Title: Test-retest reliability of perceptions of the neighborhood environment for physical activity by socioeconomic status

Citation: Turrell G, Haynes M, O'Flaherty M, **Burton NW**, Giskes K, Giles-Corti B, Wilson LM. Test-retest reliability of perceptions of the neighborhood environment for physical activity by socioeconomic status. *Journal of Physical Activity & Health*, 2011; 8(6): 829-840.

Running head: Reliability of neighborhood perceptions

Manuscript type: Original research

Keywords: epidemiology, environment, measurement, physical activity, special needs populations

Abstract

Background: Further development of high quality measures of neighborhood perceptions will require extensions and refinements to our existing approaches to reliability assessment. This study examined the test-retest reliability of perceptions of the neighborhood environment by socioeconomic status (SES).

Methods: Test and retest surveys were conducted using a mail survey method with persons aged 40-65 years (n=222, 78.2% response rate). SES was measured using the respondent's education level and the socioeconomic characteristics of their neighborhood of residence. Reliability was assessed using intra-class correlations (ICC) estimated with random coefficient models.

Results: Overall, the 27 items had moderate-to-substantial reliability (ICC = 0.41 – 0.74). Few statistically significant differences were found in ICC between the education groups or neighborhoods, although the ICCs were significantly larger among the low SES for items that measured perceptions of neighborhood greenery, interesting things to see, litter, traffic volume and speed, crime, and rowdy youth on the streets.

Conclusion: For the majority of the items, poor reliability and subsequent exposure misclassification is no more or less likely among low educated respondents and residents of disadvantaged neighborhoods. Estimates of the association between neighborhood perceptions and physical activity therefore are likely to be similarly precise irrespective of the respondent's socioeconomic background.

INTRODUCTION

Social-ecological theory posits that physical activity (PA) is influenced by a complex array of interacting factors operating at many levels, including the individual, family, friends and peers, and the neighborhood.[1] During the 1980's and 1990's research directed at understanding the determinants of PA focused on the individual.[2, 3] In more recent times, the research effort has shifted to investigating the neighborhood environment and the role that it plays in PA.[4, 5] Integral to this shift has been the development of conceptual frameworks that seek to identify characteristics of the neighborhood which facilitate or dissuade PA.[6, 7] A large number of potential factors have been delineated: in broad terms these pertain to the functionality and convenience of neighborhood infrastructure (e.g. intersections, sidewalks), crime and safety (e.g. youth on the streets, unsecured dogs, streetlights, traffic control), aesthetic surroundings (e.g. parks, graffiti and rubbish, tree coverage), and the accessibility and availability of facilities and services (e.g. shops, recreation centers, bike paths). Over 100 published studies have investigated the association between peoples' perceptions of these (and other) neighborhood factors and their influence on PA [8] and most observe a significant association. [9-12]

Knowledge of the relationship between perceptions of the neighborhood environment and PA is at a relatively nascent stage [8] and future advances in our understanding need to be supported by ongoing research into the psychometric properties of the perceptual measures.[13] A key aspect of this work is the assessment of measurement error, and this has typically been undertaken on the basis of test-retest reliability studies.[8,13] In the context of this work, reliability is an indicator of the consistency and reproducibility of responses to items and scales that measure neighborhood perceptions. Low reliability represents a threat to knowledge because reporting errors are likely to be associated with a higher probability of exposure misclassification, which in turn reduces the predictive capacity of the measure.[14, 15] To date, test-retest studies have tended to focus on the reliability of neighborhood perceptions for samples as a whole, although some have compared reliabilities across racial/ethnic groups,[16-18] males and females, [17, 19] and urban and rural areas.[20] Many of these studies used original or adapted items from the Neighborhood Environment Walkability Scale (NEWS) or the updated tool (ANEWS),[21] and reliability is often estimated using the Intra-Class Correlation (ICC). A recent review of these studies [8] reported that test-retest reliabilities for measures of neighborhood perceptions are usually between 0.6 – 1.0 in magnitude: according to Landis and Koch's[22] scale for assessing the strength of reliability coefficients, this indicates substantial or near-perfect agreement.

Further development of high quality measures of neighborhood perceptions will require extensions and refinements to our existing approaches to reliability assessment. It has been suggested that we need to assess reliability among population sub-groups with high levels of physical inactivity and increased risk of chronic disease.[8, 18] In this present paper we respond to this call by examining the reliability of neighborhood perceptions across different socioeconomic groups and residents of advantaged and disadvantaged neighborhoods. No published study appears to have focused on this issue; however, there a number of compelling reasons for doing so. First, compared with their higher status counterparts, people of low socioeconomic status (SES) and residents of disadvantaged neighborhoods make less use of their neighborhood environments for leisure time PA [23-25] hence the reliability of their neighborhood perceptions may be lower. Second, socioeconomically disadvantaged persons, especially those with less formal education, may have lower levels of literacy and comprehension,[26,27] hence they may find it more difficult to read and understand each survey item thus increasing the scope for inconsistent (i.e. less reliable) responses. Third, population-based samples frequently under-represent respondents from disadvantaged backgrounds and neighborhoods: these groups are often under-enumerated in our sampling frameworks,[28] they have higher rates of survey non-response [29] and when they do respond they are more likely to provide incomplete data (i.e. higher item non-response).[30] This under-representation is likely to result in a socioeconomically truncated sample that underestimates the magnitude of socioeconomic variability in PA in the population. Importantly, if disadvantaged groups are also less reliable in their responses to survey items then this will compound the extant non-response biases and further reduce the capacity of our models to produce accurate estimates of socioeconomic differences in PA.

The aim of this study is to assess the test-retest reliability of perceptions of the neighborhood environment by socioeconomic status. The assessment is based on data collected as part of a test-retest study for the HABITAT (**H**ow **A**reas in **B**risbane **I**nfluence Heal**T**h and **A**c**T**ivity) project. HABITAT is a multilevel longitudinal (2007-2011) investigation of change in PA among mid-aged (40-65 years) men and women living in Brisbane, Australia. Details of the HABITAT main project have been published elsewhere.[31]

METHODS

The test-retest study was awarded ethical clearance by the Queensland University of Technology Human Research Ethics Committee (Ref. No. 3967H).

Sampling

A multistage probability sampling design was used to select a stratified random sample of Census Collection Districts (CCD) (proxy neighborhoods), and from within each CCD, all dwellings containing at least one resident aged 40-65 years.

