temple.edu (R.B. Taylor).

[1] The broken windows idea has been used to support alternative policing strategies, including broken windows scholars' preferred approaches (e.g., community policing) and other approaches (e.g., zero tolerance policing; or stop, question, and frisk) as well. Scholars have questioned (Kelling, 2015) how well these latter strategies align with broken windows theorists' intentions.

http://dx.doi.org/10.1016/j.jcrimjus.2017.06.004
Received 3 May 2017; Received in revised form 15 June 2017; Accepted 15 June 2017
Available online 29 June 2017

social incivilities; the subjectivity of incivilities perceptions and their context-dependent nature; conceptual overlap with crime and other constructs) and measurement (e.g., discord between subjective and objective assessments of incivilities).[2] The current study focuses on three key interrelated issues, the second of which has not been a focus in criminology: (1) whether physical and social incivilities are empirically distinct; (2) the subjectivity of incivilities perceptions by examining whether individuals of different demographic backgrounds use certain item response categories similarly when they share the same underlying level of the incivilities perceptions (i.e., are certain incivilities indicators biased and, if so, is this due to item-wording or definitional issues in the construct across groups?); and (3) the multilevel nature of the incivilities construct and similarities and/or differences in measurement and structural relations across levels. Building on a rich body of research that has addressed some of these issues individually, the present work provides critical new insights and advances the literature by recognizing the important *interconnectedness* of these issues which, for example, leads us to investigate whether physical and social incivilities factors emerge at neither, one, or two levels of analysis.

To do so, we exploit a probability sample of 1622 residents nested within 66 Baltimore neighborhoods and employ multilevel structural equation modeling. Specifically, we use multilevel confirmatory factor analysis (MLCFA) models to explore whether incivilities constitute one (i.e., combined) or two (i.e., physical and social) constructs at both the individual and neighborhood levels of analysis, the degree to which each indicator is reflective of individual vis-à-vis neighborhood incivilities, and how pure or strong indicators are of factors at each level of analysis. We further build on the MLCFA models by incorporating covariates to examine differential item functioning (DIF) using Multi-Level Multiple Indicators and Multiple Causes (MLMIMIC) models. This approach allows us to identify and characterize items that may be functioning differently across key demographics including age, gender, and/or race. In performing our item analysis, we also obtain insights into how precisely the incivilities indicators are measuring individuals' perceptions across the latent continuum of incivilities factor(s) at both levels of analysis. Finally, we put all the pieces together in a multilevel structural equation model to examine associations between covariates, fear of crime, and the incivilities construct(s) across each level of analysis, while accounting for DIF in items. To explicate the motivation behind our research, we begin by reviewing studies examining incivilities constructs (with attention paid to distinctions between physical and social indicators), the subjective nature of incivilities perceptions and whether demographics influence how they are perceived, and multi-level analyses.

2. Distinguishability of social and physical incivilities

By the early 1980s, scholars were distinguishing different types of incivilities, and separating the social from the physical (Skogan-& Maxfield, 1981; Taylor, Shumaker, & Gottfredson, 1985, p. 263). Generalizing this distinction, Skogan (1992, p.4) suggested that "physical disorder refers to ongoing conditions" which includes things such as unkempt lawns, trash-filled lots, and abandoned buildings, whereas "social disorder appears as a series of more-or-less episodic events" which includes behaviors that can be witnessed or experienced, such as seeing teens hanging out on street corners and public drunkards or experiencing insults, rowdy neighbors, or sexual harassment. To be fair, however, there are statements in the literature that invite confusion. For instance, Skogan (1992) discussed graffiti and vandalism as evidence of social disorder. The actual acts of vandalism and graffiti themselves itself could be considered a social disorder but, unlike catcalling or insulting remarks by neighbors, there are clear physical signs of disorder left behind from vandalism and graffiti. It has been argued that drawing the distinction is important because social and physical disorder may have different causes and variable effects (Matthews, 1992; Skogan, 1986, 1992; Skogan & Maxfield, 1981; Taylor & Hale, 1986; Taylor & Schumaker, 1990). Empirical work in this area, however, is mixed; some studies found support for separating the two constructs in empirical models (Ross & Mirowsky, 1999; Taylor, 1999), while others argued that both types load onto one broader, underlying assessed incivilities construct at the streetblock level (Taylor et al., 1985) or underlying perceived disorder construct at the individual level (Ross & Mirowsky, 1999; Xu, Fiedler, & Flaming, 2005).

Taylor (1999) relied on data from Baltimore to examine changes in survey-based physical and social incivilities between 1982 and 1994. Aggregated to the neighborhood level (30 neighborhoods), Taylor employed exploratory principal components analysis and found both physical and social incivilities components.[3] Other scholars, however, take issue with the concept of distinct physical and social incivilities factors. These arguments are both theoretical and empirical. For example, Xu et al. (2005), in their study of community policing of crime and disorder in Colorado, conflated physical and social disorders in their measurement model and made a conceptual case for doing so. They argued that blending the two constructs makes sense since a broader measure of disorder captures perceived global health of a community, as a community may show signs of particular indicators but not be in decline overall.[4] This logic, however, obfuscates the theoretical importance of measuring distinct facets of disorder and their potential impacts on outcomes ranging from increased fear, crime, reduced outdoor activity and consequently worsened health. As an illustration, using data from the Project on Human Development in Chicago Neighborhoods (PHDCN), Molnar, Gortmaker, Bull, and Buka (2004) found that social disorder, such as public substance use, was associated with significant reduced physical outdoor activity among children and adolescents but the same was not found with physical disorder. Theoretically, this seems tenable as parents' may be more concerned about their children playing outdoors in the presence of drug users than they would be in the presence of properties with unkempt lawns. In short, blending the two types of indicators may lead to empirical models where significant associations may be masked and, thus, inefficient policy recommendations may be advanced. Of course, a concern with all studies using PHDCN data is that those neighborhoods, defined by identifying clusters of demographically similar census units, ignore extant neighborhood boundaries in the locale. It seems worth exploring neighborhood patterns for these items using neighborhoods that instead are grounded in local history and organizations. We will do that here.

Using a probability sample from the 1995 Survey of Community, Crime, and Health, Ross and Mirowsky (1999, p. 424) tested the distinctions between types of neighborhood disorder and concluded:

"On the whole, social and physical aspects of perceived disorder indicate one underlying concept. Many of the physical aspects of a neighborhood, such as graffiti, noise, vandalism, dirt, and grime, are indicators of the breakdown of social control. They are clear cues to residents that people are involved. Thus the distinction between social and physical disorder is not clear-cut."

Their empirical tests indicated two underlying concepts related to disorder: disorder and physical decay. In the first analysis, they conducted a unidimensional factor analytic model that contained fifteen items related to disorder, including social and physical incivilities, crime, and items related

[2] We have chosen to call visible social and physical indicators of neighborhood problems in line with Broken Windows as "incivilities." This is based on Hunter's (1978) scholarship showing that "disorder" is a broader social condition of some neighborhoods that can manifest incivilities. Recent scholarship, however, has used incivilities, disorder, and neighborhood problems interchangeably. "Incivilities" is used over "disorder" in this study because it is less ambiguous given Hunter's work. The one exception is that we use the term disorder when reviewing others' work that uses the disorder terminology.

[3] Though the vandalism indicator is theoretically categorized as a physical incivility, empirically it showed to have moderate loadings on both components.

[4] Empirical analyses, offered as a secondary justification for their combined measure, found inadequate discriminant validity between the physical and social dimensions (Xu et al., 2005, p. 163).

to social order (mostly the semantic opposites of disorder items). This model showed less than adequate fit and modification indices indicated several correlated error terms with respect to certain physical incivilities items. As such, they specified a second factor they termed "decay", which captured six of the "purest" physical disorder items. Their final model containing two factors—disorder and decay—in addition to one factor accounting for item wording (agreement bias), showed good fit to the data. Most items loaded on the disorder factor, two loaded on the decay factor, and four loaded on both. Ross and Mirowsky identified noise—usually a concept classified as a social disorder—as a physical disorder and results showed that the noise item loaded more strongly on the broader disorder factor (that captures social disorder) than on the physical decay factor (0.522 vs. 0.106). Ultimately, Ross and Mirowsky concluded that although disorder and decay are highly related to one another, they are distinct concepts. Social disorder, according to their analyses, is subsumed within the broader disorder factor and does not emerge as its own independent factor.

In sum, many researchers have distinguished between physical and social incivilities (disorder) (Matthews, 1992; Sampson & Raudenbush, 1999; Skogan, 1992; Skogan & Maxfield, 1981; Taylor, 2001; Taylor & Hale, 1986), but others have suggested that this conceptual distinction is not always necessary (Ross & Mirowsky, 1999; Xu et al., 2005). Work in this area has yielded mixed results, with some showing evidence of separate constructs of incivilities (Ross & Mirowsky, 1999; Taylor, 1999) and some showing evidence for combining both into one broader incivilities construct (Ross & Mirowsky, 1999; Xu et al., 2005). Empirically, what is clear from prior research is that social and physical incivilities are highly correlated. What remains unclear, however, is whether the degree of association is such that adequate discriminant validity between the types of incivility can be achieved. Our view is that the answer to this question necessitates careful consideration of the multilevel nature of the incivilities construct—a new contribution to the literature and point to which we return shortly.

3. Subjectivity of perceiving incivility items

Individuals rating the conditions of their neighborhood have been considered as either "informants," who report on underlying variation in neighborhood incivilities in a more-or-less objective fashion, or "critics," who view environmental conditions through individual lenses (Skogan, 1992, p. 53). Recent scholarship has challenged the notion that people see and identify signs of incivilities similarly (Harcourt, 2009; Sampson & Raudenbush, 2004).[5] Consistent with studies of other criminological constructs such as deterrence (Stafford & Warr, 1993) and peer delinquency (McGloin & Thomas, 2016), recent research on incivilities finds that incivility perceptions are not solely a function of objective accounts of incivilities (Franzini, Caughy, Nettles, & O'Campo, 2008; Hipp, 2010). Instead, incivilities are socially constructed—at least in part—with some people, and some people in some places, defining incivilities as more problematic than others. Incivilities perceptions are shaped by demographic factors such as age, race, sex, and certain features at the neighborhood level such as race and class composition (Franzini et al., 2008; Hipp, 2010; Jackson, 2004; Sampson & Raudenbush, 2004; Wallace, 2011; Wickes, Hipp, Zahnow, & Mazerolle, 2013). Kubrin (2008) and others appear correct in noting that the reliance on "objective" measures of incivilities is potentially problematic. Doing so ignores the reality that different people ascribe different meanings to symbols or to people in their environments (Harcourt, 2009) and that perceptions are context-dependent (Taylor, 1999). Perhaps it is not surprising, then, that reactions or behavioral adaptations to objective incivilities are not uniform across demographics.[6]

The idea that a group may be more (or less) bothered by physical and/or social incivilities than their demographic counterpart warrants further consideration than prior research has given it; this is because accurately quantifying group differences in incivilities perceptions hinges on using unbiased scale items or, in other words, items that have similar response probabilities across groups when individuals have the same underlying incivilities perceptions. The central problem is that observed associations could reflect true differences in unbiased incivilities scales and/or measurement bias in one or more scale items across demographical lines, the latter of which is known as differential item functioning. Assessing DIF means identifying the extent to which groups of individuals (e.g., younger vs. older) *experience* more incivility and *understand or define* the construct differently.

