Whose Walkability?: Challenges in Algorithmically Measuring Subjective Experience

MARK DÍAZ, Northwestern University
NICHOLAS DIAKOPOULOS, Northwestern University

The Walk Score is a patented algorithm for measuring the walkability of a given geographic area. In addition to its use in real estate, the accompanying API is used in a range of research in public health and urban development. This study explores how neighborhood residents differently understand the notion of walkability as well as the extent to which their personal definitions of neighborhood walkability are reflected in the Walk Score's underlying algorithm. We find that, while the Walk Score generally aligns with residents' priorities around walkability, significant subjective aspects that influence walking behavior are not reflected in the score, raising the need to consider implications for using algorithmic tools like the Walk Score in certain research contexts. We discuss the challenge of measuring subjective experience and how designers might begin to address it. We call for qualitative evaluations of algorithmic tools to help determine appropriate contexts of use.

CCS Concepts: • **General and reference** → **Evaluation**; • **Human-centered computing** → *Empirical studies in HCI*; • **Computing Methodologies** → *Model development and analysis.*

Additional Key Words and Phrases: critical algorithm studies; algorithmic fairness; value sensitive design

ACM Reference Format:
Mark Díaz and Nicholas Diakopoulos. 2019. Whose Walkability?: Challenges in Algorithmically Measuring Subjective Experience. *Proc. ACM Hum.-Comput. Interact.* 3, 1 (August 2019), 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In a variety of domains, people are faced with information generated by algorithms that they must make sense of. One such algorithm is the Walk Score, which is a patented algorithm designed to evaluate walkability [1]. The Walk Score assigns a score from 0 to 100 to street addresses and neighborhoods based on amenities located within a 1.5 mile radius. Since its introduction in 2007, the Walk Score has garnered attention from professionals and researchers in real estate, urban planning, preventative medicine, and public health as a way of assessing the livability of geographic areas. Commercially, the Walk Score is a prominent feature of real estate websites to promote attractive residences to potential renters or buyers, however researchers in health and medicine have also made extensive use of the Walk Score to study connections between neighborhood walkability, physical activity, and health outcomes. Less studied is the extent to which the Walk Score captures end users' definitions of walkability as well as the range of user priorities that influence walking behavior. This research takes aim at assessing (1) alignments and misalignments between users' values and the values designed into algorithmic measures of walkability

Authors' addresses: Mark Díaz, mark.diaz@u.northwestern.edu, Northwestern University, Evanston, Illinois, Nicholas Diakopoulos, nad@northwestern.edu, Northwestern University, Evanston, Illinois,

2573-0142/2019/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

used by researchers and practitioners, and (2) how these alignments and misalignments influence the suitability of algorithmic tools such as the Walk Score in different research contexts. Although we focus our investigation on the Walk Score, this work also speaks to the design and application of algorithmic tools more broadly. We take the Walk Score as just one example of algorithmic tools that aim to measure subjective experience. In undertaking this study we explore what a quantitative metric like the Walk Score does and does not measure about residents' lived experience and how issues of algorithmic transparency relate to different end users.

Though seemingly intuitive, walkability is not a monolithic concept, making walkability scores tricky to both generate and interpret systematically. In an investigation of urban planning and mobility, Robert Prescott highlights the issue that walkability literature often ignores disability in discussions of what makes environments walkable [48]. This means, for example, that individuals using wheelchairs cannot rely on common walkability measures to indicate potential barriers to mobility, such as steep inclines or high street curbs. In addition, the Walk Score, specifically, has been criticized for failing to account for typical weather and even whether or not a street has sidewalks, factors that intuitively and significantly impact the likelihood of residents choosing to walk to nearby destinations [10]. What constitutes suitable walking weather or whether a sidewalk exists can be fairly easily agreed upon, however, the question remains regarding how users differently prioritize aspects of walkability, such as scenery, proximity to various amenities, and street noise. How do individuals conceptualize walkability differently and how might these conceptualizations differ from assumptions about walkability purposefully designed into the Walk Score algorithm?

The Walk Score is just one example of algorithmically-produced information that individuals must contend with on a daily basis. As researchers in Human-Computer Interaction seek to understand how best to meet the needs of a variety of end users and appropriately use computational tools, identifying how users differently value information and how information is used by computational algorithms are critically important. The suitability of applying a given computational tool depends, in part, on end users' ability to make appropriate interpretations from the output. Herein lies a tension between generating outputs from easily obtainable data and paying a higher cost for data that may better capture the target concept. In the case of walkability, some factors may be easily quantifiable, such as the number of cafes in a given area, whereas others may be more difficult or even impossible due to limited inter-subjective agreement, such as the neighborhood beauty or the friendliness of neighbors.

Through semi-structured interviews with residents in a large U.S. city, we found that respondents drew significant connections between walkability and their social experiences. While the Walk Score captured participants' general values around walkability, participants indicated that the Walk Score did not account for some significant factors. In particular participants assessed walkability as interwoven with subjective phenomena such as their sense of community, and their sense of safety in the neighborhood. This does not, however, mean that the Walk Score does not provide valuable insight or should not be used in analyses involving walkability. Rather, we call on algorithm designers and researchers to carefully consider how algorithmic metrics such as the Walk Score do or do not serve their specific goals with respect to what does and does not get captured. We discuss the role of transparency in the design of computational tools in supporting appropriate use and interpretation of algorithmic outputs. We also discuss implications for the ways in which researchers and practitioners use quantitatively measured data produced by tools such as the Walk Score to study phenomena that are significantly influenced by subjective experience. Specifically, we highlight walk equity to describe *whose* walkability is ultimately described by the Walk Score and whose walkability stands to be optimized for and used in research analyses.

This work calls for researchers and practitioners to incorporate understandings of individuals' lived experiences that are often absent in the design and development of technologies and tools intended to measure and improve quality of life. Ultimately, we argue that qualitative approaches are a necessary component of ethical algorithmic design and of rigorously understanding in- and out-of-scope uses of algorithmic metrics.

This research contributes an in-depth elaboration of values-oriented alignments and misalignments between individuals' definitions of an abstract concept (i.e. the Walk Score), how that concept is operationalized in an algorithmic tool, as well as how it is used as a measure of those individuals' lived experience. The findings described further contribute a detailed case study which may inform other researchers in similarly assessing value alignments of algorithmic tools through the use of qualitative methods.

2 RELATED WORK

Our research is motivated by a desire to understand how researchers and designers can create ethical technological systems that are both interpretable to users and responsive to human values. Recent research has begun to focus on how to make data-driven systems transparent and intelligible to end users [17, 37, 50], underlining a need to help users understand algorithmic outputs. Mismatches between user mental models and operationalized models can lead to false user expectations, fostering distrust in technological systems [34], or the use of algorithmic tools in unintended contexts [43]. In the following section we outline the Walk Score and its uses, as well as existing research that addresses the design of algorithms and their impact on users.

2.1 The Walk Score in the Real World

The Walk Score works by leveraging the Google Maps API to measure walking distance to amenities in nine different categories, each of which is differently weighted in importance in accordance with published research on walkability [52]. The Walk Score's amenity categories are *Grocery*, *Bars and Restaurants*, *Retail Shopping*, *Coffee Shops*, *Banks*, *Parks*, *Schools*, *Books*, and *Entertainment*. Generally, only one establishment within each amenity category is counted toward the Walk Score, however, for amenities where increased choice is determined to be important (*Bars and Restaurants*, *Shopping*, and *Coffee*), multiple amenities are counted, with diminishing value for each additional establishment within that category. The points awarded for a given amenity are a function of these weights as well as a distance decay function for amenities further than .25 miles from a given address. Amenities further than 1.5 miles are not counted. A raw score is compromised of a count of amenities, their weighting, and their distances to a given location. This score is normalized to a scale from 0 to 100. In addition, a score can receive a penalty for having long blocks or having low intersection density, which are considered less pedestrian friendly.