Sampling of neighborhoods: As at the 2001 census, Brisbane consisted of 1654 CCDs: these are the smallest administrative units used by the Australian Bureau of Statistics (ABS) for the collection of census data, and in 2001 each CCD contained an average of 205.6 (SD 77.7) occupied private dwellings. The 1654 CCDs were assigned a socioeconomic score using the ABS' Index of Relative Socioeconomic Disadvantage (IRSD).[32] IRSD scores for CCDs are based on census data and reflect attributes such as the proportion of low income families and individuals with limited educational attainment, the unemployment rate, and the extent of the workforce in relatively unskilled occupations (among others). The CCDs were ranked by their IRSD score to form a distribution that was divided into deciles. One CCD was then randomly sampled from the first, fifth, and tenth decile and these are hereafter referred to as low, middle, and high socioeconomic status (SES) neighborhoods respectively. Table 1 compares the three neighborhoods on the basis of 2001 census data, and as would be expected from the stratified sampling design, their socioeconomic profile differed markedly.

TABLE 1 ABOUT HERE

Sampling of dwellings and individuals: Using data provided by the Australian Electoral Commission (AEC) all dwellings in each neighborhood that contained at least one resident aged 40-64 years were identified, and one resident within this age-range was selected using simple random sampling (low SES=183, middle SES=123, high SES=113, total=419). In Australia, voting is compulsory for persons aged 18 years and over, thus AEC data provides near-complete coverage of the resident population. Except for age, no other inclusion/exclusion criteria were used to identify and select respondents.

Questionnaire and data collection

A structured self-administered questionnaire was developed that asked respondents about their neighborhood, participation in PA, attitudes, social support, and demographic and socioeconomic characteristics. The questionnaire was administered between October and November 2006 using a mail survey method developed by Dillman.[33] A primary approach letter was mailed one week in advance of the questionnaire advising potential respondents of the study and encouraging their participation. This was followed a week later by a package containing a personalized cover letter (addressed to the respondent by name and hand-signed by the study's chief investigator), the questionnaire, a pre-addressed, pre-paid reply envelope, and a small gratuity (5 x $1.00 lottery tickets). A thank-you/reminder post-card was sent to respondents after seven days, followed by (where necessary) another copy of the questionnaire and a final reminder letter (up to five contacts were made). On receipt of the respondent's questionnaire it was date-stamped and a second identical questionnaire (again with a personalized cover letter and gratuity) was mailed two weeks later.

Response rates

Of the 419 questionnaires that were mailed, 13 were subsequently deemed ineligible for a range of reasons (e.g. resident deceased, no longer at address) which reduced the in-scope sample to 406. For the test-component of the study the response rate was 70.0% (n=284/406). The sex-distribution of those who did not return their test-component survey did not differ significantly from those who returned the survey (chi-square 0.080, p=0.777). No other demographic data were available for comparison. For the high, middle, and low SES neighborhoods, the test-component response rates were 75.5%, 76.7%, and 61.9% respectively. At re-test, the 284 surveys were mailed and 227 (79.9%)

were returned. Of these, 5 were deemed ineligible (identification number removed, different person completed the second questionnaire) thus the final usable sample comprised 222 respondents (78.2% response rate, n=222/284).

Measures

Neighborhood perceptions: these were assessed using 27 modified items from ANEWS [21] grouped according to four broad constructs: suburb surroundings (7 items), streets and footpaths (7 items), traffic (5 items), and crime and safety (8 items). Each item was scaled using Likert-type response options that ranged from "Strongly disagree" (1), "Neither agree nor disagree" (3) to "Strongly agree" (5). The original ANEWS items were modified in a variety of ways to suit the local context by changing the spelling of words to be consistent with Australian-English, replacing the term 'neighborhood' with 'suburb', and using Australian vernacular (e.g. replacing sidewalk with footpath, crosswalks with pedestrian crossings, and transit stop with bus stop or train station).

Neighborhood- and individual-level SES: Neighborhood SES was measured on the basis of the IRSD, as previously described. Respondents SES was measured using their education-level: no post-school qualifications; vocational/diploma (trade or business certificate, apprenticeship, associate or undergraduate diploma); and bachelor degree or higher (bachelor, post-graduate diploma, Masters, or doctorate).

Assessment of test-retest reliability

Test-retest reliability has often been assessed using the ICC calculated on the basis of a one-way analysis of variance (ANOVA). [16, 20, 34, 35] The use of a one-way ANOVA assumes that the data come from a simple random sample that contains only two primary sources of variation in peoples' responses to survey items: across the two time-points, responses vary either between- or within-individuals (the sum of which constitutes the total variance). In the context of these studies, an ICC quantifies the proportion of the total variance that is attributed to differences between-individuals, thus large ICCs indicate high consistency and reproducibility of responses to survey items within-individuals, and by extension, a low probability of exposure misclassification. In this present study, for comparability purposes we also derive ICCs to quantify the percentage of total variance that arises from between-individual differences, but only after partitioning-out additional variation due to between cluster (neighborhood) differences. We used a complex two-level sample design that generated three primary sources of variation – within individuals, between individuals, and between neighborhoods – hence we estimated the ICC using random coefficient (mixed) models to partition-out the latter source of variation when computing between-individual variation. Importantly, had we estimated ICCs using a one-way ANOVA with the neighborhood clustered data (as some previous studies have done) the coefficients in most instances would be upwardly biased due to the aggregation of both between-neighborhood and between-individual variance. The reasons for this are illustrated in Table 2, where ICCs are estimated for one of the neighborhood perception items under four different modeling scenarios.

TABLE 2 ABOUT HERE

First, we derived an ICC using a one-way ANOVA (Model 1) and then using a two-level variance components mixed effects regression model (Model 2): each approach yielded an ICC of identical magnitude (ICC=0.59). Both methods however are potentially problematic as they fail to recognize that the data structure contains three levels of variation: within- and between-individuals, and between-neighborhoods. The analysis was then extended by specifying a three-level mixed model that allows for variation between neighborhoods (Model 3): with this variation separately identified the ICC is re-estimated as 0.51. In effect, Models 1 and 2 produced ICCs that were over-estimated because the between-individual variance encapsulated variance that was more appropriately attributed to between-neighborhood variation. This said however, specifying a 3-level mixed model with only three neighborhoods was found to be unsuitable for two reasons. First, assumptions about the distributional properties of the level 3 random term were likely to be violated, and second, the models sometimes failed to converge when the between-neighborhood variance was close to zero. For these reasons, we estimated the ICCs using a two-level mixed model with neighborhood included as a fixed effect (Model 4): as can be seen in Table 2, Models 3 and 4 are equivalent in terms of their capacity to estimate the ICC when the number of level 3 units is small.

When interpreting the ICCs, we use Landis and Koch's [22] scale of strength for reliability coefficients: Poor (0.0-0.20), Fair (0.21-0.40), Moderate (0.41-0.60), Substantial (0.61-0.80), and Almost Perfect (0.81-1.00).

