Assessing the validity of facilitated-volunteered geographic information: comparisons of expert and novice ratings

Kelly Kalvelage, Michael C. Dorneich, Christopher J. Seeger, Gregory J. Welk, Stephen Gilbert, Jon Moon, Imad Jafir & Phyllis Brown

GeoJournal
Spatially Integrated Social Sciences and Humanities

ISSN 0343-2521

GeoJournal
DOI 10.1007/s10708-017-9781-z

Springer

GeoJournal
DOI 10.1007/s10708-017-9781-z

Assessing the validity of facilitated-volunteered geographic information: comparisons of expert and novice ratings

Kelly Kalvelage · Michael C. Dorneich · Christopher J. Seeger · Gregory J. Welk · Stephen Gilbert · Jon Moon · Imad Jafir · Phyllis Brown

Abstract Facilitated-voluntary geographic information (f-VGI) is a promising method to enable systematic collection of data from residents about their physical and social environment. The method capitalizes on ubiquitous mobile smartphones to empower collection of geospatially-referenced data. It is important to evaluate the validity of user-generated content for use in research or program planning. The purpose of this study was to test whether the aggregated environmental ("bikeability") ratings from novice community residents converges with ratings from experts using a robust research-based, paper audit-tool (the established Pedestrian Environment Data Scan (PEDS) tool). Equivalence testing statistically showed overall agreement between the composite ratings of bikeability within the novice group. Agreement in categorical ratings between novices and experts were examined using the summary agreement index, which showed substantial agreement across the 10 locations rated by 11 novices using an f-VGI mobile application and four experts using PEDS; variability depended on the nature of the specific questions asked. Results reveal overall substantial agreement between novice and expert ratings for both composite scores and individual categorical ratings. However, additional research is needed to refine the methodology for use in formalized research applications.

Keywords Facilitated-volunteered geographic information · Built environment · Bikeability · User-generated content · Validity

K. Kalvelage · M. C. Dorneich · S. Gilbert
Human Computer Interaction, Iowa State University, Ames, IA, USA

M. C. Dorneich (✉) · S. Gilbert
Department of Industrial and Manufacturing Systems Engineering, Iowa State University, 3004 Black Engineering Building, Ames, IA 50011-2164, USA
e-mail: dorneich@iastate.edu

C. J. Seeger
Department of Landscape Architecture, Iowa State University, Ames, IA, USA

G. J. Welk
Department of Kinesiology, Iowa State University, Ames, IA, USA

J. Moon · I. Jafir · P. Brown
MEI Research, Ltd, Edina, MN, USA

Introduction

Systematic evaluations of physical and social environments are critical for advancing public health research (Sallis et al. 1998; Humpel et al. 2002). Chronic diseases that plague society are strongly influenced by lifestyle behaviors, and these, in turn, have been directly attributed to the environments in which we live. For example, the high prevalence of obesity, one of the most significant public health

Published online: 26 May 2017

problems facing the U.S (Flegal et al. 2010; Ogden et al. 2010), has been widely attributed to an obesogenic environment that favors too much eating and too little physical activity (Ewing et al. 2006; Feng et al. 2010; Lee et al. 2011). Research has specifically demonstrated that environmental factors explain underlying disparities in the population, including variability in physical activity behaviors (Berrigan et al. 2006; Frank et al. 2008), dietary behaviors (Zenk et al. 2009) and the prevalence of obesity (Frank et al. 2008; Gordon-Larsen et al. 2006; Heinrich et al. 2008). These insights have been made possible by reliable systematic evaluations of the built environment, which link assessments about aspects of the environment to their geospatial location (Boarnet et al. 2006). A number of powerful geospatial tools have been developed to evaluate local food (Lytle 2009; McKinnon et al. 2009; Sweeney et al. 2015) and physical activity environments (Brownson et al. 2009), but a major limitation is that the detailed onsite documentation of the environmental characteristics is both time and resource intensive (Keast et al. 2010; Story et al. 2009).

Recognizing the need for valid, reliable, and efficient methods to collect data about the built environment, researchers have looked to public data collection as a viable alternative to expert audits (Chatzimilioudis et al. 2012; Fritz et al. 2011; Heipke 2010). Public data collection based on user-generated content methods using citizens offers promise for replacing formal data collection strategies that rely on trained professionals to collect data on environments (Brabham 2008). The idea is that a large number of untrained citizens may converge to the same answer or provide insights on a problem that an individual or an expert may be unable to solve (Surowiecki 2005).

The collection of user-generated geospatially-referenced data has been termed volunteered geographic information (VGI), and it has specifically demonstrated utility for compiling geospatially-referenced data about the built environment (Goodchild 2007). The impacts of these methods have been dramatically enhanced by the ubiquitous use of mobile smartphones and broad acceptance of social media and engaging mobile applications. For example, WikiMapia combines geographically-referenced user-generated content to map and describe all of the geographical objects in the world (Koriakine and Saveliev 2016).

User-generated content and VGI have been used successfully as a less expensive means of acquiring detailed geographic information, and can be deployed on short notice (Goodchild and Li 2012). However, the type of information volunteered by the public is inherently random and at their discretion. It is common for an urban planner, designer, or researcher to need a specific form of geospatial data. To advance these goals, it is necessary to guide or facilitate participants' collection of more specific and useable types of data. Using "facilitated" volunteered geographic information (f-VGI) enables the collection of geospatial data that are guided and prompted more directly by researchers to address specific objectives (Seeger 2008). For example, a customized audit instrument could be coupled with VGI to facilitate data collection and evaluation of consumer perceptions about a new recreation path or the safety of a specific park or area within a park. Participants could either be guided to go to a specific location to complete coding or could receive prompts to provide feedback when they enter a certain location. While the feasibility of VGI applications is well documented, it is not clear whether public data collection carried out by untrained users, or novices, is sufficiently valid to be useful for research applications. Therefore, the purpose of this paper was to define and determine the validity of f-VGI collected data to evaluate aspects of the built environment.

