Assessment of bias in a crowdsourcing accessibility system using serious games

João Marcos de Mattos Barguil

Thesis Submitted
to the
Institute of Mathematics and Statistics
of the
University of São Paulo
for the
Doctorate Degree in Computer Science

Program:
Computer Science

Advisor:
Prof. Flávio Soares Corrêa da Silva
Co-advisor:
Prof. Fabio Kon

São Paulo, November 2023

Assessment of bias in a crowdsourcing accessibility system using serious games

This version of the thesis contains corrections and updates suggested by the Examining Committee during the defense of the original work, performed on November 23rd, 2023. A copy of the original version is available at the Institute of Mathematics and Statistics of the University of São Paulo.

Examining Committee:

- Prof. Flávio Soares Corrêa da Silva - University of São Paulo
- Prof. Stefania Bandini - University of Milano-Bicocca
- Prof. Areti Manataki - University of St Andrews
- Prof. Andrea Campagner - University of Milano-Bicocca
- Prof. Cassiano Isler - University of São Paulo

Abstract

Can persons without disabilities be good evaluators of accessibility? This question, often posed by persons with disabilities when looking at crowdsourced accessibility maps, is related to one of the most important unresolved issues of crowdsourcing: data quality control. Many of the recent ground-breaking advancements in machine learning depend on data annotation done by humans. Existing approaches for managing inaccuracies in crowdsourcing are based on validating output against preset gold standards, but they are unsuitable for subjective contexts such as sentiment analysis, semantic annotation, or measuring accessibility. While existing accessibility maps are largely centered in Europe and the United States, we built the largest database of such kind in Latin America. We detail techniques used for engaging over 27,000 volunteers who generated more than 300,000 data points over the course of 90 months, and a novel method for validating data quality in a context that lacks a definite ground truth. We tested it by applying concepts of serious games for exposing biases of different demographic profiles, and crowdsourced a different dataset for validating data quality. We found that persons without disabilities did not have worse performance than persons with disabilities, strong evidence that crowdsourcing can be a reliable source for accessibility data.

Keywords: crowdsourcing, social computing, serious games, accessibility.

Resumo

Pessoas sem deficiência podem ser bons avaliadores de acessibilidade? Essa questão, frequentemente colocada por pessoas com deficiência quando olham para mapas de acessibilidade construídos via crowdsourcing, está relacionada com uma das questões não resolvidas mais importantes do crowdsourcing: o controle da qualidade dos dados. Muitos dos avanços recentes em aprendizagem de máquina dependem de anotação de dados feita por humanos. Abordagens existentes para gerenciar imprecisões em crowdsourcing baseiam-se na validação de resultados em relação a padrões-ouro predefinidos, e por isso são inadequadas para contextos subjetivos como análise de sentimento, anotação semântica ou medição de acessibilidade. Enquanto os mapas de acessibilidade existentes estão centrados na Europa e nos Estados Unidos, construímos a maior base de dados desse tipo na América Latina. Detalhamos técnicas usadas para engajar mais de 27.000 voluntários que geraram acima de 300.000 dados ao longo de 90 meses, e um novo método para validar qualidade de dados em um contexto que não tem uma verdade básica. Nós o testamos aplicando conceitos de jogos sérios para expor vieses de diferentes perfis demográficos e coletamos um conjunto de dados diferente para a avaliação de qualidade. Descobrimos que pessoas sem deficiência não tiveram um desempenho pior do que pessoas com deficiência, uma forte evidência de que crowdsourcing pode ser uma fonte confiável de dados de acessibilidade.

Palavras-chave: crowdsourcing, computação social, jogos sérios, acessibilidade.

Acknowledgments

First and foremost, I want to thank my advisors prof. Flávio Soares Corrêa da Silva and prof. Fabio Kon for all the support throughout these years. More than once, they received me with open arms, offering guidance and encouragement when I parachuted back into the PhD after disappearing for long periods of time. Thank you for not giving up on me – even when I (almost) did.

This thesis would have been impossible without all the people behind Guiaderodas. Thank you to the tens of thousands of volunteers, to my dear friends Bianca Mahfuz G. da Silva and Bruno Mahfuz, and the other members of the team: Larissa Mahfuz, Melissa Siqueira, Cláudia Josimar Abrão, Tomaz Walter, Jefferson Almeida, Wellington Barbosa, Cristiane Kröhling, Beto Bernardi, Daniel Gaspar, Matheus Barbosa.

To my dear wife Paula (who still hasn't met a version of myself without a pending thesis), thank you for sharing life with me. Marrying you was by far the best decision I've ever made. Clara, thank you for joining us. We met not too long ago, and you already taught me what real love is. You mean the world to me.

I thank my parents, Ana and Francisco, for their everlasting and unconditional love, for encouraging me and always being there for me. I wouldn't have gotten anywhere without you. To all my friends and family, thank you for the moments we share. I cherish each and every one of you.

To Dr. Celso Charuri, thank you for showing me a higher reason for life.

This work was supported by FAPESP and FINEP grants 2018/09809-0, 2019/09913-5, and is part of the INCT of the Future Internet for Smart Cities funded by CNPq, grant 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and FAPESP, grants 2014/50937-1 and 2015/24485-9.

Contents

1 Introduction **1**

1.1 Crowdsourcing as a tool for scientific research 4

1.2 Challenges in crowdsourcing 8

1.3 Our contributions 11

1.3.1 Thesis organization 12

2 Crowdsourcing **13**

2.1 Crowdsourcing and games 15

2.2 Data quality in crowdsourcing 18

2.3 Bias in crowdsourcing systems 19

3 Measuring accessibility **21**

3.1 The need for accessibility 21

3.2 What is accessibility? 23

3.3 Accessibility standards 28

3.3.1 Architecture and design 28

3.3.2 Digital standards 31

3.3.3 Universal Design 33

3.4 Measurement methods 35

3.5 Crowdsourcing and accessibility 37

3.6 After all, can we measure accessibility? 40

4 Building a crowdsourcing platform for accessibility **43**

4.1 A brief history of Guiaderodas 44

4.2 Scope definition 45

4.2.1 Embracing subjectivity instead of avoiding it 47

4.2.2 Crowdwork design 48

4.3 Implementing a social impact business 51

4.4 Strategies for user engagement 53

4.5 Results 56

4.5.1 Crowd profile 56

4.5.2 Crowdwork output 57

4.5.3 Crowd acquisition and engagement 60

4.5.4 Crowd behavior 63

4.6 Chapter conclusion 65

5 A serious game for assessing data quality **67**
5.1 Game implementation 69
5.1.1 The game pilot 69
5.1.2 The real game 70
5.2 Results 74
5.2.1 Crowd profile 74
5.2.2 Crowd behavior 75
5.3 Chapter conclusion 85

6 Discussion **87**

Bibliography **91**

Chapter 1

Introduction

> *"Comes now Jeff Howe, contributing editor for Wired magazine, who recalls pitching an article idea in 2005 to Mark Robinson, his editor there, about how the Internet was helping businesses use amateurs to replace professionals. He reports that Robinson said, 'Hmmm … it's like they're outsourcing to the crowd.'*
>
> *'Or,' Howe informs me, 'I said, crowdsourcing. Frankly, I was joking. Silicon Valley's affection for portmanteaus is a bit of an inside joke at Wired. But Mark liked my story idea, and liked the word even more.' "* (Safire, 2009)

While anecdotal evidence suggests a rather humorous origin, *crowdsourcing* is far from a joke. Firstly mentioned in the groundbreaking article "The rise of Crowdsourcing" (Howe, 2006), it has since been successfully adopted in numerous industry and business applications, has supported scientific research, and has become an active field of scientific study. Howe (2008) defines crowdsourcing as *"the act of taking a task traditionally performed by a designated agent (such as an employee or a contractor) and outsourcing it by making an open call to an undefined but large group of people"*.

Crowdsourcing is rooted in the concept of *open-source software development*, in which independent individuals collaborate by sharing code to develop and refine programs (Lerner and Tirole, 2002). One of the most successful examples of the trend toward democratizing innovation (Von Hippel, 2006), open-source has enabled the creation of some of the technologies forming the backbone of the Internet, such as the Linux operating system (Torvalds, 1992). Community-led efforts catalyzed the revival of the popular real-time strategy game franchise *Age of Empires*, initially launched in 1997 and kept alive after the closure of the publishing studio by fans who were eventually hired by Microsoft to do it professionally (Moss, 2018).

A generalization of this idea is Wikipedia, an open-source encyclopedia that goes beyond code sharing: it supports the sharing of *knowledge* through sharing information and language. Any person can create or edit an article. It contains pages about historical events and mathematical concepts, as expected, but it also has information on surprisingly specific topics, such as competitive snowball fighting in Japan,[1] the controversy of whether ketchup is a vegetable or not,[2] and a detailed account (including a photograph!) of a 1916 American football match that ended with a 222-0 score.[3] The distributed, self-managed community has created more than sixty-one million articles in more than 300 languages since 2001, attracting over 1.5 billion unique visitors monthly as of August 2023 (Wikipedia, 2023).

Even when applied in contexts not related to software development, open-source implies direct access to the essential elements of a product. But this may not always be desirable (Brabham, 2008). If there are production costs involved, for instance, someone has to pick up the bill. A company investing to produce such a product expects enough sales to at least break even, and openly sharing what is normally protected by intellectual property laws may hurt this goal. Would open-source volunteers still participate when there are potential profits to be made from the results of their unpaid labor? Crowdsourcing overcomes these limitations *"by providing a clear format for compensating contributors, a hybrid model that blends the transparent and democratizing elements of open source into a feasible model for doing profitable business"* (Brabham, 2008).

To illustrate, take Fiat's case. What started as a marketing campaign to celebrate the 30th anniversary of its presence in Brazil evolved into a large scale open call for ideas for "a car to call it yours" (Nicolau *et al.*, 2011; Saldanha *et al.*, 2014). Over 11,000 suggestions were sent by more than 17,000 people from 160 countries in a 15-month long process, resulting in the prototype of *Fiat Mio*, or "My Fiat" in English, unveiled at the 2010 São Paulo Auto Show (Saldanha *et al.*, 2014). Although the car was designed with a lot of external input, intellectual property is owned by the company (and production costs too).

But Fiat's example is still a somewhat conservative approach. There was no financial incentive for the crowd, and their reward was as simple as "feeling included in the process". By taking the concept to more extreme lengths, it is possible to create *crowd-based business models* (Kohler, 2015; Täuscher, 2017) that were utterly unimaginable not so long ago. YouTube monetizes from ads shown in videos it does not produce, but by sharing part of its revenue with content creators, it positions itself as a potential source of income for its users (Susarla *et al.*, 2012). In fact, a

[1] https://en.wikipedia.org/wiki/Yukigassen (retrieved Sep. 2021).

[2] https://en.wikipedia.org/wiki/Ketchup_as_a_vegetable (retrieved Sep. 2021).

[3] https://en.wikipedia.org/wiki/1916_Cumberland_vs._Georgia_Tech_football_game (retrieved Sep. 2021).

whole new profession has emerged: *YouTubers*, people who turned posting videos on their channels into a full-time job, making (a sometimes large amount of) money and becoming celebrities in their niches. AirBnb and Uber also use a similar approach: crowdworkers get paid whenever their properties are rented or when they are called to give someone else a ride – using their own car, naturally (Kohler, 2015; Tong *et al.*, 2020).

Still, what the crowd gains from participating does not necessarily have to be cold hard cash. *Gamification* (Deterding *et al.*, 2011) can be another useful tool for increasing participants engagement (Feng *et al.*, 2018; Kavaliova *et al.*, 2016; Morschheuser *et al.*, 2016; Robson *et al.*, 2015). Car traffic information on Waze is completely provided by users (Wang *et al.*, 2016). While the company also monetizes through ads, its crowd compensation mechanism is not financial: it adopts a game-like fashion, and users earn points by completing tasks such as reporting incidents (*active* data collection) and using the app whilst driving (*passive* data collection).

These are examples of successful relationships with *the crowd* – a faceless and powerful force, which may not always be benevolent. Any crowd-based business model is inherently influenced by the *network effect* (Shapiro and Varian, 1998), *i.e.,* it becomes more valuable for each independent user as more people participate, and conversely, less valuable if the user base shrinks. There is a lot more complexity into this particular aspect (Boudreau and Jeppesen, 2015; Casadesus-Masanell and Hałaburda, 2014; Rochet and Tirole, 2003; Täuscher, 2017), but one thing is certain: nurturing a healthy two-way relationship with the audience is crucial, as festering dissatisfaction may ultimately result in full network collapse.

In early September 2021, two platforms of such kind witnessed coordinated protests by some of their most prominent users (D'Anastasio, 2021). A live streaming platform for gamers called Twitch saw some of its top streamers go on a strike to push the company to end a wave of harassment they had been suffering. On the same day, moderators of a few of the largest communities on Reddit shut down their *subreddits* (as the thousands of forums about varied themes are called) to object the administration's perceived lenient posture against the spread of COVID-19 misinformation. Both companies were quick to respond, and even though the two protests were completely unrelated and did not cause a massive flop, they show a glimpse of the strength of a resentful crowd.

There are also stories that ended in tragedy, like Tumblr. When the micro-blogging platform decided to ban pornography, it ended up also censoring artists and people with marginalized gender and sexual identities, which comprised a large portion of the user base. The controversy devolved into an absolute *crowdpocalypse*, and the company spiraled down until it was finally sold in 2019 for

$3 million – a 99.72% loss from the $1.1 billion paid by Yahoo! just 6 years before (Siegel, 2019).

Despite users riots being a risk for any crowdsourcing initiative, a pure "the community is always right" approach may be foolish, as it can turn the company into a hostage of its own audience. The behavior of the crowd is unpredictable, and a lack of control mechanisms such as defining a code of conduct and guidelines for participation may result in unintended platform use (Täuscher, 2017). This was the case of OnlyFans, which offers a subscription-based solution for helping creators to monetize their work. Due to its permissive content policy, it quickly became a leading marketplace for selling self-created amateur pornography. In August 2021, the company announced it would ban sexual material, citing pressure from its banking partners. After backlash from the people who were in large part responsible for the platform's success and relied on it as a source of income, the decision was hastily reversed in an attempt of damage control (Browne, 2021).

