

# An Audit Tool for Longitudinal Assessment of Health-Related Characteristics of Urban Neighborhoods: Implementation Methods and Reliability Results

Bonnie Ghosh Dastidar (✉ [bonnieg@rand.org](mailto:bonnieg@rand.org))

RAND Corp

Gerald Hunter

RAND Corp

Jennifer C Sloan

RAND Corp

Rebecca L Collins

RAND Corp

Andrea S Richardson

RAND Corp

Wendy Troxel

RAND Corp

Natalie Colabianchi

University of Michigan

Tamara Dubowitz

RAND Corp

---

## Research article

**Keywords:** audit tool, built environment, direct observation, longitudinal assessment, neighborhood environment, reliability

**Posted Date:** February 13th, 2020

**DOI:** <https://doi.org/10.21203/rs.2.23465/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on October 7th, 2020. See the published version at <https://doi.org/10.1186/s12889-020-09424-8>.

# Abstract

**Background:** Improving the neighborhood environment may help address chronic disease and mortality. To identify neighborhood features that are predictors of health, objective assessments of the environment are used. Multiple studies have reported on cross-sectional assessments of health-related neighborhood features using direct observation. As study designs expand to better understand causation and predictors of change, there is a need to test whether direct observation methods are adequate for longitudinal assessment. To our knowledge, this is the first study to report on repeated direct observations of the neighborhood environment.

**Methods :** The Pittsburgh Hill/Homewood Research on Neighborhood Change and Health (PHRESH) study conducted longitudinal assessments in two low-income, African American neighborhoods at three waves (years 2012, 2015, 2017). The PHRESH audit tool is a modification of earlier validated tools, with an emphasis on environment features relevant for physical activity, sleep, and obesogenic behaviors. Trained data-collector pairs conducted direct observations of a 25% sample of street segments in each neighborhood. At each wave, we audited a sub-sample of street segments twice and assessed reliability using percentage inter-observer agreement and krippendorff's alpha statistics. Stability of these items was assessed as exhibiting moderate or high agreement at every time point.

**Results :** Across waves, a majority (81%) of the items consistently demonstrated moderate to high agreement except for items such as public/communal space, amount of shade, sidewalk features, number of traffic lanes, garden/flower bed/planter, art/statue/monument, amount of trash, and physical disorder. The list of items with poor agreement includes features that are easy to miss (e.g. flower bed/planter), hard to assess from outside (e.g. public/communal space), or may change quickly (e.g. amount of trash).

**Conclusion:** In this paper, we have described implementation methods, reliability results and lessons learned to inform future studies of change. We found the use of consistent methods allowed us to conduct reliable, replicable longitudinal assessments of the environment. Items that did not exhibit stability are less useful for detecting real change over time. Overall, the SSA direct observation tool is an effective and practical instrument to detect change in the neighborhood environment.

## Background

Neighborhoods are important for health [1-4]. In fact, the neighborhood environment has been linked to multiple health outcomes including sleep, mental health, cardiovascular risk, and mortality [5-9]. Certain features (e.g. sidewalks) may directly encourage active transportation and physical activity [10-15] and others (e.g. street lighting, noise) may impact sleep [16-18], which, in turn, may influence chronic diseases [19, 20]. Residents of low-income and racially/ethnically segregated neighborhoods share a disproportionate burden of chronic disease [21], as well as limited access to resources, which could

contribute to poor health [22-24]. Improving the neighborhood environment holds promise for addressing health-related behaviors associated with chronic disease and mortality [25].

Micro or granular features of the neighborhood (e.g. street lighting) may affect residents' experiences more directly than macro-level features (e.g. residential density), thus providing stronger links with health behaviors [26-28]. Also, micro-level features are more easily modified than macro-level features. For example, it takes less time and money to repair a sidewalk than to change the land-use mix of a community. While there are multiple approaches for collecting detailed assessments of micro-features of neighborhoods [29-34], direct observation using audit tools is the preferred approach because it allows for systematic observation of detailed or granular features [27]. Mixed findings regarding the relationship between micro features of the environment and health outcomes could be due to differences in measurement approaches across studies. An increased interest in the *loca* environment for public policy has led to increased emphasis on the rigorous development, implementation and validation of audit tools for direct observation.

In a comprehensive review, Brownson et al (2009) [27], described multiple audit tools for direct observation of the physical environment [27]. These tools shared some common content including one or more measures of: land use (e.g., presence and type of housing); streets and traffic (e.g., traffic volume); sidewalks; bicycling facilities; public space/amenities (e.g., presence of benches); architecture or building characteristics (e.g., building height); parking and driveways (e.g., parking garage); maintenance (e.g., litter); and indicators of safety (e.g., graffiti). Other features less consistently assessed are noise levels, or health promotion supports (e.g., billboards promoting physical activity) [27]. Existing audit tools have been used for one-time examinations of the neighborhood environment. As designs expand to better understand causation and predictors of change, there is a need to test whether audit tools are adequate for longitudinal assessment.