This study examined the utility of a customizable mobile application designed to support a variety of f-VGI applications for evaluating the environment. The assumptions of f-VGI described above were specifically evaluated using the widely discussed environmental need to evaluate the "bikeability" of communities as a way to promote active transportation and recreation (Clifton et al. 2007). Over the last fifteen years, a number of data collection instruments have been developed to assess associations between the built environment and walking and biking behaviors (Keast et al. 2010; Pikora et al. 2002), but one of the most established tools is the Pedestrian Environment Data Scan (PEDS) tool. This tool captures standards related to bikeability and has proven to be a very reliable method for data collection by trained raters, or experts (Emery et al. 2003; Moudon and Lee 2003; Turner et al. 1997).

The associations between consumer ratings using a simplified audit tool and experts using the PEDS tool provides a good framework for evaluating the utility of f-VGI applications for research on the built environment.

Related work

Volunteered geographic information (VGI)

Goodchild (2007) defined the term Volunteered Geographic Information to describe a phenomenon taking place where untrained citizens, through the use of new mobile and spatial technologies, could create data sets of information. These data sets could be a result of their own experiences, preferences, observations, or general knowledge (Seeger 2017). While user-generated content had been broadly implemented in various forms over the years, the unique component of this form of citizen participation was the incorporation of detailed geospatial information (Sui et al. 2012). Elwood et al. (2012) describe the opportunity for a new form of 'geographic information' to be created by citizen scientists. Critics have voiced concern over the validity of this so called "crowd-sourced" data and have worked to identify the mechanisms to limit or overcome user-generated inaccuracies (Goodchild and Li 2012; Flanagin and Metzger 2008).

The benefits of the citizen science movement have been well documented (Dickinson and Bonney 2012). As a whole, VGI provides a platform for which to conduct this work. Examples of this have included the collection of air pollution, hydrologic measurements, and sidewalk and bike routes (Fienen and Lowry 2012; Stevens and D'Hondt 2010; Schlossberg et al. 2007).

Audit tools

Environmental audits are used by a wide variety of professions, including the design and planning professions, to identify where correctable problems might occur. A commonly used and recognizable audit is a home energy audit that can identify where a house is losing energy and money. In the planning and landscape architecture profession, audit tools occur in the form of post-occupancy evaluations that identify if citizens use and participate within a site as the designer expected (Bordass and Leaman 2005). Planners, transportation officials, environmental researchers, and health professionals interested in understanding the relationships between physical inactivity and health related issues will often use any of a variety of audits (Schlossberg et al. 2015; Schlossberg 2006). These audits are often designed to assess the infrastructure supporting biking and walking networks for the purpose of either identifying the "relationship between urban form and pedestrian mobility" (Schlossberg 2006) or to identify locations where improvements to the network could be made.

One of the best-known audit tools, called SPACES, was developed over fifteen years ago and was used to inventory the characteristics of segments of a roadway (Schlossberg et al. 2007; Pikora et al. 2002). Since that time a variety of audit instruments have been developed. Another common audit tool is the Pedestrian Environment Data Scan (PEDS) (Clifton et al. 2007). This audit tool was used as the basis for this project and is discussed further in a later section in the paper. The incorporation of GPS enabled technology and smartphones has allowed audit tools to have increased spatial accuracy that can then be used to better visualize the information collected from the audit using GIS (Schlossberg et al. 2007).

Evaluation of reliability and validity

Data collected on the built environment needs to be valid and reliable (Saelens et al. 2003). Data validity is determined by comparing data provided from a group of raters with a "gold standard" (e.g. U.S. Census) for agreement. If a gold standard is not available, data provided by experts can serve as a proxy. Raters are considered experts if they have been trained or have extensive experience in the related field (Emery et al. 2003). Agreement between groups of raters can be evaluated via an agreement index (Ƙ) tailored to group comparisons (Vanbelle and Albert 2009). In one study to assess the validify of a two new survey instruments (Emery et al. 2003), a gold standard had not been previously established so a small group of expert raters served as the proxy. The experts document walkability and bikeability at several locations. These data were compared to data collected by two data collectors using the suitability

assessment instruments for the same locations. Data are considered valid if the individual raters are in agreement with the experts, or there is no presence of incompatibility between them (Emery et al. 2003).

Data reliability focuses on data consistency, dependability, and repeatability within a group (Cunningham et al. 2005). Data reliability does not use a gold standard for comparison. A study conducted by Comber et al. (2013) examined VGI data reliability using land cover data. A group of raters rated several locations independently. Data are considered reliable if all ratings for a location were in agreement (Comber et al. 2013; Foody and Boyd 2012). Reliability agreement has traditionally been measured using percentage agreement, kappa (Fleiss 1971; Cohen 1960), alpha (Krippendorff 1970), and intraclass correlation coefficient (ICC; Fleiss and Cohen 1973; Kraemer 1981). These agreement ratings are designed to examine group heterogeneity by providing a chance-corrected measure of agreement between raters. These agreement ratings examine agreement between two independent raters, between several raters, or between an isolated rater and a group of raters (Boarnet et al. 2006).