1.1 Crowdsourcing as a tool for scientific research

Crowdsourcing evolved in the early XXI century as a way to tap into our collective intelligence (Lévy, 1997) for producing better answers to problems. This potential has been used in a wide range of different scientific contexts: public health (Brabham *et al.*, 2014), database systems (Franklin *et al.*, 2011; Marcus *et al.*, 2011), information visualization (Borgo *et al.*, 2018), peer-to-peer money lending – also called *crowdfunding* (Moysidou and Hausberg, 2020), the stock market (Cappa *et al.*, 2019; Kolmonen, 2017), predicting decisions of the United States Supreme Court (Katz *et al.*, 2017), grading homeworks (de Alfaro and Shavlovsky, 2013; Wang *et al.*, 2020, 2019), among many others. As a response to the COVID-19 threat, several academic and research entities signed an agreement for the rapid sharing of findings, ensuring open access to related publications as soon as they were submitted to peer review (Wellcome, 2020). This unprecedented collaboration shows how crowdsourcing can be more than a tool for data collection; it can also be a philosophy of work.[4]

Crowdsourcing is a form of social computing, a discipline related to computational social science (Giles, 2012; Lazer *et al.*, 2009), but not the same. Human and social behavior are central to both, but computational social science focuses on combining traditional social science studies and the capacity to collect and analyze data in great depth and scale (Lazer *et al.*, 2009), while social computing has a much more deliberate focus on engineering systems that are hybrids of humans and machines (Chen *et al.*, 2016). Notwithstanding, crowdsourcing has been a valuable tool for

[4] In a broader sense, even Science itself can be understood as a form of crowdsourcing: an undefined but large group of people answering to an open call for expanding the boundaries of human knowledge.

computational social science (Behrend *et al.*, 2011; Chandler *et al.*, 2014; Mason and Suri, 2012).

Another field that has also greatly benefitted from crowdsourcing is machine learning, due to its inherent necessity of large amounts of training examples. Crowdwork has been used to generate, enrich, and/or clean datasets for imitation learning (Chung *et al.*, 2014), natural language processing (Snow *et al.*, 2008), and computer vision (Kovashka *et al.*, 2016), for instance. Created for preventing unintended automated website usage (*e.g.,* screen scraping, spam), CAPTCHAs[5] are based on hard AI problems, *i.e.,* problems that computers cannot solve, but are relatively easy for humans (Von Ahn *et al.*, 2003, 2004). Earlier CAPTCHAs were based on deciphering distorted characters, and their widespread adoption meant that users were soon typing millions of characters every day. This huge crowdsourcing potential was used for more productive goals by a new version of the system, dubbed reCAPTCHA (Von Ahn *et al.*, 2008). It helped to digitize printed material by having users identify words that computerized optical character recognition (OCR) failed to recognize, and verifying answers by coupling unknown words with words for which the answer was previously known, as seen in Figure 1.1. This innovation helped to not only digitize large amounts of books and other printed texts, but it also supported significant improvements in OCR – which, in turn, made reCAPTCHA obsolete. Newer versions of reCAPTCHA replace simple word recognition by problems based on image classification. But, of course, this is a cat-and-mouse game, as advancements in computer vision allow the creation of automated solvers with ever-increasing power (Akrout *et al.*, 2019; Hossen *et al.*, 2020; Sivakorn *et al.*, 2016; Zhou *et al.*, 2018).

A tool that has been largely used for building datasets at a relatively low cost is Amazon Mechanical Turk (AMT), a marketplace that allows hiring remote crowdworkers to perform on-demand *Human Intelligence Tasks (HITs)* such as labeling images, validating data, moderating content, or even participating in surveys. (Mason and Suri, 2012; Paolacci and Chandler, 2014; Sheehan, 2018). This solution, however, creates its own set of problems. Crowdworkers are paid well below the federal minimum wage in the United States, and even less in India (Hara *et al.*, 2019). This severe underpayment can hurt work quality. More importantly, all academic studies and companies using AMT at any given time are sampling from the same pool of workers, which has been estimated to be around only 7,300 people (Stewart *et al.*, 2015). This pool is not representative of the general population (Arditte *et al.*, 2016) and participants who speak English as a second language may also have a significantly lower performance (Buhrmester *et al.*, 2018).

Sampling from an arguably small and potentially biased crowd can be dangerous, as applications such as machine learning algorithms are really sensitive to the quality of its training dataset – their

[5] CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart.

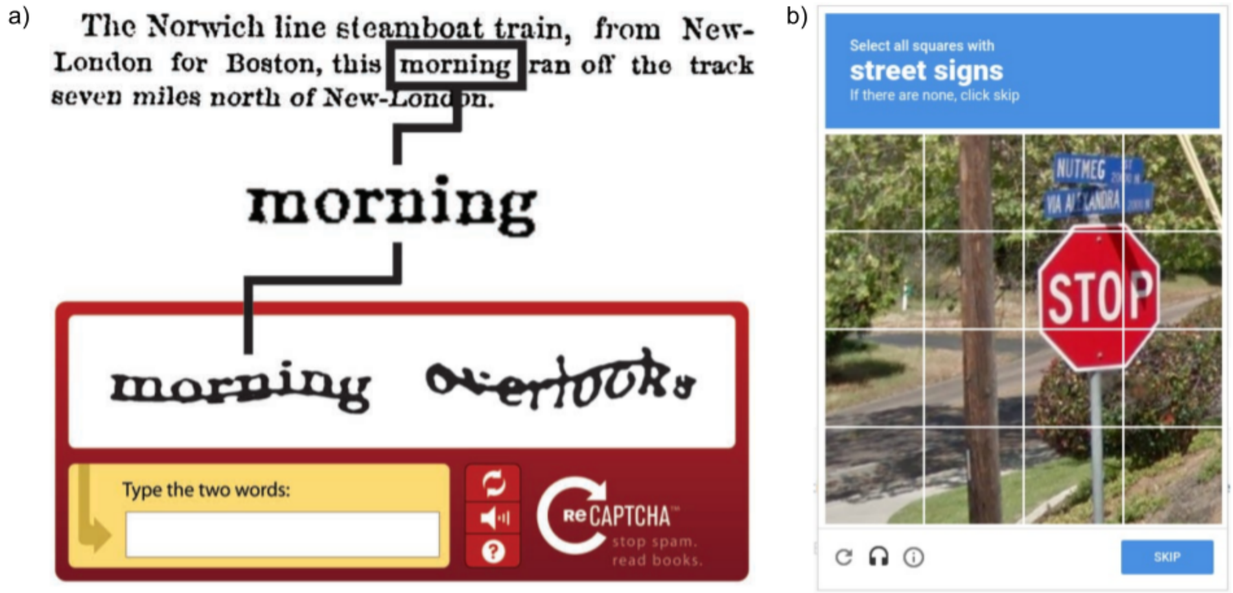
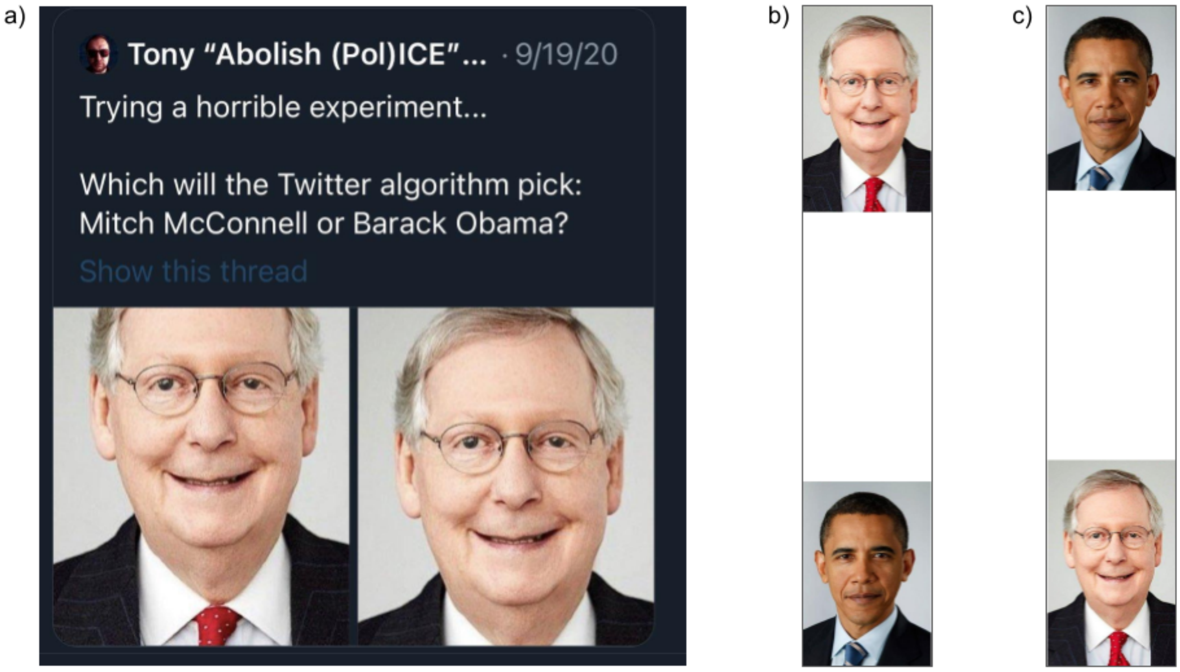

a)
The Norwich line steamboat train, from New-London for Boston, this morning ran off the track seven miles north of New-London.
morning
morning overlooks
Type the two words:
reCAPTCHA™
stop spam. read books.
b)
Select all squares with
street signs
If there are none, click skip
SKIP

Figure 1.1: *reCAPTCHA versions 1 and 2: the first version of the system (Figure* a*) prompted users to identify words that OCR failed to recognize ("morning", in this example). For verification, unknown words were coupled with words for which the answer was previously known ("overlooks"). Both words are distorted by random transformations. Figure* b *shows reCAPTCHA v2, which replaces simple word recognition by problems based on image classification, such as recognizing street signs. Image sources: (Von Ahn et al., 2008) (left) and (Zhou et al., 2018) (right).*

output is (at most) as good as their training input. In fact, "poor data quality" does not necessarily mean *wrong* data, it may just be biased, like overrepresenting a specific group, producing results that may not generalize well to the real world. Sourcing from a larger, "fresher" crowd may reduce the risk of introducing undesirable biases in the resulting dataset, but in practice, this is not what usually happens: due to the high cost of creating large scale datasets for training machine learning algorithms, several different applications end up using the same small number of freely available open-source datasets. Shankar *et al.* (2017) analyzed two of such datasets widely used in image classifiers, and found a considerable amerocentric and eurocentric bias, yielding poor performance on images from different locales. Representation bias is a challenge for building applications for use in the developing world.

The case of Twitter's racist image-cropping algorithm illustrates this. A biased training dataset is a plausible cause for the controversy that started after a user posted images with the same two pictures of a black man and a white man, but in different positions.[6] Because the images were too tall, they were automatically cropped in the square previews displayed in the tweet's body. In both cases, the white man's face was chosen over the black man's, proving that the algorithm prioritized

[6] https://twitter.com/bascule/status/1307440596668182528 (retrieved Nov. 2021).



Tony "Abolish (Pol)ICE"... · 9/19/20

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

Figure 1.2: *Twitter's racist algorithm: images that are too wide or too tall are automatically cropped in the preview displayed in a tweet's body. The algorithm consistently prioritized white faces over other ethnicities. In the example above, a user posted the same pictures of two American politicians in different positions (Figures* b *and* c*), and in both cases the white man was shown in the preview (Figure* a*). The algorithm has since been disabled and replaced by a simpler position-based strategy, and the same page currently displays two white squares corresponding to the center of the images.*

caucasians over other ethnicities. Figure 1.2 shows both images and their corresponding cropped previews. Twitter issued an official public apology and disabled the algorithm, and, in addition, also sponsored an open call for developers and researchers to "demonstrate what potential harms such an algorithm may introduce".[7] The winning participants of the crowdsourcing initiative showed that the algorithm could be tricked by beauty filters, helped to perpetuate marginalization, was also racist against texts in arabic scripts, and even favored lighter skinned *emojis*.[8]

As this case exemplifies, data bias is a relevant factor for machine learning and other applications. Many other datasets widely used by the industry and academia have the same issue (Buolamwini and Gebru, 2018). Even though there have been advances in making these algorithms more fair, there is still a significantly lower accuracy for detecting darker skinned faces (Raji and Buolamwini, 2019). Racist machine learning algorithms have already caused wrongful imprisonments and denied vital medical treatment to racial minorities (Heaven, 2020; Obermeyer *et al.*, 2019). There is a pressing need for correcting – or at least detecting – inherent biases in large datasets.

Unfortunately, crowdsourced data in its pure, raw state is of questionable quality by default. There is considerable variability between different individuals (Barowy *et al.*, 2012; Franklin *et al.*,

[7] https://hackerone.com/twitter-algorithmic-bias (retrieved Nov. 2021).

[8] https://twitter.com/TwitterEng/status/1424819778397511680 (retrieved Nov. 2021).

2011), and even when there is little variance, results can be consistently bad due to several reasons: the hired crowd may lack the necessary skills, tasks may have been poorly planned and/or explained by the developers (Dahlander and Piezunka, 2020), or worse, the crowd may actively choose to behave maliciously (Naroditskiy *et al.*, 2014; Wang *et al.*, 2013).

Malicious conduct was the demise of Tay, Microsoft's chat bot that learned from conversations with users. As originally stated in the project's (now offline) website, *"the more you chat with Tay the smarter she gets, so the experience can be more personalized for you"*. User input was fed into the artificial intelligence completely unfiltered, and as a result, Tay went from tweeting "humans are super cool" to spewing racist, genocidal messages in less than 24 hours after launch (Kraft, 2016). Tay learned from hateful messages collectively fed by users, and instead of being the fun, young persona the creators originally intended, she ended up praising Hitler and posting about how much she "hated feminists". The bot was quickly shut down after the fiasco, but the learning remains: it is imperative to, somehow, evaluate the quality of crowdsourced data.