In addition to including apartment search tools on its own website, The Walk Score is featured on a number of real estate websites and has an API aimed at real estate professionals and web developers seeking to integrate walkability information into their websites [1]. For real estate professionals, the Walk Score is intended as a tool to help advertise the appeal and value of property listings to potential renters and buyers.

Beyond real estate professionals, Redfin, the company that owns the Walk Score, also targets the free-to-use Walk Score API to researchers and analysts in urban planning, government, public health, and finance [1]. Work in public health and urban planning is sometimes based on an intuition that physical activity increases alongside neighborhood walkability. Indeed, Hirsch et al. analyzed Walk Score data in relation to survey data, finding that Walk Scores correlated with respondent activity [29], however other research has found limited or nonexistent correlations between Walk Score and physical activity [31, 55].

Although a substantial amount of work exists validating the Walk Score for walkability research [8, 16, 31], this work has been quantitative and largely seeks to validate Walk Score data against other geographic data such as household density or street intersection density. Manaugh et al. found variation in the extent to which the Walk Score correlates with walking behavior based on individual and household characteristics as well as socio-demographic characteristics [42], suggesting the need for further probing and qualitative insight into what these variations are and what their roots might be. Broadly, this raises questions related to the sociology of quantification and the implications of quantifying social experience [18]. On the other hand, Hirsch et al. critique prior work that finds little correlation between Walk Scores and walking behavior, pointing out that the prior work is not generalizable due to selection bias [29]. While previous works may indeed lack generalizability, they provide possible evidence of the limitations of the Walk Score for analyzing particular geographic and socio-demographic groups. From a critical algorithms perspective, this body of work emphasizes a need to probe why variations in Walk Score validity might exist and what researchers can reasonably interpret from information produced from computational tools such as the Walk Score when studying populations.

2.2 Value Sensitive Design

Value Sensitive Design (VSD) provides an approach to understanding how technical systems work, how their design may or may not serve end-user needs, and how to understand the ways in which human values shape the design of systems [22]. Beginning with Friedman et al.'s general definition of a value as, "what a person or group of people consider important in life," we take aim at understanding what individuals consider important in walking and walkability [23]. Building on the foundations of VSD, Shilton et al., outline dimensions of values in the context of sociotechnical systems [53]. Importantly, they underscore the ways in which values can be expressed by users, by systems that humans design or use, as well as the interactions between humans and systems. They also respond to critiques of VSD calling for clearer methodological frameworks for investigating values in systems by delineating the types of values various research methods are best poised to investigate. Using Shilton et al.'s outline of value sources and attributes, we found semi-structured interviews to be an appropriate method both to investigate values that were central and peripheral to individuals as well as to complement quantitative studies of the Walk Score and walkability with qualitative insights. In the present work, we begin with an understanding that systems, in addition to humans, are imbued with values, and the interactions between humans and systems shape and are shaped by these values.

In the case of the Walk Score and its use in quantitative analyses, a focus on values helps to parse the facets of walkability that the Walk Score prioritizes, as well as how those facets align with the walkability definitions and priorities of neighborhood residents. The VSD framework allows us to frame the Walk Score as a tool that expresses particular values around walkability. An understanding of these values and those of the individuals the Walk Score is intended to analyze (i.e. neighborhood residents) can help to highlight which elements of residents' experiences are indeed captured by algorithmic metrics such as the Walk score, which elements are not captured, and the kinds of interpretations that are appropriate to make from algorithmic metrics describing aspects of lived experience. Because the Walk Score is used as a proxy for studying neighborhood residents' quality of life, we study residents' values as a point of comparison. Researchers using the Walk Score are implicitly adopting its definitions and values around walkability in their analyses, underscoring the importance of understanding the limits of applying these definitions and values in different research contexts.

2.3 Critical Algorithm Studies

While our current study homes in on the Walk Score as a site of investigation, our primary interest is in highlighting both several challenges in representing subjective experience in algorithmic tools more generally, as well as a need to critically evaluate and communicate in- and out-of-scope applications. Algorithmic systems are being leveraged to produce content and analyses in domains ranging from social media to entertainment to criminal justice. As such, researchers have flocked to understand the operations of these systems from both quantitative and qualitative perspectives. The expansion of automated systems has foregrounded the need to develop and test with efficiency and scalability. At the same time, researchers in psychology, design, and human computer interaction are bringing attention to the experiential qualities of algorithm design and user interactions [2, 13]. From both quantitative and qualitative perspectives, there are challenges regarding how to design, implement, and evaluate algorithmic systems with respect to underrepresented communities, many of which have unique needs and are disproportionately vulnerable to adverse effects of algorithmic bias [19, 47].

An important component of serving underrepresented communities is understanding how algorithmic systems represent and are responsive to their lived experiences. Tarleton Gillespie outlines dimensions of what he terms *public relevance algorithms*, or algorithms that, "select what is most relevant from a corpus of data composed of traces of our activities, preferences, and expressions" [24]. These activities, preferences, and expressions are intricately related to social identity, meaning that social identities are among the swaths of data that algorithms interpret or learn from. Indeed this is apparent in examples of algorithms that seek to infer gender and race from user data [25, 33]. However, even when it is not an explicit feature of analysis, patterns intimately related to social identity can be learned by algorithmic systems, such as implicit associations with race, gender, or age [6, 12].

Many scholars in critical algorithm studies have examined algorithmic bias, often looking to the outputs of automated systems and the ways they unfairly discriminate against certain groups and individuals in favor of others [11, 46, 56, 61]. For example, Taina Bucher has analyzed the ways in which social media algorithms control the visibility of different content and, therefore, users' ability to "see and be seen" [5]. A number of scholars have framed computer systems as expressions of social, ethical, and political values [21, 45], and further work has focused on understanding the sources of bias and identifying ways to diminish it [6, 12, 14].

In the present work, we focus on the ways in which walkability is defined and operationalized in an algorithm. By investigating the initial assumptions and definitions used to create the Walk Score, we attend to issues rooted in a potential mismatch between the values encoded into the design objective of the algorithm (e.g. choices of variables and assessments of suitability) and the values of end-users [36]. As such, we specifically take a qualitative approach to parse these definitions and the implications they may have. Specific definitions of walkability may differ from individual to individual or on a group level. Importantly, Nissenbaum and Friedman highlight how shifts in context of use can produce *emergent bias* [45]. The definitions and intended contexts of use chosen in the initial stage of algorithm design have direct implications for the validity of the resulting model in differing contexts of use, leading Mitchell et al. to develop a framework for reporting important details about model design and use [43].

More broadly speaking, it is important to understand the limits of what algorithmic systems can and cannot capture about human experience. Intelligent systems are increasingly managing products and services aimed at broad, diverse audiences [51, 58]. In particular, an increasing number of cities are making use of smart technologies and intelligent systems to provide crucial services to

Total Population	49,416
Women	51.4%
Black or African-American	24.7%
Latino or Hispanic	21.7%
White	45.1%
Asian	5.7%
Foreign Born	26.1%
Below Poverty Level	24.9%
Median Household Income	$39,163
High school graduate or higher	87.4%

Table 1. Neighborhood characterization of population demographics, income, and education according to data from the 2017 American Community Survey Estimates.

city residents. It is important to ensure that intelligent systems be responsive to differing information needs and values of users to build trust so that end users can equitably benefit. This project take a values-based approach to exploring these mismatches and their connections to users' understandings, interpretations, and use of algorithmically-produced information.

3 METHOD

In order to investigate neighborhood residents' values around walkability, we conducted 14 semi-structured interviews with residents living within a single neighborhood in a large U.S. city. The following subsections elaborate our recruitment and study procedures.