DIF can result from item wording that leads to inconsistent interpretation across groups. This can occur if the words used in an item have alternative meanings due to cultural differences or lack clarity in one group but not another. In an Item Response Theory (IRT) framework (see Embretson & Reise, 2000), commonly employed in psychology and educational testing, a biased item would be one that measures ability (e.g., math performance) with a question that contains vocabulary that would be differently understood across groups. If DIF exists and item wording is not to blame, DIF can represent fundamental differences in the very definition of the construct across groups. As constructs of orderly and disorderly are argued to be inherently subjective (Harcourt, 2009), it is therefore possible that the very definitions of incivility indicators vary across groups. Because underlying levels of the construct (e.g., social incivilities) are controlled, significant DIF for an item can signify that a certain group (e.g., older) may need to see more (or less) of that particular incivility in order to have comparable overall incivilities perceptions to their demographic counterparts (e.g., younger).[7]

Importantly, despite the sizable amount of recent work using scales to demonstrate variation in incivilities perceptions along demographic lines, the degree to which commonly used incivilities *indicators* exhibit DIF is unknown. Identifying the degree of DIF is crucial for identifying poorly worded items that may need deletion or revision and/or to identify the precise items contributing to definitional differences in the incivilities construct(s) across groups. Modeling DIF is also important because analyses that fail to account for DIF when assessing group differences in the construct or its relationships to external variables could be misleading (Fleishman, Spector, & Altman, 2002); this is especially the case for smaller scales, as in those with few indicators per construct (see Teresi, Ramirez, Jones, Choi, & Crane, 2012). With many studies employing five or fewer indicators per incivility dimension (Hipp, 2016; McGarrell, Giacomazzi, & Thurman, 1997; Perkins & Taylor, 1996; Ross & Mirowsky, 1999; Sampson & Raudenbush, 2004; Xu et al., 2005), the need to assess the role of item-level DIF is particularly important. In short, the present study draws explicit attention to the potential problem of DIF in incivilities research and clarifies its consequences—small or large—for our understanding of demographic differences in subjective perceptions of incivility.[8]

4. Individual perceptions and neighborhood contexts

In roughly a fifteen-year span from 1975 to 1990, theoretical

[5] The basis for this challenge has theoretical roots in work from much earlier times (Mead, 1934).

[6] There are many exceptions where perceived incivilities have been studied, however. (See, among others, Hinkle & Yang, 2014; Link et al., 2017; Taylor, 1994; Wallace, 2011; Yang & Pao, 2015).

[7] Put differently, all else being equal, two groups that have similar perceptions on true DIF items will actually have *different* overall perceptions. As a non-criminology example, it is commonly known that an item like "I cry a lot" in depression scales functions differently across men and women such that women are more likely to endorse the item, even when they have the exact same depression levels. That is, controlling for depression, women are more likely to cry. Thus, all else being equal, men and women who cry equally, given the existence of DIF, would have different overall levels of depression with males being higher. Ignoring DIF can thus confound item bias with trait differences.

[8] A general focus on item-level analysis and measurement also permits an assessment of the varying degree of precision in measurement for individuals that span the spectrum of perceived incivilities factors. Should physical and social incivility separate at the individual level it would also permit an assessment whether certain demographic factors exhibit DIF for neither, one, or both types of incivility.

models of incivilities evolved from a sole focus on psychological outcomes (e.g., see Wilson, 1975; Garofalo & Laub, 1978) to a predominant focus on ecological processes (e.g., see Skogan, 1992). The implications of this general theoretical shift upward toward macro level units of analysis imply substantial between-neighborhood variation in incivilities and that ecologically-based—as opposed to psychologically-based—indicators might be preferable (Taylor, 1999). However, as previously discussed, assessed neighborhood indicators, obtained through assessments (see Sampson & Raudenbush, 1999; Taylor et al., 1985), neglect resident subjectivity and related within-neighborhood (individual) psychosocial processes, and can be subject to issues of observer bias (Hoeben, Steenbeek, & Pauwels, 2016). Indeed, recent empirical work suggests that individual perceptions of disorder remain relevant (Hinkle & Yang, 2014; Hipp, 2010; Wallace, 2011; Yang & Pao, 2015) and the oft-cited Broken Windows Theory acknowledges both levels of analysis (Wilson & Kelling, 1982).

Several recent studies have measured incivilities at the individual level—either as a predictor or outcome—without a similar assessment of the neighborhood level contribution in the construct. Hinkle and Yang (2014) predicted individual-level perceptions of social disorder among a sample of New Jersey residents; Hipp (2010) assessed perceived social and physical disorder in fixed effects models (i.e., models that controlled for each household cluster); and Wallace (2011), using a similar fixed effects approach, tested the impact of routine activities and neighborhood attachment on individual disorder perceptions. At the same time, several studies have taken the opposite approach and have focused solely on the neighborhood level impacts and correlates of incivilities. Sampson and Raudenbush's (1999) well-known study from Chicago, building on pioneering earlier streetblock on-site assessments of land use, and physical and social incivilities by Taylor and Perkins and colleagues (Taylor et al., 1985; Perkins, Meeks, & Taylor, 1992), gauged both perceived incivilities and assessed features by driving up and down streets and aggregating indicators to the census tract cluster level. Similarly, Taylor (1996) again assessed physical deterioration—an aspect of fundamental or “pure” physical disorder—among Baltimore neighborhoods using a principal component score based on structured on-site assessments of streetblocks.

A smaller number of studies have measured incivilities at both the individual and community levels. Using resident surveys, Perkins and Taylor (1996) examined the nexus between incivilities and fear of crime by exploiting multiple indicators of both social and physical incivility at both the individual level and aggregated to the streetblock level. In addition, Sampson and Raudenbush (2004) measured signs of disorder at two levels of analysis. Their survey data measured perceived social and physical disorder as the outcome. They predicted these outcomes with block-level data on physical and social disorder from systematic social observation and area characteristics as well as various individual and group level demographic variables. Underscoring the critical importance of disentangling within- and between-neighborhood effects in perceived disorder, findings revealed that the effect of race (i.e., Black relative to White) operated in opposite directions across levels of analysis (Sampson & Raudenbush, 2004).

While informative, the studies that have employed multi-level measures of incivilities have utilized a hierarchical linear modeling framework and have employed manifest (as opposed to latent) variables. This observed variable approach assumes that each indicator is an equally good measure of the construct—which may not be true—and precludes any item-level assessment of whether indicators function differently across demographic groups after accounting for the incivilities construct itself. Importantly, these studies also do not investigate how factor structures may vary or show uniformity across levels. In sum, despite theoretical and empirical reasons for treating incivilities perceptions as a multilevel construct, more studies have measured incivilities at the individual level than any other, and the literature has not addressed the possibility that the factor structure of incivilities may vary across levels and that differential item functioning may exist in certain incivilities indicators for one or more demographic groups. To fill these gaps in the literature, the present study complements path-breaking research on perceptions of incivilities at the individual and neighborhood levels of analysis by assessing the implications of unaddressed measurement issues for multilevel investigations into the sources and consequences of perceived incivilities. We now outline our research questions.

5. Research questions

Intersecting the focal issues of whether physical and social incivilities are separable, the degree to which items exhibit differential item functioning and how this affects conclusions about associations and the subjectivity of perceptions across demographic groups, and the multi-level nature of the construct, the current study addresses the following research questions: (1) Are physical and social incivilities distinguishable at the individual and/or neighborhood levels of analysis?; (2) To what extent do incivilities indicators exhibit differential item functioning across age, gender, and race?; (3) Do indicators provide more information, or in other words more precise measurement, for individuals and neighborhoods with particular levels of (physical and/or social) incivilities?; (4) What is the effect of demographic effects on (physical and/or social) incivilities, after adjusting for DIF?; and finally, (5) If incivilities are indeed distinguishable at either level of analysis, are there dissimilar relationships between these constructs and measures of fear during the day and at night?

6. Data and measures

The data used in this study derived from a probability sampling design of residents in Baltimore, MD. Originally collected to examine the relationships among crime, residents' attitudes, physical and social deterioration, and neighborhood structure, the data collection effort had a particular emphasis on measuring neighborhood-level incivilities and responses to them. As such, they are ideal for our purposes in the current investigation. As others have demonstrated (Hinkle & Yang, 2014; Kubrin, 2008; Swatt, Varano, Uchida, & Solomon, 2013), measuring incivilities using surveys of residents, as opposed to “objective” assessments, holds certain advantages. The main advantage vis-à-vis the Broken Windows thesis is that *perceiving* signs of incivility is the key factor that triggers other social and psychosocial processes, including increased risk perceptions, fear, and community withdrawal.

Among 236 neighborhoods in Baltimore, 66 were randomly sampled in 1982 for inclusion in the study. It is important to note that the neighborhoods were defined ecologically in 1979, using input from extant community organizations at the time and from local district planners (Taylor, Brower, & Drain, 1979). These units had at the time an ecological validity not available in other neighborhood level samples in other cities. They were used as the basis for the 1980 Census Neighborhood Statistics program. Households in these neighborhoods were selected for the survey through multistage random sampling techniques (for more details on data collection, see Taylor, 1994; Taylor et al., 1985). Within each neighborhood, eight census blocks were randomly selected, and one side of each block was randomly selected if it included residential telephone listings and did contain solely apartments. Interviews were completed with 1622 heads of household (88% by phone, 12% in-person). The initial response rate was 87%.

The sample was 56% White and 44% African American, 33% male and 67% female, and had a median age of 46. Fifty-five percent reported being married or living as married, and the median number of years of formal education was 12. The average income was between $20,000 and $25,000 in 1982 dollars.

6.1. Measures

6.1.1. Neighborhood disorder

6.1.1.1. Social incivilities items. Five items were used as indicators of social incivilities. Measured on a Likert scale from "not a problem," to "somewhat of a problem," to "a big problem," respondents were asked the degree to which the following represented problems in their neighborhood: (1) "groups of teenagers hanging out", (2) "people who say insulting things or bother other people when they walk down the street", (3) "bad elements moving in", (4) "people fighting and arguing", and (5) "the amount of noise in the area".

6.1.1.2. Physical incivilities items. The following five items were used as indicators of physical incivilities using the same Likert scale: (1) "vandalism, like people breaking windows or spray painting buildings", (2) "vacant housing", (3) "people who don't keep up their property or yards", (4) "litter and trash in the streets", and (5) "vacant lots with trash or junk".

While we report detailed psychometric analyses below, Cronbach's alpha for the combined incivilities scale is 0.87. For social and physical items alone it is 0.82 and 0.75, respectively.

6.1.2. Covariates

Age was originally measured in years and is dichotomized using the median as the cut-point (1 = 46 and under, 2 = 47 and up). *Gender* is a dichotomous measure (1 = Male, 2 = Female). *Race* is a dichotomous measure (1 = White, 2 = African American).

6.1.3. Dependent variables

Fear of crime was measured using two items originally conceived in the National Crime Victimization Survey (NCVS): "How safe would you feel being out alone in your neighborhood during the day?" and "…at night?" Likert responses categories ranged from "very safe" to "very unsafe."

7. Methodology: multilevel sem

To address the study's research questions, we employ multilevel structural equation modeling, including the estimation of confirmatory factor analytic (MLCFA) and multilevel multiple indicator and multiple causes (MLMIMIC) models.

7.1. MLCFA model to assess multilevel factor structure

MLCFA seeks to explain common variance in survey items from one or more latent factors at two levels of analysis. When data are nested, such as in neighborhoods, a chief advantage of MLCFA compared to other factor analytic approaches (e.g., single level EFA and CFA, hierarchical latent variable analysis) is the ability to model and compare different factor structures at each level of analysis. In the present context, MLCFA permits the determination of whether individual perceptions and/or aggregated perceptions of neighborhood incivilities are usefully separated into physical and social components. Technical details of MLCFA models (see Dunn, Masyn, Johnston, & Subramanian, 2015; Muthén, 1991) and MLMIMIC models for DIF detection (see Kamata & Vaughn, 2011) have been previously described in detail. Here, we elect to review conceptual aspects of these models with emphasis on applied analytic procedures.