The Pittsburgh Hill/Homewood Research on Neighborhood Change and Health (PHRESH) study leverages a natural experiment design, comparing an intervention and a control neighborhood, to evaluate whether neighborhood improvements benefit residents' health [8, 24, 35]. Between 2011 and 2018, the intervention neighborhood received about \$200 million in publicly-funded investments, while the comparison neighborhood received approximately \$48 million in publicly-funded investments. Efforts involved physical infrastructure modification (i.e., lengths, names, traffic patterns) and construction of streets, housing and landscaping. To systematically document change, we conducted multiple direct observations of the neighborhood environment over a 5-year period with an emphasis on features that may impact physical activity or sleep.

Of the existing audit tools, four were comparable to ours with respect to detail, content and data collection approach: Systematic Pedestrian and Cycling Environmental Scan (SPACES) [36]; Analytic Audit Tool and Checklist Audit Tool (SLU) [37, 38]; Systematic Social Observation protocol [29] and Pedestrian Environment Data Scan (PEDS) [39]. Two of these studies reported that 70% of items had kappa statistics [40] above .40, one reported average reliability of .87, while the fourth study reported high inter-observer agreement of 75% or greater [27]. Longitudinal studies may encounter pitfalls if these audit tools are not reliable over time. Mismeasurement can obscure meaningful differences, while systematic bias can produce spurious findings. In this paper, we describe the implementation methods, lessons learned, and stability of reliability estimates from PHRESH longitudinal assessments of the neighborhood environment at three time points over five years. Our findings can help inform future studies of changes to the built and social environment.

## Methods

### Context

PHRESH is an ongoing study of two low-income and predominantly African American communities in Pittsburgh, PA. We are examining features of the built and social environment that correlate with health, as well as documenting to what extent changes impact residents' health and well-being, diet, exercise, sleep, heart, and cognitive health. The PHRESH study follows over time a cohort of households and their built and social environment to evaluate these questions. Details on study design have been described elsewhere [41, 42]. To systematically measure neighborhood-level change, we conducted detailed assessments of the physical and social environment at three different timepoints (2012, 2015 and 2017). We modified the Bridging the Gap/Community Obesity Measures Project (BTG-COMP) Street Segment Observation form [43-45], which draws from validated instruments used by other major studies assessing neighborhood features correlated with walking and overall physical activity [36, 38, 39, 46-48]. All study protocols were approved by the organization's Institutional Review Board.

### Audit Tool

The PHRESH Street Segment Audit (SSA) tool is a detailed assessment of neighborhood-level physical and social features related to health behaviors, with an emphasis on physical activity and sleep. As seen in **Table 1**, our tool includes (i) **Land use mix** capturing diversity of land use, (ii) **Physical activity (PA) facility** to include spaces for play or physical activity; (iii) **Walking/cycling environment** including presence of sidewalks, shoulders and bike lanes; (iv) **Safety signs** including traffic calming and control features; (iv) **Amenities and litter** including features that make a segment appealing, as well as pedestrian friendly, as well as two subjective assessments (perceived safety of walking; perceived attractiveness for walking) to complement the objective assessments. To the existing BTG-COMP audit tool, we added

**Environment** (e.g. trees, cliffs/ravines) and **Gathering places** (e.g. restaurants, barbershop, church). In the last data collection round (2017), we added **Social disorder** items (e.g. presence of police, people selling illegal drugs); a single item on **Noise pollution** and **Physical disorder** items (e.g. amount of beer or liquor bottles, abandoned cars), as they have been shown to be related to health behaviors such as sleep [49-51]. See Supplemental Table 1 for a full list of items.

### Street Segment Selection

Due to homogeneity among street segments within a concentrated geographic area and to reduce costs, we decided to audit a random, representative sample from each of the study neighborhoods. To draw a representative sample, we constructed a complete listing ( $n=2,027$ ) including all segments within a quarter mile of the neighborhood boundaries. The listing was compiled using a geographic shapefile provided by ESRI (ESRI, 2011) and the City of Pittsburgh, with additional accuracy checks of the shapefile conducted using updated Google map images, on-the-ground inspections and any feedback from data collectors. The decision to draw a random 25% sample was informed by an earlier published study [52]. Therefore, 511, 585 and 586 segments were sampled in 2012, 2015 and 2017, respectively. Whenever possible, a street segment was followed over time. The planned change in the study neighborhoods from the start were expected to affect the nature and existence of some streets. Thus, we established consistent rules to address such changes. Specifically, if a sampled segment did not exist at a follow up wave, a randomly selected segment from the same neighborhood served as replacement. If a sampled segment was bisected, both parts were included. If a segment was lengthened, the new attributes (including revised length) of the segment were recorded for follow up audits.

### Data collection

Audits were conducted by data collectors familiar with the neighborhoods, who were trained by an experienced trainer. Training consisted of three parts: (i) in-class presentations including examples and photographs (figure 1) with discussions about highlighted characteristics to look for; (ii) field practice on 'live' street segments around the training site; and (iii) a certification exercise where the data collectors and the trainer independently rated the same street segment, and compared ratings to test the data collector's understanding of the tool, observation skills, and data recording technique. Data collectors were given a comprehensive manual with the safety protocol and detailed description of audit tool items accompanied with photo examples (figure 2), and a summary sheet responding to common questions asked (e.g. FAQ). Each street segment was audited by a team of two data collectors (hereafter, DC pair), which is shown to improve reliability of ratings [39]. The DC pair walked the street segments together and made a single joint rating for each item, with discussions to resolve disagreements about proposed ratings in real time. A field coordinator oversaw data collection and assigned data collectors to street segments using maps. In each year, audits were conducted between August and October.