Analysis

Of the 222 test-retest surveys that were returned, two had missing data at both time-points for education, and another provided no data about date of birth: these were excluded, which reduced the analytic sample to n=219.

The analyses were undertaken in three stages using STATA/SE 10.0 for Windows. First, we estimated whole-sample ICCs using each of the four methods indicated in Table 2. This offers a methodological contribution by illustrating the potentially biasing effects on the ICC of not adjusting for neighborhood-clustering. For each of the ICCs that were estimated using the neighborhood-adjusted mixed model, we used a bootstrap procedure[36] to derive 95% confidence intervals (CI). Specifically, 1000 random samples were drawn from the test-retest analytic sample and for each sample an ICC was calculated, and the standard error of the resultant sampling distribution of ICCs was used to estimate 95% CIs.

Second, using a neighborhood-adjusted mixed model we estimated an ICC for each education group to assess whether people with varying levels of education differed in their consistency and reproducibility of responses to each survey item. Preliminary analyses revealed that education was significantly associated with age ($p \leq 0.029$) and with many of the neighborhood perceptions; however, education was not associated with sex ($p \leq 0.129$). Given this, ICCs for each education group were estimated without and with adjustment for respondent's age and the results were similar, so only the latter are reported.

Third, when estimating ICCs for each separate neighborhood, it wasn't possible (appropriate) to include a neighborhood term in the model specification. Hence, we initially derived ICCs using a two-level variance components analyses with no terms in the fixed part of the model. This was subsequently extended by adjustment for education and age, as each was significantly associated with neighborhood SES ($p \leq 0.001$ and $p \leq 0.025$ respectively). Sex was not associated with neighborhood SES ($p \leq 0.454$).

Differences between the education groups and between the neighborhoods in the magnitude of their ICCs were examined in a number of ways. First, they were compared on their mean ICC for each item on the basis of t-tests (using variances derived from the bootstraps). Second, the distribution of the 27 ICCs was examined using box-plots to compare the education groups and neighborhoods in terms of their medians and inter quartile ranges. Third, we tested for an overall difference in the median of the ICCs between each pair of education- and neighborhood-contrasts (e.g. low vs. high education, low vs. middle neighborhood SES) using a non-parametric Wilcoxon sign-test (two-sided).

RESULTS

Table 3 compares the socio-demographic characteristics of the test-retest sample with those who did not return their retest survey. Respondents with low levels of education were more likely to have not returned their retest survey. The two groups did not differ significantly (at the conventional $p \leq 0.05$) by sex, neighborhood SES, country of birth, self-rated health, PA, or age. There was an increased tendency for non-Australian born respondents to have not returned their retest survey, especially those from a non-English speaking country.

TABLE 3 ABOUT HERE

Table 4 presents ICCs for the test-retest sample using the four different modeling scenarios. As shown, analyses that ignore the inherent neighborhood clustering within the data (i.e. one-way ANOVA and a two-level mixed model) systematically over-estimate the ICCs when compared with the three-level model and the neighborhood-adjusted two-level model: these latter two approaches tend to produce ICCs of similar magnitude. For some of the neighborhood perceptions, the ICCs were identical (or nearly so) irrespective of which of the four modeling approaches was used. This occurred when the magnitude of the between-neighborhood variance for the perceptual item was small.

The ICCs produced by the neighborhood adjusted mixed models ranged from 0.41 to 0.74 (mean 0.62, SD 0.09). For each of the four neighborhood constructs, the magnitudes of the ICCs were as follows: suburb surroundings 0.51 – 0.68 (mean 0.60, SD 0.06), streets and footpaths 0.51 – 0.72 (mean 0.62, SD 0.08), traffic 0.52 – 0.64 (mean 0.58, SD 0.05), and crime and safety 0.41 – 0.74 (mean 0.66, SD 0.11).

TABLE 4 ABOUT HERE

Table 5 presents ICCs by education-group: there were few statistically significant differences in the mean ICCs between the groups, however when these were evident the ICCs tended to be largest among the low or middle educated. The box-plot for ICC by education group (Figure 1) shows that the median value of the 27 ICCs was higher for the low educated group overall, and that the inter-quartile range (i.e. the middle 50% of the distribution) was above the median value for the high and middle educated groups. The Wilcoxon sign-test however indicated that the median values for the low, middle, and high education groups were not significantly different.

TABLE 5 AND FIGURE 1 ABOUT HERE

Table 5 also presents ICCs by neighborhood SES: again there were only a small number of statistically significant differences between the neighborhoods in their mean ICCs for each item. When differences were observed however, the ICC tended to be largest for the middle and low SES neighborhoods. Figure 1 shows that the median ICC for the 27 items was largest for the low SES neighborhood, intermediate for the middle SES neighborhood, and smallest for the high SES neighborhood: none of the pair-wise differences in median values were statistically significant.

DISCUSSION

This present study assessed the reliability of 27 adapted items from the ANEWS tool[21] and the findings contribute to the measurement literature in a number of important ways. We found that the ICCs for the total sample ranged from 0.41 - 0.74 (mean 0.62, SD 0.09). Based on Landis and Koch's scale of strength for reliability coefficients[22] 29.6% (n=8) of the items showed 'moderate' reliability (0.41 – 0.60) and 70.4% (n=19) 'substantial' reliability (0.61 – 0.80). None of the ICCs were classified in the 'poor' (0.0

– 0.20), ‘fair’ (0.21 – 0.40) or ‘almost perfect’ (0.81 – 1.0) categories. In terms of the four neighborhood constructs, items relating to crime and safety showed the highest average reliabilities, followed by items pertaining to streets and footpaths, suburb surroundings, and traffic. It is unclear why the reliability coefficients for the crime and safety items were higher than for the other constructs (or alternatively, why the ICCs for the other constructs were lower): there are however, a number of possible reasons. First, perceptions of crime and safety may be more robust and less susceptible to modification by the test-retest process, which is somewhat akin to a ‘perceptual intervention’ for highly visible features of the neighborhood such as greenery, or the quality of streets and footpaths. For example, as a result of completing the test survey respondents may have become more sensitized to the physical characteristics of their neighborhood and notice aspects of their environment that they were previously unaware of, thus producing a different response on the retest survey, resulting in lower reliabilities. Second, the lower reliability coefficients for items measuring local traffic and suburb surroundings may reflect the changeable and/or subjective nature of these items. The level of traffic in a neighborhood for example may vary according to time of day. Moreover, in another study[37] we found that even trained auditors had difficulty assessing neighborhood aesthetics and the size of street trees, presumably because of the subjective nature of these constructs. Third, the higher reliabilities for the crime and safety items may have been a chance occurrence: the fact that previous test-retest studies have not found that ICCs are noticeably higher for crime and safety items tends to support this interpretation.