We begin by estimating within- and between-level item correlations and intraclass correlation coefficients to examine whether there is evidence that incivilities should be modeled as a multilevel construct. Next, we estimate alternative multilevel unidimensional and correlated factors confirmatory factor analytic models specifying either one (combined incivilities) or two (physical and social incivilities) factors at the within (W) and/or between (B) neighborhood levels. Four possible factor structures are compared: 1W-1B; 2W-2B; 1W-2B; and, 2W-1B. For instance, the 1W-2B model specifies one combined incivilities factor at the within level (i.e., individual level) but separate, correlated physical and social incivilities factors at the between level (i.e., neighborhood level).[9] The alternative MLCFA models are compared using standard fit criteria including comparative fit index (CFI), Tucker-Lewis Index (TLI), root mean square error of approximation (RMSEA), and within- and between-level standardized root mean residuals (SRMR), the latter of which provides insight into fit at each level. Conventional cutoffs indicating good model fit include: CFI and TLI > 0.95; RMSEA < 0.06; and SRMR < 0.08 (see Hu & Bentler, 1999). We re-estimate the models using computationally slower maximum likelihood estimation to obtain BIC values to aid comparison of non-nested models. BIC differences > 10 indicate very strong support for one model over another (Raftery, 1995).

7.2. MLMIMIC model for DIF detection and item information

Following identification and interpretation of the best fitting MLCFA model, we use MLMIMIC models to probe for differential item functioning (DIF) across age, gender, and race at the within level of analysis.[10] For example, we ask whether young and old individuals have different response probabilities to any items even though they are within the same neighborhood and, importantly, share the same level on the incivilities factor. The general framework for DIF detection with MIMIC models was advanced some thirty years ago (Muthén, 1985, 1988, 1989) and has several advantages.[11] Most relevant to the present study, MIMIC modeling permits DIF detection with multidimensional constructs and allows for the inclusion of multiple covariates. MIMIC models have been shown to have similar power as IRT-based methods for the detection of uniform DIF (Stark, Chernyshenko, & Drasgow, 2006; Woods, 2009). We employ the MLMIMIC approach for the detection of uniform DIF by adding three demographic covariates to the multilevel model at the within level, which are permitted to influence both the factor(s) and the individual items. DIF is detected if a covariate has a direct effect on an item after controlling for levels of the factor(s). Conceptually, if an item is without bias, a covariate (e.g., age) should only influence item response through an underlying factor. In short, the pathways from the covariate(s) to a factor capture differences in factor means across groups, whereas direct pathways from the covariate(s) to the items represent DIF.

Kamata and Vaughn (2011) illustrate an intuitive approach to multilevel DIF detection using the free baseline method with selection of an anchor through preliminary DIF assessment.[12] To guard against

[9] For the MLCFA analyses, we employ robust weighted least square (WLSMV) estimation and model the polytomous survey items as categorical. In one set of analyses, we set the scale by fixing an arbitrary loading for each factor to 1; this permits a significance test for variance in incivilities construct(s) at the within and between levels. This is useful for determining whether construct(s) at each level of analysis have significant variation and may therefore be used to predict external variables such as crime/crime rates. We also obtain the standardized solution, which is achieved by setting the scale through fixing factor variances to 1 and freeing all factor loadings. This parameterization is useful for examining factor loading strength.

[10] In general, there are two options for detecting DIF within a CFA framework which include multiple group analysis and MIMIC modeling. Multiple group analysis allows for a more complete assessment of measurement invariance including tests for factor loadings, item thresholds, and residual variances; however, multiple group analyses are far more complex within a multilevel context and are estimated as multilevel finite mixture models with known classes (see Asparouhov & Muthen, 2012). Moreover, one shortcoming of the multiple group approach is the inability to easily examine multiple grouping variables, such as age, sex and race, within the same model.

[11] The chief disadvantage is that it detects uniform DIF only, although latent factor by covariate interactions have been used to test for non-uniform DIF in a MIMIC framework in single level analyses (see Woods & Grimm, 2011). Uniform DIF refers to differences in thresholds (or "difficulty" in IRT terms) whereas non-uniform DIF refers to differences in factor loadings (or "discrimination" in IRT terms). See Kamata and Bauer (2008) for a research note on the connection between factor analysis and item response theory.

[12] Within a unidimensional MIMIC framework, two key alternative analytic strategies exist to detect DIF, including the "constrained baseline" and "free baseline" methods (see

inflated DIF detection with constrained baseline model approaches, we employ the "DIF-free-then-DIF" iterative procedure to identify anchors and test for DIF (Wang, Shih, & Sun, 2014; see also, Woods, 2009). The procedure is labor intensive and first employs a constrained baseline method to test each item for DIF using all other as anchors. Constrained approaches test each item individually for DIF using the other items as anchors. Items flagged as having DIF are removed from the potential anchor set (i.e., their loadings are freed) and a partially constrained baseline model is then estimated to flag additional items with DIF. This process is repeated until a subsequent iteration detects no additional DIF. The next step in the procedure involves ranking the non-DIF items based on the likelihood ratio to aid selection of anchors (Wang et al., 2014; Woods, 2009). However, in the multilevel context, we found this to be impractical given the need to employ computationally burdensome maximum likelihood estimation to obtain the required statistic. As a more practical alternative, therefore, we rank items based on their Wald Statistic, with lower scores indicating the best non-DIF anchors (Liu, 2011). For each covariate, we select one item as a reference for each factor at the within level.[13] Thus, there would be three referent items (i.e., one for each covariate) if there is one within factor but six referent items (i.e., two for each covariate) if there are two within factors. Selected anchors are then used to test the remaining items for DIF using the free baseline method, which allow all direct effects from covariates to items to be estimated with the exception of referent non-DIF items.

With DIF items identified through the above procedures, we estimate three alternative MLMIMIC models and compare them with the Satora-Bentler chi-square test. Of interest, we compare a no-DIF model (i.e., the MLCFA) to the DIF-only model, the former being nested in the latter. We fully expect the no-DIF model will fit significantly worse than the DIF-only model. Then, we compare a DIF-only model to the free-baseline model, with the former being nested in the latter. In this case, we fully expect the DIF-only model to fit no worse than the free baseline model and therefore to be preferred on the basis of parsimony.

We illustrate DIF in the DIF-only model using item characteristic curves, which are graphical representations that link specific "abilities" or latent factor scores to item response probabilities (see Embretson & Reise, 2000). For instance, an individual with a latent factor score of +1 (or 1 standard deviation above the mean) might have a 0.6, 0.3, and 0.1 probability of responding to categories 1 (not a problem), 2 (somewhat of a problem), and 3 (a big problem), respectively. With respect to DIF, items that are more difficult for one group (relative to another) have item characteristic curves that are shifted right (i.e., up the continuum of latent factor scores). We summarize the direction and size of significant DIF effects with odds ratios. We next highlight the overall precision of measurement for each latent factor across the continuum of scores using test information curves. These provide clarity as to where individuals along the continuum of latent factor incivility scores are measured with the greatest and least measurement error, which has implications for scale improvement efforts.

7.3. MLMIMC model to assess multilevel structural relations

In the final stage of our analysis, we examine mean differences across age, gender, and race in the incivilities constructs using the MLMIMIC (DIF-only) model, which accounts for the multilevel nature and adjusts for DIF. We assess structural relations between the incivilities constructs and fear of crime during the day and at night across two levels of analysis. Collectively, this permits an assessment of the implications of considering multilevel factor structure and item bias in estimates of differences in incivilities across age, gender, and race and for associations with external variables. All models are estimated in Mplus v 7.31 and all graphs are created in Stata v14.1 using exported Mplus plot data. We report pairwise present analyses, Mplus's default for handling missing data with WLSMV estimation. We also investigate primary models using listwise deletion and the results are substantively similar (available upon request).

8. Results

8.1. Descriptives, ICCs and correlations

Table 1 contains item descriptive information including wording, purported sub-dimension, item proportions and intraclass correlation coefficients (ICC). Across all items, category proportions indicate that a majority of respondents answered "not a problem" but there is a substantial percentage of respondents who characterize various incivilities as "somewhat of a problem" or "a big problem." Specifically, on average, approximately 67% of respondents mark the lowest category, 22% the middle category, and 11% the highest category. This variation in item response is attributable to both individual and neighborhood units of analysis. For instance, the average ICC for the ten items is 0.181 with a median value of 0.157. The ICCs range from a low of 0.099 to a high of 0.283. Thus, each item has at least approximately 10% of the total variation attributable to neighborhoods and this degree of between-neighborhood variability in item responses is supportive of the need to employ multilevel analysis. At the same time, the ICC values suggest that the majority of the variation in items is due to individual differences within, rather than between, neighborhoods. Finally, the spread of the ICC values across the ten items implies that all items do not possess the same salience for understanding neighborhood context. For instance, Items B (vacant housing), J (people fighting and arguing), and F (vacant lots with trash or junk) have the highest salience for understanding between neighborhood differences in incivilities.

Table 2 contains the within level and the between level polychoric item correlations along the lower left and upper right of the table, respectively. The within-level (individual) item correlations range from 0.296 (Items B and H) to 0.681 (Items G and H), with an average item correlation equal to 0.468. The between-level (neighborhood) item correlations range from 0.528 (Items A and F) to 1, with an average item correlation equal to 0.861. It is clear that the item correlations are considerably higher at the between level. Indeed, item correlations differ in size across levels on average by 0.398, with 0.15 serving as the minimum difference (Items A and G) and 0.614 being the maximum difference (Items C and D). Comparison of a rank ordering of the correlations at each level provides further insights, since absolute loading strength might differ across levels but relative ordering could be quite similar. The item correlations themselves correlate across levels only moderately ($r = 0.44$, $p < 0.05$). In short, there is nontrivial variability in items at both levels of analysis. Moreover, the dissimilarity in interrelationships among items across the within and between levels of analysis support MLCFA, and this dissimilarity suggests a legitimate possibility that different factor structures might exist across levels.

8.2. MLCFA models

Table 3 contains fit statistics for four alternative MLCFA models. Factor structures with neighborhood incivilities specified as a single construct at the within level (Models 1 and 2) have substantially worse fit than models that have physical and social incivilities as correlated factors at the within level (Models 3 and 4). For example, Models 1 and 2, which have only one within factor, have fit statistics that indicate either poor (e.g., TLI < 0.90) or less than optimal (e.g., CFI < 0.95)

(*footnote continued*)
Wang, 2004; see also Stark et al., 2006). Constrained baseline methods are simple to apply but have higher Type 1 error rates, resulting in the detection of spurious DIF (Wang et al., 2014).

[13] This choice is critical because DIF coefficients represent the degree of DIF compared to the DIF-free item(s). Thus, the goal is to select item(s) as anchors that are in fact unbiased to facilitate parameter interpretation. While each additional item risks contamination of the referent, additional anchors provide increased power (Wang, 2004).

Table 1
Item descriptives and intraclass correlation coefficients.