Methods

The study adopts a measurement agreement perspective to examine the agreement within novice ratings, and between novice and expert ratings of the built environment to test the utility of f-VGI applications. While agreement between individuals is important, the main focus is on the overall agreement between the aggregated reports of the novices and the experts (i.e. criterion validity) and not in the agreement between individuals themselves (i.e. inter-rater reliability). Reliability is a pre-requisite of validity so it is important to understand the inter-rater reliability of the responses from the various individuals for rating different segments or settings. The composite "user-generated content" value represents the group rating and this is compared to the expert raters to evaluate criterion validity. Thus, the specific objective of the study is to evaluate if the crowd-sourced f-VGI approach provides an equivalent indicator provided by experts trained to collect data on the built environment. Agreement was evaluated for both the overall bikeability score as well as for individual items.

Participants

A total of 44 participants (28 male, 16 female) volunteered to participate in this study—40 novice raters and four expert raters. All participants recruited had some regular biking experience, since the study required biking. Novice participants ranged in age from 18 to 72 years (M = 31.7), and biked 1.8 days per week (SD = 1.5) for an average of 55 min (SD = 33) per biking session. Expert participants were 23–43 years of age (M = 30), and biked 1.0 days (SD = 0.8) per week for an average of 70 min (SD = 42) per day. Novice participants had no experience assessing bikeability nor were they trained on such assessments. Expert participants had professional experience with geospatial assessments of bikeability.

Task and materials

The validated Pedestrian Environment Data Scan (PEDS) audit tool was used as the basis for the evaluation. The PEDS tool was designed to capture a range of built environmental aspects related to walking and bicycling (Clifton et al. 2007). PEDS was used in two forms—pen and paper, used by expert participants, and electronic, used by novice participants. The move from paper to electronic modality is assumed to have no effect on location ratings in and of itself, but rather it may reduce transcription errors in recording (Clifton et al. 2007).

The electronic version was deployed using a mobile application, based on the PiLR ecological momentary assessment (EMA) app. PiLR EMA is a cross platform (iOS and Android) app that allows researchers to create customized content (i.e. surveys) from a web interface and then to deploy the app with the custom content to participants without requiring any custom programming. Data is automatically uploaded to a server and access to the collected data is provided via the same web interface. The electronic survey used in this study was based on the PEDS tool. The electronic survey was modified to only include the relevant questions from the PEDS tool related to bikeability to reflect the type of information researchers would want to know and users would find valuable, and were ordered in a more logical approach. The questions were re-worded to a question–answer format while maintaining the

initial question content, number and wording of response options, and intent (see Fig. 1).

Both novice and expert collection groups were treated exactly the same except for the level and type of training they received. Novice participants were trained to use the mobile device. Specifically, they were trained on the interface to make sure that they could answer the survey questions, change their answers if they wished to do so, and navigate between questions. No training was given on how to decide on a rating for a location. Experts were specifically trained on how to use the PEDS tool using the accompanying PEDS protocol training manuals. The PEDS training manuals provided detailed procedures on how to assess the variations in the built environment for each question and guidance on how to select the appropriate response category. For example, to answer the question "this location is attractive for bicycling", the manual says "response … should answer the question 'would you want to bike this segment?' This includes finding the area aesthetically pleasing and existence of destinations" (Clifton et al. 2007). Novices answered this question solely on their personal definition of "attractive".

After training, novice and expert participants were given the primary task: bike to assigned locations and rate them. All ten locations were given to expert participants and three randomly assigned locations to novice participants. Thus, each location was rated by between 11 and 13 novice participants and four experts. Expert participants took the PEDS manual with them for reference when completing the primary task.

Ten bike path locations were selected along formally defined routes within a community. The locations were chosen based on a variety of path qualities, such as type (sidewalk, paved trail, footpath, and street), material (gravel, dirt, brick, concrete, and asphalt), and condition (good, fair, and poor). Four examples of locations are shown in Fig. 2.

At all locations, participants were asked the same five questions about the bikeability of the area (segment type, material type, condition, attractiveness and safety for biking). Seven of the ten locations included additional questions that were applicable to the location (i.e. is the path along a road, distance from the road, are there buffer zones and what type, and are there markings and what type), so the range of questions per location ranged from 5 to 13. Each question was associated with a different number of response options ranging from two options to five. Table 1 summarizes the question type, number of responses, and total number of questions for each of the 10 locations.

Data analyses

The experiment used a between-subjects design taking into account the heterogeneity within each group. The independent variable was Experience at two levels: novice and expert. The analyses were designed to evaluate agreement within the novice

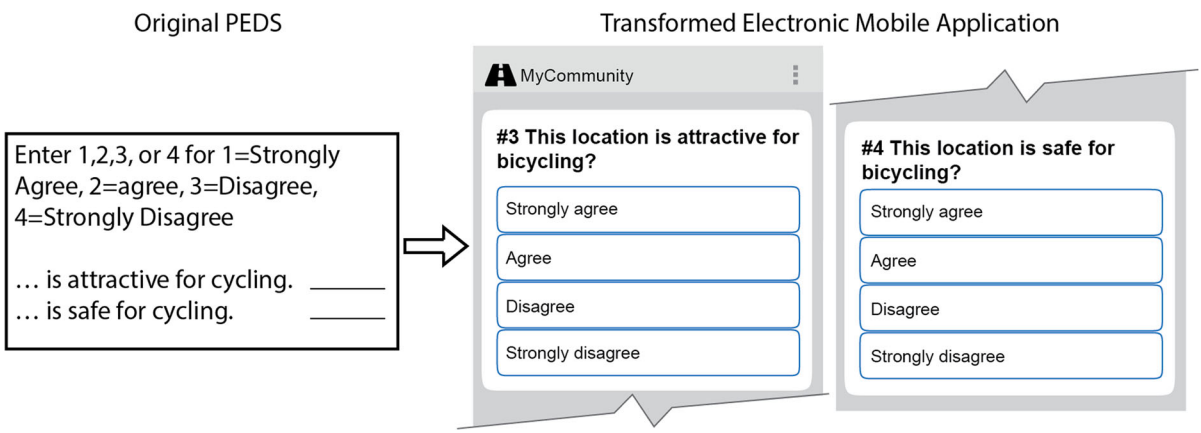

Fig. 1 Sample question transformed from pen and paper PEDS survey to mobile application survey