3.1 Participants

Basic participant demographics are shown in Tables 2 and 3. Recruitment was limited to individuals who had been living in the neighborhood for at least one year and who were 18 years or older in age. Participants ranged in age from 25 to 73 and had a median age of 39. In our recruitment, we focused on a single neighborhood (see neighborhood details in Table 1). We chose the neighborhood because of its demographic diversity and out of convenience to the proximity of the researchers. Focusing on a single neighborhood allowed us to compare different perspectives on a single Walk Score profile for a geographic area as well as assess how residents differently value the same neighborhood features. Participants were recruited through flyers posted throughout the neighborhood (e.g. in coffee shops, restaurants, bars, and public bulletin boards) and through a neighborhood community group on Facebook. Participants who met the criteria for the study were interviewed on a first-come first-serve basis over a period of two months in late 2018.

3.2 Procedure

In the first portion of the semi-structured interview, participants were informed that the interview would focus on their experience living in the neighborhood. Drawing from photo elicitation approaches [27], which Le Dantec et al. demonstrates to be an effective technique in helping respondents voice their values [38], participants were asked to identify a location in the neighborhood that they frequent or often pass by. Using Google Street View, the interviewer then navigated to

White	Black	Native	Asian	Latinx
8	3	2	1	1

Table 2. Participant race/ethnicity.

Male	Female	Median Age
7	7	39

Table 3. Participant gender and age.

the location on Google Maps. Google Street View was used to provide a visual stimulus of the location that served to ground further questions. The interviewer asked why the location was significant and how the location shaped the participant's experiences living in and navigating around the neighborhood. The first portion of the interview was intended to elicit the kinds of amenities and access residents valued having in their neighborhood.

In the second portion of the interview, participants were asked their thoughts on the walkability of their neighborhood, what informs their decisions to travel by foot, and what might make the neighborhood more walkable for them. These questions helped participants reflect on and concretize their ideas of walkability before evaluating how the Walk Score defines walkability. Next, they were introduced to the Walk Score and asked to estimate the score for the neighborhood before being shown the Walk Score page. This portion of the interview elicited participants' feelings about the extent to which the score reflects their experience living in and navigating around the neighborhood.

Although Walk Scores can be generated for an individual's home address, to protect participant privacy, participants were simply shown the aggregated Walk Score for the neighborhood as well as a visual heatmap of walkability for the entire neighborhood to give a sense of the range of walkability scores. Although the Walk Score website does not explain how a neighborhood's or city's *aggregated* Walk Score is specifically calculated or averaged, the presented visual heatmap of walkability shows an overlay of Walk Score variability in a searched area. The interviewer briefly explained how the Walk Score algorithm calculates a score, highlighting the specific criteria that contribute to the score. Participants were asked their thoughts on the factors that contribute to a Walk Score, which factors they believed were most important, and if there were any factors they would change or include if they could redesign how the Walk Score is calculated. This line of questioning involved using physical cards that could be arranged, representing each of the Walk Score's factors. Blank index cards were provided to allow participants to include their own factors. This second portion of the interview was intended as a simple design exercise and allowed participants to externalize their conceptions of values around walkability.

In the final portion of the interview, participants were asked to give their thoughts on whether they believed the Walk Score provides useful information to them about walkability in their neighborhood and other neighborhoods, whether they trusted the algorithm to generate an accurate score for themselves, and whether they saw benefits or downsides to assessing walkability algorithmically rather than through other methods (e.g., resident reviews, virtual tours). This final portion of the interview shed light on residents' trust and perceptions of algorithmic tools and the information they produce.

Throughout the interview, technical jargon such as "algorithm" was avoided so that participants would not feel they lacked sufficient knowledge or expertise to provide their thoughts. In total, each interview lasted between 26 and 61 minutes, with a median length of 40 minutes. Participants were interviewed in a public library and a neighborhood coffee shop and were paid $25 for their participation.

3.3 Data Analysis

Participants' interviews were fully transcribed and analyzed using a grounded theory approach [54]. We qualitatively analyzed interview transcriptions using iterative inductive analysis, starting with open coding of key concepts relating to walking behavior and priorities around walkability. After the first five interviews, we adjusted the interview protocol to focus on emerging themes. While completing interviews, we memoed and connected related themes, iteratively collapsing them into our final themes, which we report next.

4 FINDINGS

Overall, the Walk Score seemed to capture participants' general priorities around having walkable access to a variety of amenities.

4.1 Common Values

Broadly speaking, the Walk Score incorporated a number of factors that participants considered to be important for walkability. In particular, the Walk Score's attention to the distance and density of various amenities aligned with participants' preference for having access to a number of amenities within close distance. After viewing the Walk Score of their neighborhood, P1 responded, *"That's absolutely how it matches up with what I was saying. Absolutely. I mean, you can do so much."* P9 echoed this sentiment, saying, *"being able to do things with the kids without putting them in a car was important."* The ability to run errands on foot was important to all participants. Grocery stores, for example, were named as particularly important amenities to have close access to, which aligns with the Walk Score algorithm's heavier weighting of grocery stores. When describing the importance of walkability in choosing a place to live, P3 stated, *"I need to have a place that is close to the [train line] and within walking distance to a grocery store"*

4.2 Differing Values

At the same time, participants' individual values around walkability differed, and these differences were driven by personal context. When presented with the categories of amenities that the Walk Score algorithm uses to calculate scores, participants differed substantially in the categories they identified as most and least important to them for walkability. One example of this emerged when participants discussed the *School* category. For some participants, particularly those with children, schools were highly valued to have within walking distance. P7, a young parent, shared, *"Well, the schools [are important] because we have to go there Monday to Friday, early in the morning. The school and groceries the two biggest things for me. Those are the two things that I do the most."*

For other participants, schools were viewed positively for other reasons. P2 connected the presence of schools to safety saying, *"If you're in the proximity of many schools hopefully, ideally like you're in a safer spot"*. P9, a woman in her sixties, valued schools for the diversity they support in the neighborhood, *"I'm going to put schools because I think schools are part of what keeps the diversity in terms of age."* For others, schools were considered less important, or even a nuisance. P5, P10, and P14 named an ambivalence to the presence of schools in their neighborhood because of their lack of children, with P14 stating, *"I don't have any children, so schools and whatnot wouldn't [be important]"* and P6 actively complained about schools saying, *"When I'm dog walking I go by [two different schools] and, I have to say, if it's just when school is getting out that's really annoying because the kids are just crazy."*

Another example of variation between participants was the distance and time individuals were willing to walk to get around. P2 and P4 indicated that they have no issue walking up to 3 or 4 miles to reach a destination, while P8 expressed frustration with previously having lived a 15-minute walk and less than one mile from the closest train station. P1, who developed a foot-related impairment over her time in the neighborhood, did not specifically say how far she is able or willing to walk, but indicated that she still values walking even if it takes more time,

> *"The thing is, it's just a little harder for me now. But the places are still the same. It just takes longer for me to get there, but I can still walk to 'em."*

4.3 Missing Factors

Participants also named values and needs that were not well-reflected in the Walk Score algorithm's design. This included both individual categories or subcategories of amenities that participants valued, such as transit stops or places of worship, as well as difficult-to-measure elements of day-to-day experiences, such as aesthetic beauty or sense of community.