Item ID	Item wording	Incivilities type	Intraclass correlation	Category proportions		
				Not a Problem	Somewhat of a problem	A big problem
A	Vandalism, like people breaking windows or painting buildings?	Physical	0.099	0.619	0.275	0.107
B	Vacant housing?	Physical	0.283	0.801	0.144	0.055
C	People who don't keep up their property or yards?	Physical	0.115	0.556	0.344	0.100
D	People who say insulting things or bother other people when they walk down the street?	Social	0.154	0.766	0.158	0.076
E	Litter and trash in the streets?	Physical	0.196	0.593	0.274	0.133
F	Vacant lots with trash or junk?	Physical	0.238	0.761	0.140	0.099
G	Groups of teenagers hanging out?	Social	0.148	0.546	0.277	0.177
H	The amount of noise in the area?	Social	0.160	0.560	0.269	0.171
I	Bad elements moving in?	Social	0.150	0.750	0.160	0.090
J	People fighting and arguing?	Social	0.268	0.740	0.181	0.079
Average			0.181	0.669	0.222	0.109

Table 2
Within-level and between-level item correlations.

	A	B	C	D	E	F	G	H	I	J
A. Vandalism	–	0.598	0.703	0.825	0.782	0.528	0.651	0.731	0.860	0.805
B. Vacant housing	0.425	–	0.936	0.855	0.827	0.859	0.716	0.810	0.878	0.775
C. Yards	0.407	0.495	–	0.999	0.967	0.961	0.851	0.931	0.999	0.993
D. Bother	0.470	0.306	0.385	–	0.964	0.775	0.806	0.955	1.000	0.963
E. Litter	0.457	0.379	0.533	0.478	–	0.861	0.852	0.931	1.000	0.956
F. Vacant lots	0.376	0.418	0.439	0.391	0.597	–	0.730	0.859	0.880	0.839
G. Teenagers	0.501	0.332	0.457	0.598	0.512	0.426	–	0.891	1.000	0.880
H. Noise	0.410	0.296	0.393	0.505	0.531	0.443	0.681	–	1.000	1.000
I. Bad elements	0.383	0.356	0.531	0.519	0.515	0.484	0.550	0.555	–	1.000
J. Fighting	0.461	0.375	0.415	0.611	0.480	0.410	0.577	0.577	0.647	–

Notes: Within-level (lower left); Between-level (upper right).

Table 3
Fit statistics for multilevel confirmatory factor analytic models.

Model	Factor structure	χ^2	df	RMSEA	CFI	TLI	SRMR	BIC[a]
1	1W-1B	381.542	70	0.052	0.916	0.891	0.057 (w) 0.043 (b)	22,087.549
2	1W-2B	378.994	69	0.053	0.916	0.890	0.057 (w) 0.043 (b)	22,058.541
3	2W-1B	235.633	69	0.039	0.955	0.941	0.042 (w) 0.043 (b)	21,991.906
4	2W-2B	233.716	68	0.039	0.955	0.941	0.042 (w) 0.043 (b)	21,979.500

[a] Statistics obtained from empirical maximum likelihood estimation.

fit. BIC values also suggest Models 1 and 2 fit substantially worse than Models 3 and 4. Focusing on the two models that separate incivilities into its physical and social components at the individual level, fit is good and quite similar overall (see Models 3 and 4). RMSEA, CFI, TLI and SRMR within and between level fit statistics are identical out to three decimal places across these models. While traditional fit indices suggest both models fit the data well, the BIC statistic provides evidence in favor of Model 4 over Model 3, meaning that two factors at the between level might be preferred to one factor at the between level. Yet, the correlation between the latent social and physical incivilities factors at the neighborhood level in Model 4 (2W-2B) reveals lack of discriminant validity ($r = 0.98$, $p < 0.001$). The correlation between physical and social incivilities at the individual level is very strong ($r = 0.840$, $p < 0.001$), but its value indicates acceptable discriminant validity between these constructs.[14] The 2W-1B factor structure provides good fit to the data and is selected over the 2W-2B factor structure on the basis of acceptable discriminant validity and parsimony.[15] The unstandardized solution for the 2W-1B model suggests that there is significant factor variation in these constructs. Specifically, results indicate significant individual level variation in physical incivilities ($s^2 = 0.739$, $p < 0.001$) and social incivilities ($s^2 = 1.083$, $p < 0.001$), as well as significant neighborhood level variation in combined incivilities ($s^2 = 0.109$, $p < 0.05$).

Table 4 contains the standardized factor loadings for physical and social incivilities at the within level and for combined incivilities at the between level. The average item loading on the physical incivilities

[14] Cutoff for acceptable discriminant validity is $r < 0.85$ (see Kline, 2005, p. 73).

[15] Alternative one and two factor single level models adjusting for clustering provided good fit to the data by traditional fit indices. The two factor model was clearly preferred to the one factor model. Importantly, BIC scores for the single level one-factor model (22,387.012) and the single level two-factor model (22,284.085) were considerably higher (indicating worse fit) than for the MLCFA models permitting 2 factors at the within level. Interestingly, the correlation between physical and social disorder in the single level analysis was a bit higher (0.87) than the MLCFA putting it above conventional cutoffs for adequate discriminant validity.

Table 4
Standardized factor loadings for the 2W-1B MLCFA model.

Item ID	Within-level		Between-level
	Physical	Social	Combined
A. Vandalism	0.652		0.758
B. Vacant housing	0.578		0.856
C. Yards	0.691		1.000[a]
D. Bother		0.721	0.975
E. Litter	0.773		0.977
F. Vacant lots	0.677		0.862
G. Teenagers		0.808	0.872
H. Noise		0.773	0.973
I. Bad elements		0.772	1.000[a]
J. Fighting		0.783	0.986
Avg. loading	0.674	0.771	0.926

All loadings statistically significant ($p < 0.001$).
($\chi^2 = 239.910$, df = 71, CFI = 0.954, TLI = 0.942, RMSEA = 0.038).
[a] Residual variance set to zero.

(0.674) factor is lower than the average loading on the social incivilities factor (0.771). Moreover, the spread of loadings is slightly wider for physical incivilities (0.578–0.773) as compared to social incivilities (0.721–0.808). For both factors, however, loadings are substantial and statistically significant ($p < 0.001$). Average factor loadings on the combined incivilities factor at the neighborhood level are considerably higher (0.926) than loadings at the within level. Not surprisingly then, all factor loadings at the between level are statistically significant as well ($p < 0.001$).[16] The square of standardized loadings indicates the explained variance in the item or, more technically, the variance explained in the continuous latent response that underlies the categorical item (which is set to one in factor analysis of ordered polytomous data). For instance, 42.5% of the within-level variance in vandalism (Item A) is explained by physical incivilities, whereas 57.3% of the between-level variance in the item is explained by combined incivilities. It is useful to keep in mind that approximately 10% of the total variation in the vandalism item is at the neighborhood level (see Table 1). Thus, while a lower percentage of the variation in Item A is explained at the within level, this still represents a greater overall proportion of the total item variance in the underlying continuous latent response variable explained, since 90% of the variation is at the within level.

In sum, the 2W-1B model fit the data well and provides adequate discriminant validity among the individual level factors. Thus, the answer to the first research question is that physical and social incivilities are separable at the within level but not at the between level. All factors have significant factor variation, suggesting utility as exogenous or endogenous variables in multilevel structural equation models. The 10 items reflect considerably more within level variation in incivilities than between level variation as indicated by the ICCs (Table 1). However, the standardized loadings (and their squares) suggest that items are purer indicators of combined incivilities at the neighborhood level than they are indicators of physical and social incivilities at the individual level (Table 4). Comparing the individual level factors, the social incivilities items are purer indicators of its respective factor than are the physical incivilities items.

8.3. MLMIMIC models

8.3.1. Differential item functioning

We now add covariates to the 2W-1B MLCFA model to estimate a series of MLMIMIC models to inspect and adjust for DIF at the individual level. Table 5 summarizes the results from the multistage "DIF-free-then-DIF" detection method.[17] The far right of Table 5 identifies the final set of items exhibiting DIF as detected from the free baseline MLMIMIC for DIF detection with DIF-free anchors. Six items, as opposed to eight, are found to exhibit DIF. Items A, B, and F are confirmed to function differentially across race. More specifically, odds ratios showed that African Americans with similar factor levels in physical incivilities perceptions score systematically lower on the vandalism item (Item A) and systematically higher on the vacant housing and vacant lots items (Items B and F). Items D, H, and I, which are all social incivilities items, are confirmed to operate differently across age. Older individuals with similar factor levels in social incivilities perceptions score systematically lower on the bother item (Item D) and systematically higher on the noise and bad elements items (Items H and I). No items were found to be functioning differentially across gender.

A Satora-Bentler Chi-square test for decrement in fit of nested models reveals that a no-DIF model, as compared to the DIF-only model in which is it nested, fits the data significantly worse ($\Delta\chi^2 = 63.97$, df = 6, $p < 0.05$). Thus, the model that accounts for the six DIF effects is preferred over the model that assumes no DIF. We also compute a second Satora-Bentler Chi-square test that compares the fit of the free baseline DIF model to the DIF-only model. Constraining the non-DIF pathways to zero does not result in worse fit ($\Delta\chi^2 = 16.80$, df = 18, $p > 0.05$), suggesting the more parsimonious DIF-only model is preferable to a model that estimates all possible DIF effects. In short, the DIF-only model is the preferred model because it accounts for the DIF across racial groups in the physical incivilities construct and the DIF across age in the social incivilities construct at the within level of analysis, while also excluding unnecessary direct pathways between the covariates and items that are not statistically significant.

Fig. 1 shows item characteristic curves (ICCs)—which provide a useful way to visualize uniform DIF—for the six biased items. For instance, the ICCs for Item I show they are shifted left for older individuals as compared to younger individuals. The "bad elements moving in" item is therefore less difficult for older people to endorse; in other words, older people are *more* concerned with bad elements moving in. When individuals have the same social incivilities factor levels, Item I functions differently for younger and older individuals, which might signify differences in defining what exactly a "bad element moving in" actually is. As another example, Item A ("Vandalism, like people breaking windows or painting buildings?") is more difficult for African Americans to endorse than Whites, even after accounting for underlying physical incivilities factor levels. Hence, a Black individual with average physical incivilities perceptions (latent score = 0) has approximately a 0.69, 0.26, and 0.05 probability of rating the vandalism item not a problem, somewhat of a problem, and a big problem, respectively. A similar White individual's corresponding probabilities to the same item are approximately 0.50, 0.39, and 0.11. Put simply, on average, Blacks rate vandalism as less of a problem than similar Whites in their neighborhoods.

The above evidence along with details of the ICCs reveal two important insights, which are useful for answering the second research questions. First, while 30% of the items exhibit DIF across race and age, the DIF for each covariate does not occur systematically in one direction and therefore item bias will offset to an extent at the scale level. Hence, Fig. 2 shows the ICCs are not shifted in the same direction across all

[16] Not an uncommon occurrence in multilevel factor analysis, two error variances at the between level were set to zero to avoid negative residual variances and, therefore, these loadings were exactly equal to 1.

[17] Using constrained baseline DIF detection to remove DIF-items from the potential anchor set, results indicate that eight direct effects should not be constrained to zero for anchoring purposes. Items A, B, and F are flagged as functioning differently for Whites and African Americans; Items D, H, and I are flagged as functioning differently across age; and Items H and J are flagged as having DIF across gender. Freeing these constraints, Iteration 2 identified no additional items as exhibiting DIF. Recall, the reason to not stop here is to control Type 1 Error, though we certainly expect many of these identified items will ultimately be found to exhibit DIF. The center of Table 5 contains the Wald statistics used to rank items and select anchors. The six items with the lowest Wald statistics are selected such that each covariate has an anchor for each individual level factor. The anchors are denoted by an asterisk.

Table 5
Results of DIF-free-then DIF detection procedure.