Fig. 2 Examples of four locations demonstrating a *gravel path* (location H), *concrete street* (location D), *brick trail* (location B), and *concrete sidewalk* (location E)

Table 1 Question type, number of responses, and total number of questions per location

Question type	# Response options	Number of questions by locations									
		A	B	C	D	E	F	G	H	I	J
Segment type?	5	1	1	1	1	1	1	1	1	1	1
Material type?	5	1	1	1	1	1	1	1	1	1	1
Condition of path?	3	1	1	1	1	1	1	1	1	1	1
Attractive for biking?	4	1	1	1	1	1	1	1	1	1	1
Safe for biking?	4	1	1	1	1	1	1	1	1	1	1
Along road?	2		1			1		1	1		1
Distance from road	3					1			1		
Buffer?	2								1		
Type of buffer?	2								5		
Marking?	2	1			1						
Type of marking?	2				4						
Total number of questions per location		6	6	5	10	7	5	6	13	5	6

group and between the novice and expert collection groups, as well as overall bikeability scores.

Evaluation of overall bikeability score

The composite bikeability scores were used to compare overall agreement between novice ratings and expert ratings. Bikeability scores were calculated for each location across all questions for that location. The response values were assigned using the standard PEDS tool values with the poorest response option (e.g. path is in poor condition, no buffer zone is present, the path is along a road) assigned a 0, and the best response option (e.g. path is in good condition, buffer zones are present, the path is not along a road) assigned a maximum value between 1 and 4 depending on the number of response options (e.g. 0–1 for No/Yes, or 0–3 for four options on a scale from "strongly agree" to "strongly disagree"). There is a maximum of 13 (depending on location) possible features to rate on scales that range from 0–1 to 0–4. Question responses were summed for each participant

and averaged to produce a novice collection group bikeability score and an expert collection group bikeability score for each location.

Equivalence testing was used to statistically examine the overall agreement between bikeability ratings from the novice group and the expert group. Standard statistical tests (e.g. ANOVA and t-tests) are focused on evaluating differences, but non-significant differences cannot be assumed to imply equivalence or agreement. With equivalence testing, it is possible to empirically determine whether the novice estimates are statistically "equivalent" to the ratings from the expert. Using ideas suggested by Robinson et al. (2005), the equivalence across multiple locations was tested by regressing the mean estimates from the novices against the mean reference values from the experts. This allowed us to evaluate overall equivalence across the set of 10 different locations rather than individual locations. Robinson et al. (2005) suggested regression-based equivalence regions of ±10% of the reference mean for the intercept and ±10% of the slope of 1 that would be expected for equal means on the two measures (i.e. 0.9–1.1 for a hypothesized slope of 1). The calculated 10% equivalence zone for the intercept was ±3.3 units, a zone capturing 22% of the overall 15 point response range (i.e. ±3.3/15 = 22%). Equivalence was established based on whether the entire 90% confidence interval for the regression slope was contained in this interval. Site-by-site equivalence tests were performed separately using equivalence zones defined based on the mean of expert ratings ±10%.

Evaluation of individual survey items

Because each location was rated by a different set of novice raters, a one-way random effects ANOVA model was used to calculate the ICC values for the individual items. ICC can be interpreted as follows: anything <0.2 indicates a poor agreement, 0.21–0.40 a fair agreement, 0.41–0.60 a moderate agreement, 0.61–0.80 a substantial agreement, and above 0.80 a perfect agreement (Shrout and Fleiss 1979).

The agreement index (Ƙ) (Vanbelle and Albert 2009) was used to examine the agreement in individual items between the novice collection group and the expert collection group. The agreement index defines two groups to be in perfect agreement (Ƙ = 1) if they have the same probability of classifying each location in the Ƙ-categorical scale. Agreement index is calculated for the items that use the same categorical scale. Therefore, the agreement index is calculated for each question across all locations. To normalize the number of ratings by novices for each location, the location with the fewest number of ratings was used. For locations that had more than eleven, ratings were randomly selected and dropped. Thus the analysis was performed with 11 novice ratings per location.

The agreement index is calculated using the observed proportion of agreement (P_O), the mean probability of agreement expected by chance (P_E), and the maximum proportion of agreement as shown in Eq. (1).

$$Ƙ = \frac{P_O - P_E}{P_M - P_E} \quad (1)$$

The agreement index is expressed from −1.0 to 1.0 where 1.0 represents perfect agreement, 0.0 represents the absence of agreement. Anything <0.0 represents agreement less than that expected by chance. While there is no consensus on what constitutes an acceptable agreement, the agreement index can be considered an extension of Cohen's kappa (Vanbelle and Albert 2009). Therefore, the Kappa scale is followed, which is interpreted in the same way as the ICC.