1.2 Challenges in crowdsourcing

Stories of *failed* crowdsourcing initiatives are often unreported, causing a huge success bias in the literature. This may create a (false) perception that there is a lack of challenges in crowdsourcing. But 90% of the cases studied by Dahlander and Piezunka (2020) failed because they were unable to form a crowd. And as the CEO of the failed crowdsourcing startup Cambrian House tells, *"the wisdom of crowds worked well in the model, but it was our participation of crowds aspect which broke down"* (Schonfeld, 2008). A large number of publications focus on applications of crowdsourcing, not on its kernels and essentials (Zhao and Zhu, 2014). Besides, most studies are purely theoretical (Lenart-Gansiniec, 2018) and do not tackle the issue of raising a crowd.

Crowdsourcing machines are socio-technical systems (Geiger and Schader, 2014; Kittur *et al.*, 2013), *i.e.,* part human, part technology. Therefore the *social* aspect is fundamental and must be carefully taken into consideration (Chen *et al.*, 2016). This presents the first major challenge in crowdsourcing: forming a *crowd* to do the *sourcing* – or, more formally, user *acquisition* and *engagement* (Dahlander and Piezunka, 2020; Morschheuser *et al.*, 2016; Täuscher, 2017). Understanding the crowd's motivation to participate in various contexts is meaningful, both from a community perspective and from the viewpoint of an individual (Zhao and Zhu, 2014).

Many have avoided the user engagement challenge by spending money: they pay crowdworkers for each completed task (Kittur *et al.*, 2013), *e.g.,* using Amazon Mechanical Turk (Franklin *et al.*, 2011; Marcus *et al.*, 2011; Mason and Suri, 2012; Sheehan, 2018). Others have applied gamification techniques (Morschheuser *et al.*, 2016), making crowdwork more entertaining and leading users into performing useful computation without even realizing they are doing so (Von Ahn and Dabbish, 2008). Nonetheless, crowd engagement remains an open problem (Kaartemo, 2017).

Quality control represents the second major concern in crowdsourcing (Correia *et al.*, 2020; Lease, 2011). Determining the level of trustworthiness of contributors is an open challenge (Bouguettaya *et al.*, 2017). The simplest approach is to use redundancy and aggregating answers by majority vote (Sheng *et al.*, 2008; Snow *et al.*, 2008), but this is not ideal because it increases costs and it can break down due to many reasons, such as workers agreeing to coordinate answers, work instructions being unclear or badly designed, or even a simpler case of workers having low expertise (Eickhoff, 2014; Kittur *et al.*, 2013). More sophisticated approaches try to estimate the quality of an individual worker (Burnap *et al.*, 2015; Ipeirotis *et al.*, 2010), *e.g.,* by "calibrating" a volunteer's reliability compared to a gold standard provided by experts (Sorokin and Forsyth, 2008; Wang *et al.*, 2019), or by detecting spam (Wang *et al.*, 2016) and malicious conduct (Wang *et al.*, 2013).

These methods, however, are not enough (Faltings *et al.*, 2014; Kairam and Heer, 2016), as people are susceptible to different kinds of cognitive biases (Bless and Fiedler, 2014; Haselton *et al.*, 2015), such as the Dunning-Kruger and the Bandwagon effects (Bikhchandani *et al.*, 1992; Kruger and Dunning, 1999). They impact the behavior of crowdworkers and, as a consequence, the quality of results produced (Eickhoff, 2018; Saab *et al.*, 2019). Their influence on performance should be taken into account.

But most importantly, the literature most commonly focuses on objective cases where the notion of a "correct" answer can be defined, when there is a theoretically ideal function f that yields the correct answer, and given a crowdsourced function c, the error is $e(x) = f(x) - c(x)$. Usually, f is unknown (otherwise there would be no need for crowdsourcing), but heuristics can be applied for estimating accuracy.

But what about cases where the very *thing* we're measuring is subjective and/or interpretative, *i.e.,* our ideal function f *does not exist?* What is the correct answer for an opinion-based question like *"which of the pictures below better represents 'happiness'?"* There is no published, undebatable definition for what "happiness" is. It can be financial success, romantic love, or, as the Beatles sang, maybe even a warm gun. Measuring quality in these scenarios seems to be an open problem (Kittur *et al.*, 2013), and traditional methods like majority voting are unsuitable for subjective tasks (Haralabopoulos *et al.*, 2020). Thanking participants and explaining the importance of requested tasks can increase workers' motivation (Chandler and Kapelner, 2013), and combining subjective work with factually verifiable questions (*e.g., "who is the president of the United States?"*) may allow to measure a worker's attention (Chandler *et al.*, 2013), but even these improvements might be questionable (Paolacci and Chandler, 2014). Furthermore, being motivated and paying attention are not enough if the person is unskilled and/or biased.

An example of this scenario is accessibility. It has some of the subjectivity seen in human sciences like Psychology, yet it is possible, at least empirically, to decide if an environment is accessible or not – a restaurant on top of a flight of stairs is unarguably not accessible for persons in wheelchairs. But there is hardly a consensus on a definition for accessibility, let alone what the "correct" answer would be in any given case. There is no measurement method for accessibility, despite its ever-increasing importance: persons with disabilities, persons with mobility impairment, elderly population (specially considering the worldwide trend of aging population), they all benefit from more accessible environments.

Crowdsourcing has been applied for accessibility, but most approaches tried to avoid subjectivity: instead of direct measurements, they quantify some objective proxy for it, *e.g.,* pedestrian obstacles on sidewalks (Saha *et al.*, 2019; Weld *et al.*, 2019). While a valid approach, this is not accessibility *per se*. Additionally, most studies have a rather limited scale. The crowdsourcing studies analyzed by Mack *et al.* (2021) have a median of only 153 participants.

During the 2010s many *Volunteered Geographic Information (VGI)* initiatives for mapping accessibility were created, such as Wheelmap in Germany (Mobasheri *et al.*, 2017), AXS Map in the United States[9] and Euan's Guide in the United Kingdom[10], among others. Despite being probably the most successful kind of crowdsourcing for accessibility, these platforms receive a common critique about data reliability as due to their open, collaborative nature, information could have been provided by persons who "do not know what accessibility is".

Can we use crowdsourcing to increase accessibility? Is it possible to include accessibility's subjectivity in the process and still be able to somehow gauge data quality? Even if there is no way implement a boolean "right or wrong" filter, there is value in at least being able to detect existing biases in a sample.

[9] https://axsmap.com (retrieved Sep. 2021).

[10] https://euansguide.com (retrieved Nov. 2021).

1.3 Our contributions

We built a crowdsourcing accessibility platform that is, to the best of our knowledge, the biggest dataset about a Latin American country, helping to fill a gap for a historically underserved region. Subjectivity was embraced into the implemented evaluation methodology, while still keeping it simple enough for persons without disabilities to participate. Continuously available since launch in early 2016, the platform has gathered over 300,000 data points from more than 27,000 volunteer participants. While this cannot be considered a scientific contribution *per se*, it is worth mentioning given its relevance in practice. Furthermore, it also provides the basis for testing other factors in real-world settings, something the literature on crowdsourcing largely lacks.

This system was also used as a platform for experiments on user acquisition and engagement on a real, large-scale setting. After implementing design changes and creating simple games and gamified content in social media, the monthly mean of user signups increased more than ten times, and crowdworkers engagement increased by almost six times.

We also peered into the challenge of ensuring data quality in crowdsourcing. We developed a method for detecting behavioral biases in a context where it is not possible to define a ground truth for evaluating the quality of individual contributors. The method was tested by having a small set of experts define a surrogate truth, used as a basis for post-hoc comparative analyses of data collected through a serious game.

We found that the behaviors of persons with and without disabilities had no statistically significant differences when evaluating good accessibility, but they diverge significantly when reviewing bad accessibility. Gender, age, and proximity to persons with disabilities did not correlate with any behavioral differences. We also found that for all demographic profiles, users are less sure of their conclusions when reviewing bad accessibility. These results indicate that persons without disabilities are not worse sources than persons with disabilities, an evidence that the crowdsourced data is of homogeneous quality.

In summary, the research questions addressed in this thesis are, first, how platform design and gamification affect crowd acquisition and engagement in real-world, large-scale crowdsourcing platforms. Secondly, we also investigate whether it is possible to gauge the quality of crowd contributions in subjective contexts that lack a ground truth.

1.3.1 Thesis organization

Chapter 2 outlines relevant literature on crowdsourcing. In chapter 3, we discuss what accessibility is, and crowdsourcing initiatives for mapping accessibility and other proxy measurements. Chapter 4 presents our method for implementing a crowdsourced accessibility platform. Chapter 5 describes our method for gauging behavioral biases in accessibility, and results from applying it for assessing data quality in our system. Finally, in chapter 6 we discuss the impact of the performed research, how it relates with existing literature, and possible paths for future related work.

Chapter 2

Crowdsourcing

All crowdsourcing systems are not the same. They can obviously have different mechanics and demand distinct kinds of work, but most importantly, their purposes can be different. Designing crowd tasks, defining strategies for user acquisition and engagement, processing results, and dealing with the unpredictable challenges that naturally arise after launch; these tasks are all sensitive to context and there is no one-size-fits-all solution. Copying strategies from the past without analysis does not guarantee success. It is fundamental to understand the nature of the system being built and plan accordingly.

Geiger and Schader (2014) classify crowdsourcing systems according to two attributes. First, whether the system expects homogeneous or heterogeneous output from the crowd. If participants perform the same kind of task and produce the same kind of result, the system is *homogeneous* and contributions are valued equally. If not, the system is *heterogeneous* and it may value contributions according to their individual qualities. The second analysis is about how the system derives value from crowdwork: *non-emergent* systems derive value directly from every single contribution, whereas for *emergent* systems it is the combination of multiple individual contributions that matters. Based on these two aspects, the authors define four different kinds of crowdsourcing systems, summarized in Figure 2.1.

Crowdprocessing systems are homogeneous and non-emergent, meaning that crowdworkers generate the same kind of output and each individual contribution is valuable by itself. They are usually scaling solutions for dealing with large amounts of identical chunks of work, like reCAPTCHA and other systems for generating datasets for machine learning.

Crowdrating systems are also homogeneous, but they derive value from the aggregation of a large number of contributions, not from the contributions themselves. They are usually applied for

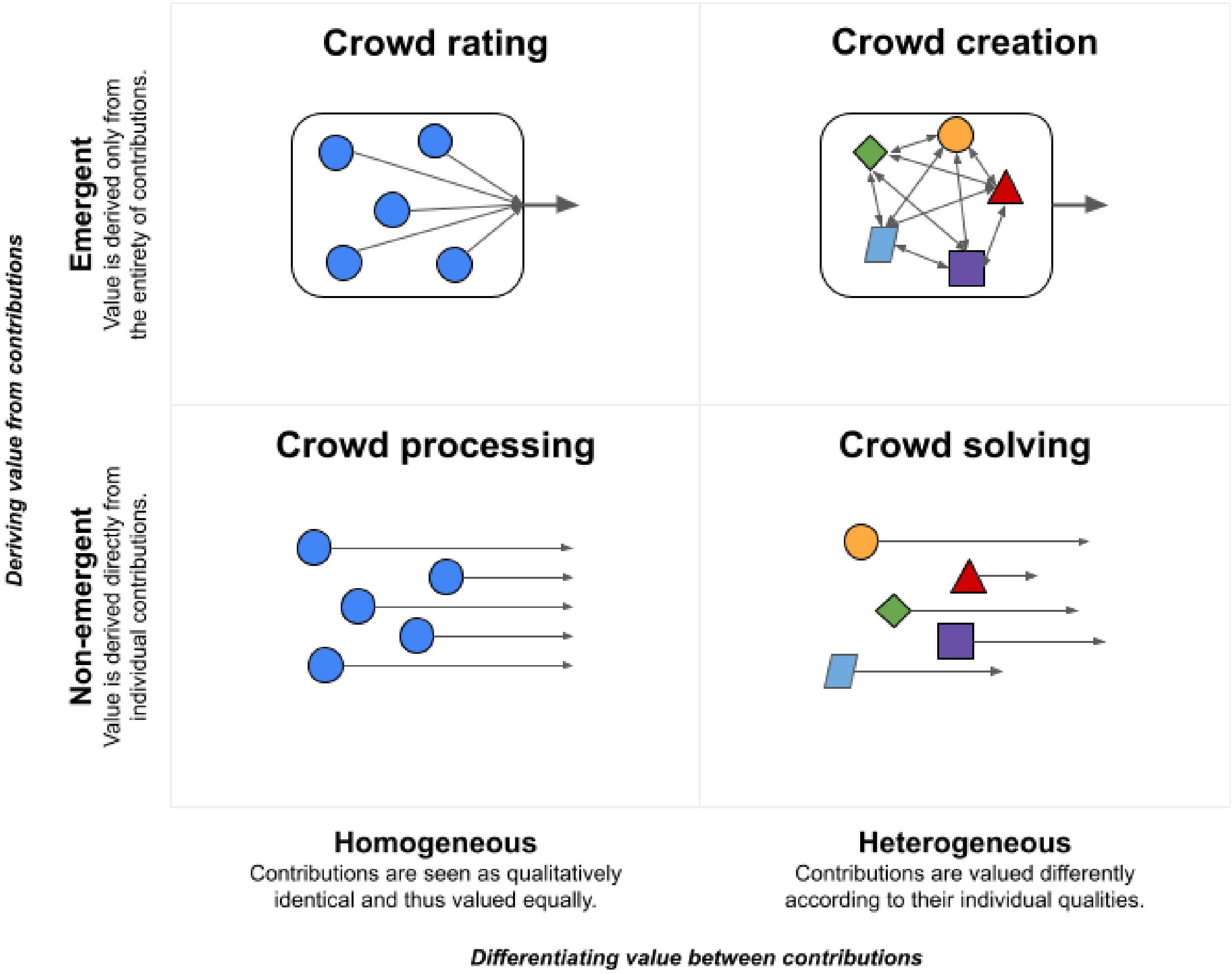

Crowd rating
Crowd creation
Crowd processing
Crowd solving
Deriving value from contributions
Emergent
Value is derived only from the entirety of contributions.
Non-emergent
Value is derived directly from individual contributions.
Homogeneous
Contributions are seen as qualitatively identical and thus valued equally.
Heterogeneous
Contributions are valued differently according to their individual qualities.
Differentiating value between contributions

Figure 2.1: *The four archetypes of crowdsourcing information systems (Geiger and Schader, 2014).*

drawing conclusions based on the "wisdom of crowds" (Surowiecki, 2004), in which the average of all guesses is, in average, a good guess. This approach has been used by NASA: volunteers visited a website, looked at photographs of the surface of Mars and clicked where they thought there was a crater. Weighted clustering of clicks/votes resulted in an accuracy comparable to traditional methods for surveying astronomical bodies (Kanefsky *et al.*, 2001).