Item	DIF in item (elimination as anchor)		Wald ranking of DIF-free items			Final DIF
	Iteration 1	Iteration 2	AGE	GENDER	RACE	(OR)
A. Vandalism	R	R	2.059	0.023	–	R (0.450)
B. Vacant housing	R	R	0.094	1.263	–	R (1.993)
C. Yards	–	–	0.027[a]	0.017[a]	0.092	–
D. Bother	A	A	–	0.344[a]	0.205	A (0.622)
E. Litter	–	–	0.736	0.387	0.092[a]	–
F. Vacant lots	R	R	0.844	1.809	–	R (1.630)
G. Teenagers	–	–	0.017[a]	2.822	0.020[a]	–
H. Noise	A, G	A, G	–	–	0.140	A (1.695)
I. Bad elements	A	A	–	1.128	2.538	A (1.701)
J. Fighting	G	G	0.017	–	0.582	–

DIF detection in item: A = age; G = gender; R = race; OR = Odds Ratio.
[a] Selected referent items (DIF constrained to zero).

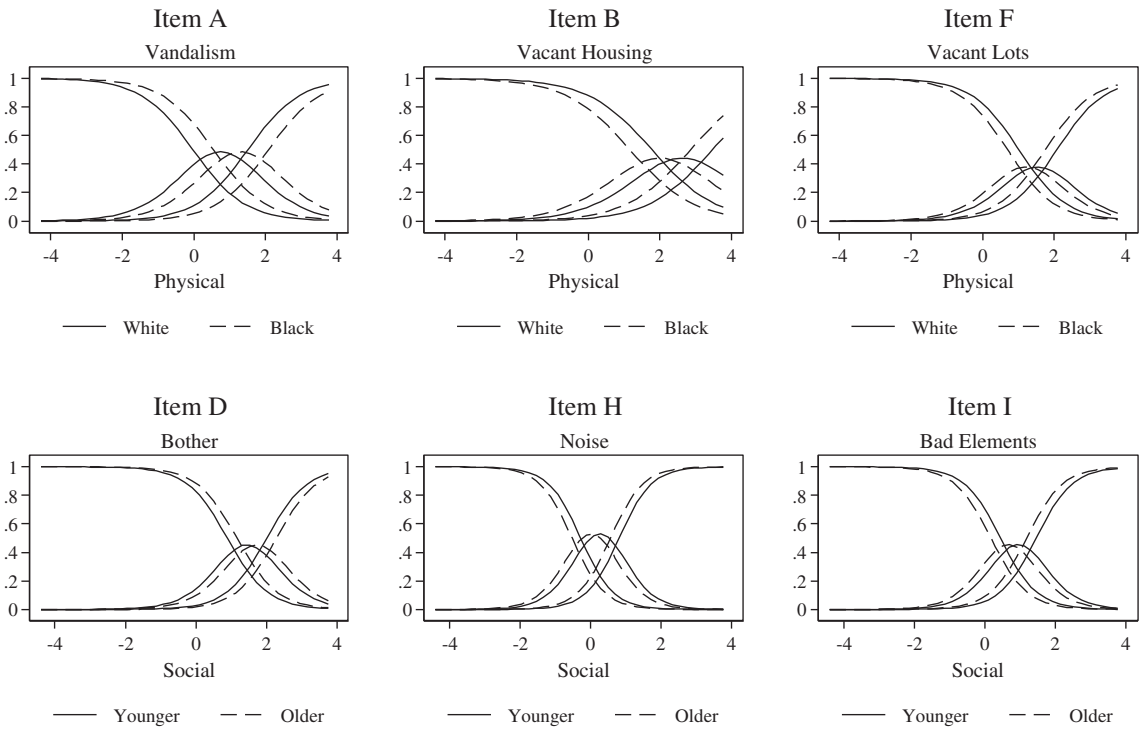

Fig. 1. Item characteristic curves for items with DIF.

items that exhibit DIF for a particular covariate. Second, group differences in the ICCs are not especially large; still, there is nontrivial and statistically significant uniform DIF in these items. Thus, individual DIF effects appear small and do not accumulate systematically in one direction, suggesting that the consequences of DIF may be relatively minor. At the same time, the possible impact of DIF is more important for scales containing a smaller number of items per construct (see Teresi et al., 2012). Given three out of five items for each latent factor at the individual level are found to exhibit DIF for the same covariate, DIF should be adjusted when estimating race differences in physical incivilities and age differences in social incivilities. In addition, DIF itself may provide critical insights into the very nature of subjectivity in perceptions of incivilities, which we address at length in the discussion section.

8.3.2. Total information curves

Fig. 2 shows partial total information curves (PTIC) for physical (5 items) and social (5 items) incivilities at the within level and the total information curve (TIC) for combined incivilities at the between level (10 items). There are interesting differences in the information obtained from scale items for the three latent factors. Items D, G, H, I, and J provide a greater amount of information on social incivilities than Items A, B, C, E, and F do for physical incivilities. The peaks of the information curves occur roughly one standard deviation above the mean, which suggests that the items collectively measure individuals in this area with a greater level of precision. The neighborhood level combined incivilities information curve is shifted even further to the right with the items providing the greatest amount of information for neighborhoods with very high incivilities levels (i.e., roughly two standard deviations above the mean). The answer to the third research question is that items provide less information about individual perceptions of physical and social incivilities and combined neighborhood incivilities at the low end of the continuum. Therefore, the items do not distinguish between individuals (and neighborhoods) well for those who do not perceive much incivilities.

8.3.3. Multilevel structural relationships

Using the MLMIMIC model that accounts for the unique factor

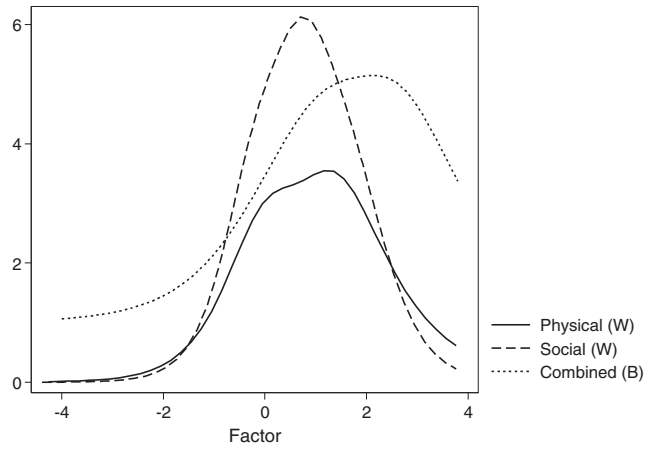

Fig. 2. Information curves for individual and neighborhood incivilities factors.
Notes: W = Within level; B = Between level.
Combined Incivilities TIC includes all 10 items; Physical PTIC includes 5 items; Social PTIC includes 5 items.

structure at each level and adjusts for DIF in six items, we now examine whether these covariates are related to incivilities at the individual and neighborhood levels (research question 4). Specifically, we test whether age, gender, and race are related to physical and/or social incivilities at the individual level, and assess whether average age, proportion female, and proportion Black are related to combined incivilities at the neighborhood level. In addition, we examine whether physical and social incivilities are differentially related to perceptions of safety during the day and at night within neighborhoods, and whether combined neighborhood incivilities are associated with aggregated perceptions of safety during the day and at night between neighborhoods (research question 5).

Fig. 3 contains results from the full multilevel structural equation model (MLMIMIC) with outcome variables. Reported are fully standardized coefficients for continuous covariates (STDYX) and partially standardized coefficients for the effects of binary covariates at the within level of analysis (STDY). Only significant pathways are pictured and DIF pathways are omitted for pictorial simplicity. Overall, model fit statistics suggested good fit to the data ($\chi^2 = 210.23$, $p < 0.01$, CFI = 0.983, TLI = 0.977, RMSEA = 0.018) with better fit at the individual level ($SRMR_{Within} = 0.045$) than the between level ($SRMR_{Between} = 0.071$). Beginning with results at the between neighborhood level, findings indicate that the proportion of African Americans in a neighborhood is significantly associated with combined incivilities ($p < 0.05$), but that proportion old and proportion female are not. For a one standard deviation increase in the proportion of African American residents, combined neighborhood incivilities is expected to increase by nearly 60% of a standard deviation ($\beta_{RACESTDYX} = 0.58, p < 0.01$). At the neighborhood level of analysis, then, neighborhoods plagued by high levels of combined incivilities tend to be those that have a greater proportion of African American residents. The gender and age structure of neighborhoods, however, did not have significant associations with combined neighborhood incivilities levels. Neighborhoods with higher levels of combined incivilities were associated with increased levels of fear during the day and at night ($p < 0.01$).[18] Finally, neighborhood racial composition had a statistically significant association with both fear measures ($p < 0.05$), but neighborhood age and gender composition were not significantly associated with either fear measure.

Turning attention to the within, or individual, level of analysis, age and race have statistically significant associations with both physical and social incivilities ($p < 0.05$). Compared to younger individuals in similar neighborhoods, average levels of physical incivilities are 0.18 standard deviation units lower and average levels of social incivilities are 0.40 standard deviation units lower for older individuals ($p < 0.05$).[19] Controlling for race and gender, the age relationship is over twice as strong on social incivilities as compared to physical incivilities. Compared to Whites in similar neighborhoods, African Americans perceived significantly less physical and social incivilities ($p < 0.05$). Specifically, physical incivilities perceptions are on average 0.463 standard deviation units lower and social incivilities perceptions are 0.418 standard deviation units lower for African Americans, relative to Whites. Within neighborhoods, men and women did not report significantly different levels of physical or social incivilities.

Controlling for the covariates, social incivilities was significantly associated with both fear during the day and fear at night ($p < 0.01$), but physical incivilities was not significantly associated with either fear outcome. Age and gender were significantly associated with fear during the day ($p < 0.01$), and all three covariates were significantly associated with fear at night ($p < 0.01$). Within neighborhoods, older individuals and women were both significantly more likely to experience fear during the day/night, as compared to younger individuals and males, respectively. Probit indexes summarizing the relationships between the demographic factors and fear indicate that these relationships were more pronounced at night. For instance, average differences between men and women were 0.380 standard deviation units for daytime fear but 0.612 standard deviation units for nighttime fear. Africans Americans and Whites did not have significant differences in perceptions of fear during the day, however, African Americans reported significantly less fear during the night than Whites ($p < 0.01$).

Considering the findings across levels, results indicate neighborhoods with a higher proportion of minority residents experience significantly greater amounts of combined incivilities, but within neighborhoods, African Americans perceive less incivilities as compared to Whites in similar neighborhoods (see also Sampson & Raudenbush, 2004). Thus, the between and within level effects of race on the incivilities constructs operate in different directions across levels of analysis, a fact that is obscured in single level models.[20] Considering the additive effects of the covariates at the individual level combined with insights from the neighborhood level analysis, older White females living in communities with higher concentrations of minorities would tend to experience the greatest levels of fear. Fear during the day and at night also tends to be greatest for those who perceived higher levels of social incivilities and find themselves in neighborhoods with more combined incivilities. As a final point worth mentioning, demographics explained a greater proportion of the between-level variance in incivilities as compared to the within-level variance. Specifically, the between-level exogenous variables, effectively proportion African American, explained 35% of the variance in combined neighborhood incivilities. At the individual level, however, the covariates collectively explained only 6% and 8% of the variance in physical and social incivilities, respectively.

[18] In an additional analysis (available upon request), we created a two-item latent fear factor which revealed substantively similar findings as reported herein. The coefficient of combined neighborhood disorder on the two-item neighborhood fear latent factor was 0.552, as opposed to 0.630 for daytime and 0.465 for nighttime fear.