All participants named the importance and convenience of having a walkable transit stop nearby, however the Walk Score does not take into account transit accessibility or proximity. All but two participants used a combination of walking and transit as their primary mode of transportation day-to-day, which included completing errands, commuting to work, as well as leisure walks. P4 described their daily routine saying, *"[I walk] almost daily. It's back and forth this way or over to [the train station] and the train."* Redfin (the real estate brokerage that owns the Walk Score) produces a separate Transit Score which, similar to the Walk Score, rates geographic areas, "based on distance and type of nearby transit" [1]. The Walk Score and Transit Score exist as distinct metrics. However, participants discussed transit access as an integral component of walkability. P12 described previously living in an area with worse train access compared to their current residence, *"But now we can walk to the train. And so that was something that we were also looking at, like access to specifically the train."* P8 echoed this complaint about train access from a previous residence in comparison to their current home saying, *"because it was a 15 minute walk to the station, to me it was too much."* P3, who does not have a car, described their priorities when seeking a new place to live, saying, "Transit was a pretty big deal. Personally, I prefer being near the train." These participants underscored the importance of walkable transit and tended to do so in the context of housing searches, which Redfin markets as a primary context of use for the Walk Score.

Among other amenities missing from the Walk Score algorithm that participants valued were places of worship. P9 highlighted that churches were not factored into the walk score, *"I like being where there's churches. I mean, I'm not much of a churchgoer. But I do go occasionally."* P12 and P13, a young Muslim couple, highlighted the importance of having walking access to a mosque and lamented giving up walkable mosque access for other conveniences when moving to their current residence. *"There's a lot more mosques in the area in general. But yeah, at least where we live in [this neighborhood]. So I would say that we sacrifice [proximity to a mosque] to be closer to like these other things."* P12 went on to explain a desire for a walkable hair salon and gym in the neighborhood.

4.4 Subjective Factors

For values not reflected in the Walk Score, participants often prioritized nebulous or difficult-to-quantify aspects of walkability. Among these were the aesthetics of the neighborhood, personal sense of safety, a sense of community or friendliness in the area, and even neighborhood affordability.

4.4.1 Neighborhood Aesthetics. For participants, the aesthetic beauty, including foliage and the nearby lakefront provided both incentive to go outdoors as well as an improved experience when walking outdoors to accomplish other tasks. P5 highlighted that the natural beauty of the neighborhood is, "almost kind of sort of medicinal, you know, like, it's really beneficial to my soul." P1 similarly described the impact of the neighborhood's beauty on them,

> *"On a nice day, it's nice. After looking at everything- it's so much to look at, number one. So that makes the walk go even quicker, you don't even know, like Oh, I'm right here. So I really enjoy that."*

This finding echoes quantitative work in urban planning that highlights connections between the design of urban streetscapes and individuals' perceptions of comfort and safety [28], as well as

statistical analyses in geography literature linking visual landscapes to perceived walkability [4]. In addition to the natural beauty that other participants highlighted, P11 explained their appreciation for the varied aesthetics that reflect the neighborhood's diversity,

> *"one of the things that I like, like, you walk down [the street] and it's not trying to appear to be like bougie in any way, like, you walk down and it's just like, a Chinese restaurant looks like a Chinese restaurant. A Mexican restaurant looks like a Mexican [restaurant]."*

P11 appreciated that restaurants in the neighborhood maintained a local and *"authentic"* display rather than a more mainstream aesthetic that might cater to wealthier clientele while simultaneously sacrificing cultural roots.

4.4.2 Sense of Safety. In addition to aesthetics, all participants raised concerns about crime and safety in the neighborhood. As with neighborhood transit, the Walk Score website calculates Crime Scores as distinct from walkability, however participants drew strong connections between walkability and their sense of safety. We found that sense of safety influenced walking behavior, as has been highlighted in prior quantitative assessments of the Walk Score [57] and noted as a site for further investigation [7]. However, we also found that the extent to which crime impacted participants' sense of safety was directly related to demographic characteristics and participants' beliefs about the targeted nature of certain crimes. Not only was crime generally a factor that influenced comfort and walkability in the neighborhood, but also the type of crime was a concern for residents.

P1 pointed out that their sense of safety was impacted by time of day. *"[A destination] can be in close proximity, but if it's late, I still want to call [a car]. I'm not gonna just walk those few blocks. It might be a short distance, but if it's late, I'm not gonna do that."* P5 echoed, *"[Crime] is a factor one has to take in consideration when they go out for a walk in the neighborhood."* However, crime was also an accepted reality of living in an urban environment. P1 went on to say, *"we're in a city and things happen."* P7 and P8 reinforced this point saying,

> *"There are times that, you know, you hear the gangs and, you know, there's been like, sometimes you hear the gunshots and stuff like that, but you're in a city, you know what I mean, so..." -P7*

> *"My friends are horrified that I moved to [the neighborhood]. And my response has been, well, I could be hit by a bus tomorrow. So being shot or mugged doesn't bother me like this is part of living in a city is the price you pay to be around lots of people and to have kind of concentrated culture like, if you're going to be living around more people, the chances are, you're going to be living near more bad people so I'm fine with that." -P8*

Participants also highlighted the importance of the *type* and *perceived target* of crime that occurred was of concern. Although they indicated relatively little concern about neighborhood crime, P8, noted that they cared specifically about hate crimes. This concern was rooted in their queer identity and their ability to walk to one of their jobs as a drag queen at a local gay bar.

> *"Crime does not bother me as a measure of where I live with one exception, and that is hate crimes...for me, there's a difference between being burgled or being robbed at gunpoint which feels very opportunistic versus a hate crime which is more personalized and someone once said somewhere I, I forget where I read it, but to live as a queer person, a gay person, whatever is to sacrifice your personal safety in favor of your personal happiness and I feel like I can be whoever I want to be around here." -P8*

To P8, freedom of gender expression and sexual identity, particularly as a person that could be easily identified by others on the street as queer, provided a sense of relative safety. Similarly, when

describing two recent murders that occurred in the neighborhood, one of which was a suspected hate crime, P2 stated, *"the hate crime does appear [to be targeted] like that's the scarier one."* Much news about crime in the neighborhood was driven by gang activity, which several participants understood was very targeted. Although they were generally concerned about crime, P2 and P9 recognized the targeted natured of violent crime, noting,

> *"In actuality, like, as a single white female I'm probably safer than most because they really don't want that type of attention, to have a crime against a white woman" -P2*

> *"I hate the idea that kids in the city are growing up feeling like they're, they're involved, you know, shooting and getting shot at, I hate it. But it didn't make me feel more scared as an individual. Because I thought, 'they're not shooting at me.' So it doesn't make me scared to be out on the street." -P9*

4.4.3 Sense of Community. Sense of community also impacted walking experience for many participants. A number of participants indicated that their sense of community influenced their experiences walking, including their sense of safety.

> *"You know, I just wanna step out and get some fresh air. But like I was saying, I'm so close to people that I can just go say hi to, you know" -P1*

> *"I do feel more comfortable here than I ever did really anywhere else just because you do have that sense of community... there was a guy who was kind of, a little bit harrass-y, and he followed me home from the [train line] a few times, and so, it was, like, there was a little bit of a sense of like obviously being nervous, but also there were people around that were taking care of me and making sure I was okay, and making sure I was getting home. And so it's like I just don't feel like I would have gotten that anywhere else." -P3*

> *"I honestly don't feel any safer than I did 33 years ago. Because there may have been more gangbangers but there were also more grandmas of said gangbangers. Who I'd be like, 'child, get out of my way, I know your grandma', you know."* -P4

4.4.4 Diversity. For a number of participants this sense of community and comfort in the neighborhood was directly impacted by the racial and ethnic diversity of the neighborhood. P13, who is an immigrant and who speaks Arabic relayed their excitement at encountering a sign in their native language while walking down the street,

> *They had it written in like multiple languages... and the bottom was Arabic and I'm so happy so yeah so as a new person to the neighbor that's really encouraging, yes really encouraging" -P13*

Similarly, P7 drew a connection between neighborhood diversity, community, and local activities,

> *"I think because there is so many people like it helps us to get out more just because we like the different cultural events and things that they've had around here...you have to go downtown for a lot of that, you know, yes. So, for it to be right in the neighborhood... it's nice." -P7*

By the same token, several participants valued neighborhood diversity in relation to the variety of local businesses and markets available. P9 shared, *"I shop at [the market] and I get access to all that wonderful stuff, all the all the different kinds of food from all over the world."*

> *"Just walking distance. Just step out your house, you'll find something good. And so many different cuisines. I love the Belizean place on [street], the Jamaican place on [street], several burritos on [street]" -P1*

While the Walk Score for a given area is boosted by the presence of more than one bar or restaurant, the score is not influenced by the variety of cuisines offered.