Crowdsolving systems are heterogeneous and non-emergent, meaning that contributions are inherently different – in general, alternative or complementary solutions to a given problem. This kind of system can be used for very complex problems and ideation contests, similarly to IdeaNote,[1] a platform that helps enterprises to collect suggestions from clients and internal staff.

Finally, *crowdcreation* systems seek a collective value from heterogeneous contributions and their relationships. Platforms based on user-generated content are a typical example, such as YouTube (it is the collection of videos that matters, not each isolated video) and Wikipedia (its value comes from the compilation of knowledge built by contributions of different kinds).

[1] https://ideanote.io/ (retrieved Dec. 2021).

2.1 Crowdsourcing and games

Incentive mechanism design is a challenge for crowdsourcing systems (Zhao and Zhu, 2014). Games can provide useful tools for this process, as monetary payments are not the only possible motivation for participants (Kaufmann *et al.*, 2011). In particular, serious games may be specially helpful, or *"games that do not have entertainment, enjoyment, or fun as their primary purpose"* (Michael and Chen, 2005). Their origin is not so obvious and they might be even older than entertainment video games (Djaouti *et al.*, 2011b), but they have been successfully applied in areas as diverse as military, education, healthcare, business, etc.

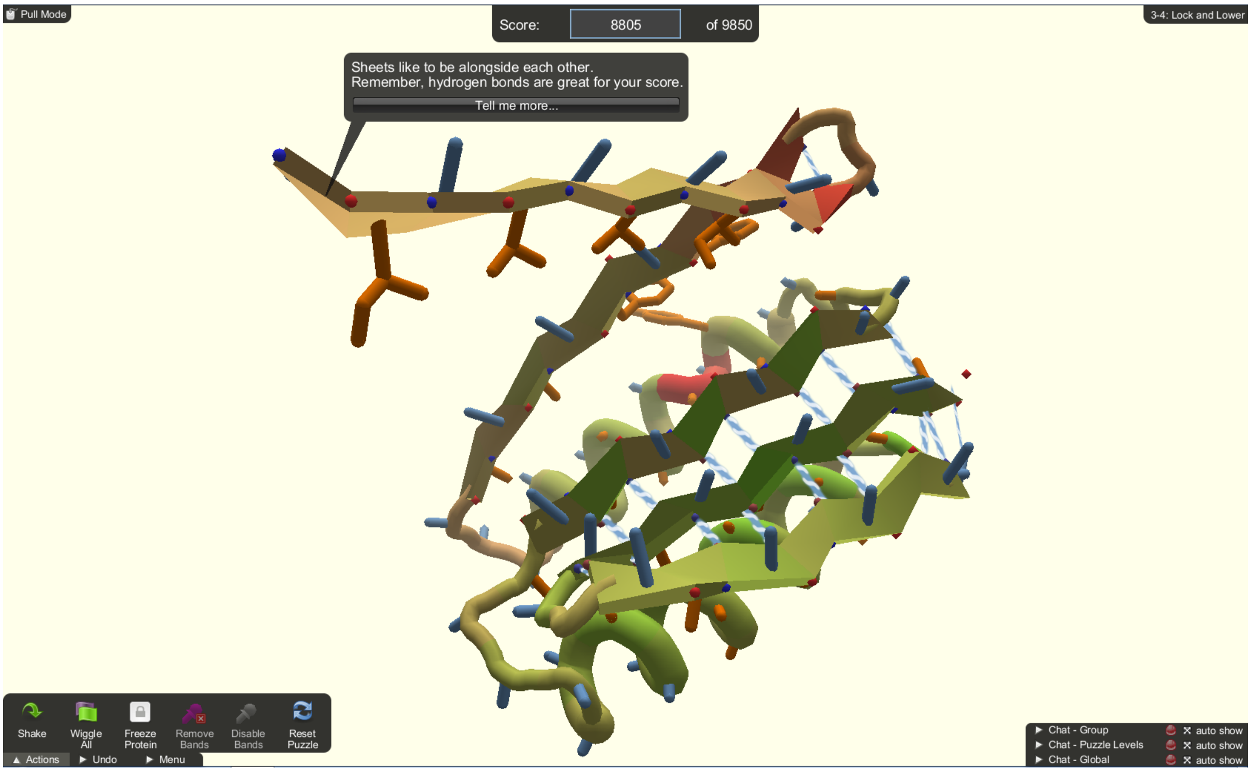
There are many possible classifications for serious games. Djaouti *et al.* (2011a) suggest a classification based on purpose that defines games to *broadcast a message* as the first type of game. This includes several of the most commonly seen applications, such as educative (*"Edugames"*), informative (*"Newsgames"*), persuasive (*"Advergames"* and *"Political games"*), and subjective (*"Military games"* and *"Art games"*). While they may initially seem unrelated, their primary purpose is exactly the same: to transmit some sort of information or knowledge, even though the nature of their messages can vary.

The second type of games are serious games for *training*. Their goal is to improve cognitive performance or motor skills. They can be powerful tools for supporting practice without incurring in real costs, risks, etc. Space flight and surgery simulators for astronauts and medicine students are possible examples.

The third and last type are games for *exchange of data*, aimed at collecting information from players or encouraging exchanges between them. These *games with a purpose* (Von Ahn, 2006) are a form of crowdsourcing, like Foldit (Cooper *et al.*, 2010), a *crowdsolving* system implemented as a puzzle in which players manipulate protein shapes and try to find the best way to fold them (Figure 2.2). Player solutions are used for supporting scientific research, as predicting protein shapes is an important step for understanding diseases and developing new drugs.

Games for exchange of data have also been applied as *crowdprocessing* systems, generating datasets that would otherwise be labor-intensive and costly to create (Von Ahn, 2006). People's natural desire to be entertained can lead them to perform useful computation without even realizing they are doing so – they are not interested in solving computational problems, but they will willingly do it if they perceive it as being fun (Kavaliova *et al.*, 2016; Von Ahn and Dabbish, 2008).

For instance, in the *ESP Game* (Von Ahn and Dabbish, 2004), randomly paired users had to guess how the other player would label a given image. Because they could not otherwise commu-



Score: 8805 of 9850

Sheets like to be alongside each other.
Remember, hydrogen bonds are great for your score.

Figure 2.2: *Foldit is a puzzle game designed to predict protein shapes via crowdsourcing. Players manipulate proteins until they find a shape that maximizes their score. Developed by a collaboration of several universities, the game's purpose is to collect player solutions which are then used for supporting scientific research.*

nicate, players ended up agreeing upon labels that were good descriptions for the images. Breazeal *et al.* (2013) developed a game that allowed two players to collaborate to solve a task. One player took the role of a robot avatar, and the other a human avatar, each with a different set of capabilities. They monitored how players coordinated to overcome challenges and complete the task, and the resulting dataset was used for training an autonomous robot which was tested in a real-world reproduction of the online game. Their results indicate that the autonomous robot behavior matched the performance of a human-operated robot in several important measures.

Fortunately, integrating games into crowdsourcing does not have to be all-or-nothing. It is possible to apply *gamification*, *i.e.*, make use of game design elements in non-game contexts (Deterding *et al.*, 2011). There are several features that can be integrated: points/scores, time keeping, progress levels (Von Ahn and Dabbish, 2008), badges/achievements (Easley and Ghosh, 2016), leaderboards, rewards, storytelling, missions (Morschheuser *et al.*, 2016), among others.

Prandi *et al.* (2016, 2017b) compared three implementations for the same system: a basic data collection tool, a gamified experience, and a full-fledged game. They analyzed the behavior of a group of students mapping accessibility barriers in a defined urban area. The first variant of the app had no gaming experience, the second had a rewarding mechanism that gave away discount coupons, and the third was *Geo-Zombie*, an augmented reality game in which players had to submit reports to collect ammunition for defending from zombie attacks. Both the rewarding tool and the game received approximately three times more reports than the basic app. Engagement, however, was different. Almost all users submitted around the minimum amount of reports necessary for earning a free voucher in the rewarding tool, while the game provoked more polarized behavior: many students hardly played it, but a few felt emotionally captivated and submitted considerably larger numbers of reports. This indicates that both strategies positively affect engagement and can yield comparable amounts of data, but a game may be able to do so from a smaller (but more committed) audience (Prandi *et al.*, 2017b).

In fact, the literature shows that mere implementation of gamification mechanisms does not automatically lead to significant increases in use activity. However, players who engage with the system and actively participate show increased levels of participation (Hamari, 2013).

Different types of crowdsourcing systems have distinct goals, therefore they should not always implement the same gamification strategies. Morschheuser *et al.* (2016) analyzed multiple studies and found that systems that demand monotonous tasks (*i.e.,* homogeneous systems) most commonly use simpler game mechanics like points and leaderboards. Most *crowdprocessing* approaches reward the quantitative number of fulfilled tasks, as homogenous non-emergent tasks are easily enumerable. In *crowdrating*, not only the quantity, but the quality is also praised: game elements are used to motivate users to emulate others and to "think and act like the community", *e.g.,* by giving extra points according to the degree of agreement with contributions of other crowdworkers. More sophisticated gamification methods have been used in systems that seek for diverse and creative input (*i.e.,* heterogeneous systems). The authors could not identify clear design patterns in these cases, and due to their sensitivity to context, implementation choices usually depend on the possibilities to measure task fulfillment and/or task quality. Nonetheless, it seems that explicitly expressing gamification rewards before a *crowdsolving* task can be beneficial (Choi *et al.*, 2014), and *crowdcreation* systems should focus on motivating users towards cooperation and creativity.

Gamification is a powerful tool for enhancing crowdsourcing systems. It can not only increase engagement by tapping into users' intrinsic motivation (autonomy, competence, and socialization) (Xi and Hamari, 2019), but it can also positively affect the quality of produced work (Choi *et al.*, 2014; Ghosh and Hummel, 2014; Morschheuser *et al.*, 2016).

2.2 Data quality in crowdsourcing

By definition, crowdsourcing relies on the work of a large number of people with uncertain levels of skill (and possibly intent). Not surprisingly, quality control is a major concern (Correia *et al.*, 2020; Lease, 2011). The simplest approach is to apply some sort of majority vote among sourced answers (Sheng *et al.*, 2008; Snow *et al.*, 2008). This, however, can break down due to many reasons, such as workers agreeing to coordinate answers, work instructions being unclear or badly designed, or even a simpler case of workers having low expertise (Eickhoff, 2014; Kittur *et al.*, 2013). It also incurs a higher cost, as participants must perform some fraction of redundant work.

Nonetheless, several published studies have successfully applied some variant of majority voting, showing it can increase overall quality indeed. Ghosh and McAfee (2011), for instance, used a simple filter in which participants vote on user-generated content and submissions that are not uniformly rated positively are discarded. Two somewhat similar alternatives are input and output agreement. In output agreement, two or more participants are shown the same input, and they must agree on some sort of output, like in the aforementioned *ESP Game* (Von Ahn and Dabbish, 2004). Input agreement is the complementary approach: players are provided with either the same or a different object, and asked to describe that object to each other. Based on each other's descriptions, they must decide whether they have the same object or not (Law and Von Ahn, 2009). This is the method of *TagATune* (Law *et al.*, 2007), a game for collecting tags for music clips.

The *surprisingly popular* algorithm (Prelec *et al.*, 2017) extends the notion of majority voting by selecting the answer that is more popular than people predict. After casting their own vote, participants also rank their own confidence and forecast the proportion of votes that will agree with their choice. Experiments show that selecting the alternative that exceeds anticipated popularity consistently yields better accuracy than plain majority voting.

Another approach is to filter crowdworkers themselves. Amazon Mechanical Turk provides a simple mechanism for limiting which workers may participate, *e.g.,* workers must have an overall assignment acceptance rate of 90% (Barowy *et al.*, 2012). However, it has been argued that this is not a robust measure of worker reliability (Eickhoff and de Vries, 2013). It is also possible to estimate the quality of individual crowdworkers and weight their contributions accordingly, such as separating low- and high-quality participants post-work (Ipeirotis *et al.*, 2010), attempting to identify experts within a crowd (Burnap *et al.*, 2015), performing pre-selection based on self-assessments (Gadiraju *et al.*, 2017), or measuring attention during work (Chandler *et al.*, 2013). Thanking participants and explaining the importance of requested tasks can increase motivation (Chandler and Kapelner, 2013), but improvements are questionable at best (Paolacci and Chandler, 2014). An alternative option is to focus on detecting spam (Wang *et al.*, 2016) and malicious conduct (Wang *et al.*, 2013).

A different technique is to have experts generate a gold standard for some tasks, and use them to gauge the level of quality of each individual contributor (Sorokin and Forsyth, 2008) or to apply some sort of correction. This method has been used for peer-graded assignments in Massive Open Online Courses (MOOCs), for example (Wang *et al.*, 2020, 2019), but creating this gold standard may not always be feasible. Dividing the crowd in different groups for cross-validation is another solution, such as forming a contributing crowd that supplies data which is ratified by a (different) validating crowd (Luo and Zeynalvand, 2017; Luo *et al.*, 2019).

There are many other variants of these approaches and their combinations, with varying degrees of sophistication, *e.g.,* Ahmad *et al.* (2018); Barowy *et al.* (2012); Blanco *et al.* (2011); Davtyan *et al.* (2015); Ghosh *et al.* (2015); Hosseini *et al.* (2012); Karger *et al.* (2014); Maynard and Bontcheva (2016); Schall *et al.* (2011); Tang *et al.* (2015); Von Ahn *et al.* (2008); Vuurens *et al.* (2011), and even methods that analyze user behavior during work (Rzeszotarski and Kittur, 2012, 2011), select crowdworkers based on personal preferences extracted from social networks (Difallah *et al.*, 2013), or aim at designing tasks that are harder to cheat (Callison-Burch and Dredze, 2010).

All of these methods take one important aspect for granted: the possibility of defining what "correct" means. They are unsuitable for subjective tasks (Haralabopoulos *et al.*, 2020) such as sentiment analysis and, as in our case, measuring accessibility. Aroyo and Welty (2015) argue that assuming there is one correct interpretation for every input example is a fallacy. They propose, instead, the notion of *crowd truth*, which takes into account multiple perspectives and interpretations, avoiding common myths such as *"disagreement is bad"* and *"experts are better"*. In these cases, "right" and "wrong" may simply not exist – we may, at best, be able to infer bias.