[19] We elected to dichotomize age because many investigations into differential item functioning use categorical covariates and we believe these might be more familiar to readers. However, the MLMIMIC model can be estimated with age treated as a continuous covariate and results from a model which does just this indicates a pattern of findings that are similar. Specifically, within neighborhoods, a one standard deviation increase in age results in a 0.11 standard deviation decrease in physical disorder perceptions ($p < 0.05$) and a 0.254 standard deviation decrease in social disorder perceptions ($p < 0.05$). In both this supplemental analysis and the findings based on the dichotomous age variables, the coefficient for social disorder is roughly 2.3 times the coefficient for physical disorder.

[20] A single level two-factor CFA model revealed no statistically significant effects of race on social or physical disorder.

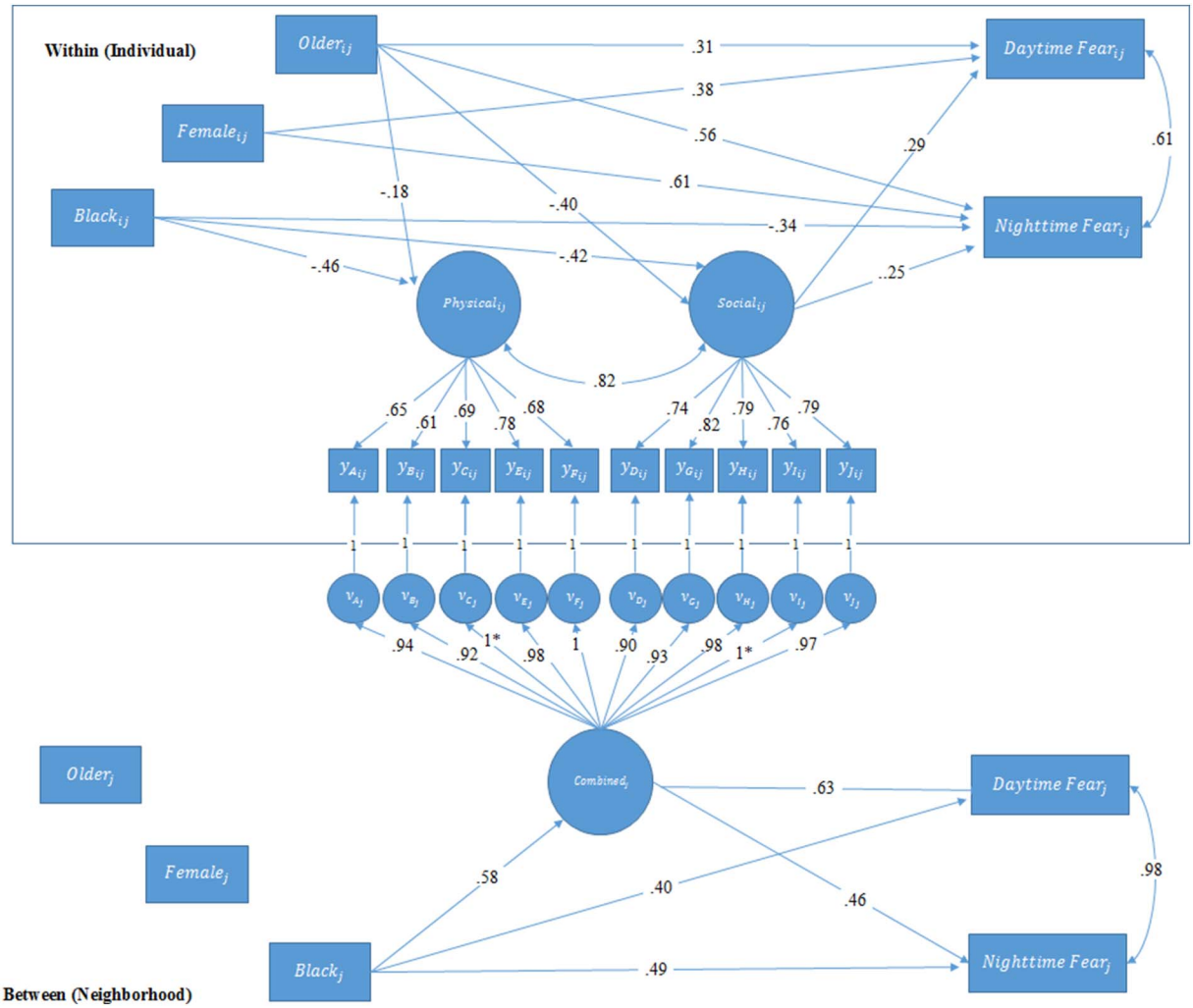

Fig. 3. Multilevel structural equation model summarizing significant structural relationships.
Notes:
* Item residual variance set to zero.
χ^2 = 210.233, df = 150, CFI = 0.983, TLI = 0.977, RMSEA = 0.018.
$R^2_{Physical(W)}$ = 0.06; $R^2_{Social(W)}$ = 0.08; $R^2_{Combined(B)}$ = 0.35;
Fully standardized (STDYX) coefficients reported. Exceptions are effects of within-level exogenous variables (i.e., Older, Female, Black) on endogenous variables, where partially standardized (STDY) coefficients are reported.
For pictorial simplicity the errors, non-significant structural pathways, and significant DIF pathways are not drawn.

9. Discussion

The current study is the first of which the authors are aware that employs appropriate analytic techniques—MLCFA and MLMIMIC models—to determine both whether incivilities emerges as one or two constructs at the individual and/or neighborhood levels of analysis and whether certain items function differently across age, gender and race when controlling for overall levels in the latent constructs. Moreover, using a multilevel structural equation model and adjustment for DIF, the current study examined whether covariates were significantly related to incivilities at individual and neighborhood levels of analysis and determined to what extent covariates and incivilities construct(s) were associated with daytime and nighttime fear at both levels of analysis. Several important empirical findings emerge from the study with important implications for measuring incivilities, advancing theory and implementing strategies to curb incivilities and fear.

9.1. Key findings and implications

9.1.1. Divergent factor structures

By eliminating confounding of individual and neighborhood levels of analysis, results supported two factors (physical and social incivility) at the individual level but only one factor (combined incivility) at the neighborhood level. When measuring incivilities at the neighborhood level, there is no utility in separating physical and social aspects of incivilities. Put simply, the places where you find high levels of physical incivilities are the same places you find high levels of social incivilities. This complements in interesting ways the less sophisticated study of Taylor et al. (1985) that obtained one social and physical incivilities component at the streetblock level based only on assessed social and physical features, and extends that unitary idea to perceptions of incivilities at the neighborhood level. As neighborhoods are compared on their incivilities, both social activities and physical aspects contribute to an overall categorization. Within neighborhoods, however, individuals' perceptions or experiences with physical and social incivilities in their neighborhood can be usefully separated, even though these factors are, perhaps not surprisingly, still highly correlated. Our multilevel investigation on the factor structure of incivilities may help to explain why some research has suggested physical and social incivilities should be combined into one construct (Xu et al., 2005), yet other research has advocated for separating these either into a broader incivilities factor and physical decay, or social and physical incivilities (Ross & Mirowsky, 1999; Taylor, 1999).

Importantly, our results suggest aggregating individual perceptions does *not* result in an "isomorphic" construct; in other words, there is dissimilarity in factor structure and thus construct meaning across levels of analysis (e.g., see Bliese, 2000). Perhaps this should not be surprising as there are two common reasons for non-isomorphism, which include clustering of individuals in neighborhoods by attributes and perceptual determinants that have both individual and neighborhood sources (see Bliese, 2000; Bliese, Chan, & Ployhart, 2007). We know both to be true (e.g., see Krivo, Peterson, & Kuhl, 2009; Sampson & Raudenbush, 2004). What these findings imply, therefore, is that incivility at the neighborhood level is not simply a summary of individuals' social and physical incivilities perceptions. Rather, due to nonrandom distribution of individuals in communities, combined incivility may serve in part as a proxy for clustering of attributes in communities (e.g., concentrated disadvantage or poverty). Indeed, poverty has been shown to be strongly linked to neighborhood level disorder (Sampson & Raudenbush, 2004). Similarly, shared group characteristics, such as willingness of neighbors to correct unsightly and disorderly conduct in the community, contribute to incivility ratings alongside one's own perceptions and experiences with physical and social incivility, thereby opening the possibility that combined incivility at the neighborhood level also serves as a proxy for broader community-level processes such as collective efficacy (e.g., Sampson, Raudenbush, & Earls, 1997). This may explain why there is a lack of discriminant validity between social and physical incivility at the neighborhood level and, thus, the emergence of combined neighborhood incivility.

9.1.2. DIF: race and physical disorder, age and social disorder

While no items were found to function differently across gender, three of five physical incivilities items across race and three of five social incivilities items across age were found to exhibit DIF. That is, different types of people used response categories differently, even after controlling for overall levels in the latent constructs. The "vacant housing" and "vacant lots with trash or junk" items (Items B and F) are less difficult for Blacks to endorse than for Whites. With similar underlying levels of physical incivilities perceptions, this means that Blacks score these two items as more of a problem than Whites on average. With respect to the vacant housing indicator, there may be interpretation differences in the meaning of "vacant housing" that explain this discrepancy. Definitional issues with vacant housing could stem from the fact that vacancy may be seen as less of a problem when houses are actively being sold on the market and former owners have simply moved out prior to the sale; boarded-up vacant dwellings, something more common in disadvantaged areas, signal something more permanent and problematic than new neighbors. This might be more of an across-neighborhood explanation, though perhaps differences in prior experiences with vacant housing may cut across demographic lines thereby biasing interpretation of the item at the individual level. Absent definitional issues, alternatively, the results might imply that Blacks are desensitized to the incivility of vacant housing as they have to score these items higher to have comparable levels of overall physical incivilities perceptions. The trash-filled vacant lots indicator (Item F) operates in a similar fashion and here item wording and interpretation is less obviously problematic. To have similar overall physical incivilities perceptions as Whites, trash-filled vacant lots must be scored as more of a problem for Blacks implying a desensitization of this indicator as well. It follows that if Blacks reported similar problems with trash-filled vacant lots in their neighborhoods as Whites, all else being equal, Blacks would have lower overall perceptions of physical incivility.

The "vandalism, like people breaking windows or spray painting buildings" indicator (Item A) is more difficult to endorse for African Americans, which could signal a difference in the underlying definition of graffiti. One possible explanation is that graffiti is more likely to be perceived by Blacks as having an artistic element, and therefore is more difficult to report as a "problem." Scales measuring neighborhood incivilities then should pay careful attention to possible differences in item meaning that may occur due to racial or cultural definitional differences that could develop from longstanding racial and socio-economic stratification. This is especially the case when trying to root out the subjective nature of incivilities across groups as biased items can hamper these efforts.

Compared to younger individuals, older individuals find the item "People who say insulting things or bother other people when they walk down the street" (Item D) more difficult to endorse as a problem, but the items "The amount of noise in the area" and "Bad elements moving in" (Items H and I) less difficult to endorse as problems. Interestingly, while DIF across race is all related to physical incivilities, DIF across age is found to all be related to social incivilities, which could reflect generational differences in item meaning. For instance, definitions of what constitutes "noise" and "bad elements" may differ between older and younger individuals, even when accounting for underlying perceptions in social incivilities. Also, younger and older individuals with similar incivilities perceptions marked Item D differently and this does not obviously appear to be due to item wording or interpretation. In this case, DIF might reflect a lack of understanding of trendy insults among the older population or that there is a level of respect granted to older individuals that is not afforded to younger individuals. These possible generational differences should be considered when devising items tapping into social aspects of incivilities.