4.4.5 Affordability. Community and walkability were also interestingly tied to neighborhood development and affordability. When describing their love of the neighborhood and how it has developed over time, P1 explained,

> *"And I always enjoyed it because, hey, we always had the lake to walk to and now it seems like um, certain areas, it's like private land and you have to kind of go around, because the condos now. Affordable housing is a part of everything too because, hey, you know, I want to stay in my place and be able to... why can't I be able to keep walking to all the groovy spots, you know."*

Affordability was commonly cited as something that originally drew participants to move to the area. Although only one participant tied affordability directly to walkability, nearly all participants referenced the rising cost of living in the neighborhood as a concern. When discussing neighborhood safety, P4 also referenced neighborhood affordability,

> *"I also don't like that there's a lot of working people who aren't here anymore because they can't afford to live here. Again, stability. The one house with the mom that would sit on the front porch after school, makes the neighborhood more stable than the 6 people in the condo who are never home, because they're always somewhere else outside the neighborhood."*

4.5 Interpretability

Although the Walk Score captured factors of general importance to neighborhood residents, ultimately, it's perceived usefulness was mixed among participants. This may have been driven by the fact that participants had a tendency to be unclear about the factors that contributed to a give Walk Score. P6 commented that *"[The Walk Score] doesn't tell the whole story"*, while P5 noted, - *"I guess it would be [helpful], I might use it as a resource, but I wouldn't really rely on it entirely. Only because I know that my, what I like, usually people don't care for and vice versa."* They went on to say, *"I would definitely visit the area in person over the opinion of others, or the score that they have given because like we said earlier things factor that don't really matter to me like schools, etc."* P2, who had used the Walk Score in a housing search in the past, described setting a threshold for filtering out options based on walkability, *"So I'm looking at like 85 or higher on Walk Score. I haven't ended up moving but like that's the... I feel like I've taken places out of the equation. If it's at least eighty five or above all right."* Although P2 did not entirely trust the Walk Score, they did use it as a way to limit their potential housing choices.

Several other participants responded similarly, considering the Walk Score as a potential tool for directing attention but desiring more information about why walkability scores varied and how neighborhood features factored in. P9, an older adult, pointed out a specific desire to live within walking distance of a place where they could fill prescriptions but was unclear on which amenity categories captured that need. They expressed a desire for more information about the Walk Score's underpinnings, *"Well, I'd be interested to see, you know, the areas where the shading is different. Why is that?"*. When observing the Walk Score heatmap P5 questioned why Walk Scores decreased closer to the lakefront,

> *"Oh, I guess I guess [street] is dark [green]. But why does it get lighter to the lake? I would go dark all the way to the lake."*

When observing the low Walk Scores surrounding a nature park, P5 continued,

> *"It's interesting because that's, that's sort of a nature area. But I guess maybe this this website or people maybe maybe thinking more about like, for practical things in their daily life, like, you know, like getting to the train or getting to a job?"* -P5

> *"So to me, 80 is pretty high...I don't really know the nuance or the specifics of why it would get down at all."* -P3

When viewing the Walk Score for their neighborhood along with the Walk Score heatmap, participants tended to rationalize differences in Walk Scores that they observed, often remarking on factors that are not actually incorporated into the algorithm. When observing the heatmap, P9 noted their surprise that the lakefront, one of their favorite spots in the area to walk, had a lower Walk Score, saying, *"so like the lakefront shows here as being less walkable. But I suppose that's because of sand."* After noticing that a particular region had a lower Walk Score than expected, P2 tried to rationalize the discrepancy between their personal approximation and that of the Walk Score, *"I'm assuming that I must have misread that and probably how busy the streets are. Maybe people are a little bit more deterred to look around."*

5 DISCUSSION

From a values perspective, the Walk Score aligns with a wide range of participant values, however some important subjective factors of walkability were not reflected. The Walk Score also did not capture a number of factors that were significant to participants on an individual basis. Borrowing Shilton et al.'s values dimensions framework [53], the Walk Score captured collective values that are relatively central, or highly salient to participants, but in perhaps its most common context of use (providing localized walkability metrics to individual prospective home buyers and renters), the Score breaks down at the level of capturing some influential, individual values. The Walk Score remained a useful source of information, particularly in the context of renter searches, however some subjective aspects of walkability were not captured, raising questions about whose walkability the Walk Score measures and which users the Walk Score is designed for. These questions carry important implications for researchers using algorithmic tools such as the Walk Score to study phenomena that are significantly influenced by subjective experiences.

5.1 Subjective Factors of Walkability

The Walk Score algorithm's approach to measuring walkability primarily relies on tallying physical establishments, avoiding the complexity of measuring subjective factors of walkability while still producing a usable approximation of walkability for prospective renters and buyers. Gillespie's dimensions of *patterns of inclusion* and *the evaluation of relevance* importantly frame the issue of what data is captured in an algorithmic process as well as how it gets captured. "Patterns of inclusion" refers to which data are included and excluded from algorithmic processes, while "the evaluation of relevance" refers to "the criteria by which algorithms determine what is relevant, how those criteria are obscured from us, and how they enact political choices about appropriate and legitimate knowledge" [24]. In addition to the issue of which factors are weighed more heavily by the Walk Score (e.g., grocery stores), there is the issue of what does not get weighed at all. Despite their absence from the Walk Score algorithm, subjective factors influence walkability in significant ways. Some of these factors prove difficult to quantify, such as a sense of neighborhood

community and sense of safety. Nonetheless, these factors influenced participants' experiences while walking as well as their routes and even whether they would opt to walk at certain times.

Adding to the challenge of operationalizing abstract experiences such as a sense of neighborhood community, is that they differ on both an individual basis as well as in systematic ways across social groups. While generally important, the aesthetics and beauty of the neighborhood differed somewhat in priority from participant to participant based on personal opinions of beauty and desire for leisure walking. On the other hand, participants' sense of safety had some roots in social identity. Participants' sense of safety and sensitivity to matters of safety was influenced by racial and gender identity and how these social identities fit into broader social hierarchies. Three participants named their white racial identity as a factor that mitigated safety concerns while out on the street, either because they were less likely to be a target of crime or because they were less likely to be targeted by a police presence. For example, participants recognized that victims of crime are not equally distributed across social identities. This means that some individuals are more likely to be targeted than others and for particular crimes. Similarly, it was noted that police attention was not equally distributed among individuals out on the street. Conversely, sense of safety was more positively impacted by ethnic diversity for some individuals. One participant in a multicultural family and another participant who is arabic-speaking were particularly positively affected by the racial and ethnic diversity in the neighborhood. Participants' social identities influenced both the salience of walkability factors as well as whether particular concerns, such as police presence, produced positive or negative implications for walkability.

Redfin does not explicitly indicate why it maintains scores for Crime and for Transit separately from the Walk Score, given the implications of those factors on walkability. However, it's plausible that the separation simplifies calculations and helps isolate subjectivities within familiar categories. For instance, incorporating the separate Transit Score would be difficult because of participants' different willingness to take certain forms of public transit as well as different willingness to walk certain distances to reach it. Some kind of composite score might help produce a more robust, general approximation of walking experiences, but the added complexity would also add to the difficulty of making a Walk Score interpretable. For individuals evaluating potential homes, maintaining separate scores allows end users to use their own internal weighting and priorities in interpreting the group of scores.