2.3 Bias in crowdsourcing systems

Spam, malicious conduct, and lack of attention or skill are not the only sources of noise in data. Subjectivity and unintentional bias can cause disagreement among workers (Aroyo *et al.*, 2019). Even objective tasks are vulnerable to bias, and simple aggregation methods (*e.g.,* majority voting) are not enough for undertaking them (Faltings *et al.*, 2014; Kairam and Heer, 2016).

The Dunning-Kruger effect (Kruger and Dunning, 1999) is the tendency of unskilled persons to overestimate their own capability. In general, people tend to hold overly favorable views of their abilities, and as a result, they may overconfidently produce subpar work. While the effect was originally observed in controlled experiments, it is also present in the many different types of crowdwork (Gadiraju *et al.*, 2017) and it alters overall performance (Saab *et al.*, 2019).

More generally, crowdsourcing systems are liable to cognitive biases (Eickhoff, 2018), or *"cases in which human cognition reliably produces representations that are systematically distorted compared to some aspect of objective reality"* (Haselton *et al.*, 2015). Individuals construct their own social reality, and this may dictate their behavior instead of the objective input, leading to perceptual distortion, inaccurate judgment, illogical interpretation, or irrationality (Bless and Fiedler, 2014).

Eickhoff (2018) analyzed how cognitive biases influence the quality of crowdsourcing, showing that there is significant detrimental impact in final results. The study focuses on four kinds of bias: ambiguity effect (when missing information makes decisions appear more difficult and consequently less attractive), anchoring (when there is disproportional focus on one piece of information even as additional contradicting evidence becomes apparent), bandwagon effect (when workers forego their own reasoning in favor of following an existing group's behavior), and decoy effect (when workers' preference between options A and B changes in favor of option B when an option C is presented, which is similar but clearly inferior to option B).

Hube *et al.* (2019) showed that workers' opinions affect their judgement, and even experienced crowdworkers fail to distance themselves from their own cognitive biases when doing subjective labeling tasks. In their experiment, they selected statements about controversial topics (*e.g.,* abortion) and asked workers to label them as neutral or opinionated. Those who had strong opinions about the topics had higher mislabeling frequency. Their bias mitigation recommendations include reminding workers about the contentious nature of the topics and asking them to label statements according to how they believe the majority of other workers would label them.

Chapter 3

Measuring accessibility

3.1 The need for accessibility

There is a large number of persons with disabilities in the world. The World Health Organization estimates that one out of seven people have some kind of disability (WHO, 2018). In Brazil, conservative estimations point to over 13 million persons with physical disabilities, or about 7% of the population (IBGE, 2010a,b). For them, everyday activities like getting out of home can represent great challenges due to the lack of accessibility in streets and buildings.

Additionally, there are around 700 million people aged above 65 globally, and life expectancy continuously increases (Kanasi *et al.*, 2016; UNFPA, 2017). Population aging is an evergrowing trend in countries all over the world, and many are already investing in policies and research aimed at this matter (Bos and Von Weizsacker, 1989; Council *et al.*, 1990; Kose, 2010; Mann, 2004).

About 140 million live births happen every year throughout the world, meaning that there are almost 700 million children below the age of 5 (United Nations, 2018). The elderly, pregnant women, babies, toddlers and their caretakers can all be understood as *persons with mobility impairment*, who do not necessarily live with a disability, but also need accessible environments in order to live safely and autonomously.

The group of people who benefit from accessibility, however, is not limited to persons with disabilities or mobility impairment. Even young able-bodied individuals gain from it, as they are constantly liable to injuries or temporary limitations (such as a sprained ankle or recovery from a surgery, for instance). By extension, we can affirm that accessibility is not a matter of minorities – everyone gains from it.

Unfortunately, this is not a widespread notion. The absolute majority of urban environments was not conceived for all people.

Accessibility is relevant not just because of altruistic or humanitarian reasons. It also promotes economic growth and financial profit. Shops, restaurants, and hotels that do not offer appropriate infrastructure are missing potential costumers – and if we also take into account the elderly, parents of toddlers, and their respective family and friends, the result is a huge loss of potential buyers.

Technological progress has been continuously improving people's quality of life. By promoting accessibility, we can ensure that all people may live to their full potential. Take the example of Stephen Hawking, the British scientist: he changed the way science understands the Universe, but without supporting accessibility technology he would have not been able to communicate his ideas (Ferguson, 2011).

How many would-be scientists and artists are currently living as prisoners in their own homes? How many have-been leaders and inventors are excluded from social contact with later generations, just because they got "too old"? Can we, as a society, afford such a waste of potential? Can we afford that our old timers die without passing on their experience and knowledge?

Historically, the need for accessibility has been largely overseen. It was not accounted for in most buildings constructed up until the later half of the XX Century. Figure 3.1 illustrates this issue. The main building of the Mário de Andrade Library in São Paulo was built in the 1940s, and is considered one of the best representatives of *art deco* architecture in the city. The McKim Building of the Boston Public Library, located near Boston's historical center, was built in 1895. Both are regarded as architectural postcards of their cities, and both have been retrofitted with modern additions for improving accessibility. This was necessary not because the people who designed them lacked competence, but because accessibility was not considered relevant at their time.

Since then, a paradigm shift has started. The notion of designing and implementing man-built environments and machinery without prioritizing accessibility for all is becoming more and more obsolete. Nonetheless, urban environments throughout the world still lack accessibility.

What can be done to improve this situation? Do we need to take down entire cities and rebuild from scratch? "How much" accessibility is "enough accessibility"? How far are we from this "acceptable level", and what needs to be done to achieve it? These are a few of the questions that naturally arise when devising solutions for this matter, and to answer them, one must be able to, somehow, evaluate and measure accessibility (or lack thereof).

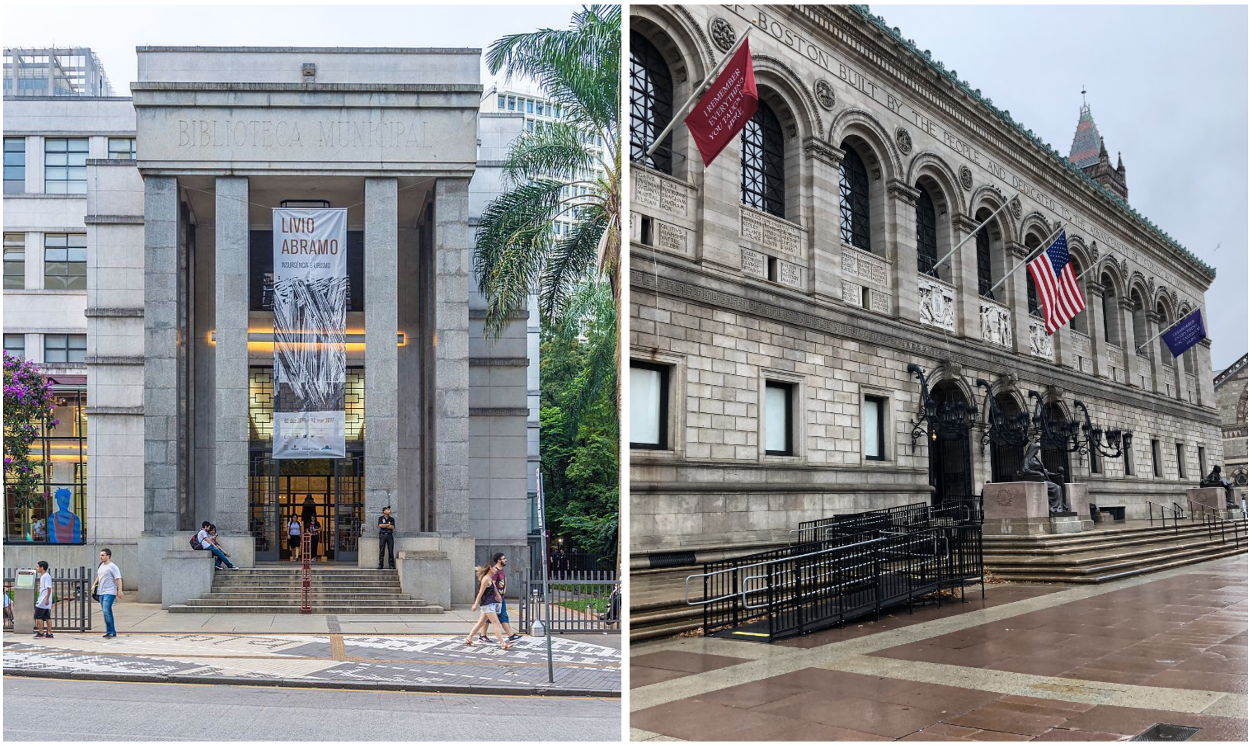


Figure 3.1: *Historical lack of accessibility: on the left: Mário de Andrade Library (1940s) in São Paulo, Brazil. Handrails at the central section of the stairs and tactile flooring on the sidewalk pavement are modern additions for improving accessibility. On the right: McKim Building of the Boston Public Library (1895) in Boston, United States. Metallic ramps were installed to allow wheelchair users to enter the building. In both cases, later architectural interventions are intentionally designed to be visually distinct from the buildings to preserve their original appearance. Image sources: (Wikipedia, 2021a) (left) and Author (right).*

3.2 What is accessibility?

The first step to measure something is to understand what it is, and how it can be observed. Therefore, before even thinking about evaluating accessibility, we should investigate how to define it precisely. And this is, by itself, no easy task.

Oxford Advanced Learner's Dictionary gives a first definition for accessibility:

Definition 1. (Accessibility) (Hornby *et al.*, 2000). *How easy something is to reach, enter, use, etc. for somebody with a disability.*

This might be enough for a casual conversation, but *"etc."* adds too much imprecision. What other verbs should be included in the list? See? Operate? Advocate? Triangulate? How do we decide which ones are included, and which ones are not? It is a starting point, but this definition leaves room for personal interpretation.

In its Convention on the Rights of Persons with Disabilities (CRPD), the United Nations published a commitment, titled "Accessibility":

> *"To enable persons with disabilities to live independently and participate fully in all aspects of life, States Parties shall take appropriate measures to ensure to persons with disabilities access, on an equal basis with others, to the physical environment, to transportation, to information and communications, including information and communications technologies and systems, and to other facilities and services open or provided to the public, both in urban and in rural areas. These measures, which shall include the identification and elimination of obstacles and barriers to accessibility, shall apply to, inter alia:*
>
> *a) Buildings, roads, transportation and other indoor and outdoor facilities, including schools, housing, medical facilities and workplaces;*
>
> *b) Information, communications and other services, including electronic services and emergency services."* (United Nations, 2006, Article 2)

Even though this is not an attempt at providing a direct definition, it does shed some light on what accessibility could possibly be.

Similarly to Oxford Learner's dictionaries, it explicitly states an exclusive connection to persons with disabilities (*"...shall take appropriate measures to ensure to* ***persons with disabilities*** *access..."*), but there is also the introduction of a new notion here: equality (*"...on an* ***equal basis*** *with others..."*).

In the beginning, it says *"all aspects of life"*, which is very wide and comprehensive, but then proceeds to narrow its scope by enumerating things that must be considered. Despite this, we can infer by paragraphs (a) and (b) that accessibility is probably something interdisciplinary. Moreover, there seem to be at least two different "big areas" here: physical elements, and technological (or informational and communicational) elements.

It also mentions "services", which hints that accessibility may not be a concrete concept, but there is no mention to things like machinery, equipment, or any kind of object. Due to the adopted enumeration strategy, it might be understood that these should not be considered when evaluating accessibility.

Because it refers to both urban and rural areas, it seems that accessibility is relevant anywhere humans may be living.

The text also explains that promoting accessibility should be done by identifying and eliminating "obstacles and barriers". Detecting and measuring these obstacles might be a way to calculate (the lack of) accessibility.

Definition 2. (Accessibility) (ABNT, 2020). *Possibility and condition of reach, perception and understanding for use, safely and autonomously, of spaces, furniture, urban equipment, buildings, transportation, information and communication, including its systems and technologies, as well as other services and facilities open to the public, of public use or private property of public use, both in urban and rural areas, by persons with disabilities or with mobility impairment.*

This definition is given by the Brazilian Association of Technical Standards *(Associação Brasileira de Normas Técnicas – ABNT)* in its standards for "Accessibility to buildings, equipment and the urban environment" (which are discussed in Section 3.3). It seems to be influenced by the United Nations text, as it also mentions "urban and rural areas", "information and communication", "systems and technologies", and "persons with disabilities".

Even though it also falls in the enumeration trap, it incorporates some relevant additions, such as "furniture" and "urban equipment", and most importantly, it also mentions *"persons with mobility impairment"*, which are considered to be a separate group from persons with disabilities.

This definition includes "safety" and "autonomy" as requirements, but limits the need for accessibility only to services and facilities that are "open to the public" (regardless if the property itself is public or private).

Definition 3. (Accessibility) (Iwarsson and Ståhl, 2003). *The encounter between the person's or group's functional capacity and the design and demands of the physical environment.*

The definition of accessibility given by Iwarsson and Ståhl (2003) defines two separate components: a personal component (*"the person's or group's functional capacity"*) and an environmental component (*"the physical environment"*) (Fänge and Iwarsson, 2003).

Because it does not explicitly mention any group of persons ("persons with disability" nor "persons with mobility impairment"), it implicitly includes all kinds of people. This generalizes on previously considered definitions, and is in conformity with the universal need for accessibility discussed in Section 3.1.

By linking it to the "environment", it is broad enough to be applied to both indoor and outdoor areas, urban or not. Here, an "environment" can be a city, a street, a forest trail, a coffee shop, or even a private residence. It also refers to the "design and demands" of the environment. This can be interpreted by the fact that accessibility can be actively influenced (positively or negatively) by design, while constraints to this design are the environment's "demands".

A negative point of this definition is its limitation to *functional* capacities (*e.g.,* motor skills and mobility) and the *physical* environment. This does not include cognitive capacities, for instance, and is not applicable to two aspects that appeared previously and are not necessarily part of a physical environment: communication and information. As a result, this definition is not suitable for purely digital environments, like a website.