For theoretical reasons, the bulk of past work in this area has relied on survey-based perceptions of incivilities. Given the current revelation involving perceptions and DIF—in conjunction with new work suggesting that systematic social observation done properly has key benefits (Hoeben et al., 2016)—perhaps there has been an overreliance on survey assessments of incivilities at the expense of on-site assessments conducted by multiple trained raters. At the least, differential item functioning should be accounted for in empirical investigations, especially those that seek to identify and characterize group differences in the incivilities construct(s). For example, while accounting for DIF did not change the conclusions regarding the direction of effects or the statistical significance of coefficients, the significant effect of age on social incivilities was found to be 21% larger after taking the DIF into account. Finally, given that DIF across age occurs exclusively for social incivilities items and DIF across race occurs exclusively for physical incivilities items, it would be particularly worthwhile to conduct focus group interviews to confirm or refute our speculative, though logical, sources of DIF. Understanding the origins and sources of DIF in types of incivilities items is useful for scale improvement.

9.1.3. Item information greatest at high incivility

Results indicate that the items provide greater information for individuals and neighborhoods with higher standardized scores on the construct(s) (see Fig. 2). Thus, individuals with higher levels of physical and social incivilities are measured more precisely than individuals with lower levels. Similarly, neighborhoods with higher levels of combined incivilities have less measurement error than neighborhoods with lower levels. On the one hand, it is good that individuals and neighborhoods that have more problematic levels of incivilities are measured more precisely because these may be of most interest to those studying incivilities and fear of crime. That said, the lack of very low difficult (or easy) items suggests that both individuals with lower physical and social incivilities perceptions and neighborhoods with lower combined incivilities are not measured very precisely. As such, differentiating between neighborhoods that have very low incivilities as compared to low incivilities is less precise, though having more refined measurement may be a worthwhile goal. For instance, it could be fruitful to single out a community with the lowest neighborhood incivilities level for case study or in-depth qualitative study to identify reasons or ways that the community wards off incivilities. Doing so with high precision would require greater information at the lower end of the incivilities continuum.

Our findings suggest that scale developers might consider adding items with lower difficulty, so long as they were consistent with conceptualizations of the physical and social aspects of incivilities—items that do not compromise face or content validity. Along these lines, any efforts to improve incivilities measurement might keep in mind physical incivilities items are less pure indicators as compared to social incivilities items at the individual level. Related, fine-grained measurement at lower levels of perceived incivilities might be improved by considering revisions to response options. For instance, four categories (e.g., no problem, slight problem, moderate problem, large problem), as opposed to the three employed in this measure, could provide additional precision in measurement at the lower ends of the construct continuums.

9.1.4. Multilevel differences in predictors and outcomes of incivilities

The effects of covariates on incivilities provided interesting differences across levels of analysis. For example, race has significant effects on incivilities across levels, but in different directions. Neighborhoods with higher concentrations of African Americans experience greater levels of combined neighborhood incivilities. Interestingly, however, within neighborhoods, African Americans perceive significantly less physical and social incivilities, even after adjusting for items exhibiting DIF, as compared to Whites in similar neighborhoods. This finding confirms and builds on other work in this area (Hipp, 2010; Sampson & Raudenbush, 2004). One possible explanation is that the general high exposure to incivilities among African Americans has generally desensitized this racial group to certain social and physical signs of incivilities relative to their demographic counterpart. In line with some research (Carvalho & Lewis, 2003; Franzini et al., 2008; Sampson & Raudenbush, 2004), our multilevel investigation challenges the common assumption that "…people recognize disorder when they see it, and uniformly want something done about it" (Xu et al., 2005, p. 156). There are critical implications of subjectivity in incivilities perceptions for order maintenance policing. As Kubrin (2008) has pointed out, there exists a "fine line" between aggressive order maintenance policing and perceptions of harassment by residents. Within neighborhoods, order maintenance policing activities might be less welcomed among racial minorities as opposed to majority group members due to differences in perceptions of the underlying problem (or prior treatment by the police). This points to the importance of having citizens working with police to identify the issues most pressing in their communities. To some citizens in neighborhoods, the most pressing issues may indeed be alleviating physical and/or social incivilities, in other communities it may be gun violence, and still in others procedural justice during police-citizen encounters.

While age, sex, and race and aggregated demographics are useful for predicting physical and social incivilities at the individual level and combined incivilities at the neighborhood level, respectively, most of the variation in the incivilities factors at both levels is left unexplained. Additional attention is needed to explain how and why individuals perceive the "incivilities" they encounter in their neighborhood in the way that they do (Franzini et al., 2008; Hipp, 2010; Jackson, 2004; Sampson & Raudenbush, 2004; Wallace, 2011; Wickes et al., 2013). Indeed, objective and perceptual measures share different degrees of discriminant validity with crime and criminological constructs (Taylor, 1999). Understanding why this is the case is an area of inquiry in need of much empirical attention and research may need to move beyond a focus on sociodemographic factors to psychological constructs such as implicit bias, optimism, and pessimism. In this vein, we see recent work examining the disconnect between direct and indirect measures of peer delinquency (e.g., Boman & Ward, 2014; McGloin & Thomas, 2016; Young, Barnes, Meldrum, & Weerman, 2011) as a potentially useful heuristic for exploring differences between objective and perceived measures of incivilities.

Social incivilities significantly influence individual perceptions of fear during the day and night but physical incivilities do not. Combining physical and social incivilities at the within level of analysis would obscure these insights.[21,22] When implementing policies to reduce differences in fear across neighborhoods, holistic efforts—possibly via "coproduction" with the community (Braga et al., 2015)—aimed at reducing incivilities may be warranted. It is helpful to keep in mind that the legacy of Broken Widows aligns much more closely with community policing endeavors that with more aggressive policing practices (e.g., stop, question, and frisk) (Kelling, 2015). Further, approaches to reduce neighborhood incivilities can, and probably should, pull from a variety of sources including crime prevention through environmental design (e.g., strategic placement of trashcans), community programs (e.g., senior lawn and home maintenance assistance, community clean-ups), and infrastructure investment (e.g., recreation centers for teenagers). Within neighborhoods, the findings imply that initiatives and interventions targeting adverse, chronic social conditions may be best situated to tackle the issue of resident fear. Thus, while quality of life is likely improved with community clean-ups, restoration of vacant houses, and improved maintenance of properties, fear may not be significantly reduced by these strategies as compared to those addressing local social issues. This conclusion, however, is under the assumption this relationship flows in the hypothesized direction from incivilities to fear, which may not be tenable (Link, Kelly, Pitts, Waltman-Spreha, & Taylor, 2017).

9.2. Limitations and future directions

The current study has certain limitations that warrant consideration and help to guide future research. First, while the current study provided a detailed investigation into differential item functioning and the multilevel factor structure of perceived incivilities, it did not intend to, nor speak to, recent debates concerning whether incivility is indistinguishable from crime (Gau & Pratt, 2008). While some prior studies have documented little utility in splitting physical and social incivilities, the current MLCFA analyses suggest there is utility in separating these, albeit at the individual level of analysis only. We would suggest future research employing MLCFA and MLMIMIC analyses consider whether crime and incivilities can be separated at different levels of analysis.[23] Second, our data are cross-sectional and thus our findings are subject to concerns of incorrect temporal ordering. Indeed, using longitudinal data and cross-lagged models, Link et al. (2017) recently found that fear influenced incivilities perceptions, but not the other way around. Given the complexities identified across levels with our multilevel structural equation models, we encourage the exploration of causal ordering issues at both levels of analysis to see if they operate similarly for social and physical incivilities at the individual level and for combined incivilities at the neighborhood level. The remainder of the study's findings, especially at the individual level, are fortunately not subject to this concern as demographic factors are antecedent to incivilities. Moreover, the use of cross-sectional data is not problematic for conclusions from the MLCFA models per se, although employing longitudinal data would permit an assessment of

[21] It is worth noting that a single-level observed variable analysis employing ordinal logistic regression found that physical and social disorder both influence fear during the day and at night (results available upon request), which suggests that important differences may emerge when individual and neighborhood levels in disorder are both considered in MLMIMIC models. In addition, single level analyses regressing social and physical disorder on the covariates indicated small, positive effects of race on physical disorder ($p > 0.05$) and social disorder ($p < 0.05$).

[22] To our knowledge, Mplus v7.31 provides no formal tests for multicollinearity. VIF assessment in Stata v14.1 of observed physical and social disorder measures did not reveal problems with multicollinearity (VIFs = 1.31).

[23] We would encourage future research to consider conceptual differences in individual indicators too. For instance, on the one hand, "litter and trash in the streets" can be a sign of crime (i.e., littering); on the other hand, it can be a sign of a city's trash removal services which leave much to be desired and residents' unwillingness to remedy the problem.

longitudinal measurement invariance and, if achieved, an assessment of the degree of stability in perceived incivilities at both the individual and neighborhood levels over time. We find such questions incredibly intriguing but beyond the scope of the current study. Third, we might encourage future multilevel research on perceived incivilities to assess both DIF and factor structure using smaller units of analysis (e.g., see Weisburd, Hinkle, Braga, & Wooditch, 2015; Steenbeek & Kreis, 2015), which is a general trend in incivilities research (Weisburd, Bernasco,- & Bruinsma, 2009; Skogan, 2015).

Finally, there may be concerns that the data we employed are dated. Underlying this concern is a suspicion the results might not be replicable. Whether these results can or cannot be replicated is not a study limitation per se; it is an external validity question and thus an outstanding empirical question (Taylor, 1994, p. 164). Thus, we of course recommend scholars seek to replicate the current findings with alternative data sources. But the "dated" quality of the current data set also represents, in this instance, a tremendous strength. The neighborhoods in question emerged from a community organization and community planner input process designed to maximize the ecological validity of identified neighborhoods for the entire city of Baltimore (Taylor et al., 1979). These units were adopted for the 1980 Decennial Census Neighborhood Statistics Program in Baltimore City. Further, the neighborhood definition process predated the survey itself by only three years. Since neighborhood boundaries can change over time (Taylor, 2001, p. 303–354) this means that the residents in this survey were sampled from functioning neighborhoods whose spatial boundaries aligned closely with current dynamics on the ground. We are not aware of *any* publicly available data set with ecologically valid neighborhood units, samples drawn close in time to the ecological definition process, a large number of neighborhood units, and a large number of perceived incivility indicators. Given the interests here in discriminating neighborhood vs. individual functioning, the data set used here is appropriate.

The present study set out to address the controversy surrounding the meaning and measurement of "incivilities" by using resident perceptions to identify incivilities constructs at individual and neighborhood levels of analysis, assess differential item functioning and subjectivity in perceptions, and examine multilevel relationships between perceived incivilities factor(s), demographic covariates, and fear of crime. This work has provided new windows into the divergent multilevel factor structure and subjectivity of incivilities perceptions, thereby tackling some of the more pressing criticisms surrounding the arguably broken construct (see Kubrin, 2008). Still, further construct validity assessments of incivilities are warranted as these investigations provide important "implicit assessment of theory" (see Sullivan & McGloin, 2014, p. 447). Additional research is also needed to provide a better understanding of how and why people form their perceptions of their neighborhood. Both of these tasks will enable more appropriate and informative tests of the incivilities thesis and the consequences of incivilities for individual and group quality of life.

References

Asparouhov, T., & Muthen, B. (2012). Multiple group multilevel analysis. *Mplus Web Notes*, *16*, 1–45.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein, & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*(4), 551–563.

Boman, J. H., & Ward, J. T. (2014). Beyond projection: Specifying the types of peer delinquency misperception at the item and scale levels. *Deviant Behavior*, *35*(7), 555–580.