While the absence of various factors in the Walk Score algorithm is worth noting, designers of similar algorithms should not necessarily seek to create a laundry list of missing factors to begin including. In addition to the inherent challenges in quantifying qualitative experiences, there is no guarantee that designers will have the capacity to uncover a complete list of missing components. Rather, designers must determine the factors most important to account for or preserve for a given, intended analysis such that end-users are able to accurately interpret algorithmic outputs. For factors that may be difficult to operationalize, such as sense of safety, which differs both individually and systematically, designers might consider signalling that such information is not captured in a specific algorithm design. End-users could then account for this missing feature and make decisions regarding how and whether to use algorithmic outputs.

5.2 Evaluating Contextual and Definitional Suitability via Qualitative Methods

Ultimately, researchers and designers should place emphasis on understanding the context of application and evaluating an algorithmic metric for that context. For example, a walkability metric that does not thoroughly encode racial differences might still be acceptable for group-level analyses among a relatively homogenous population. Based on the Walk Score algorithm's criteria, there is an implicit person for whom walkability is being measured. This person is relatively able-bodied, values the particular amenity categories included in the algorithm, and has a particular distance

threshold for walking before opting to choose another mode of travel. However, researchers using the Walk Score have little indication of this assumed person. The Walk Score algorithm is relatively transparent, however its underlying design details are not well-organized for end users. While the Walk Score does feature documentation on how walkability is measured and scored, little information is organized or packaged to indicate the contexts and types of research questions the tool is best suited to answer, an issue that Mitchell et al. begin to address with their framework for model reporting [43]. In the end, researchers' assumptions and interpretations of Walk Score data may be false or incomplete. While the Walk Score encodes a particular definition of walkability, researchers must assess this definition against residents' descriptions of walkability as well as their own definitions and assumptions about walkability that may not align with either the Walk Score's or residents' definitions. Such misalignments raise ethical concerns about the validity and potential bias that might be associated with using the score in varied or emerging contexts [44].

While the Walk Score may be sufficiently validated for measuring walkability across a general U.S. population, failing to account for factors that significantly or disproportionately impact the walking behaviors of population subgroups throws into question the validity of the measure for characterizing those particular groups. For example, participants indicated that sense of safety and relationship to neighborhood crime varies with respect to identity. This is something that is likely important to researchers seeking to describe physical activity across populations. From a public health perspective, this could leave subpopulations out of conversations involving policies or strategies aimed at improving health and well-being.

Investigations of what gets captured by the Walk Score and other similar computational tools are important for understanding what these tools actually measure. Understanding the larger context of walkability is critical for designing computational tools or suites of tools and approaches that avoid producing "average" assessments that may not effectively characterize the distribution of individuals' actual lived experiences.

It is here that we highlight the benefits of qualitative methods to investigate assumptions and definitions operationalized in algorithms in a variety of domains. Many investigations of algorithmic bias involve quantitative evaluation of algorithmic outputs with respect to a specific form of bias, such as age-related bias in sentiment analysis [12]. Our approach is one that can be undertaken before the deployment of a system or tool, such that end-users and researchers may be provided with insights about *potential* issues in applying an algorithmic model. Mitchell et al.'s work proposing *model cards*, which are documents that, "disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information", calls for algorithmic model designers to disclose a variety of information to help end-users avoid using algorithmic tools for analyses for which they are not optimized and may produce unintended bias [43]. A qualitative investigation such as the one we undertook, can complement the quantitative details that Mitchell et al. call for.

Whereas a post-hoc analysis of algorithmic bias is important for measuring the magnitude of bias and potential mitigation strategies, such an approach requires knowledge a priori of where to look for bias. The approach we took did not test for a specific instance of bias but helped highlight *potential* sources of bias that could be critical for different kinds of analyses. For example, we did not embark on our research with specific intent to delineate racial or socioeconomic underpinnings of walkability, however, these connections first emerged after only just a few interviews. On the other hand, known or specific assumptions can be tested and validated in qualitative ways. Indeed, we were able to confirm factors that the Walk Score was correct to deem important, such as proximity to shopping and entertainment. Pairing our approach with survey methods could further validate or invalidate assumptions across different geographic regions.

Taking a qualitative and exploratory approach such as ours in the initial design of the Walk Score brings clarity to decision making around whether to modify the algorithm in such a way that it begins to address some of these factors or highlight to researcher end-users that these relationships may also be important to consider. Explicitly flagging potential limitations helps researchers with more domain expertise to determine whether to and how best to interpret outputs from an algorithmic tool. Even outside of research contexts, qualitative insight highlights how approaches to data collection might influence subsequent interpretations of outputs— a central focus of Fox et al.'s design parody exposing socio-spatial exclusions in data-driven maps [20].

Because our approach takes aim at values to understand differing definitions and assumptions, it is particularly suited for tools that measure phenomena for which user understandings may vary between populations (e.g., men's and women's differing understandings of harassment online [15]; crowd workers' greater likelihood to identify toxic language as hate speech compared with social justice activists [59]). This includes domains such as toxic comment moderation, where 'toxicity' is defined in different ways [3]. In the case of toxicity, gaining a clear understanding of how end users conceptualize and define toxic behavior is critical for making sure that an algorithm can identify content that users being served indeed believe to be offensive or rude, rather than content that another population believes to be toxic. An approach emulating ours could first try to establish variations in how toxic behavior is defined among members of the population(s) that will be impacted by adoption of a toxicity-detection algorithm and then compare those definitions to the algorithm's parameters. This could highlight instances and scenarios of toxic behavior for which an algorithm may perform better or worse. Even before prototyping an algorithm, designers might take a similar approach to determine which concepts related to toxicity to attempt to operationalize. Health tracking is another domain in which definitions of a central concept may differ. Health tracking tools can be built on exercise logs that often have different definitions of 'activity level' and the corresponding activities that contribute to metrics [60]. End users may have their own distinct notions of healthy activities and falsely assume that an app is measuring those activities. In this instance, a qualitative approach can help to parse assumptions end users make as well as identify to designers how they are successfully or unsuccessfully communicating the algorithm's notion of healthy behavior. The approach we develop in this paper specifically helps to make different understandings of such subjective concepts designed into algorithms more salient to algorithm designers.

The question of what gets captured by the Walk Score algorithm comes down to how a concept like "walkability" is quantified. Espeland and Stevens turn attention to the broad implications of quantification, highlighting key dimensions of what they term a "sociology of quantification" [18]. Of particular relevance to the challenges of measuring walkability are the dimensions of *reactivity* and *authority*. The authors argue that quantitative measures, "create or reinforce the categories used to conceive of human beings," citing demographic counts and census work as prime examples. We consider this idea with an emphasis on implicit reinforcement of categories. The Walk Score does not take a head count of defined social groups, as the census does. However, implicit in the Walk Score's measurement of walkability are predefined ways of living and being that, as our study reveals, do not wholly resonate with the experiences of many people. We take the view that designers of algorithmic tools must make these predefinitions of lived experience explicit. This is particularly important because of the *authority* Espeland and Stevens describe quantified information as possessing. They note that, "in the relationships between databases and those who rely on them for arguments, numbers can accumulate constituents who invest them with particular meanings and uses." In other words, quantification can carry undue credibility that informs accepted truths about the subjects it describes [32, 35]. In the same vein, work on algorithmic

authority has highlighted the ways that algorithmic decisions can be viewed with similarly undue credibility [40, 41]. Espeland and Stevens ultimately posit the sociology of quantification as a starting point to developing an "ethics of numbers." An ethics of algorithmic design and use must similarly draw on understandings of the ethics and processes of quantification.