Results

Evaluation of overall bikeability score

The key outcome of interest in this evaluation of f-VGI is whether the bikeability ratings from novices agree with those of the experts. The average of the 118 estimates from the novices and the 40 experts for each of the 10 test locations are plotted in Fig. 3. The standard errors for individual ratings are displayed with whisker plots around the points for both the experts (x direction) and novices (y direction). The dark shaded 90% confidence interval (CI) around the regression line includes the line of identity (i.e. slope of 1, depicted as the heavy gray line), demonstrating that the observed regression line is not significantly different from this line. Because the shaded CI band of the regression line is also completely within the dotted-line equivalence band (i.e. ±3.3), the results

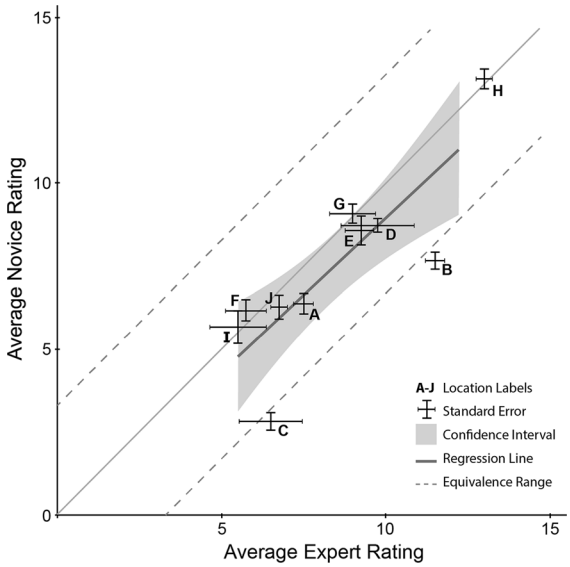

Fig. 3 Average bikeability ratings from novices and experts for each of 10 test locations. The *error bars* represent standard error. The shaded confidence interval band of the regression line is completely within the *dotted line* equivalence band (i.e. ±3.3 units), thus demonstrating statistical equivalence at this level

also demonstrate statistical equivalence at this level. These results demonstrate that there is 90% confidence that the novice raters have scores within 3.3 units (i.e. within 22%) of the experts. However, as is apparent in the figure, there is some variability in the degree of agreement across the individual locations.

Evaluation of agreement in individual items

Table 2 shows the ICC for the novice collection group (ICC_{Novice}) for each individual item as well as the overall agreement index that reflects agreement with the ratings from the experts. The overall $ICC_{Novice} = 0.71$, demonstrates substantial agreement but there was some variability across items. On a per question basis, ICC_{Novice} results show four items (material, along road, distance, markings) above 0.80 (perfect agreement), three items (segment, condition, safe) between 0.61 and 0.80 (substantial agreement), and three items (marking types, buffer zones and types, attractive) between 0.41 and 0.60 (moderate agreement).

Table 2 also shows the overall agreement index and the corresponding proportion of observed agreement (P_O), expected agreement (P_E), and maximum agreement (P_M) used to calculate the overall marginal probabilities of classifying the items. To calculate P_E for each question, an n by n table was generated using the marginal frequencies across experts and novices, where n equaled the number of response options for the question. Using those values, the frequency of agreement in which n measurements agree over the total number of ratings results in the expected agreement P_E. The average agreement index over the ten locations was $\kappa = 0.64$, demonstrating substantial agreement between the novice and expert collection groups. Three items (distance, marking types, and buffers and types) have an agreement index above 0.80 (perfect agreement), three items (material, condition, and markings) are between 0.61 and 0.80 (substantial agreement), two items (segment and along road) are between 0.41 and 0.60 (moderate agreement), and two items (attractive and safe) are below 0.40 (fair agreement).

Discussion

To advance research with user-generated f-VGI methods it is critical to evaluate the underlying assumption that a collection of individuals can provide assessments that would converge to provide ratings that would replicate or emulate perspectives from a trained expert using a validated evaluation tool. The results support the utility of the f-VGI coding since we observed statistical equivalence for the overall assessment of bikeability scores between the novices and experts. There are relatively large confidence intervals around the estimates but this may be due to the relatively small sample of novices as well as some inherent variability among the experts. The results based on the equivalence testing indicate that there is 90% confidence that scores from novices are within 3.3 units of the bikeability score of the experts. The band would likely be much tighter since the plot shows very good alignment and agreement for seven of the 10 locations that all fall on (or near) the line of identity. Agreement was not as high for three of the other items and this influenced the overall fit of the regression line and led to larger bands. It is not clear why there is variability across locations but it may be due to the fact that experts may have focused on different aspects of the environment than the novices due to their training in the field. The major difference in ratings for these

Table 2 Intraclass correlation coefficient results for each question was aggregated across all locations for the novice and expert collection groups

Question	ICC_{Novice}	Novice agreement assessment	P_O	P_E	P_M	κ	Expert/novice agreement assessment
Segment	0.63	Substantial	0.625	0.298	0.871	0.571	Moderate
Material	0.92	Perfect	0.780	0.272	0.963	0.735	Substantial
Condition	0.69	Substantial	0.634	0.364	0.778	0.652	Substantial
Along road	0.86	Perfect	0.755	0.482	1.000	0.526	Moderate
Distance	0.83	Perfect	0.909	0.477	1.000	0.826	Perfect
Markings	1.00	Perfect	0.875	0.500	1.000	0.750	Substantial
Marking types	0.47	Moderate	0.850	0.508	0.920	0.767	Substantial
Buffers and types	0.53	Moderate	0.964	0.540	1.000	0.922	Perfect
Attractive	0.56	Moderate	0.368	0.310	0.579	0.217	Fair
Safe	0.627	Substantial	0.445	0.303	0.679	0.378	Fair
Overall	0.71	Substantial				0.635	Substantial

locations appears to be in subjective aspects of the environment (safety and attractiveness). Therefore, it may not be a problem for cases in which more objective coding is used.