Comparing the whole-sample ICCs found in this study with previous studies is difficult for a range of reasons: the wording and level of measurement of the items often differed; the studies varied in their method of data collection (telephone, mail survey, or face-to-face); the time-lags differed between administering the test and retest surveys; and the samples on which the studies were based ranged from the general population,[20] a group of African university students,[19] and ethnically and racially diverse women.[16,17] This heterogeneity notwithstanding, the size of the whole-sample ICCs for many items were broadly similar across some of the studies. For example, Evenson and McGinn,[17] Brownson et al[20] and this present study reported ICCs that ranged as follows: interesting things to see (0.61-0.64); tree cover along footpaths/sidewalks (0.49-0.54); attractive buildings (0.64-0.68); availability of footpaths/sidewalks on streets (0.72-0.77); four-way intersections (0.51); hilly streets (0.52-0.53); footpath maintenance (0.68-0.69); heavy/lot of traffic (0.65-0.67); speed of traffic (0.62-0.65); exhaust fumes from motor vehicles (0.50-0.63); amount of crime in neighborhood (0.61-0.72); unsecured dogs (0.63-0.67); safety from crime at night (0.69-0.73); and safety from crime during the day (0.31-0.49). Arguably, the magnitude of most of the ICCs from each of these studies is within acceptable levels of reliability according to Landis and Koch's scale of strength for reliability coefficients.[22] Importantly, the fact that these broadly similar ICCs are found despite varying social and cultural contexts and the use of differently worded and measured items and modes of survey administration points to the general robustness of peoples' perceptions of their neighborhood environments; and it attests to the stability of the ANEWS items and their suitability for public health monitoring and surveillance. This said however, it is worth noting that some of the ICCs for the neighborhood perception items (albeit the minority) varied markedly across the studies. For example, the reliability of the ANEWS item 'The streets in my neighborhood do not have many, or any, cul-de-sacs (dead-end-streets)' was considerably lower in the Brownson et al[20] study (ICC 0.18) than in the present study (ICC 0.63) that removed the double negative and used the simpler statement 'Many streets in my suburb have cul-de-sacs (dead-end-streets)'. Items that use double negatives may confuse respondents, and future studies may wish to use both items in order to assess whether the ANEWS measures could be further enhanced with this minor modification.

This is the first-known study to have compared socioeconomic groups and neighborhoods in terms of their test-retest reliabilities for items measuring neighborhood perceptions, although previous studies have examined sub-group differences by race/ethnicity, rurality, and gender. Evenson et al[16] observed that ICCs tended to be larger for white women compared with women of Latino, African American, or Native American origin. Brownson et al[20] found that some “blocks of questions” were more reliable for urban than rural residents, and Evenson and McGinn[17] and Oyeyemi et al[19] reported that reliabilities tended to be higher for males. The average difference in the reliabilities between these sub-groups was usually only small-to-modest in magnitude and sometimes item specific, however, the differences provided an important context and precedent for this present study of socioeconomic differences in test-retest reliabilities. We originally hypothesized that people of low SES and residents of disadvantaged neighborhoods would exhibit lower reliabilities than their more advantaged counterparts because of lower levels of literacy and comprehension,[26,27] and lower rates of leisure time PA in the neighborhood.[23-25] For the majority of the items however, there was no convincing evidence that the reliabilities differed systematically by education group or neighborhood SES. It therefore seems that higher rates of survey and item non-response among socioeconomically disadvantaged respondents (and the problems of sample bias that this often causes) will not, in most cases, be compounded by concomitant lower reliabilities. Moreover, exposure misclassification associated with low reliability seems to be no more or less likely among low educated respondents or residents of disadvantaged neighborhoods, hence estimates of the association between neighborhood perceptions and

PA are likely to be similarly precise irrespective of the respondent's socioeconomic background.

Contrary to our hypothesis, the few statistically significant socioeconomic differences that we found showed that the ICCs tended to be larger among respondents from low educated groups and/or low SES neighborhoods. This was apparent for items that measured perceptions of neighborhood greenery, interesting things to see, litter or rubbish, traffic volume and speed, crime, and rowdy youth on the streets. The reasons for the larger reliabilities for these constructs are unknown; however, the socioeconomically disadvantaged are more likely than the advantaged to walk for transport[38] and thus may be more exposed to the negative aspects of these neighborhood characteristics (e.g. sparse neighborhood greenery, fewer interesting things to see, more crime) resulting in a greater consistency and reproducibility of their responses and hence larger ICCs.

In sum, our findings in relation to socioeconomic differences in test-retest reliabilities are not unlike the pattern of findings reported in studies that compared reliabilities between racial/ethnic groups, urban and rural areas, and men and women. In the majority of cases the magnitude of the ICCs were similar (and acceptable) irrespective of population group however there appears to be some items (or groups of items) that have more or less saliency depending on the sub-group; and where this occurs, estimates of reliability are commensurately higher or lower depending on the particular context or circumstances. Further, given the stability of many of the ICCs despite substantial between-study differences on numerous design and measurement issues, there is no obvious reason why our findings pertaining to SES would not apply to a non-Australian population: future research should investigate this issue.

This paper made an important methodological contribution to the assessment of measurement error by showing that when estimating ICCs using a neighborhood based cluster sample it was necessary to take account of any between-neighborhood variation else the reliability coefficients may be upwardly biased. This was illustrated by comparing ICCs calculated using a one-way ANOVA (which doesn't partial-out the neighborhood variation) and a neighborhood-adjusted two-level mixed model. The former technique consistently produced ICCs that were larger than that derived using the latter procedure. Given the recent proliferation of neighborhood-based studies of PA (and health more generally) [39] the use of mixed-models to accurately estimate test-retest ICCs will become increasingly necessary.

Study limitations

First, although the study achieved a high overall response rate at test (70.0%) and re-test (78.2%), low educated respondents and those from the most disadvantaged neighborhood were less likely to have returned their test-retest survey hence the socioeconomic profile of the analytic sample was likely to be truncated vis-à-vis the general population. Second, the analyses may have been under-powered at the both the individual- and area-levels and as a result many of the between-group and between-neighborhood differences failed to reach statistical significance. Third, as the sample comprised only three neighborhoods it is not possible to generalize to the wider population of Brisbane neighborhoods. Differences between the neighborhoods in the magnitude of their ICCs may simply reflect each area's unique and idiosyncratic features, and not necessarily the features of advantaged and disadvantaged neighborhoods more generally. Finally, as there were only three neighborhoods we were required to use a two-level mixed model with neighborhood included as a fixed effect to estimate the ICCs for each education group. Arguably, this is a sub-optimal approach to modeling a data structure that has three sources of variation. This said however, with a small number of neighborhoods it seems to matter little whether the ICC is estimated using a three-level mixed model (where the between-neighborhood variance is treated as a random term) or a two-level model with adjustment for neighborhood as a fixed (average) effect. Future test-retest studies employing mixed effect models to estimate ICCs should explore these issues further using a larger number of neighborhoods.