## Reliability testing

A random sub-sample of street segments was subject to reliability testing ( $n=60$  in 2012, 2015;  $n=100$  in 2017). Each segment in the reliability sample was audited twice within a one-week period. Different DC pairs conducted the two ratings, so that no individual rated the same segment twice. The two ratings were also matched on day and time in 2017 because these factors were considered important for the new physical and social disorder items (see Table 1) added to the 2017 audit tool. Our reliability statistics were chosen to accommodate the response categories used in the SSA tool. About half the items had three response categories (“neither”, “either”, “both sides of the street”), while the rest were mostly binary noting whether a feature was present or absent in that street segment. A few items (e.g. physical disorder) had more than three response categories (e.g. none, a few (1-3), some (4-6), a lot (7 or more)).

Reliability analysis included calculation of prevalence, percentage inter-observer agreement (hereafter, PO) [53, 54] and krippendorf’s alpha (hereafter, KA) [55-58]. Reliability statistics including KA are sensitive to base or prevalence rates. Therefore, while the KA is more rigorous and indicates whether agreement exceeded chance levels, we computed the PO statistic as a supplemental index of interrater reliability for all items. PO indicates the proportion of street segments where DC pairs were in exact agreement (e.g. both rated as “no” for the same street segment). For figure 3, we used the following classification for PO: PO > 90% indicates excellent agreement, PO between 75% and 90% indicates good agreement, and PO < 75% combines moderate and fair to poor agreement [59, 60]. Consistent with prior research, KA > .75 indicates excellent agreement, KA between .4 and .75 indicates intermediate to good agreement, and KA < .40 indicates poor agreement [61]. The reliability statistics can tell us whether an audit tool item has good to excellent agreement at a *single* time point. On the other hand, items with good to excellent agreement at *every* timepoint demonstrate stability, making them appropriate to detect change.

## **Reliability Results**

KA or PO statistics, with color-coding to indicate level of agreement, are displayed in Fig. 3. For most items, we report KA; where items are very common or rare, we report PO. In 2012, 93.8% of items had excellent (62.5%) or good (31.3%) agreement. In 2015, 91.3% of items had excellent (83.8%) or good (7.5%) agreement. In 2017, 83.5% of items had excellent (55.7%) or good (27.8%) agreement. When assessing stability across waves, 81.4% (79 out of 97) of items had good to excellent agreement at every timepoint, making them sufficiently reliable to detect change. Prevalence statistics are shown in supplemental Table 1.

Twelve of 14 Land use mix items had good to excellent agreement while two items (public/communal spaces, other land use) had poor agreement at all waves. Five out of 6 Environment items had good to excellent agreement across waves, while one item (“do trees shade sidewalk?”) had poor agreement at one of the three waves. Inspection of the individual raters’ responses suggests that raters seemed to have

difficulty in choosing “some” versus “many” as a response. For all 8 items in the PA facility category, there was uniformly excellent agreement at each wave.

There were 20 items in the Walking/Cycling environment category. Within the sub-category “Intersection and Crossing” including four items (traffic light, pedestrian signal at traffic light, stop sign, marked crosswalk), all had good to excellent agreement at every wave. Of the 8 items in the sub-category “Street features”, four showed good to excellent agreement at every wave. Another three items (“street and sidewalk buffer”, “continuous sidewalk”, “sidewalk continuous at both ends between segments”) showed poor agreement at one of the waves, while a fourth item (“curb cuts or ramps missing at crossing points”) exhibited consistently poor agreement at every wave. The four items in the sub-category “Traffic features” (“traffic circle/roundabout”, “speed hump/table”, “median with traffic island”, “curb extension/bulb-out”) and the two cycling environment items demonstrated good to excellent agreement at every wave. The other two items in Walking/Cycling environment (street type, number of traffic lanes), showed poor agreement at either one or two of the timepoints.

There were five items in the Safety signs category; all were reliably assessed at every wave. 12 out of 16 Amenities and litter items had good to excellent agreement at every wave. Two items (“art or monument”, “garden bed/planter”) showed poor agreement at one of the three waves, while a third item (“amount of trash/litter on street”) showed low agreement at every wave. Of the two more general assessments made by raters (“perceived safety”, “attractiveness of segment for walking”), only one (“perceived safety”) had poor agreement in one wave. Also, PO was excellent for 7 of the 8 items in the Physical activity facility category, and poor for 1 item (“other gathering place”) at two of the three waves.

For 17 items in three categories, we cannot assess agreement at multiple time points because they were only measured in 2017. A single, ordinal item in Noise pollution (with 4 response categories: “no”, “a little”, “some” or “a lot of pollution”) demonstrated good agreement. Seven of the 8 Social disorder items had excellent agreement (PO statistic > 90%) while one item (“adults loitering, congregating, or hanging out”) had poor agreement (PO < 75%). Three of the 8 Physical disorder items (“discarded cigarette butts”, “garbage, litter, broken glass”, “buildings with broken windows”) had low agreement while the other five had good or excellent agreement.