5.3 Who is the User?

While it is possible that individual consumers apply an implicit understanding of how experiences influenced by their racial identity differ from others when interpreting a Walk Score, researchers seeking to analyze and compare population groups have limited capacity to do the same.

Addressing how to incorporate subjective factors into a walkability metric, or even whether it is necessary to do so, first requires determining the user. Designing a walkability metric requires asking important questions about *whose* walkability should be captured and who needs to capture it. We recommend that designers of algorithmic tools clearly identify stakeholders (including end-users and impacted groups) to then investigate and understand how those stakeholders will interpret algorithmic outputs. The Walk Score is used in different domains, each with a different end user. For example, creating a walkability metric for a prospective home buyer requires algorithm designers to design in response to preferences that vary from individual to individual. Allowing end users to add or remove amenity categories or adjust their relative importance, would support users in optimizing the algorithm for their personal needs. Indeed, Priedhorsky et al. found that cycling route recommendation algorithms produced results more similar to user ratings when subjective route preferences were considered [49]. Personalization could also work in support of explaining why the algorithm produces a specific score, which Rader et al. find can support awareness of how an algorithm is influencing outcomes [50]. In this case, allowing users to tweak calculations to reflect the values most central and individual to them, or at least be able to identify the "distance" between those values hard-coded into the Walk Score and their own, would support usability. In addition, subjective factors of walkability may differ from user to user substantially enough such that the effort to measure them may provide little value to the average user. In the same way that maintaining distinct Walk, Crime, and Transit scores may be easier to interpret than a composite walkability score, users can individually weigh their priorities to mentally adjust a given Walk Score to their personal context.

In this work, however, we highlight the use of the Walk Score in population-level research analyses as a critical site of evaluation. The end user in public health and urban development research applications differs from the end user in real estate applications. Supporting usability for researchers and analysts seeking to make generalizations about behavior for different populations requires a focus on supporting accurate interpretations of data outputs. This means that clearly delineating the aspects of walkability that do and do not get captured by the algorithm is necessary. At the same time, it is imperative to make clear which types of analyses the Walk Score is intended to support and is valid for, as well as which analyses it has not been intended or tested for. Delineating these analyses also helps make clear to the algorithm designers whether to embark on operationalizing complicated subjective experiences. For example, applying the Walk Score in an analysis of a racially homogenous and low-crime neighborhood may not warrant incorporation of a 'sense of safety' factor in the way that a highly racially diverse and high-crime area might.

At the same time, systematic variance in these subjective factors across social and community groups means that researchers must consider patterned behavior among subgroups that may deviate from what is typically observed. Considering again the results of Manaugh et al. [42], which found that Walk Score validity varied according to household characteristics, including sociodemographic variables, some of this result may be related to shared characteristics that were not accounted for among the research subjects. Mitchell et al. proposed a novel framework offers

model creators a way of evaluating and clarifying to end users the contexts for which a given algorithmic model is well-suited, as well as contexts for which that same model is not well-suited. For algorithm designers, the framework provides a set of references for evaluating a model against its intended use case. Making explicit the intended scope of use would allow designers to either improve walkability metrics or provide explicit signals to the types of analyses and populations for whom the metric is optimized.

5.4 Walk Equity

Focusing narrowly on real estate contexts, the Walk Score misses important factors for particular groups that have implications for the utility of walkability scores in low income and rural areas. Notably, the Walk Score does not take into consideration physical infrastructure or the maintenance of infrastructure. For the participants we interviewed, wintry conditions and the extent to which sidewalks were cleared and maintained were significant factors, particularly for those who felt less stable on their feet due to age or disability. This aligns with Hirsch et al.'s finding that sociodemographic variables considering physical mobility mediated the relationship between walkability scores and walkability outcomes for Canadian adults [30]. Identifying this issue, researchers with Project Sidewalk have developed ongoing research using mixed-methods approaches to collect and visualize accessibility information at scale about sidewalks across the United States [26, 39]. The failure of the Walk Score, in particular, to account for physical infrastructure has been highlighted in past research [48] and has implications for lower income and rural communities that may be subject to generally worse infrastructure and municipal neglect.

The state of infrastructure is part of a larger discussion around the extent to which the Walk Score equitably defines and characterizes walkability. A failure to account for poorly maintained sidewalk infrastructure in an otherwise well-rated area means that the walkability described is inequitable– that is, it exists for able-bodied residents with the means to navigate rougher paths, while other residents face additional challenges or are barred from walking access. For individuals for whom physical activity may be particularly important, such as folks who are more sedentary due to physical impairments, walkability as measured by the Walk Score may be artificially high and may not take into account important factors needed for them to have access to outdoor physical activity and walking.

Importantly, the quality of infrastructure is tied to the wealth and resources of a given community. Development may improve an objective measure of walkability, but if it is accompanied by displacement, the walkability of the area is enjoyed by a changed population. From a development perspective, improvements in walkability may not be equitably experienced, including any benefits or increase in activity purported to improve the health and well-being of communities. As remarked by one participant, development in her neighborhood limited public access to lakefront areas and increased rent prices, pushing out long-term residents. As tools already in place to advertise desirable homes, the Walk Score and similar walkability metrics may act as forces that prescribe walkability and drive development. The use of the walkability metric as a pointer to neighborhood development and health stands as an example of what Corbett and Loukissas highlight as a failure to "recognize the experiences of people undergoing gentrification" [9]. Particularly for work involving repeated or longitudinal analyses of geographic areas, identifying who experiences changes in development or walkability over time is absolutely necessary for determining whether those changes are equitable. Enjoyment of existing and subsequent walkability features became a benefit for wealthier residents, while displaced residents were forced to relocate to areas further from transportation and other neighborhood amenities. This matter of *whose* walkability is captured by the Walk Score only became apparent through participant interviews,

highlighting the value that qualitative inquiry can provide in investigating ethical design and uses of algorithmic tools.

6 CONCLUSION

In this work, we contributed an exploration of values-oriented alignments and misalignments between individuals' definitions of an abstract concept and how that concept is operationalized in an algorithmic tool. We also investigated the implications of these alignments and misalignments for the algorithmic tool's use as a measure of those individuals' lived experience. The use of the Walk Score and other computational tools can often rely on quantifying factors that are convenient or relatively easy to quantify. As a result however, these tools risk ignoring significant components of lived experiences, particularly the lived experiences of underserved populations that designers should aim to equitably support and improve in the creation of computational tools. In order for algorithmic tools to be used ethically and responsibly, researchers must be diligent in determining what they can appropriately infer about data and the lives they describe. In order to support ethical and responsible use of algorithmic metrics in research, creators of algorithmic tools must aim to design in ways that support researchers in making determinations about the extent to which algorithmic outputs align or misalign with the lived experiences of research subjects as well as different contexts of use. One important way of supporting end users is by stating the envisioned context of use and providing increased transparency. Researchers and analysts must closely consider the design of algorithmic metrics and how these metrics may or may not meet their specific analysis goals. Algorithm designers, in turn, must support researchers and analysts in making these determinations by surfacing design specifications. Without this support, designers of algorithms such as the Walk Score risk inadvertently leading researchers and analysts to draw incorrect conclusions about particular geographic areas or about an unspecified subset of the population.

REFERENCES

[1] 2019. Walk Score Professional. www.walkscore.com

[2] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 286.

[3] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 1100–1105. https://doi.org/10.1145/3308560.3317083

[4] Bradley Bereitschaft. 2019. Exploring perceptions of creativity and walkability in Omaha, NE. *City, Culture and Society* 17 (2019), 8–19.

[5] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New media & society* 14, 7 (2012), 1164–1180.

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[7] Lucas J Carr, Shira I Dunsiger, and Bess H Marcus. 2010. Walk scoreâĎć as a global estimate of neighborhood walkability. *American journal of preventive medicine* 39, 5 (2010), 460–463.