Conclusion

Although a large number of studies show that peoples' perceptions of their neighborhood environment are associated with their levels of PA, there is an ongoing need to continually improve the reliability of items and scales that measure neighborhood perceptions. Doing so will improve the predictive capacity of our models, add to our understanding of how the neighborhood environment influences behaviour, and ultimately, lead to the development and implementation of effective (and cost efficient) interventions to increase rates of PA in the population.

Acknowledgements

We gratefully acknowledge Ms Kathryn Jordan and Ms Robyn Baguley for their work as Project Managers on HABITAT.

Funding source

The HABITAT study is supported by project grants from the National Health and Medical Research Council (NHMRC) (Nos. 339718, 497236). GT and BGC are supported by NHMRC Senior Research Fellowships (Nos: 390109 & 503712 respectively) and NB is supported by a Heart Foundation Research Fellowship (No: PH08B3905) and an NHMRC Capacity Building Grant (No: 252977).

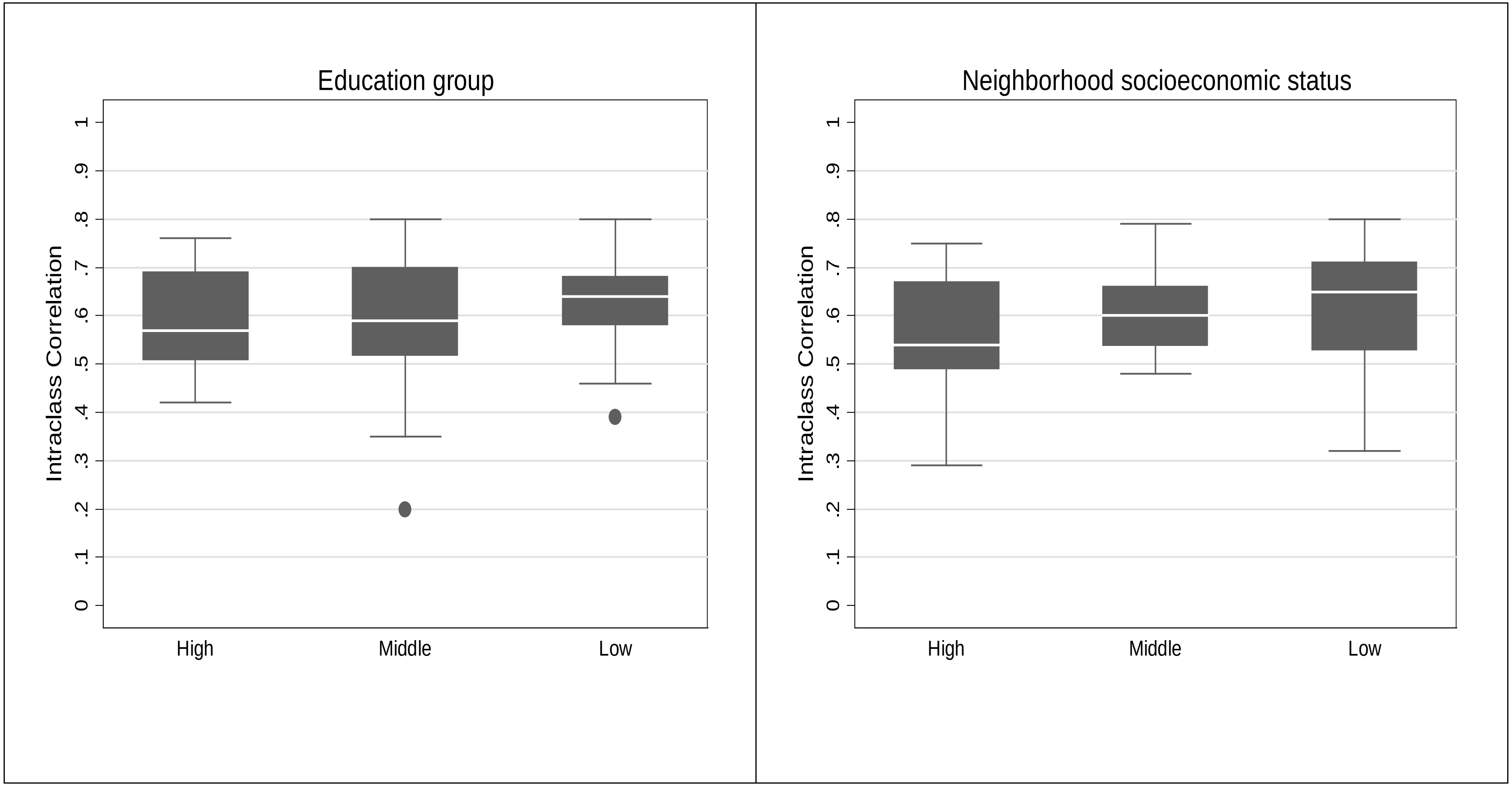
References

1. Sallis J, Cervero R, Ascher W, Henderson K, Kraft M, Kerr J. An ecological approach to creating active living communities. *Annu Rev Public Health.* 2006;27:297-322.
2. Trost SG, Owen N, Bauman AE, Sallis JF, Brown WJ. Correlates of adult's participation in physical activity: review and update. *Med Sci Sports Exerc.* 2002;34(12):1996-2001.
3. Sallis JF, Kraft K, Linton LS. How the environment shapes physical activity: a transdisciplinary research agenda. *Am J Prev Med.* 2002;22(3):208.
4. Fuzhong L, Fisher KJ, Bauman A, Ory MG, Chodzko-Zajko W, Harmer P, et al. Neighbourhood Influences on Physical Activity in Middle-Aged and Older Adults: A Multilevel Perspective. *J Aging Phys Act.* 2005;13:87-114.
5. Giles-Corti B, Timperio A, Bull F, Pikora T. Understanding physical activity environmental correlates: increased specificity for ecological models. *Exerc Sport Sci Rev.* 2005;33(4):175-181.
6. Pikora T, Giles-Corti B, Bull F, Jamrozik K, Donovan R. Developing a framework for assessment of the environmental determinants of walking and cycling. *Soc Sci Med.* 2003;56(8):1693-1703.
7. Ramirez L, Hoehner C, Brownson R, Cook R, Orleans T, Hollander M, et al. Indicators of activity-friendly communities - An evidence-based consensus process. *Am J Prev Med.* 2007;31(6):515-524.
8. 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(4S):S99-S123.
9. Saelens BE, Sallis J, Frank LD. Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. *Ann Behav Med*. 2003;25(2):80-91.
10. Humpel N, Marshall AL, Leslie E. Changes in neighborhood walking are related to changes in perceptions of environmental attributes. *Ann Behav Med.* 2004;27(1):60-7.
11. Owen N, Humpel N, Leslie E, Bauman A, Sallis JF. Understanding environmental influences on walking: review and research agenda. *Am J Prev Med.* 2004;27(1):67-76.
12. Duncan MJ, Spence JC, Mummery WK. Perceived environment and physical activity: a meta-analysis of selected environmental characteristics. *Int J Behav Nutr Phys Act.* 2005;2:11.
13. Oakes JM, Masse LC, Messer LC. Work group III: Methodological issues in research on the food and physical activity environments: addressing data complexity *Am J Prev Med.* 2009;36(4S):S177-S181.