## **Discussion: Lessons Learned**

PHRESH is an ongoing study of two low-income and predominantly African American urban communities in Pittsburgh, PA. To assess whether neighborhood-level changes impact residents' health and well-being, diet, exercise, sleep, heart, and cognitive health, we conducted three assessments of the physical and social environment in the two neighborhoods over a period of five years (2012–2017). These data were examined to identify correlates of, and the extent to which neighborhood-level changes, affected obesogenic behaviors such as physical activity, sleep, and heart health. Our neighborhoods are representative of other urban, disadvantaged areas that may benefit from improvements in the environment, which is a promising approach in the fight against chronic diseases. To our knowledge, this

is the first study to conduct systematic longitudinal observations of the environment. In this paper, we have provided guidance on sampling, data collection and reliability assessment. In addition, we include results of repeated reliability testing to determine whether this audit tool and its standard set of items have enough stability across time to detect change.

As with any research endeavor, sampling was a critical step. Earlier work had demonstrated that a 25% sample of street segments produced valid estimates of the built environment [52]. When assessing neighborhood-level change, one difficulty is that these changes can modify the underlying street network. Our experience suggests that secondary sources of data may include non-negligible errors potentially due to lags in updating secondary databases. Whenever feasible (e.g. in a compact environment), we recommend careful verification of any available listing of street segments in the neighborhoods to ensure high accuracy. It is equally important to update the street network at each assessment wave to understand the degree of change in the street network. To reflect ongoing changes in the street network, we carefully identified and sampled new street segments at each wave. When sampling new segments, there should be systematic rules in place. For instance, when an entire street segment was demolished, should the replacement come from the same geographic area or be sampled entirely at random? Should a newly bisected street count as two new streets, or as the same street segment from a prior wave? A changing street network also made segment-level panel analysis difficult; instead, it was more reasonable to identify a stable unit of analysis (e.g. a residential buffer of each study participant) for assessing change.

Thorough and consistent training of data collectors at every wave was also important. Ideally, each round of training should employ the same methods and trainer to avoid systematic biases in ratings across waves. During training, it was important to balance classroom learning with 'live' practice. In the classroom, the use of visuals (e.g. photographs) to demonstrate each item worked well. Field practice focused on individual sections of the audit tool and presented a variety of observations. It was important to budget enough time to allow data collectors to discuss questions/situations with the trainer. Thus, the training schedule needed to be flexible to allow extra time for such hard-to-assess items. Furthermore, we found field practice to be the most valuable part of training. When selecting individual data collectors, attention to detail was an important individual trait. Early on in PHRESH, we also integrated a community engaged research framework to ensure its longevity and success [41].

Assessment of (inter-rater) reliability or agreement of individual SSA items helped identify which items performed well at a single timepoint, and across time. A majority of SSA items (81%) had high reliability. Low agreement indicated items that were difficult to rate objectively or with a single observation. For example, "amount of litter" or "adults loitering, congregating or hanging out" may vary even over a short window of time (e.g. a few hours or a day). In the case of trash, we re-assessed agreement for a small subset of street segments in the reliability study where two observations were conducted within hours of each other. However, the agreement for trash or litter did not improve. Items with substantial temporal variation may require multiple ratings (> 2) to fully capture more meaningful levels of variation. Certain items (e.g. perceived safety) are inherently subject to interviewer interpretation, and demonstrated lower

agreement, as might be expected. Certain neighborhood features were likely easy to miss across an entire block or street segment (e.g. bar on a single window, cigarette butts on the ground; garden bed/planter), or difficult to assess from the outside (e.g. public/communal space, vacant building) as was necessary according to the audit protocol.

Given the study findings, we can suggest the types of items that may be able to capture change. Consistent with previous research, more subjective measures are less reliable than more objective (observable) ones [39]; dichotomous ratings have higher reliability than ordinal response scales (although the greater number of response categories may be valuable for providing finer distinctions). Large, visible items (e.g. buildings, traffic signs) were consistently reliable. While sidewalks are an important feature of the walking environment, sidewalk conditions can change quickly over a city block, making it challenging to rate consistently. Also, rare/low prevalence features (see supplemental Table 1) did not lend themselves well to KA testing. For example, the only Gathering places in the study neighborhoods with prevalence above 5% were churches. If the low prevalence items were readily identified, the PO statistic showed consistency in endorsing their absence.

While some features of the environment may change, there are also features that should be time invariant. Yet, when we compared slope (“flat”, “slight hill”, “steep hill”) across years for a sub-group of street segments with three years of complete data, 22% of the segments had different values although slope is unlikely to change. Also, 10% of street segments were endorsed as having art/monument in 2012, while only 2% of segments had art/monument three years later (2015.) which may point to some confusion over what constitutes art. Therefore, we recommend the use of only those SSA items with consistently good to excellent agreement at every wave to assess change. Future studies may be able to improve the measurement of these less reliable items through particularly detailed and intensive training or procedures (e.g., mapping out a visual area into a grid to more systematically inspect for broken windows), clearer rules and examples for determining whether something is a communal space, or by the addition of a “cannot determine” category to the form. Even subjective ratings might be improved if anchored through training or explicit item instructions (e.g. 1 = a place where you would not feel physically at risk of violence from another person if walking alone in daylight, etc.), and use of multiple raters to reduce individual rater idiosyncracies.