The f-VGI evaluation of bikeability at a given location involved a series or items. Therefore, it is also important to look at the variability in responses to the individual items since that captures inter-rater reliability. The computed ICC values show substantial agreement within the novice ratings but some variability across the individual items. This suggests that some items were be rated quite consistently while others revealed more individual variability in ratings. Considering each question for the novice collection group separately, results indicate 100% of the ICC_{Novice} are at least in moderate agreement, 70% were at least in substantial agreement, and 40% were in perfect agreement. The variability in agreement found here is specific for evaluations of bikeability but they also provide insights about other applications of f-VGI. The nature of the items and the type of wording would be important considerations in using f-VGI to promote more reliable responses and more stable means for crowd sourced estimates.

The inter-rater reliability is important to characterize but the main focus of the paper is on evaluating the overall agreement between the aggregated reports of the novices and the experts (i.e. criterion validity). Based on the overall agreement index we observed substantial overall agreement between the novice and expert group ratings of bikeability. Furthermore, on a per question basis, 80% of the questions were in at least moderate agreement, 60% were in at least substantial agreement, and 30% were in perfect agreement. The two items that are below 0.41 are the subjective questions asking if the path is safe and attractive for bicycling.

While there was not "perfect agreement," a relative comparison between the novice and expert collection groups showed moderate to perfect agreement at least 80% of the time. With larger samples of novice participants, the results would possibly continue to converge toward the more refined expert ratings. It is important to note that the PEDS tool itself shows an overall percentage agreement of 80% when comparing across experts for agreement (Clifton et al. 2007). Thus, there is always some natural variability in ratings even within and among expert participants using more refined and comprehensive tools. Further, the subjective questions do follow the same trend in the PEDS tool validation indicating a substantially lower percentage agreement than the other items. Pairing the agreement index results with both ICC results imply that novice community members are able to achieve the same agreement level within their ratings as an expert group using a validated survey.

Conclusion

The study demonstrated the utility of an innovative data collection method that capitalizes on the geospatial capabilities in smartphones to facilitate consumer/resident evaluation of their environments. The concept

is based on established VGI methods that enable participants to record geospatially-referenced experiences, observations or personal perspectives using mobile-web applications (Goodchild 2007). However, it extends these methods towards facilitated-VGI (f-VGI) applications (Seeger 2008) by mobilizing and empowering citizens to conduct and log assessments of their environment using specific criteria needed to assess specific research questions. A unique advantage of this approach is that it opens up new opportunities for citizens to become part of the change process by coding and prioritizing needs they see in society (Schlossberg et al. 2012; Buman et al. 2013). The much-heralded VGI pothole finder application worked because people had a vested interest in documenting the need for street repairs in their neighborhood (Berkowitz 2010; Kwarteng et al. 2014). Similarly, citizens could become engaged in coding aspects of the obesogenic environment if they knew that these efforts could promote or drive change. Projects like this, involving active transportation, are currently being implemented in communities across the country (Seeger et al. 2014; Schlossberg et al. 2012).

This study investigated the possibility of novice f-VGI being a valid substitute for expert data collected using traditional methods. The goal of any assessment of the built environment is to achieve a consensus rating. The development of valid data collection techniques is important to ensure that decisions are made with valid and agreed upon data. Leveraging the community to assess their own built environment allows researchers to gather data from people who may be invested in the outcome.While community members may be invested in improving their built environment, they typically are not experts in generating ratings to help assess the built environment. Typically it takes trained experts, using validated audit instrument, to assess an environment. However, this study has demonstrated that using a f-VGI mobile application could guide community members to providing ratings that resulted in a meaningful data set that is just as valid as one generated by an expert. This opens up the possibility that anyone, when properly guided, can provide valid data. The basis for guidance is what determines if an agreement can be defined or not. The key is not about the method of collection; it is about how the user is guided to provide the information.

Publicly collected data can improve researchers' ability to collect a wider scope of data by mobilizing and empowering citizens to conduct and log assessments of their environment. Incorporating f-VGI capability with these methods guides the users in collecting data on specific criteria necessary to assess specific research questions. The results of this study indicate a strong potential for f-VGI to be a valid substitute for traditional methods collecting data related to the built environment. These results add to the existing literature on data validity of user-generated content, but expand on defining validation as the agreement between two groups. Given the small sample size, further work needs to be done to see how big of a sample size is needed to ensure a reliability that matches expert assessment, and if it is possible to estimate the marginal value of gathering additional data (where continuing to gathering additional data beyond a certain point is of no actual use and a waste of time and resources).

Finally, to realize the power of f-VGI, an infrastructure needs to be built to allow researchers to quickly author surveys, push them to targeted users, collect the data, and process it for use with other data. In fact, it is in the merging and visualizing of primary and secondary data that researchers can start to answer research questions. Future work includes the development of a platform to combine and visualize f-VGI data with existing publically available data for analysis.

Acknowledgements Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health Under Award Number HHSN26120140 0034C. The authors would like to thank Dr. Phillip Dixon, a statistician in the Department of Statistics at Iowa State University, and Paul Hibbing, a Ph.D. student at the University of Tennessee for their contributions and assistance with conducting the equivalence testing analyses for the paper and for making the associated figure. The analyses provided a useful way to summarize the agreement between the novices and experts across multiple locations.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Berkowitz, B. (2010). *Using GPS to tag potholes*. http://en.seeclickfix.com/.

Berrigan, D., Troiano, R. P., McNeel, T., DiSogra, C., & Ballard-Barbash, R. (2006). Active transportation increases adherence to activity recommendations. *American Journal of Preventive Medicine, 31*(3), 210–216.