14. Lin SS, Glaser SL, Stewart SL. Reliability of self-reported reproductive factors and childhood social class indicators in a case-control study in women. *Ann Epidemiol.* 2002;12:242-247.
15. Sallis JF, Johnson MF, Calfas KJ, Caparosa S, Nichols JF. Assessing perceived physical environmental variables that may influence physical activity. *Res Q Exerc Sport.* 1997;68(4):345-51.
16. Evenson KR, Eyler AA, Wilcox S, Thompson JL, Burke JE. Test-retest reliability of a questionnaire on physical activity and its correlates among women from diverse racial and ethnic groups. *Am J Prev Med.* 2003;25(3Si):15-22.
17. Evenson KR, McGinn AP. Test-retest reliability of a questionnaire to assess physical environmental factors pertaining to physical activity. *Int J Behav Nutr Phys Act* .2005;2(7).
18. Brownson RC, Eyler AA, King AC, Shyu Y, Brown DR, Homan SM. Reliability of information on physical activity and other chronic disease risk factors among US women aged 40 years or older. *Am J Epidemiol.* 1999;149(4):379-391.
19. Oyeyemi AL, Adegoke BOA, Oyeyemi AY, Fatudimu BM. Test-retest reliability of IPAQ environmental-module in an African population. *Int J Behav Nutr Phys Act.* 2008;5(38).
20. Brownson RC, Chang JJ, Eyler AA, Ainsworth BE. Measuring the environment for friendliness toward physical activity: a comparison of the reliability of 3 questionnaires. *Am J Public Health.* 2004;9:473-483.
21. Cerin E, Saelens BE, Sallis JF, Frank LD. Neighborhood Environment Walkability Scale: validity and development of a short form. *Med Sci Sports Exerc.* 2006;38(9):1682-91.
22. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
23. Giles-Corti B, Donovan RJ. Socioeconomic status differences in recreational physical activity levels and real and perceived access to a supportive physical environment. *Prev Med.* 2002;35(6):601-611.
24. van Lenthe FJ, Brug J, Mackenbach JP. Neighbourhood inequalities in physical inactivity: the role of neighbourhood attractiveness, proximity to local facilities and safety in the Netherlands. *Soc Sci Med.* 2005;60(4):763-775.
25. Kavanagh AM, Goller JL, King T, Jolley D, Crawford D, Turrell G. Urban area disadvantage and physical activity: a multilevel study in Melbourne, Australia. *J Epidemiol Community Health.* 2005;59(11):934-940.
26. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G. Indicators of socioeconomic position (part 1). *J Epidemiol Community Health.* 2006;60(1):7-12.

27. Berkman LF, Macintyre S. The measurement of social class in health studies: old measures and new formulations. *IARC Scientific Publications.* 1997;138:51-64.
28. Turrell G, Najman JM. Collecting food-related data from low socioeconomic groups: How adequate are our current research designs? *Aust J Public Health.* 1995;19(4):410-416.
29. Turrell G. Income non-reporting: implications for health inequalities research. *J Epidemiol Community Health.* 2000;54:207-214.
30. Burton NW, Turrell G, Oldenburg B. Item non-response in a population-based study of physical activity. *J Phys Act Health.* 2004;1(4):345-61.
31. 31. Burton NW, Haynes M, Wilson LA, Giles-Corti B, Oldenburg BF, Brown WJ, et
32. al. HABITAT: A longitudinal multilevel study of physical activity change in midaged
33. adults. *BMC Public Health.* 2009;9(1):76.
34. 32. Australian Bureau of Statistics. Socio-economic Indexes for Areas, Australia, 2001. Canberra; 2003. Report No.: Cat. No. 2039.0.
35. Dillman DA. Mail and Internet Surveys: the Tailored Design Method Chichester: Wiley; 2000.
36. Evenson KR, McGinn AP. Test-retest reliability of adult surveillance measures for physical activity and inactivity. *Am J Prev Med.* 2005;28(5):470-478.
37. Ogilvie D, Mitchell R, Mutrie N, Petticrew M, Platt S. Perceived characteristics of the environment associated with active travel: development and testing of a new scale. *Int J Behav Nutr Phys Act.* 2008;5:32.
38. Mooney CZ, Duval RD. Bootstrapping: a non-parametric approach to statistical inference Newbury Park, CA; 1993.

Figure 1: Overall magnitude of intra-class correlations by education group [a] and neighborhood socioeconomic status [b]



a. p-value for difference in median ICC between low vs high education (p = 0.169), low vs middle (p = 0.248) and middle vs high education (p= 0.578)
b. p-value for difference in median ICC between low vs high neighborhood SES (p = 0.169), low vs middle (p = 0.405) and middle vs high neighborhood SES (p = 0.108)

Table 1: Socioeconomic profile of the three neighborhoods in the HABITAT test-retest study [a]

	Neighborhood SES		
	Low	Middle	High
Single parent families (%)	5.7	6.3	1.6
Low income families (%)[b]	17.8	7.3	8.9
Unemployment rate (%)	10.4	4.4	3.6
Early school leavers (%)[c]	44.0	30.9	21.1

a. Source of data: Australian Bureau of Statistics Census Data (CDATA) 2001.
b. Families with income less than $21,000pa ($400 per week)
c. Proportion of the population aged 15 years and over who left school at age 15 years or less, or did not go to school.