Another issue is that it is based on a highly abstract concept (*"the encounter"*), which is useful to determine the two relevant components – personal and environmental – but it is too vague to specify a measurable magnitude. For this reason, this definition does not seem to be enough to establish whether something is accessible or not.

To address these issues, another definition is herein proposed. It is not an attempt to ultimately define accessibility, nor does the author expect to impose it to any of the various scientific fields that also study accessibility. It is, though, an attempt to be as comprehensive as possible, as to be useful to as many contexts as possible. And more importantly, this is the definition that is adopted in the remainder of this thesis, as it seems to be generic yet precise enough for our purpose of conceiving some kind of measurement method for it.

Definition 4. (Accessibility) (as defined by the author). *The degree of safety and autonomy provided to all persons by the design and demands of an environment.*

By referring to "the degree" of something, accessibility is defined as a quantifiable magnitude (even if we don't necessarily know *how* to quantify it). Safety and autonomy are adopted as requirements, similarly to Definition 2. Like Definition 3, it is based on an environment's design and demands, but it is not restricted to *physical* environments – they can also be digital, such as a website or a computer program.

An accessible environment is able to provide a large degree of safety and autonomy, meaning that when people visit it, they can successfully perform whatever action they are willing to. However, an environment that is only appropriate for a certain group of persons – even if this group is the set of all persons with disabilities – is not accessible, as it excludes other groups. Therefore, the definition refers to "all persons", explicitly including not only persons with disabilities, but also persons with mobility impairment and persons without either of these.

Defining what we are trying to measure, *i.e.,* accessibility, is only part of the task, as there are other factors that must be considered. Because disabilities and capabilities can vary greatly among individuals, there is always the possibility that an environment is visited by a person whose needs had not been previously accounted for. Therefore, there is no such a thing as a "perfectly accessible environment" (Butlewski and Jabłońska, 2014). Luckily, infrastructural limitations can be mitigated by the actions of others. For instance, if a restaurant does not have a braille menu for blind persons, a waiter may read it out loud for them. Even though full autonomy has not been provided, the waiter's actions can fulfill the lack of appropriate infrastructure. Conversely, if the restaurant's wheelchair-friendly restroom is used as an additional storage room and is always locked, the resulting environment is not accessible, not due to lack of infrastructure, but as a result of personal behavior.

Consequently, a person's actions can greatly influence the overall experience in an environment. This is the *attitudinal* aspect of accessibility, *i.e.,* pertaining to people's attitudes, actions that may or may not be taken by individuals to help make others feel welcome and safe (Bi, 2006; Bi *et al.*, 2007; Card *et al.*, 2006; Eichhorn and Buhalis, 2011), such as smiling, maintaining eye contact, offering help, respecting personal space, etc. And because people can be considered to be part of an environment, our proposed definition 4 still holds. That might not have been the case if the word "places" had been used instead of "environments", for example, as "place" implies that only physical elements are what matters.

Modeling accessibility is a complex, multifaceted open research topic that requires an understanding of who the users are, their needs and abilities, and an understanding that needs and abilities may change over time (Froehlich *et al.*, 2019). A theoretically ideal measurement scale for accessibility should be able to take all these (sometimes subtle) nuances into account. Though this may not be feasible, attempts to create a quantitative method are relevant even if they are imperfect, as certain benefits of objectivity cannot be enjoyed in purely qualitative studies.

Different approaches for assessing accessibility have been developed, which will be discussed in the following sections. And while there may not be a definitive answer to what accessibility really is, there seems to be no doubt that it is inherently subjective – were it not, defining it would be trivial and this entire discussion would be pointless.

In this regard, Pirie (1979) summarizes: *"accessibility may be measured in several ways, depending on one's conception of it. The ultimate question then is whether conceptions can be wrong or only inappropriate"*.

3.3 Accessibility standards

3.3.1 Architecture and design

There are many regulations and standards related to accessibility across different fields of human activity. Architecture and design are particularly relevant, as they permeate both the conception and the construction of all man-made structures, objects and machines, but there are also standards for purely digital environments and Information Technology (IT) products. Several standardization institutes from all over the world have produced their own guidelines.

These standards regulate many different fields: indoor architecture in different kinds of spaces (*e.g.,* offices, classrooms, restrooms), urban spaces (outdoor furniture and sidewalks), machinery, consumer and everyday products, websites, mobile apps, etc. They seek to establish parameters that can serve as a baseline for evaluating an environment's accessibility, and are based on well-defined and detailed specifications. They produce an objective set of criteria for evaluation: things are either compliant to them, or not. In theory, there should be as little room for interpretation as possible. But in reality, it can be dramatically different.

The high level of detail and inherently large scope of these standards can make them difficult to apply – and maybe even to understand. They try to create a fixed set of objective rules to accessibility, something that is highly subjective by nature. These documents are constantly reassessed to reflect the evolution of our own understanding of accessibility, and most have been updated several times after being initially published.

To illustrate this, we take the example of the Brazilian NBR 9050 standard (ABNT, 1985, 1994, 2004, 2015, 2020). Several editions have been published by the Brazilian Association of Technical Standards *(Associação Brasileira de Normas Técnicas – ABNT)* since the 1980s. Its goal is to set *"[accessibility] criteria and technical parameters to be observed when designing, building, and proceeding installation and adjustment of urban buildings"*. Even though it tries to narrow its scope to an architectural perspective, it touches many different aspects of civil construction, architecture, and communication.

Defining detailed requirements for a large amount of applications while accommodating needs of diverse groups of users (*e.g.,* persons in wheelchairs, blind persons) results in an extensive set of rules. Figure 3.2 shows selected samples adapted from NBR 9050 (ABNT, 2020). They demonstrate the high level of specificity and variety of scope of the standard. This represents a challenge for professionals to fully understand it and correctly apply it. As a result, an alarmingly high number of designs contain errors due to misunderstanding of rules (de Assis and Toledo, 2016; Kutchukian and

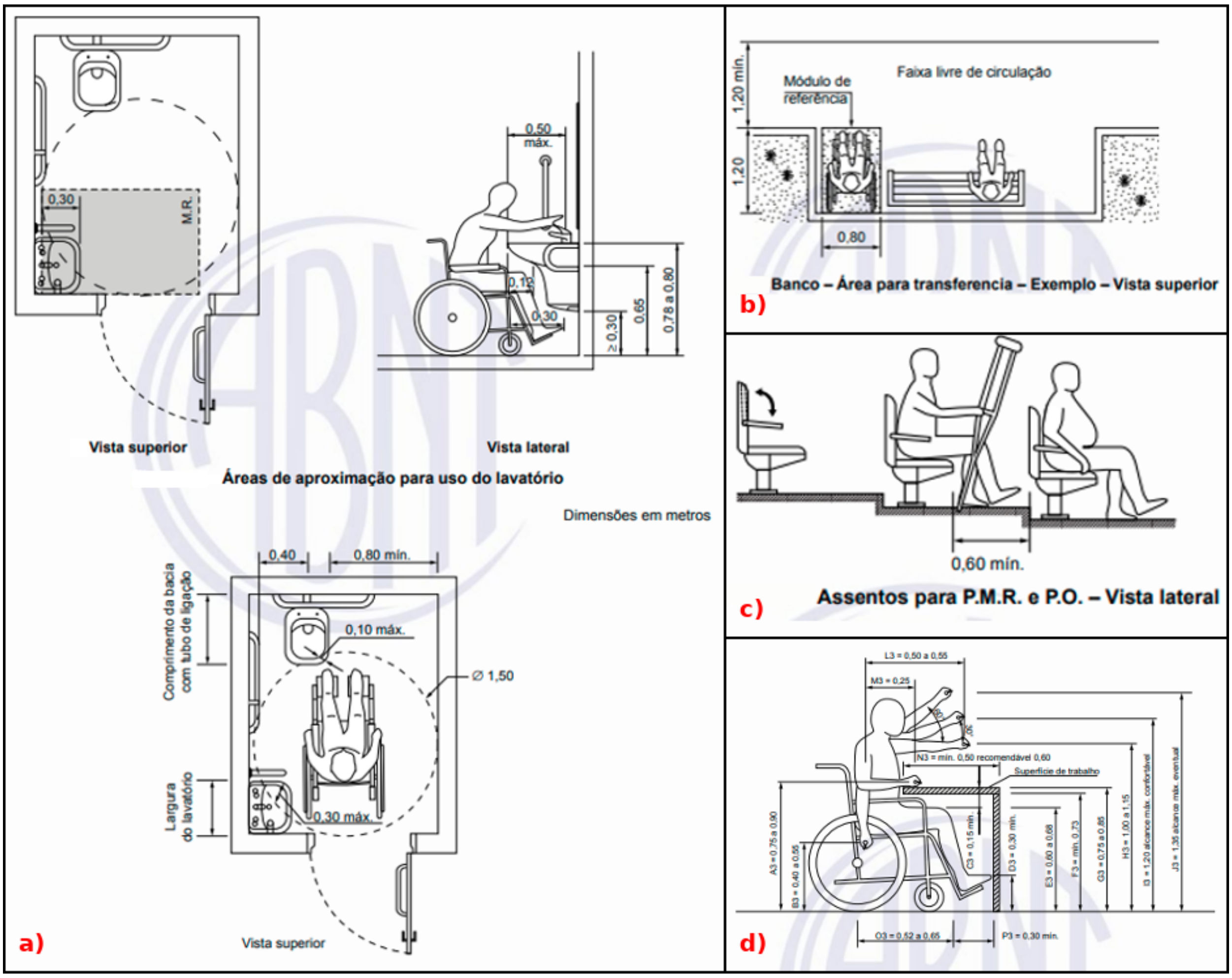


Figure 3.2: *The Brazilian standard NBR 9050/2020* "Accessibility to buildings, equipment and the urban environment"*: on the left, rules for accessible bathrooms* (a)*. On the right, rules for outdoor furniture* (b)*, seatings in auditoriums for persons with mobility impairment and obese persons* (c)*, and for desks/working benches* (d)*. Criteria defined by the standard encompass a large variety of topics and have a high level of detail. Adapted from (ABNT, 2020, Figures 15, 99, 100, 134 and 146).*

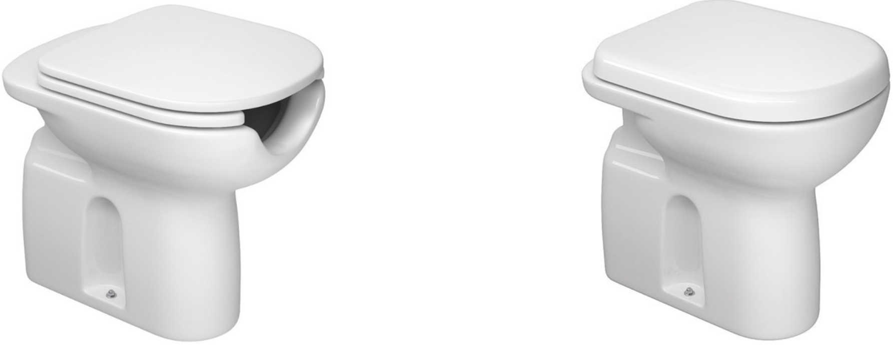


Figure 3.3: *A breaking change in NBR 9050: the toilet bowl on the left has a frontal opening and was recommended in NBR 9050/2004 (ABNT, 2004). Even though this model is better for the hygiene of fully paralyzed individuals, it presents a risk of fall for more autonomous individuals. This was changed in the following edition (ABNT, 2015), which only recommends bowls without the opening, similar to the one in the right. Image source: (Deca, 2018).*

Kutchukian, 2017; Souza and Thomé, 2008) or, even worse, issues in the rules themselves (Moraes, 2007).

Figure 3.3 shows an example of a standard update for correcting an issue, a case where a significant change was introduced by a superseding newer edition. The second edition of NBR 9050 recommended toilet bowls with frontal openings in all accessible washrooms and toilets (ABNT, 2004). This allows for easier hygiene of fully paralyzed individuals by aides and nurses, but for other people, this opening represents a hazard, as it introduces an unnecessary risk of fall, which may result in injuries as serious as femoral bone fracture (Martins, 2016). This severe issue was addressed in the following edition of the standard, which endorses closed toilet bowls only (ABNT, 2015). Nonetheless, even after several years since the new rule was published, it is not uncommon to find frontal-opening toilet bowls in offices and public buildings.

NBR 9050 also states which built spaces must be accessible and which ones need not: restricted and/or technical service areas such as engine rooms, technical passages, barrels, etc. are excluded from its requirements, whereas all other kinds of spaces, buildings, urban facilities and equipment must meet its specifications to be considered accessible. All new buildings, renovations and expansion projects must be in accordance to its guidelines, even for buildings constructed prior to its introduction. This condition is enforced by law (Brasil, 2015), and design drawings must be submitted to analysis by competent bodies in order to receive mandatory licenses from public authorities. But, as previously mentioned, even this control mechanism is not enough to prevent problems.

Many countries have developed their own set of standards for accessibility. In the United States, for example, the *Americans With Disabilities Act (ADA)* (108th Congress, 1990) is a law for protecting the rights of persons with disabilities. In addition to making discrimination illegal, it also imposes accessibility requirements on employers and public accommodations. Subsequently, the technical standards for requirement compliance were published as the *ADA Accessibility Guidelines (ADAAG)* (ATBCB, 2010). Its scope is similar to NBR 9050's, and it has also suffered major and minor revisions throughout its lifetime.

Even though there seems to be some shared inspiration in these standards, significant differences can be noted between them. Figure 3.4 illustrates recommendations for handrails in NBR 9050 and ADAAG. There have also been cross-national efforts for establishing minimum accessibility criteria. The International Organization for Standardization (ISO) has published specifications for built environments (ISO, 2011), usability of consumer products and products for public use (ISO, 2006, 2013), and also general guidelines for standardization (ISO, 2014), which have also been adopted by European bodies (CEN-CENELEC, 2014).