Boarnet, M. G., Day, K., Alfonzo, M., Forsyth, A., & Oakes, M. (2006). The Irvine–Minnesota inventory to measure built environments: Reliability tests. *American Journal of Preventive Medicine, 30*(2), 153–159. doi: 10.1016/j.amepre.2005.09.018.

Bordass, B., & Leaman, A. (2005). Phase 5: Occupancy - post-occupancy evaluation. In W. F. E. Preiser & J. Vischer (Eds.), *Assessing building performance* (pp. 72–78). Oxford: Elsevier.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: The International Journal of Research Into New Media Technologies, 14*(1), 75–90.

Brownson, R., Hoehner, C., Day, K., Forsyth, A., & Sallis, J. (2009). Measuring the built environment for physical activity: State of the science. *American Journal of Preventive Medicine, 36*(4 Suppl), S123–S599.

Buman, M. P., Winter, S. J., Sheats, J. L., Hekler, E. B., Otten, J. J., Grieco, L. A., et al. (2013). The Stanford Healthy Neighborhood Discovery Tool: A computerized tool to assess active living environments. *Amercian Journal of Preventative Medicine, 44*(4), e41–e47. doi: 10.1016/j.amepre.2012.11.028.

Chatzimilioudis, G., Konstantinidis, A., Laoudias, C., & Zeinalipour-Yazti, D. (2012). Crowdsourcing with smartphones. *IEEE Internet Computing, 16*(5), 36–44. doi: 10.1109/MIC.2012.70.

Clifton, K. J., Livi Smith, A. D., & Rodriguez, D. (2007). The development and testing of an audit for the pedestrian environment. *Landscape and Urban Planning, 80*(1–2), 95–110. doi:10.1016/j.landurbplan.2006.06.008.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1), 37–46.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., & Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation, 23*(1), 37–48. doi: 10.1016/j.jag.2012.11.002.

Cunningham, G. O., Michael, Y. L., Farquhar, S. A., & Lapidus, J. (2005). Developing a reliable senior walking environmental assessment tool. *American Journal of Preventive Medicine, 29*(3), 215–217. doi: 10.1016/j.amepre.2005.05.002.

Dickinson, J., & Bonney, R. (Eds.). (2012). *Citizen science: Public participation in environmental research*. Ithaca: Cornell University Press.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers, 102*(3), 571–590. doi:10.1080/00045608.2011.595657.

Emery, J., Crump, C., & Bors, P. (2003). Reliability and validity of two instruments designed to assess the walking and bicycling suitability of sidewalks and roads. *American Journal of Health Promotion, 18*(1), 38–46. doi: 10.4278/0890-1171-18.1.38.

Ewing, R., Brownson, R. C., & Berrigan, D. (2006). Relationship between urban sprawl and weight of United States youth. *American Journal of Preventive Medicine, 31* (6), 464–474.

Feng, J., Glass, T., Curriero, F., Stewart, W., & Schwartz, B. (2010). The built environment and obesity: A systematic review of the epidemiologic evidence. *Health Place, 16* (2), 175–190.

Fienen, M. N., & Lowry, C. S. (2012). Social. Water—A crowdsourcing tool for environmental data acquisition. *Computers & Geosciences, 49,* 164–169.

Flanagin, A., & Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3–4), 137–148. doi:10.1007/s10708-008-9188-y.

Flegal, K., Carroll, M., Ogden, C., & Curtin, L. (2010). Prevalence and trends in obesity among US adults, 1999–2008. *Journal of the American Medical Association, 303* (3), 235–241.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Fleiss, J., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*(3) 613–619.

Foody, G. M., & Boyd, D. S. (2012). Exploring the potential role of volunteer citizen sensors in land cover map accuracy assessment. In *10th international symposium on spatial accuracy assessment in natural resources and environmental science, Florianopolis, Brazil, Jul 10–13 2012* (pp. 203–208).

Frank, L., Kerr, J., Sallis, J., Miles, R., & Chapman, J. (2008). A hierarchy of sociodemographic and environmental correlates of walking and obesity. *Amercian Journal of Preventative Medicine, 47*(2), 172–178.

Fritz, S., See, L., McCallum, I., Schill, C., Perger, C., & Obersteiner, M. (2011). Building a crowd-sourcing tool for the validation of urban extent and gridded population. In B. Murgante, O. Gervasi, A. Iglesias, D. Taniar, B. O. Apduhan (Eds.), *Computational science and its applications-ICCSA 2011* (pp. 39–50). Springer.

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221. doi: 10.1007/s10708-007-9111-y.

Goodchild, M., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*(1), 110–120. doi:10.1016/j.spasta.2012.02.002.

Gordon-Larsen, P., Nelson, M., Page, P., & Popkin, B. (2006). Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics, 117* (2), 417–424.

Heinrich, K., Lee, R., Regan, G., Reese-Smith, J., Howard, H., Haddock, C., et al. (2008). How does the built environment relate to body mass index and obesity prevalence among public housing residents? *Amercian Journal of Health Promotion, 22*(3), 187–194.

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing, 65*(6), 550–557. doi:10.1016/j.isprsjprs.2010.06.005.

Humpel, N., Owen, N., & Leslie, E. (2002). Environmental factors associated with adults' participation in physical activity: A review. *American Journal of Preventive Medicine, 22*(3), 188–199. doi:10.1016/S0749-3797(01)00426-3.

Keast, E., Carlson, N., Chapman, N., & Michael, Y. (2010). Using built environmental observation tools: Comparing two methods of creating a measure of the built environment. *American Journal of Health Promotion, 24*(5), 354–361. doi:10.4278/ajhp.08063-QUAN-81.