Braga, A. A., Welsh, B. C., & Schnell, C. (2015). Can policing disorder reduce crime? A systematic review and meta-analysis. *Journal of Research in Crime and Delinquency*, *52*(4), 567–588.

Carvalho, I., & Lewis, D. A. (2003). Beyond community: Reactions to crime and disorder among inner-city residents. *Criminology*, *41*(3), 779–812.

Dunn, E. C., Masyn, K. E., Johnston, W. R., & Subramanian, S. V. (2015). Modeling contextual effects using individual-level data and without aggregation: An illustration of multilevel factor analysis (MLFA) with collective efficacy. *Population Health Metrics*, *13*(1).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(5), 275–284.

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

Gallagher, N. A., Gretebeck, K. A., Robinson, J. C., Torres, E. R., Murphy, S. L., & Martyn, K. K. (2010). Neighborhood factors relevant for walking in older, urban, African American adults. *Journal of Aging and Physical Activity*, *52*(1), 85–110.

Garofalo, J., & Laub, J. H. (1978). The fear of crime: Broadening our perspective. *Victimology*, *3*, 242–253.

Gau, J. M., & Pratt, T. C. (2008). Broken windows or window dressing? Citizens' (in) ability to tell the difference between disorder and crime. *Criminology & Public Policy*, *7*(2), 163–194.

Harcourt, B. E. (2009). *Illusion of order: The false promise of broken windows policing*. Cambridge, MA: Harvard University Press.

Hinkle, J. C., & Yang, S. M. (2014). A new look into broken windows: What shapes individuals' perceptions of social disorder? *Journal of Criminal Justice*, *42*(1), 26–35.

Hipp, J. R. (2010). Resident perceptions of crime and disorder: How much is 'bias', and how much is social environment differences? *Criminology*, *48*(2), 475–508.

Hipp, J. R. (2016). Collective efficacy: How is it conceptualized, how is it measured, and does it really matter for understanding perceived neighborhood crime and disorder? *Journal of Criminal Justice*, *46*, 32–44.

Hoeben, E., Steenbeek, W., & Pauwels, L. J. R. (2016). Measuring disorder: Observer bias in systematic social observations at streets and neighborhoods. *Journal of Quantitative Criminology*. http://dx.doi.org/10.1007/s10940-016-9333-6.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Hunter, A. (1978). Symbols of incivility: Social disorder and fear of crime in urban neighborhoods. *Paper presented at the Annual Meeting of the American Criminological Society, Dallas, TX*.

Jackson, J. (2004). Experience and expression: Social and cultural significance in the fear of crime. *British Journal of Criminology*, *44*(6), 946–966.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136–153.

Kamata, A., & Vaughn, B. K. (2011). Multilevel item response theory modeling. In J. Hox, & J. J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 41–57). New York, NY: Routledge.

Kelling, G. (2015). An author's brief history of an idea. *Journal of Research in Crime and Delinquency*, *52*, 626–629.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford.

Krivo, L. J., Peterson, R. D., & Kuhl, D. C. (2009). Segregation, racial structure, and neighborhood violent crime. *American Journal of Sociology*, *114*(6), 1765–1802.

Kubrin, C. E. (2008). Making order of disorder: A call for conceptual clarity. *Criminology & Public Policy*, *7*(2), 203–213.

Link, N. W., Kelly, J. M., Pitts, J. R., Waltman-Spreha, K., & Taylor, R. B. (2017). Reversing broken windows: Evidence of lagged, multilevel impacts of risk perceptions on perceptions of incivility. *Crime & Delinquency*, *63*(6), 659–682.

Liu, Q. (2011). *Item purification in differential item functioning using generalized linear mixed models (unpublished doctoral dissertation)*. Tallahassee: Florida State University.

Matthews, R. (1992). Replacing 'broken windows': Crime, incivilities and urban change. *Issues in realist criminology* (pp. 19–50). .

McGarrell, E. F., Giacomazzi, A. L., & Thurman, Q. C. (1997). Neighborhood disorder, integration, and the fear of crime. *Justice Quarterly*, *14*, 479–500.

McGloin, J. M., & Thomas, K. J. (2016). Considering the elements that inform perceived peer deviance. *Journal of Research in Crime and Delinquency*, *53*(5), 597–627.

Mead, G. H. (1934). *Mind, self, and society*. Chicago, IL: University of Chicago Press.

Molnar, B. E., Gortmaker, S. L., Bull, F. C., & Buka, S. L. (2004). Unsafe to play? Neighborhood disorder and lack among urban children and adolescents. *American Journal of Health Promotion*, *18*(5), 378–386.

Muthén, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics*, *10*, 121–132.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer, & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.

Perkins, D. D., Meeks, J. W., & Taylor, R. B. (1992). The physical environment of street blocks and resident perceptions of crime and disorder: Implications for theory and measurement. *Journal of Environmental Psychology*, *12*, 21–34.

Perkins, D. D., Brown, B. B., & Taylor, R. B. (1996). The ecology of empowerment: Predicting participation in community organizations. *Journal of Social Issues*, *52*(1), 85–110.

Perkins, D. D., & Taylor, R. B. (1996). Ecological assessments of community disorder: Their relationship to fear of crime and theoretical implications. *American Journal of*

Community Psychology, *24*(1), 63–107.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.
Ross, C. E., & Mirowsky, J. (1999). Disorder and decay: The concept and measurement of perceived neighborhood disorder. *Urban Affairs Review*, *34*(3), 412–432.
Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology*, *105*(3), 603–651.
Sampson, R. J., & Raudenbush, S. W. (2004). Seeing disorder: Neighborhood stigma and the social construction of 'broken windows'. *Social Psychology Quarterly*, *67*(4), 319–342.
Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *5328*(77), 918–924.
Skogan, W. G. (1986). Fear of crime and neighborhood change. In A. J. Reiss, & M. H. Tonry (Eds.), *Communities and crime* (pp. 203–229). Chicago, IL: University of Chicago Press.
Skogan, W. G. (1992). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. University of California Press.
Skogan, W. G. (2015). Disorder and decline: The state of research. *Journal of Research in Crime and Delinquency*, *52*(4), 464–485.
Skogan, W. G., & Maxfield, M. G. (1981). *Coping with crime: Individual and neighborhood reactions*. Beverly Hills, CA: Sage.
Stafford, M. C., & Warr, M. (1993). A reconceptualization of general and specific deterrence. *Journal of Research in Crime and Delinquency*, *30*(2), 123–135.
Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306.
Steenbeek, W., & Kreis, C. (2015). Where broken windows should be fixed: Toward identification of areas at the tipping point. *Journal of Research in Crime and Delinquency*, *52*(4), 511–533.
Sullivan, C. J., & McGloin, J. M. (2014). Looking back to move forward: Some thoughts on measuring crime and delinquency over the past 50 years. *Journal of Research in Crime and Delinquency*, *51*(4), 445–466.
Swatt, M. L., Varano, S. P., Uchida, C. D., & Solomon, S. E. (2013). Fear of crime, incivilities, and collective efficacy in four Miami neighborhoods. *Journal of Criminal Justice*, *41*(1), 1–11.
Taylor, R. B. (1994). *Research methods in criminal justice*. New York: McGraw Hill.
Taylor, R. B. (1996). Neighborhood responses to disorder and local attachments: The systemic model of attachment, social disorganization, and neighborhood use value. *Sociological Forum*, *11*(1), 41–75.
Taylor, R. B. (1999). *The incivilities thesis: Theory, measurement, and policy* (pp. 65–88).
Taylor, R. B. (2001). *Breaking away from broken windows: Baltimore neighborhoods and the nationwide fight against crime, grime, fear, and decline*. Westview Press386.
Taylor, R. B., Brower, S., & Drain, W. (1979). *A map of Baltimore neighborhoods*. Baltimore: Center for Metropolitan Planning and Research, Johns Hopkins University.
Taylor, R. B., & Hale, M. (1986). Testing alternative models of fear of crime. *The Journal of Criminal Law and Criminology*, *77*(1), 151–189.
Taylor, R. B., & Schumaker, S. A. (1990). Local crime as a natural hazard: Implications for understanding the relationship between disorder and fear of crime. *American Journal of Community Psychology*, *18*(5), 619–641.
Taylor, R. B., Shumaker, S. A., & Gottfredson, S. D. (1985). Neighborhood-level links between physical features and local sentiments: Deterioration, fear of crime, and confidence. *Journal of Architectural Planning and Research*, *2*, 261–275.
Teresi, J., Ramirez, M., Jones, R. N., Choi, S., & Crane, P. K. (2012). Modifying measures based on differential item functioning (DIF) impact analyses. *Journal of Aging and Health*, *24*(6), 1044–1076.
Wallace, D. (2011). A test of the routine activities and neighborhood attachment explanations for bias in disorder perceptions. *Crime & Delinquency*, *61*(4), 587–609.
Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, *72*(3), 221–261.
Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2014). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708.
Weisburd, D., Bernasco, W., & Bruinsma, G. (2009). *Putting crime in its place*. New York, NY: Springer.
Weisburd, D., Hinkle, J. C., Braga, A., & Wooditch, A. (2015). Understanding the mechanisms underlying broken windows policing: The need for evaluation evidence. *Journal of Research in Crime and Delinquency*, *52*(4), 589–608.
Wickes, R., Hipp, J. R., Zahnow, R., & Mazerolle, L. (2013). "Seeing" minorities and perceptions of disorder: Explicating the mediating and moderating mechanisms of social cohesion. *Criminology*, *51*(3), 519–560.
Wilson, J. Q. (1975). *Thinking about crime*. New York, NY: Basic Books.
Wilson, J. Q., & Kelling, G. L. (1982). *Broken windows*. *211*. Atlantic Monthly.
Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27.
Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361.
Wyant, B. R. (2008). Multilevel impacts of perceived incivilities and perceptions of crime risk on fear of crime: Isolating endogenous impacts. *Journal of Research in Crime and Delinquency*, *45*(1), 39–64.
Xu, Y., Fiedler, M. L., & Flaming, K. H. (2005). Discovering the impact of community policing: The broken windows thesis, collective efficacy, and citizens' judgment. *Journal of Research in Crime and Delinquency*, *42*(2), 147–186.
Yang, S.-M., & Pao, C.-C. (2015). Do we 'see' the same thing? An experimental look into the black box of disorder perception. *Journal of Research in Crime and Delinquency*, *52*(4), 534–566.
Young, J. T. N., Barnes, J. C., Meldrum, R. C., & Weerman, F. M. (2011). Assessing and explaining misperceptions of peer delinquency. *Criminology*, *49*(2), 599–630.

Jeffrey T. Ward is an assistant professor in the Department of Criminal Justice at Temple University. His areas of research include developmental and life-course criminology, juvenile delinquency, and measurement.

Nathan W. Link is an assistant professor of criminal justice at Rutgers University in Camden. His research focuses on issues in corrections and reentry, financial sanctioning, and substance abuse and mental health.

Ralph B. Taylor currently serves as a Professor of Criminal Justice at Temple University, with a courtesy appointment in Geography and Urban Studies. The National Science Foundation, National Institute of Mental Health, National Institute of Justice, National Institute of Corrections, the Open Society Institute, and others, have funded his research endeavours. He has authored or co-authored over 70 refereed journal articles, one textbook (*Research Methods in Criminal Justice*, McGraw-Hill, 1994) and three books (*Human Territorial Functioning,* Cambridge University Press, 1988; *Breaking Away from Broken Windows,* Westview Press, 2001; *Community Criminology*, New York University Press, 2015).