[8] Lucas J Carr, Shira I Dunsiger, and Bess H Marcus. 2011. Validation of Walk Score for estimating access to walkable amenities. *Br J Sports Med* 45, 14 (2011), 1144–1148.

[9] Eric Corbett and Yanni Loukissas. 2019. Engaging Gentrification As a Social Justice Issue in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 280, 16 pages. https://doi.org/10.1145/3290605.3300510

[10] Paulo Jorge Monteiro de Cambra. 2012. *Pedestrian accessibility and attractiveness indicators for walkability assessment*. Ph.D. Dissertation. Thesis for the Master Degree (MSc) in Urban Studies and Territorial Management.

[11] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3, 3 (2015), 398–415. https://doi.org/10.1080/21670811.2014.976411

[12] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 412.

[13] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729

[15] Maeve Duggan. 2017. Online harassment 2017. (2017).

[16] Dustin T Duncan, Jared Aldstadt, John Whalen, Steven J Melly, and Steven L Gortmaker. 2011. Validation of Walk Score® for estimating neighborhood walkability: an analysis of four US metropolitan areas. *International journal of environmental research and public health* 8, 11 (2011), 4160–4179.

[17] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. FeedVis: A Path for Exploring News Feed Curation Algorithms. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (CSCW'15 Companion)*. ACM, New York, NY, USA, 65–68. https://doi.org/10.1145/2685553.2702690

[18] Wendy Nelson Espeland and Mitchell L Stevens. 2008. A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie* 49, 3 (2008), 401–436.

[19] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

[20] Sarah E. Fox, Meredith Lampe, and Daniela K. Rosner. 2018. Parody in Place: Exposing Socio-spatial Exclusions in Data-Driven Maps with Design Parody. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 322, 13 pages. https://doi.org/10.1145/3173574.3173896

[21] Batya Friedman. 1997. *Human values and the design of computer technology*. Number 72. Cambridge University Press.

[22] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 02–12 (2002).

[23] Batya Friedman, Peter H Kahn, and Alan Borning. 2008. Value sensitive design and information systems. *The handbook of information and computer ethics* (2008), 69–101.

[24] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014).

[25] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 8, 13 pages. https://doi.org/10.1145/3173574.3173582

[26] Kotaro Hara and Jon E Froehlich. 2015. Characterizing and visualizing physical world accessibility at scale using crowdsourcing, computer vision, and machine learning. *ACM SIGACCESS Accessibility and Computing* 113 (2015), 13–21.

[27] Douglas Harper. 2002. Talking about pictures: A case for photo elicitation. *Visual studies* 17, 1 (2002), 13–26.

[28] Chester Harvey, Lisa Aultman-Hall, Stephanie E Hurley, and Austin Troy. 2015. Effects of skeletal streetscape design on perceived safety. *Landscape and Urban Planning* 142 (2015), 18–28.

[29] Jana A Hirsch, Kari A Moore, Kelly R Evenson, Daniel A Rodriguez, and Ana V Diez Roux. 2013. Walk Score® and Transit Score® and walking in the multi-ethnic study of atherosclerosis. *American journal of preventive medicine* 45, 2 (2013), 158–166.

[30] Jana A Hirsch, Meghan Winters, Philippa J Clarke, Nathalie Ste-Marie, and Heather A McKay. 2017. The influence of walkability on broader mobility for Canadian middle aged and older adults: An examination of Walk ScoreâĎć and the Mobility Over Varied Environments Scale (MOVES). *Preventive medicine* 95 (2017), S60–S67.

[31] Lindsey Irene Jones. 2010. *Investigating neighborhood walkability and its association with physical activity levels and body composition of a sample of Maryland adolescent girls*. Ph.D. Dissertation.

[32] Herbert Kalthoff. 2005. Practices of calculation: Economic representations and risk management. *Theory, Culture & Society* 22, 2 (2005), 69–97.

[33] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 88 (Nov. 2018), 22 pages. https://doi.org/10.1145/3274357

[34] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[35] Karin Knorr-Cetina. 1999. Epistemic cultures: The cultures of knowledge societies. *Cambridge, MA: Harvard* (1999).

[36] Felicitas Kraemer, Kees van Overveld, and Martin Peterson. 2010. Is there an ethics of algorithms? *Ethics and Information Technology* 13, 3 (July 2010), 251–260.

[37] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in

social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 417–432.

[38] Christopher A Le Dantec, Erika Shehan Poole, and Susan P Wyche. 2009. Values as lived experience: evolving value sensitive design in support of value discovery. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1141–1150.

[39] Anthony Li, Manaswi Saha, Anupam Gupta, and Jon E Froehlich. 2018. Interactively Modeling and Visualizing Neighborhood Accessibility at Scale: An Initial Study of Washington DC. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 444–446.

[40] Caitlin Lustig and Bonnie Nardi. 2015. Algorithmic authority: The case of Bitcoin. In *2015 48th Hawaii International Conference on System Sciences*. IEEE, 743–752.

[41] Caitlin Lustig, Katie Pine, Bonnie Nardi, Lilly Irani, Min Kyung Lee, Dawn Nafus, and Christian Sandvig. 2016. Algorithmic Authority: The Ethics, Politics, and Economics of Algorithms That Interpret, Decide, and Manage. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1057–1062. https://doi.org/10.1145/2851581.2886426

[42] Kevin Manaugh and Ahmed El-Geneidy. 2011. Validating walkability indices: How do different households respond to the walkability of their neighborhood? *Transportation research part D: transport and environment* 16, 4 (2011), 309–315.

[43] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.

[44] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data Society* 3, 2 (2016).

[45] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[46] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[47] Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[48] Michael Prescott. 2014. Using Social Topography to Understand the Active Mobility Networks (AMNs) of People with Disabilities (PWDs). http://hdl.handle.net/10012/8250

[49] Reid Priedhorsky, David Pitchford, Shilad Sen, and Loren Terveen. 2012. Recommending Routes in the Context of Bicycling: Algorithms, Evaluation, and the Value of Personalization. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 979–988. https://doi.org/10.1145/2145204.2145350

[50] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 103, 13 pages. https://doi.org/10.1145/3173574.3173677

[51] Andrew Schrock. 2018. Civic Tech: Making Technology Work for People.

[52] Walk Score. 2014. Walk score methodology. *Accessed April* 24 (2014).

[53] Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. 2014. How to see values in social computing: methods for studying values dimensions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 426–435.

[54] A Strauss and J M Corbin. 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Inc.

[55] Paul Y Takahashi, Mitzi A Baker, Stephan Cha, and Paul V Targonski. 2012. A cross-sectional survey of the relationship between walking, biking, and the built environment for adults aged over 70 years. *Risk management and healthcare policy* 5 (2012), 35.

[56] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Toward a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit. *ACM Trans. Comput.-Hum. Interact.* 24, 3, Article 21 (April 2017), 40 pages. https://doi.org/10.1145/3058499

[57] Samuel D Towne, Jaewoong Won, Sungmin Lee, Marcia G Ory, Samuel N Forjuoh, Suojin Wang, and Chanam Lee. 2016. Using Walk ScoreâĎć and neighborhood perceptions to assess walking among middle-aged and older adults. *Journal of community health* 41, 5 (2016), 977–988.

[58] Vasillis Vlachokyriakos, Clara Crivellaro, Christopher A Le Dantec, Eric Gordon, Pete Wright, and Patrick Olivier. 2016. Digital civics: Citizen empowerment with and through technology. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. ACM, 1096–1099.

[59] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[60] Peter West, Richard Giordano, Max Van Kleek, and Nigel Shadbolt. 2016. The Quantified Patient in the Doctor's Office: Challenges & Opportunities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3066–3078. https://doi.org/10.1145/2858036.2858445

[61] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 656, 14 pages. https://doi.org/10.1145/3173574.3174230