Koriakine, A., & Saveliev, E. (2016). *Wikimapia*. http://www.wikimapia.org/.

Kraemer, H. C. (1981). Intergroup concordance: Definition and estimation. *Biometrika, 68*(3), 641–646. doi:10.1093/biomet/68.3.641.

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70.

Kwarteng, J., Schulz, A. J., Mentz, G., Zenk, S. N., & Opperman, A. (2014). Associations between observed neighborhood characteristics and physical activity: Findings from a multiethnic urban community. *Journal of Public Health, 36*(3), 358–367.

Lee, R. E., McAlexander, K., & Banda, J. (2011). *Reversing the obesogenic environment*. Champaign: Human Kinetics.

Lytle, L. A. (2009). Measuring the food environment: State of the science. *American Journal of Preventive Medicine, 36*(4), S134–S144.

McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A., & Yaroch, A. L. (2009). Measures of the food environment: A compilation of the literature, 1990–2007. *American Journal of Preventive Medicine, 36*(4), S124–S133.

Moudon, A. V., & Lee, C. (2003). Walking and bicycling: An evaluation of environmental audit instruments. *American Journal of Health Promotion, 18*(1), 21–37. doi:10.4278/0890-1171-18.1.21.

Ogden, C. L., Lamb, M. M., Carroll, M. D., & Flegal, K. M. (2010). Obesity and socioeconomic status in adults: United States, 2005–2008. *NCHS Data Brief, 50*(51), 1–8.

Pikora, T. J., Bull, F. C. L., Jamrozik, K., Knuiman, M., Giles-Corti, B., & Donovan, R. J. (2002). Developing a reliable audit instrument to measure the physical environment for physical activity. *American Journal of Preventive Medicine, 23*(3), 187–194. doi:10.1016/S0749-3797(02)00498-1.

Robinson, A. P., Duursma, R. A., & Marshall, J. D. (2005). A regression-based equivalence test for model validation: Shifting the burden of proof. *Tree physiology, 25*(7), 903–913.

Saelens, B. E., Sallis, J. F., Black, J. B., & Chen, D. (2003). Neighborhood-based differences in physical activity: An environment scale evaluation. *American Journal of Public Health, 93*(9), 1552–1558.

Sallis, J., Bauman, A., & Pratt, M. (1998). Environmental and policy interventions to promote physical activitya. *American Journal of Preventive Medicine, 15*(4), 379–397. doi:10.1016/S0749-3797(98)00076-2.

Schlossberg, M. (2006). From TIGER to audit instruments: Measuring neighborhood walkability with street data based on geographic information systems. *Transportation Research Record: Journal of the Transportation Research Board 1982*(1), 48–56.

Schlossberg, M., Agrawal, A. W., & Irvin, K. (2007). An assessment of GIS-enabled walkability audits. *URISA Journal, 19*(2), 5–11.

Schlossberg, M., Evers, C., Kato, K., & Brehm, C. (2012). Active transportation, citizen engagement and livability: Coupling citizens and smartphones to make the change. *URISA Journal, 25*(2), 61–70.

Schlossberg, M., Johnson-Shelton, D., Evers, C., & Moreno-Black, G. (2015). Refining the grain: Using resident-based walkability audits to better understand walkable urban form. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 8*(3), 260–278.

Seeger, C. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal, 72*(3–4), 199–213. doi:10.1007/s10708-008-9184-2.

Seeger, C. (2017). Volunteered geographic information system in planning. In C. Yamu, A. Poplin, O. Davisch, & G. De Roo (Eds.), *The virtual and the real in planning and urban design: Perspectives, practices and applications*. Abingdon: Routledge.

Seeger, C., Lillehoj, C., Wilson, S., & Jensen, A. (2014). Facilitated-VGI, smartphones and geodesign: Building a coalition while mapping community infrastructure. In U. WissenHayek, P. Fricker, & E. Buhmann (Eds.), *Digital landscape architecture* (pp. 300–308). Berlin: Herbert Wichmann Verlag.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Stevens, M., & D'Hondt, E. (2010). Crowdsourcing of pollution data using smartphones. In *Workshop on ubiquitous crowdsourcing*.

Story, M., Giles-Corti, B., Yaroch, A. L., Cummins, S., Frank, L. D., Huang, T. T.-K., et al. (2009). Work group IV: Future directions for measures of the food and physical activity environments. *American Journal of Preventive Medicine, 36*(4), S182–S188.

Sui, D., Elwood, S., & Goodchild, M. (2012). *Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice*. Berlin: Springer Science & Business Media.

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Anchor Books.

Sweeney, G., Hand, M., Kaiser, M., Clark, J. K., Rogers, C., & Spees, C. (2015). The state of food mapping academic literature since 2008 and review of online GIS-based food mapping resources. *Journal of Planning Literature*, *31*(2), 123–219. doi:10.1177/0885412215599425

Turner, S., Hottenstein, A., & Shunk, G. (1997). *Bicycle and pedestrian travel demand forecasting: Literature review*. College Station: Texas Transportation Institute, Texas A & M University System.

Vanbelle, S., & Albert, A. (2009). Agreement between two independent groups of raters. *Psychometrika, 74*(3), 477–491. doi:10.1007/s11336-009-9116-1.

Zenk, S. N., Lachance, L. L., Schulz, A. J., Mentz, G., Kannan, S., & Ridella, W. (2009). Neighborhood retail food environment and fruit and vegetable intake in a multiethnic urban population. *American Journal of Health Promotion, 23*(4), 255–264.