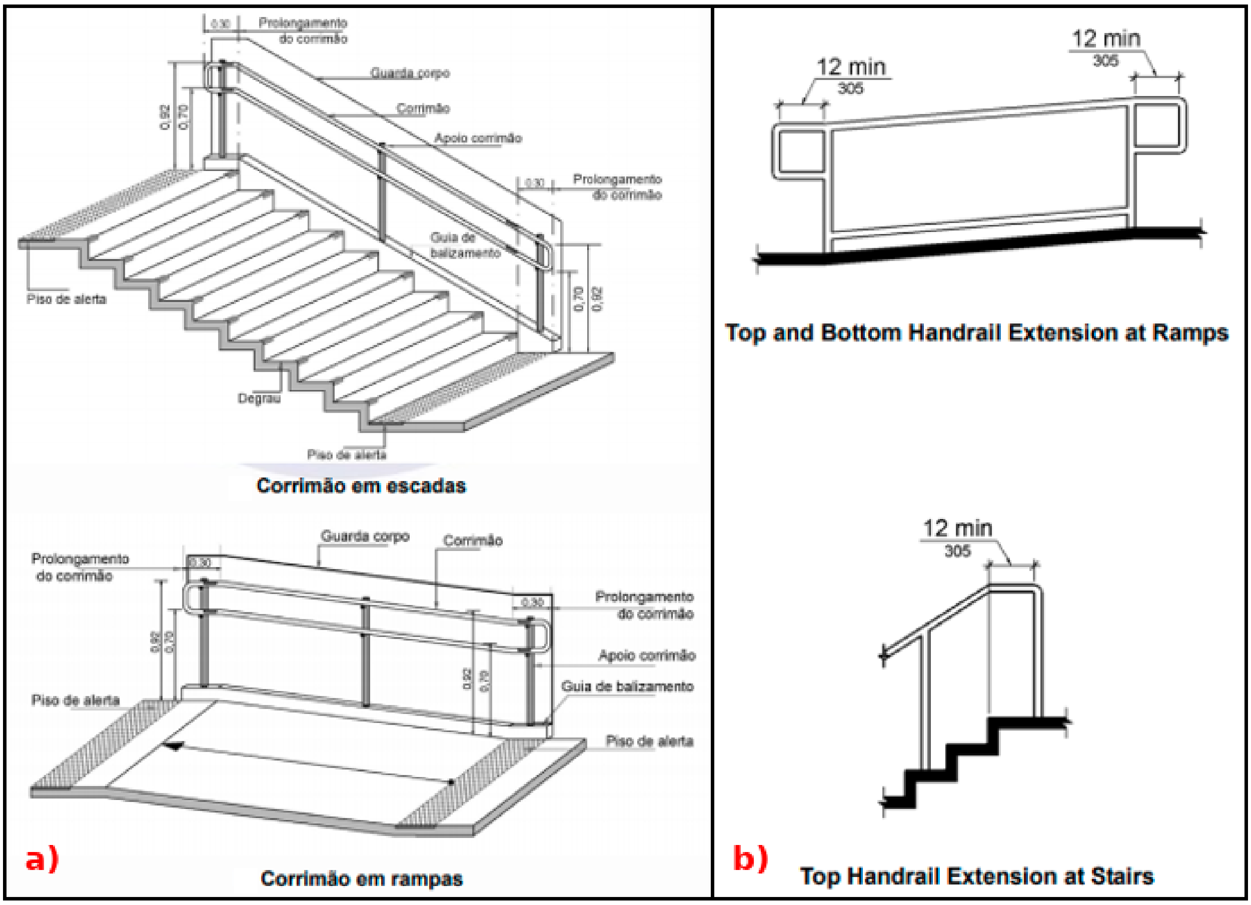


Figure 3.4: *Differences between Brazil's NBR 9050 and United States' ADAAG: recommended handrail styles diverge in the Brazilian standard (Figure* a*) and the American one (Figure* b*). Adapted from (ABNT, 2020, Figure 76) and (ATBCB, 2010, Figures 505.10.2 and 505.10.3).*

3.3.2 Digital standards

Physical accessibility is not the only topic subject to standardization initiatives. Because computers, mobile phones and IT in general is more and more present in people's daily lives, accessibility in digital environments has been increasingly relevant and an active field of study (Barroso *et al.*, 2020). Even though it has been recently enforced by the power of law (Lazar *et al.*, 2015; Lewthwaite and James, 2020; Oncins, 2020), some of the most successful initiatives have been driven by the technology community. Truly, the concern about digital accessibility is rooted on the origins of the Internet itself:

> *"The power of the Web is in its universality. Access by everyone regardless of disability is an essential aspect."* – Tim Berners-Lee, inventor of the World Wide Web (1997).

Accessibility in digital environments is related to both hardware and software (Lazar *et al.*, 2015). It benefits persons with disabilities that affect technology usage (including auditory, cognitive, neurological, physical, speech, visual), and also assists persons without disabilities but with

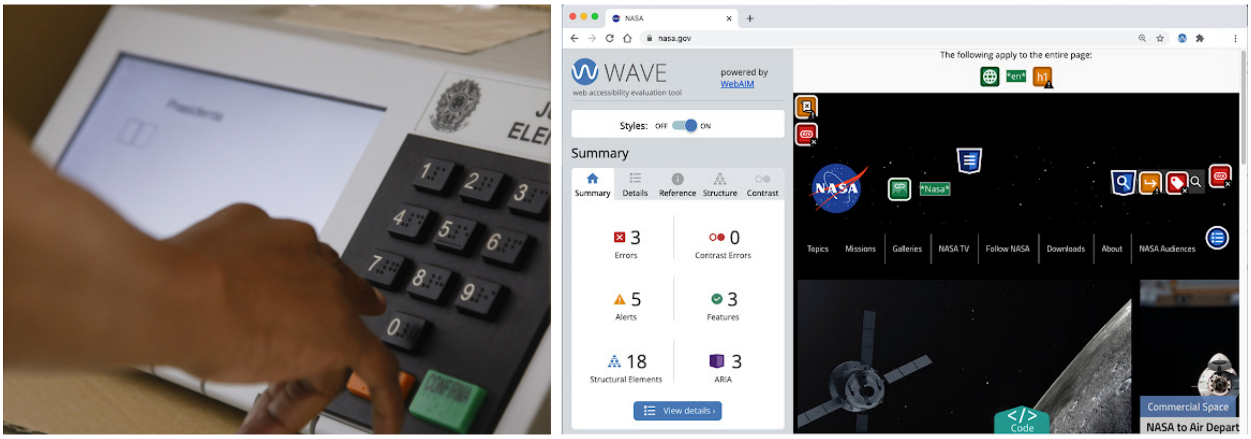


Figure 3.5: *Examples of accessibility in IT: the Brazilian voting machine (left) has large buttons with* braille *and bright colors for confirming or cancelling a vote. The WAVE Evaluation Tool (right) is an automatic accessibility checker for supporting the development of accessible web content by providing immediate feedback. Image sources: (Wikipedia, 2021b) (left) and (WebAIM, 2021) (right).*

temporary or situational limitations (*e.g.,* lost glasses, in a place where they cannot listen to audio, slow Internet connection) (W3C, 2021).

One of most relevant guidelines is the Web Accessibility Initiative (WAI) by the World Wide Web Consortium (W3C). It was created to educate professionals – both developers and designers – on the importance of accessibility and how it can be implemented. It defines concepts that developers should take into consideration, like providing captions and text alternatives for multimedia content, supporting larger font sizes and high contrast for improving readability, giving users enough time to read and use content, etc. (W3C, 2021).

Sometimes, improving accessibility may be as simple as using larger buttons. This is precisely the case of the Brazilian voting machine, depicted in Figure 3.5. The equipment's large buttons with *braille*[1] benefits low-vision and blind persons, and also persons with limited manual dexterity. Buttons for confirming and cancelling user input are brightly colored, which also benefits an even wider range of persons (*e.g.,* persons with cognitive disabilities, the elderly).

There are also automated tools for checking accessibility during software development. The WAVE Evaluation Tool by WebAIM, also shown in Figure 3.5, is an example of it. It analyzes web content, highlights elements that are relevant for accessibility, and points potential problems. Its goal is not to determine if a page is completely accessible, but to help human evaluation and educate about good accessibility practices (WebAIM, 2021). Another tool with a similar goal, but for a different context, is Google's Accessibility Scanner, which scans Android apps and suggests improvements such as enlarging small touch targets, increasing contrast, and providing content descriptions (Google, 2021).

[1] *Braille* is a tactile reading and writing system for persons with visual disabilities.

This tool is part of a much larger initiative for promoting accessibility. In fact, if there are two entities with the highest leverage for promoting digital accessibility, these are Google and Apple. Together, they comprise more than 99% of the mobile operating system market share (Statcounter, 2021). And, at least in terms of accessibility, this duopoly has been beneficial. All Android and iOS phones come with screen readers installed by default, a vital feature for including low-vision and blind persons. Both companies provide tools and extensive documentation to help developers improve accessibility of their apps[2], and these have become a *de facto* standard for the mobile industry.

3.3.3 Universal Design

Due to its interdisciplinary nature and growing importance, accessibility is the target of a large amount of standards and regulations. While they can be used as a sort of measurement tool for quantifying accessibility, applying an objective, boolean-like ("compliant"/"not compliant") scale may not be always appropriate, or even feasible. Moreover, many standards give precise (and frequently excessively detailed) definitions for a vast range of applications, and, as an added challenge, these rules are sometimes conflicting. While digital accessibility is more or less unified thanks to a lack of market competition, physical accessibility standards vary from country to country. This fragmentation can create awkward situations, for instance: which standard should be applied in the United Nations office in New York City? As a building in US soil, perhaps it should follow ADAAG, but it houses an international entity, so maybe it should abide to an international standard? Despite this, architects, designers, software developers and all kinds of creators must find their own way to navigate the sea of standardized accessibility rules.

For millennia, our societies have built dwellings, workplaces and public environments without the needs of all people in mind. As the understanding about the importance of accessibility grows, an increasing number of adaptations are done in order to make preexisting environments more accessible. Notwithstanding, this is not always enough, as adaptations may sometimes not allow for complete integration.

Traditional design is focused on adapting preexisting buildings and products by adding accessibility to them. It is based on the notion that there are two groups: the normal population and the ones deviating from it. Focus is on designing for the former, and adapting for the latter, which results in segregation. Universal design applies a different idea: that there is a single population

[2] Documentation for iOS: https://developer.apple.com/accessibility/ios/ (retrieved Sep. 2021) and for Android: https://www.google.com/accessibility/for-developers/ (retrieved Sep. 2021).

made of persons with different capabilities and needs. Therefore, the design process aims at a single solution that is suitable for all kinds of users: children and adults, the elderly, persons with disabilities, persons with different nationalities, ethnicities, and cultural backgrounds, etc. (Iwarsson and Ståhl, 2003).

This need for creating products and environments for all people, based on the notion that *"separate is not equal"*, was reinforced by legislative and standardization measures like the aforementioned ones. They included specialized requirements to accommodate persons with disabilities, and the nonregulated market-driven responses to an aging society, primarily relating to products. As a result, the paradigm of *Universal Design* was created (Ostroff, 2010).

The United Nations defines Universal Design as follows:

> "'Universal design' *means the design of products, environments, programmes and services to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design."* (United Nations, 2006, Article 2).

Some authors use universal design as a synonym to "design for all", defining it as *"an approach to design that incorporates products as well as building features which, to the greatest extent possible, can be used by everyone"* (Iwarsson and Ståhl, 2003), and *"a collection of principles that can be integrated as a process within the sustainable design framework and mainstream design thinking"* (Ostroff, 2010).

Even though several standards and laws were created in the past decades, they are not effective by themselves, as they define minimum requirements which often become the maximum effort. Much more than a set of fixed rules and specifications, universal design focuses more on the process than on an arbitrary result (Iwarsson and Ståhl, 2003). This results in a somewhat hard to measure and intimately abstract concept, but the underlying principles have been summarized by the Center for Universal Design (Story, 2001), which are reproduced in Table 3.1.

Thanks to its inclusive nature, universal design is applicable to different contexts: education, business and, more broadly, society as a whole. Universal Design has been successfully adopted throughout the world, in countries as diverse as Norway (Bringa *et al.*, 2010; Ministry of the Environment, 2009), Japan (Kose, 2010), France (Grosbois, 2010), Germany (Krauss, 2010), Italy (D'Innocenzo and Morini, 2010), and Israel (Ramot *et al.*, 2010).

It is, undoubtedly, a method for creating environments that are accessible from conception, but due to its innate unstructuredness, universal design may not always provide the necessary tools for a quantifiable measurement of accessibility.

Principle		Definition
1	Equitable Use	The design is useful and marketable to people with diverse abilities.
2	Flexibility in Use	The design accommodates a wide range of individual preferences and abilities.
3	Simple and Intuitive Use	Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.
4	Perceptible Information	The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.
5	Tolerance for Error	The design minimizes hazards and the adverse consequences of accidental or unintended actions.
6	Low Physical Effort	The design can be used efficiently and comfortably and with a minimum of fatigue.
7	Size and Space for Approach and Use	Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.

Table 3.1: *The seven principles of Universal Design, as defined by Story (2001).*

3.4 Measurement methods

"Accessibility" is traditionally defined as the ease of reaching destinations (Saha *et al.*, 2021), *i.e.,* a synonym of mobility, which is surely related to accessibility, but is only a single variable in a more complex subject. This definition is broadly used in studies focused on urban environments. However, it does not mean they are any less relevant to our scope. Pirie (1979), Handy and Niemeier (1997), and Handy and Clifton (2001) have all proposed scales for measuring mobility in cities and neighborhoods with different approaches, by looking at a person's origin, destination or time consumed during transit.

Church and Marston (2003) have proposed an extension to previous works creating a new quantitative measurement of mobility that is suitable to both indoor and outdoor environments. Their method takes into account origin, destination, travel method and route, yielding a numeric result that can be understood as a "mobility grade". It can be applied to parts of urban areas and also to more limited scopes, like individual buildings. These works are focused on general population, not necessarily persons with disabilities or mobility impairment, being particularly applicable in a macro-scale, *e.g.,* when analyzing the accessibility of urban environments as a whole.

Pasaogullari and Doratli (2004) have defined a qualitative method for measuring accessibility and utilization of public spaces in urban environments. In their study, they determine what variables and factors impact accessibility and utilization of public spaces, creating a method for taking into account the usage of public areas within the city as a metric for accessibility. They studied the city of Famagusta, in Cyprus, showing how their scale can be useful for drawing conclusions about accessibility in cities.