In our knowledge, this article is the first to conduct repeated assessments of the built and social environment to assess change. The PHRESH study’s SSA tool is reliable and practical to implement, requiring an average of 13 minutes per street segment, that trained data collectors found easy to use. While the audit tool provides rich and detailed data on environmental features, it is important to explore which of these features correlate with our health outcomes and whether these relationships are consistent over time. Due to the compact nature of our neighborhoods, we also need to test this audit tool in neighborhoods with greater variation, as certain items exhibited low or zero prevalence in the study neighborhoods. The next steps are to also to develop and to test indices that summarize features of the neighborhoods that may be predictors of the study outcomes. If valid indices of environmental features

can be derived, they will be useful in guiding public policy, urban planning and redesign in the creation of built environments that promote health.

## Conclusion

This paper presents lessons learned from repeat administrations of a comprehensive audit tool for direct observation of the built environment. Longitudinal assessments required consistency of methods and data collector training to minimize systematic differences across time. Inter-rater reliability testing conducted at each time point suggested that most items were consistently reliable and were useful to assess changes in the environment. Typically, items with poor reliability were either difficult to rate or subjective in nature, making them less useful to detect real change over time. The PHRESH-SSA tool proved to be a generally reliable and practical instrument for collecting data that trained observers found easy to use.

## References

1. Caspi CE, Sorensen G, Subramanian S, Kawachi I. The local food environment and diet: a systematic review. *Health & Place*. 2012;18(5):1172-87.
2. Giskes K, van Lenthe F, Avendano-Pabon M, Brug J. A systematic review of environmental factors and obesogenic dietary intakes among adults: are we getting closer to understanding obesogenic environments? *Obes Rev*. 2011;12(5):e95-e106.
3. Macintyre S, Ellaway A. Neighborhoods and health: an overview. *Neighborhoods and Health*. 2003;20:42.
4. Sampson RJ, Morenoff JD, Gannon-Rowley T. Assessing "neighborhood effects": Social processes and new directions in research. *Annu Rev Social*. 2002;28(1):443-78.
5. Borrell LN, Diez Roux AV, Rose K, Catellier D, Clark BL. Neighbourhood characteristics and mortality in the Atherosclerosis Risk in Communities Study. *Int J Epidemiol*. 2004;33(2):398-407.
6. Diez Roux AV, Mujahid MS, Hirsch JA, Moore K, Moore LV. The impact of neighborhoods on CV risk. *Glob Heart*. 2016;11(3):353-63.
7. Miranda ML, Messer LC, Kroeger GL. Associations between the quality of the residential built environment and pregnancy outcomes among women in North Carolina. *Environ Health Perspect*. 2011;120(3):471-7.
8. Troxel WM, Shih RA, Ewing B, Tucker JS, Nugroho A, D'Amico EJ. Examination of neighborhood disadvantage and sleep in a multi-ethnic cohort of adolescents. *Health & Place*. 2017;45:39-45.
9. Wood L, Hooper P, Foster S, Bull F. Public green spaces and positive mental health—investigating the relationship between access, quantity and types of parks and mental wellbeing. *Health & Place*. 2017;48:63-71.
10. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW, et al. Correlates of physical activity: why are some people physically active and others not? *Lancet*. 2012;380(9838):258-71.

11. Coombes E, Jones AP, Hillsdon M. The relationship of physical activity and overweight to objectively measured green space accessibility and use. *Soc Sci Med*. 2010;70(6):816-22.
12. Heath GW, Brownson RC, Kruger J, Miles R, Powell KE, Ramsey LT, et al. The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. *J Phys Act Health*. 2006;3(s1):S55-S76.
13. Mozaffarian D, Afshin A, Benowitz NL, Bittner V, Daniels SR, Franch HA, et al. Population approaches to improve diet, physical activity, and smoking habits: a scientific statement from the American Heart Association. *Circulation*. 2012;126(12):1514-63.
14. Saelens BE, Sallis JF, Black JB, Chen D. Neighborhood-based differences in physical activity: an environment scale evaluation. *Am J Public Health*. 2003;93(9):1552-8.
15. Van Cauwenberg J, De Bourdeaudhuij I, De Meester F, Van Dyck D, Salmon J, Clarys P, et al. Relationship between the physical environment and physical activity in older adults: a systematic review. *Health & Place*. 2011;17(2):458-69.
16. Cappuccino FP, Miller MA, Lockley SW, Rajaratnam SMW. Sleep, Health, and Society; From Aetiology to Public Health: Oxford Scholarship Online; 2018.
17. Johnson DA, Hirsch JA, Moore KA, Redline S, Diez Roux AV. Associations Between the Built Environment and Objective Measures of Sleep: The Multi-Ethnic Study of Atherosclerosis. *Am J Epidemiol*. 2018;187(5):941-50.
18. Laurent JGC, Allen JG, Spengler JD. The built environment and sleep. Sleep, Health, and Society; From Aetiology to Public Health: Oxford Scholarship Online; 2018.
19. Warburton DE, Nicol CW, Bredin SS. Health benefits of physical activity: the evidence. *CMAJ*. 2006;174(6):801-9.
20. Warburton DER, Bredin SSD. Health benefits of physical activity: a systematic review of current systematic reviews. *Curr Opin Cardiol*. 2017;32(5):541-56.
21. Levine JA. Poverty and obesity in the U.S. *Diabetes*. 2011;60(11):2667-8.
22. Estabrooks PA, Lee RE, Gyurcsik NC. Resources for physical activity participation: does availability and accessibility differ by neighborhood socioeconomic status? *Ann Behav Med*. 2003;25(2):100-4.
23. Gordon-Larsen P, Nelson MC, Page P, Popkin BM. Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics*. 2006;117(2):417-24.
24. Troxel WM, DeSantis A, Richardson AS, Beckman R, Ghosh-Dastidar B, Nugroho A, et al. Neighborhood disadvantage is associated with actigraphy-assessed sleep continuity and short sleep duration. *Sleep*. 2019;42(3):zsy250.
25. Sallis JF, Bauman A, Pratt M. Environmental and policy interventions to promote physical activity. *Am J Prev Med*. 1998;15(4):379-97.
26. Boarnet MG, Forsyth A, Day K, Oakes JM. The street level built environment and physical activity and walking: Results of a predictive validity study for the Irvine Minnesota Inventory. *Environ Behav*. 2011;43(6):735-75.

27. Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the built environment for physical activity: state of the science. *Am J Prev Med.* 2009;36(4 Suppl):S99-123 e12.
28. Moudon AV, Lee C. Walking and bicycling: an evaluation of environmental audit instruments. *Am J Health Promot.* 2003;18(1):21-37.
29. Caughy MO, O'Campo PJ, Patterson J. A brief observational measure for urban neighborhoods. *Health Place.* 2001;7(3):225-36.
30. Laraia BA, Messer L, Kaufman JS, Dole N, Caughy M, O'Campo P, et al. Direct observation of neighborhood attributes in an urban area of the US south: characterizing the social context of pregnancy. *Int J Health Geogr.* 2006;5(1):11.
31. Raudenbush SW, Sampson RJ. Ecometrics: toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociol Methodol.* 1999;29(1):1-41.
32. Brownson RC, Hoehner CM, Brennan LK, Cook RA, Elliott MB, McMullen KM. Reliability of 2 instruments for auditing the environment for physical activity. *J Phys Act Health.* 2004;1(3):191-208.
33. Kelly CM, Wilson JS, Baker EA, Miller DK, Schootman M. Using Google Street View to audit the built environment: inter-rater reliability results. *Ann Behav Med.* 2012;45(suppl\_1):S108-S12.
34. Twardzik E, Antonakos C, Baiers R, Dubowitz T, Clarke P, Colabianchi N. Validity of environmental audits using GigaPan® and Google Earth Technology. *Int J Health Geogr.* 2018;17(1):26.
35. RAND. PHRESH: Pittsburgh Hill/Homewood Research on Neighborhood Change and Health no date. Accessed March 21, 2019, from: <https://www.rand.org/well-being/community-health-and-environmental-policy/projects/phresh.html>.
36. Pikora TJ, Bull FC, Jamrozik K, Knuiaman M, Giles-Corti B, Donovan RJ. Developing a reliable audit instrument to measure the physical environment for physical activity. *Am J Prev Med.* 2002;23(3):187-94.
37. Brownson RC, Chang JJ, Eyler AA, Ainsworth BE, Kirtland KA, Saelens BE, et al. Measuring the environment for friendliness toward physical activity: a comparison of the reliability of 3 questionnaires. *Am J Public Health.* 2004;94(3):473-83.
38. Hoehner CM, Ivy A, Ramirez LK, Handy S, Brownson RC. Active neighborhood checklist: a user-friendly and reliable tool for assessing activity friendliness. *Am J Health Promot.* 2007;21(6):534-7.
39. Clifton KJ, Smith ADL, Rodriguez D. The development and testing of an audit for the pedestrian environment. *Landsc Urban Plan.* 2007;80(1-2):95-110.
40. Shoukri MM. Measures of interobserver agreement and reliability: CRC press; 2010.
41. Dubowitz T, Ncube C, Leuschner K, Tharp-Gilliam S. A natural experiment opportunity in two low-income urban food desert communities: research design, community engagement methods, and baseline results. *Health Education & Behavior.* 2015;42(1\_suppl):87S-96S.
42. Dubowitz T, Zenk SN, Ghosh-Dastidar B, Cohen DA, Beckman R, Hunter G, et al. Healthy food access for urban food desert residents: examination of the food environment, food purchasing practices,