Table 2: Estimation of the Intra-class correlation (ICC) coefficient under different modelling scenarios

	Model 1	Model 2	Model 3	Model 4
Item: There is lots of greenery in my suburb	One-way ANOVA	Two-level mixed model	Three level mixed model	Two level mixed model, with neighborhood included as a fixed effect
	Mean square	Variance	Variance	Variance
Between neighborhoods	--	--	0.087	--
Between individuals	0.868	0.321	0.235	0.230
Within individuals	0.222	0.222	0.222	0.222
Intra-class correlation	0.59	0.59	0.51	0.51

Table 3: Socio-demographic characteristics of the test-retest sample and respondents who didn't return the retest survey

	Test-retest sample (n=219) [a]		Non-return of retest survey (n=52) [b]		
Sex	n	%	n	%	p-value [c]
Male	95	43.4	25	48.1	0.540
Female	124	56.6	27	51.9	
Education					
High	78	35.6	11	21.2	0.004
Medium	62	28.3	9	17.3	
Low	79	36.1	32	61.5	
Neighborhood SES					
High	65	29.7	14	26.9	0.885
Medium	73	33.3	17	32.7	
Low	81	37.0	21	40.4	
Country of birth					
Australia	164	75.2	31	59.6	
Other English speaking country	20	9.2	7	13.5	
Non-English speaking country	34	15.6	14	26.9	0.074
Self-rated health					
Excellent/very good/good	173	80.1	42	80.8	
Fair/poor	43	19.9	10	19.2	0.912
How often during last six months:					
Physical activity with others in the park					
Never	91	45.5	15	28.9	
Rarely	60	28.0	18	34.6	
Once a month or more	63	29.4	19	36.5	0.195
Running or jogging					
Never	127	58.8	23	44.2	
Rarely	48	22.2	15	28.9	
Once a month or more	41	19.0	14	26.9	0.161
Cycling					
Never	153	71.2	35	67.3	
Rarely	36	16.7	5	9.6	
Once a month or more	26	12.1	12	23.1	0.081
Age (mean/sd)	52.2/7.0		50.4/6.2		0.098 [d]

a. Characteristics of respondents who returned both surveys (test and retest). For a number of the socio-demographic measures n≠219 due to missing data
b. Characteristics of respondents who did not return the retest survey.
c. p-value for chi-square test
d. p-value for two-group t-test

Table 4: Intra-class correlations (ICC) for neighborhood perceptions in the HABITAT pilot study (2006)

	N [a]	One-way ANOVA	Two-level mixed-model	Three-level mixed-model	Two-level mixed-model, with neighborhood as a fixed effect (ICC, 95% CI)
Suburb surroundings					
There is lots of greenery around my suburb (trees, bushes, household gardens)	216	0.59	0.59	0.51	0.51 (0.38, 0.63)
There are many interesting things to look at in my suburb	210	0.71	0.71	0.61	0.61 (0.54, 0.68)
There is tree cover along many of the footpaths in my suburb	215	0.58	0.58	0.52	0.52 (0.44, 0.60)
My suburb is generally free from litter or rubbish	213	0.70	0.70	0.65	0.64 (0.57, 0.72)
There are attractive buildings and homes in my suburb	215	0.73	0.73	0.68	0.68 (0.61, 0.65)
There are pleasant natural features in my suburb	215	0.78	0.78	0.61	0.61 (0.53, 0.68)
My suburb is generally free from graffiti	206	0.70	0.70	0.64	0.64 (0.55, 0.72)
Streets and footpaths					
Many streets in my suburb have cul-de-sacs (dead-end streets)	200	0.65	0.65	0.63	0.63 (0.56, 0.70)
There are footpaths on most of the streets in my suburb	211	0.72	0.72	-- [b]	0.72 (0.66, 0.77)
There are many four-way intersections in my suburb	199	0.51	0.51	0.51	0.51 (0.43, 0.59)
Many streets in my suburb are hilly	214	0.61	0.61	0.52	0.52 (0.39, 0.64)
Most footpaths in my suburb are well lit at night	192	0.70	0.70	0.70	0.70 (0.64, 0.76)
Many roads and streets in my suburb have pedestrian crossings and traffic signals	211	0.64	0.64	0.61	0.61 (0.54, 0.67)
Most of the footpaths in my suburb are well maintained (flat and even, not broken or cracked)	200	0.68	0.68	-- [b]	0.68 (0.62, 0.74)
Traffic					
In my suburb, there is usually a lot of traffic on the local streets	215	0.67	0.67	0.65	0.64 (0.58, 0.71)
The speed of traffic on most nearby streets is usually slow (50kph or less)	208	0.63	0.63	0.62	0.61 (0.54, 0.69)
There are many traffic slowing devices in my suburb such as speed humps, roundabouts	213	0.54	0.54	0.53	0.52 (0.44, 0.61)
I live on or near a main road or busy throughway for motor vehicles	216	0.53	0.53	0.52	0.52 (0.43, 0.60)
In my suburb there are a lot of exhaust fumes from motor vehicles	215	0.60	0.60	0.60	0.59 (0.51, 0.67)
Crime and safety					
There is a lot of crime in my suburb	192	0.76	0.76	0.72	0.72 (0.65, 0.78)
There are unsecured dogs in my suburb	200	0.69	0.69	0.67	0.66 (0.59, 0.73)
Children are safe walking around the suburb during the day	198	0.63	0.63	0.63	0.62 (0.51, 0.74)
Streets in my suburb are well lit at night	201	0.72	0.72	-- [b]	0.72 (0.66, 0.78)
The level of crime in my suburb makes it unsafe to on the streets at night	191	0.77	0.77	0.73	0.73 (0.67, 0.78)
There are rowdy youth on the streets or hanging around in parks in my suburb	199	0.75	0.75	0.71	0.70 (0.65, 0.76)
The level of crime in my suburb makes it unsafe to walk on the streets during the day	206	0.48	0.48	0.42	0.41 (0.26, 0.56)
In my suburb, I would feel safe walking home from a bus stop or train station at night	202	0.75	0.75	0.74	0.74 (0.67, 0.81)

a. There was a variable amount of missing or unusable data for the 27 neighborhood perception items, hence the number of cases available for analyses ranged from 191-216

b. ICC could not be estimated as model did not converge

Table 5: Intra-class correlations (ICC, 95% CI) by education-group and neighborhood socioeconomic status (SES)