- diet and BMI. *Public Health Nutr.* 2015;18(12):2220-30.
43. Kelly CM, Schootman M, Baker EA, Barnidge EK, Lemes A. The association of sidewalk walkability and physical disorder with area-level race and poverty. *J Epidemiol Community Health.* 2007;61(11):978-83.
  44. Slater SJ, Nicholson L, Chiqui J, Barker DC, Chaloupka FJ, Johnston LD. Walkable communities and adolescent weight. *Am J Prev Med.* 2013;44(2):164-8.
  45. Zenk SN, Slater S, Rashid S. Collecting contextual health survey data using systematic observation. *Handbook of Health Survey Methods John Wiley and Sons, Inc.* 2015:421-45.
  46. Day K, Boarnet M, Alfonzo M, Forsyth A. The Irvine–Minnesota inventory to measure built environments: development. *Am J Prev Med.* 2006;30(2):144-52.
  47. Emery J, Crump C, Bors P. Reliability and validity of two instruments designed to assess the walking and bicycling suitability of sidewalks and roads. *Am J Health Promot.* 2003;18(1):38-46.
  48. Slater SJ, Ewing R, Powell LM, Chaloupka FJ, Johnston LD, O'Malley PM. The association between community physical activity settings and youth physical activity, obesity, and body mass index. *J Adolesc Health.* 2010;47(5):496-503.
  49. Brownson RC, Brennan Ramirez LK, Hoehner CM, Cook RA. Analytic audit tool and checklist audit tool. 2003.
  50. Caspi CE, Kawachi I, Subramanian S, Tucker-Seeley R, Sorensen G. The social environment and walking behavior among low-income housing residents. *Soc Sci Med.* 2013;80:76-84.
  51. Sampson RJ, Raudenbush SW. Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *Am J Sociol.* 1999;105(3):603-51.
  52. McMillan TE, Cubbin C, Parmenter B, Medina AV, Lee RE. Neighborhood sampling: how many streets must an auditor walk? *Int J Behav Nutr Phys Act.* 2010;7(1):20.
  53. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol.* 1990;43(6):543-9.
  54. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol.* 1990;43(6):551-8.
  55. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas.* 2007;1(1):77-89.
  56. Krippendorff K. Estimating the reliability, systematic error and random error of interval data. *Educ Psychol Meas.* 1970;30(1):61-70.
  57. Krippendorff K. Content analysis: An introduction to its methodology, 3rd edition: Sage Publications; 2013.
  58. Krippendorff K, Fleiss JL. Reliability of binary attribute data. JSTOR; 1978.
  59. Hartmann DP. Considerations in the choice of interobserver reliability estimates. *J Appl Behav Anal.* 1977;10(1):103-16.

60. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Pract Assess Res Eval*. 2004;9(4):1-19.
61. Fleiss JL. Balanced incomplete block designs for inter-rater reliability studies. *Appl Psychol Meas*. 1981;5(1):105-12.

## Table

**Table 1. PHRESH Street Segment Audit (SSA) Tool**

<b>Audit tool section</b>	<b>Number of items</b>	<b>Sample items</b>
Land use mix	14	Housing, public/civic, office, retail, recreational, vacant.
Environment	6	Slope, slight or steep hill, number of trees, shade from trees, bars on or broken/boarded windows
Physical activity facility	8	Indoor, park, playing field, court, trail
Walking/cycling environment	20	Street type, vehicular lanes, traffic features, bike lanes, sidewalks
Safety signs	5	Bicycle or pedestrian crossing, kids at play, special speed limit
Amenities and litter	16	Neighborhood or community sign, garden/flower bed/planter, art/statue/monument, benches, drinking fountains, bus stops, trash, perceived safety while walking, attractiveness of street segment for walking
Gathering places	11	Restaurants, libraries, barbershops, churches, bars, corner stores
Social disorder	8	Presence of police, adults loitering, loud music, people smoking
Noise pollution	1	Level of noise
Physical disorder	8	Broken bottles, drug paraphernalia, graffiti, broken windows

## Declarations

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:** Not applicable.

**Availability of data and materials:** Not applicable.

**Competing interests:** The authors have indicated they have no financial relationships relevant to this article to disclose.

**Funding:** This research was supported by the National Institutes of Health (grant numbers NCI R01CA149105, NHLBI R01HL1122460, and NHLBI R01HL131531).

**Author's contributions:**

Ghosh-Dastidar: Conceptualization, Methodology, Writing - original draft;

Hunter: Data curation, Methodology, Programming, Writing – review and editing

Sloan: Data collection, Literature review, Writing – review and editing

Rebecca L. Collins: Conceptualization, Writing – review and editing

Richardson: Methodology, Writing – review and editing

Troxel: Funding acquisition, Audit tool development, Writing – review and editing

Colabianchi: Audit tool development, Writing – review and editing

Dubowitz: Funding acquisition, Audit tool development, Writing – review and editing

## **Figures**

### B4c1. Street and Sidewalk Buffer\*



■ Boundary that provides physical or psychological distance between traffic and sidewalk

### B4c1. Street and Sidewalk Buffer\*

- Can include
  - landscaped or grass strips
  - Hedges
  - Barricades
  - Fences/guard rails
  - Regularly placed street trees