	Education [a]			Neighborhood SES [b]		
	High	Middle	Low	High	Middle	Low
Suburb surroundings						
Lots of greenery	0.48 (0.31, 0.65)	0.49 (0.16, 0.83)	0.46 (0.27, 0.65)	0.29 (0.10, 0.48)	**0.54 (0.38, 0.69)***	0.47 (0.21, 0.72)
Many interesting things	0.54 (0.39, 0.68)	0.58 (0.45, 0.72)	0.65 (0.54, 0.76)	0.51(0.37, 0.65)	**0.72 (0.63, 0.81)***	**0.50 (0.36, 0.65)****
Tree cover	0.49 (0.35, 0.64)	0.52 (0.37, 0.67)	0.49 (0.34, 0.63)	0.40 (0.22, 0.59)	0.51 (0.36, 0.66)	0.53 (0.42, 0.65)
Free from litter or rubbish	0.57 (0.39, 0.75)	0.71 (0.60,0.81)	0.58 (0.43, 0.73)	0.29 (0.12, 0.47)	**0.70 (0.56, 0.85)***	0.67 (0.56, 0.78)†
Attractive buildings	0.73 (0.63, 0.82)	0.63 (0.48, 0.78)	0.66 (0.52, 0.79)	0.62 (0.44, 0.80)	0.66 (0.53, 0.79)	0.71 (0.60, 0.82)
Natural features	0.59 (0.45, 0.74)	0.52 (0.35, 0.69)	0.64 (0.53, 0.75)	0.52 (0.34, 0.69)	0.50 (0.34, 0.66)	0.67 (0.56, 0.78)
Free from graffiti	0.73 (0.63, 0.84)	0.60 (0.40, 0.79)	0.55 (0.37, 0.73)	0.63 (0.48, 0.78)	0.56 (0.36, 0.75)	0.65 (0.52, 0.78)
Streets and footpaths						
Many cul-de-sacs	0.57 (0.44, 0.70)	0.66 (0.52, 0.81)	0.61 (0.50, 0.73)	0.65 (0.53, 0.76)	0.60 (0.48, 0.73)	0.60 (0.46, 0.73)
Footpaths on most streets	0.71 (0.61, 0.80)	0.80 (0.71, 0.88)	**0.64 (0.50, 0.77)****	0.74 (0.65, 0.84)	0.64 (0.54, 0.79)	0.72 (0.62, 0.83)
Four-way intersections	0.50 (0.38, 0.62)	0.46 (0.29, 0.62)	0.49 (0.33, 0.65)	0.41 (0.26, 0.57)	0.58 (0.45, 0.71)	0.45 (0.30, 0.61)
Streets hilly	0.44 (0.16, 0.72)	0.35 (0.11, 0.60)	0.63 (0.48, 0.78)	0.68 (0.50, 0.86)	0.49 (0.28, 0.71)	0.44 (0.24, 0.64)
Footpaths well lit	0.76 (0.68, 0.84)	0.70 (0.57, 0.82)	0.63 (0.50, 0.75)	0.75 (0.64, 0.87)	**0.58 (0.44, 0.71)***	**0.74 (0.65, 0.83)****
Crossings/traffic signals	0.52 (0.39, 0.65)	0.56 (0.42, 0.71)	0.66 (0.55, 0.77)	0.56 (0.42, 0.69)	0.61 (0.50, 0.72)	0.60 (0.49, 0.71)
Footpaths well maintained	0.69 (0.61, 0.78)	0.56 (0.43, 0.69)	**0.73 (0.63, 0.84)****	0.73 (0.64, 0.82)	0.62 (0.49, 0.75)	0.67 (0.57, 0.77)
Traffic						
Lot of traffic	0.55 (0.41, 0.70)	0.65 (0.53, 0.77)	0.68 (0.58, 0.78)	0.51 (0.36, 0.65)	0.65 (0.50, 0.81)	**0.71 (0.61, 0.80)***
Speed of traffic slow	0.42 (0.26, 0.58)	0.59 (0.44, 0.73)	**0.67 (0.56, 0.79)***	0.49 (0.30, 0.67)	0.54 (0.38, 0.70)	0.66 (0.56, 0.77)
Traffic slowing devices	0.61 (0.49, 0.73)	0.46 (0.26, 0.66)	0.47 (0.31, 0.62)	0.42 (0.23, 0.62)	0.51 (0.36, 0.66)	0.57 (0.42, 0.72)
Main road/throughway	0.51 (0.40, 0.63)	0.40 (0.21, 0.58)	0.59 (0.45, 0.74)	0.53 (0.37, 0.69)	0.60 (0.47, 0.72)	0.42 (0.27, 0.56)
Lot of exhaust fumes	0.55 (0.43, 0.67)	0.52 (0.34, 0.71)	0.63 (0.48, 0.78)	0.54 (0.34, 0.74)	0.54 (0.42, 0.67)	0.64 (0.52, 0.76)
Crime and safety						
Lot of crime	0.60 (0.45, 0.74)	0.75 (0.65, 0.84)	0.72 (0.61, 0.84)	0.56 (0.41, 0.72)	0.62 (0.47, 0.76)	**0.80 (0.72, 0.88)*, ****
Unsecured dogs	0.62 (0.47, 0.77)	0.74 (0.63, 0.85)	0.64 (0.51, 0.76)	0.51 (0.32, 0.69)	0.72 (0.60, 0.83)	0.72 (0.61, 0.82)
Children are safe	0.52 (0.35, 0.69)	0.52 (0.33, 0.72)	0.71 (0.50, 0.91)	0.53 (0.29, 0.76)	0.63 (0.46, 0.81)	0.63 (0.43, 0.83)
Streets well lit	0.75 (0.67, 0.83)	0.74 (0.65, 0.84)	0.64 (0.50, 0.78)	0.71 (0.60, 0.82)	0.71 (0.61, 0.81)	0.71 (0.60, 0.82)
Unsafe streets at night	0.73 (0.63, 0.83)	0.60 (0.42, 0.77)	0.75 (0.67, 0.84)	0.71 (0.59, 0.82)	0.79 (0.71, 0.87)	0.66 (0.55, 0.77)
Rowdy youth	0.47 (0.29, 0.65)	**0.72 (0.61, 0.82)***	**0.78 (0.72, 0.85)***	0.62 (0.48, 0.76)	0.52 (0.36, 0.67)	**0.79 (0.72, 0.86)**)*, **
Unsafe streets during day	0.57 (0.35, 0.79)	**0.20 (-0.08, 0.48)***	0.39 (0.15, 0.63)	0.47 (0.32, 0.61)	0.48 (0.18, 0.78)	0.32 (0.08, 0.55)
Safe walking home at night	0.66 (0.50, 0.82)	0.70 (0.57, 0.82)	0.80 (0.72, 0.88)	0.67 (0.54, 0.81)	0.71 (0.54, 0.87)	0.77 (0.69, 0.86)

a. ICC estimated using a mixed-effects model, with adjustment for neighborhood SES and age

b. ICC estimated using a mixed-effects model, with adjustment for age and education

* Significantly different from high educated/ high neighborhood SES at p≤0.05; ** Significantly different from mid educated/mid neighborhood SES at p≤0.05