### B5d1. Street and Sidewalk Buffer\*



### B4c1. Street and Sidewalk Buffer\*



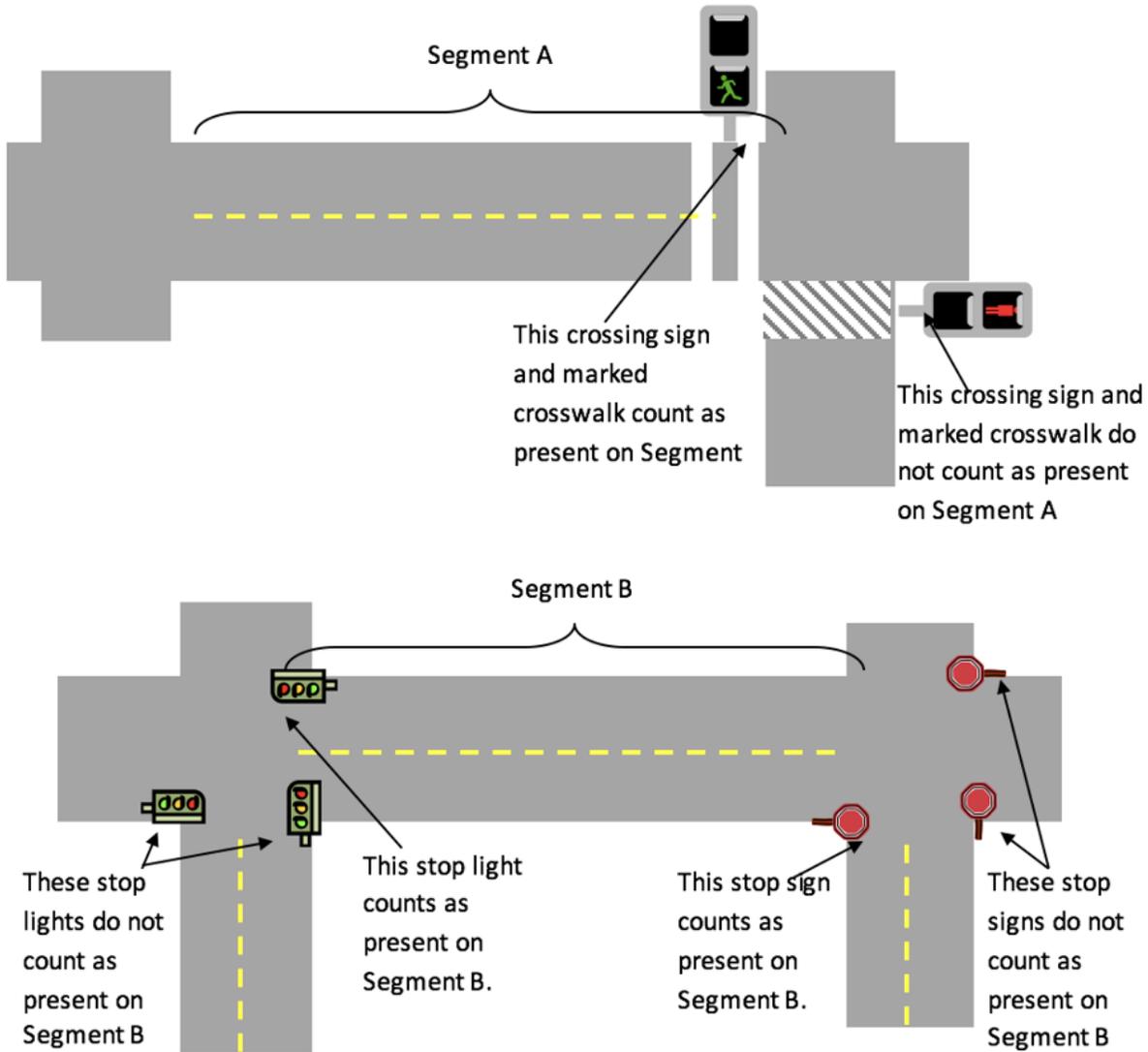
Figure 1

Example of PHRESH SSA classroom training slides.

**d. Marked Crosswalk**

Refers to a crossing point with markings for a pedestrian to cross the street segment that you are observing. These markings include painted lines, zebra striping or different road surface or paving, such as bricks. They may include flashing lights level with the street. Marked crosswalks are usually located at the end of a segment at a point of intersection but they may be present at other locations.

See the diagrams below for clarification on the items in **B6**.



**Figure 2**

Example of PHRESH SSA Training Manual

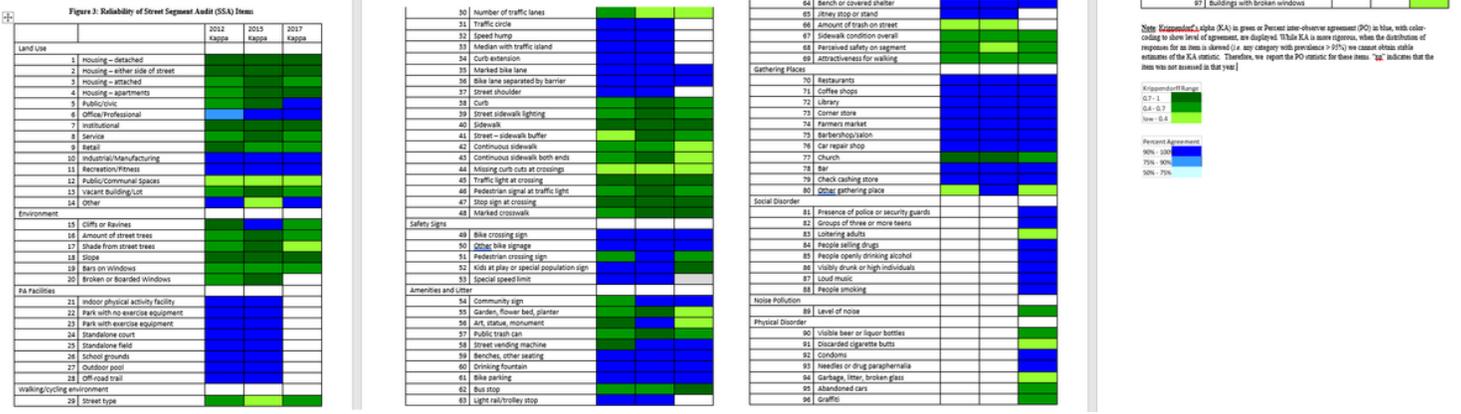


Figure 3

Reliability of Street Segment Audit (SSA) Items