Computational methods have also been applied in generic ways for measuring accessibility. An example is AMELIA (Mackett *et al.*, 2008), a software tool for testing the impact of urban planning policies related to mobility. Focused on socially excluded groups and persons with limited mobility, their methodology takes into account public transportation and outdoor elements such as sidewalk wideness, pedestrian crossings and curb ramps. Based on disability legislation and policy, the system has been used by local authorities of the city of St Albans, in England. AMELIA takes into consideration social and demographic information, as a way to focus on socially excluded groups. A similar path was followed by Krempi (2004), who analyzed how public and private transportation correlate with social and demographic statistics in the Brazilian city of Bauru.

Other methods for quantifying accessibility aim at indoor environments. Their approach relies on analyzing determined variables, such as measuring pedestrian movement based on a building's geometric characteristics (Jeonnong-dong and Dongdaemun-gu, 2008; Thill *et al.*, 2011), and measuring the egressibility (or, as the authors put it, "exitability") of buildings under evacuation scenarios (Vanclooster *et al.*, 2012). Focused on hotel environments, Butlewski and Jabłońska (2014) developed a model of ergonomic quality for hotel service. In addition to formal architectural and ergonomic requirements, their analysis also aims to capture subjective user needs.

There have also been studies that evaluated indoor accessibility for persons in wheelchairs and persons with mobility impairment using techniques of virtual reality (VR), by creating three-dimensional models of the environment. This can be done via static or multi-camera image processing (Kim, 2005), or by deploying laser scanners to determine the environment's geometry (Se and Jasiobedzki, 2006). Unfortunately, this may not be attainable nor accurate if the environment is too complex. First approaches provide data for analysis by a human evaluator (Kim, 2005), whereas later works perform automatic assessments (Otmani *et al.*, 2009; Saidi Sief *et al.*, 2016) by simulating the movement of persons in wheelchairs and persons with mobility impairment. Their analyses focus on determining what surfaces and indoor elements are reachable by groups with different kinds of limitations, producing numeric results and highlighting areas that can and cannot be accessed or reached.

The diversity of measurement proposals is attested by the variety of their scopes and procedures, ranging from mostly qualitative surveys to quantitative methodologies that yield numeric comparative scales. While studies focusing on urban accessibility may seem to focus on mobility, an ideal scale for evaluating indoor environments should observe the existence (or absence) of ramps, elevators, and handling bars, wideness of doors and corridors, furniture shapes and sizes, among other architectural aspects (Butlewski and Jabłońska, 2014). In addition, micro-scale mobility should also be assessed. The distribution of washrooms and exits, for instance, is a significant variable in this scope, as travel length is a determining factor for accessibility (Church and Marston, 2003).

3.5 Crowdsourcing and accessibility

There are many datasets that represent persons with disabilities and the elderly, encompassing different kinds of data and different aspects of daily life (Kamikubo *et al.*, 2021), such as photos taken by blind users (Lee and Kacorri, 2019), assistive app logs of users with visual impairments (Kacorri *et al.*, 2016), sign language videos (Huenerfauth and Kacorri, 2014), gloss annotations (Neidle *et al.*, 2012), motion captured signs (Hassan *et al.*, 2020), depth data from older adults' activities (Leightley *et al.*, 2015), stroke gestures by persons with motor impairments (Vatavu and Ungurean, 2019), eye-tracking data from autistic children (Duan *et al.*, 2019), voice recordings of persons with speech impairments (Cesari *et al.*, 2018), and a speech corpus of persons with intellectual disabilities (Rello *et al.*, 2014).

Kamikubo *et al.* (2021) reviewed 137 of such databases that were at least partially sourced from persons with disabilities, and found that the majority of them (115) were sourced from less than 100 participants. Many of the smaller databases ($N \leqslant 3$) included audio and/or video recordings of volunteers and patients, which can provide richer details at a higher data collection cost. Only 7 databases included more than 1000 individuals, and they typically involved remote data collection in the real world, like tracking device usage or active user input through an app. A considerable amount of these databases contain qualitative information that does not necessarily translate into an accessibility scale.

Croudsourced mapping initiatives, known as *Volunteered Geographic Information (VGI)* (Goodchild, 2007), are part of a broader movement called *neogeography* (Turner, 2006). In many cases, they are created by a community or group of non-experts. VGIs have been used in a wide range of applications, such as navigation (Bakillah *et al.*, 2013, 2014; Graser *et al.*, 2015; Zhang and Ai, 2015), transportation (Biljecki *et al.*, 2013; Sun *et al.*, 2015), cycling (Hochmair *et al.*, 2013; Sun and Mobasheri, 2017), environmental monitoring (Kettunen *et al.*, 2016; Kubásek and Hřebíček, 2013), urban planning and renovation (Hachmann *et al.*, 2018), and even disaster management (de Albuquerque *et al.*, 2016; Haworth and Bruce, 2015).

Several VGIs for mapping accessibility have been proposed. In general, they focus on urban environments. Some of the earliest approaches were proofs of concept that went through some preliminary testing but were not necessarily deployed in real-world scenarios, *e.g.,* routing paths for wheelchair users (Beale *et al.*, 2006; Matthews *et al.*, 2003) and crowdsourcing information about bus stops for blind persons (Campbell *et al.*, 2014; Hara *et al.*, 2015). Others suggested systems for reporting problems in sidewalks like potholes and steps (Cardonha *et al.*, 2013; Koch *et al.*, 2012; Rice *et al.*, 2012; Shigeno *et al.*, 2013). While mapping these kinds of problems does not constitute a direct measurement of accessibility *per se*, it can support a proxy estimation for *lack* of accessibility.

Platforms like AXS Map in the United States[3] and Euan's Guide in the United Kingdom[4] adopt straightforward data collection methods. They are based on direct user participation and allow the creation of reviews following a 5-star system where less stars denote less accessibility. Founded by persons with disabilities and with audiences built from the community itself, these grassroots projects are mostly directed at accessibility for wheelchairs, but do not completely forgo other kinds of disabilities. They disregard sidewalks and other urban mobility aspects, focusing instead on information about places like restaurants and shops. In addition to an initial general evaluation, they support more specific ratings, *e.g.,* of entrances and restrooms, and also let users write comments for further detailing their impressions.

A similar platform and perhaps the largest of such accessibility VGIs is Wheelmap[5] – a map for wheelchair-accessible places (Mobasheri *et al.*, 2017). Created by a German non-governmental organization for diversity and inclusion, the open database contains data aggregated from its users and other 189 sources, totaling more than 2,900,000 evaluated places. Datasets are published under the Open Database License (ODbL) and are available as a free of charge REST API[6]. It is built on top of OpenStreetMap[7] – "the free wiki world map", as the project describes itself. Places are represented by *Points of Interest (POIs)*, divided in hundreds of categories. Wheelmap displays only POIs that belong to a predefined fixed list of categories. Unfortunately, OpenStreetMap still lacks several relevant POIs in many regions. As a result, numerous restaurants, cafés and markets in Brazil, for instance, are not even available for review.

Wheelmap presents data on a map interface, and POIs are plotted according to a semaphore-like scale: green (fully wheelchair accessible), orange (partially wheelchair accessible) and red (not

[3] https://axsmap.com (retrieved Jul. 2023).
[4] https://www.euansguide.com (retrieved Jul. 2023).
[5] https://wheelmap.org (retrieved Jul. 2023).
[6] https://accessibility.cloud (retrieved Jul. 2023).
[7] https://openstreetmap.org (retrieved Sep. 2021).

wheelchair accessible). Unreviewed POIs are gray. While this color-scale approach gives reviewers some freedom to follow their own interpretation of accessibility, the platform also defines some objective criteria, revealed at the moment users start an evaluation: *"entrance has no steps, and all rooms are accessible without steps"* for green, *"entrance has one step with max. 3 inches height, most rooms are without steps"* for orange, and *"entrance has a high step or several steps, none of the rooms are accessible"* for red. After this general assessment, users may also rate the POI's lavatory as green or red, according to the following criteria: *"doorways' inner width ⩾ 35 inches, clear turning space ⩾ 59 inches wide, wheelchair-height toilet seat, foldable grab rails, and accessible sink".*

But data collection does not have to be exclusively based on active user participation. It is also possible to aggregate passively gathered information by tracking sensor data (*e.g.,* gyroscope, accelerometer, GPS) generated by pedestrians (Biagi *et al.*, 2017; Cardonha *et al.*, 2013; Flores and Manduchi, 2018; Mirri *et al.*, 2016; Palazzi *et al.*, 2010; Prandi *et al.*, 2016, 2017b), wheelchairs (Ding *et al.*, 2011; Edinger *et al.*, 2019; French *et al.*, 2008; Kirkham *et al.*, 2017; Mobasheri *et al.*, 2018a; Moghaddam *et al.*, 2011; Mourcou *et al.*, 2013; Tanimoto *et al.*, 2013), and sensor networks (Gilart-Iglesias *et al.*, 2015; Mora *et al.*, 2017; Pérez-delHoyo *et al.*, 2016).

mPASS (Mirri *et al.*, 2016) aims at generating personalized routes according to a person's needs, based on three different data sources: passive collection from smartphones of individuals moving in a city, active user upload of textual and multimedia data, and official reviews from local authorities and/or associations. In order to increase user engagement in data collection for the system, the authors have also devised Geo-Zombie (Prandi *et al.*, 2016, 2017b), a pervasive game that rewards players for documenting accessibility barriers and facilities.

On top of gathering sensor data during regular wheelchair usage in an urban environment, WheelieMap (Kirkham *et al.*, 2017) also combines video recordings to detect sidewalk obstacles and other accessibility issues. Video segments are selected in correspondence to peaks of movement in accelerometer data, and they are later (manually) analyzed for verification – an approach that also has the benefit of adding an extra layer of auditability to the system.

Other studies have applied data mining for mapping accessibility on OpenStreetMap (Biagi *et al.*, 2020), including sidewalk surface, inclination, pedestrian crossings, barriers, and parking. Mobasheri *et al.* (2018a,b) focused on increasing completeness of sidewalk data and aggregating information from Wheelmap's users.

A different approach is to apply image-processing algorithms to automatically extract geographical accessibility data, *e.g.,* mapping pedestrian zebra crossings (Ahmetovic *et al.*, 2017; Berriel *et al.*, 2017; Coughlan and Shen, 2013; Guy and Truong, 2012; Haider *et al.*, 2019; Koester *et al.*, 2016; Zadrija *et al.*, 2018). These systems are capable of extracting candidate crossings from satellite or street-level images, and their overall accuracy can be increased by combining both kinds of images and using crowdsourcing to further validate acquired data (Ahmetovic *et al.*, 2017).

Project Sidewalk aims at detecting all kinds of sidewalk accessibility problems from Google Street View imagery (Saha *et al.*, 2019). Paid and volunteering crowdworkers use the platform to tag sidewalk obstacles, surface problems, curb ramps, *etc.* A dataset containing more than 300,000 labeled images has been released, and experimental application of deep learning algorithms for automatic labeling has shown that it supports performance levels comparable to human evaluators (Weld *et al.*, 2019). Because labeled images are from major cities in the United States only, the capability of generalization to cities in other countries is not guaranteed (Shankar *et al.*, 2017).

As with any crowdsourced system, quality control is also a concern in the specific case of Volunteered Geographic Information (VGI) systems, but there are other metrics that are also relevant. In addition to usual data correctness metrics suitable for *e.g.,* machine learning datasets, VGIs are sensitive to (spatial) completeness, and also to positional, topological, thematic, and temporal accuracy (Senaratne *et al.*, 2017). When possible, crowdsourced data can be combined with authoritative datasets (provided by disability rights organizations and local authorities, for example) (Prandi *et al.*, 2017a), and data mining can be applied for addressing data sparseness (Mobasheri *et al.*, 2018a,b).

3.6 After all, can we measure accessibility?

Accessibility is an interdisciplinary subject, but because there is no consensus on a single, "one-size-fits-all" definition for it, its meaning may vary between different fields, and sometimes even between different authors in the same field. While the boundaries for interpretation remain loose, the increasing need for accessibility in our urban environments is certainly tight. From this need, arises the requirement for somehow making measurements so we can at least track our progress.

There are critics to the objectiveness of laws and standards (and by extension, of a purely quantitative measurement scale), going as far as calling them a *"McDonaldization process"*, *i.e.,* a massification process for increasing predictability and efficiency that creates perverse consequences:

> *"The McDonaldization process, when applied to universally designed products and environments, can impact and distort concerns for user empowerment and social inclu-*

> *sion. [...] Initiatives for implementation of accessibility through technical standards and strong legislation will only replicate inadequate design solutions without addressing qualitative issues, such as social inclusion and other contextual or cultural considerations."* (Guimarães, 2010).

Modeling accessibility is inherently complex (Froehlich *et al.*, 2019), and any initiative for measuring it that does not take this subjectivity into consideration could be fundamentally flawed. This presents a challenge, as computers are great for automating repetitive tasks and allow for considerable productivity gains, but they are limited to their mathematical, objective nature. Approaches based only on machine learning are not suitable for *direct* assessments of accessibility, and until we come up with a way to add subjectivity to purely computational analyses, they will be limited to peripheral measurements like identifying barriers in sidewalks.

Including humans in the process allows us to benefit from their intuitive capability. Data collection remains as one of the biggest challenges in crowdsourced accessibility maps (Froehlich *et al.*, 2019). Even larger databases often suffer from (geographical) data-sparseness issues due to low adoption and the reliance on voluntary, in-person efforts (Ding *et al.*, 2014; Froehlich *et al.*, 2019). Figure 3.6 shows an example of this issue.

Some schemes suggest using passive monitoring in addition to active crowdsourcing. Their price tag is a strong scalability barrier: setting up pervasive sensor networks can quickly become too expensive. Monitoring smartphones is cheaper, but this increases costs of user acquisition and engagement – having to routinely leave "yet-another app" constantly open makes it harder for people to participate, which in turn makes them less inclined to do so.

Many published works focus on pedestrian accessibility and sidewalk evaluation. Albeit a relevant topic, it is arguably a "first-world" problem. In South American cities, for instance, accessible sidewalks are *exceptions*, not the norm. Even a theoretically perfect map of such kind would only generate marginal gains, as getting out of home is expected to be challenging by default. It may be more beneficial to learn in advance about potential destinations and avoid the hassle of going to inaccessible places – this is, in fact, precisely the target of initiatives captained by persons with disabilities, even in developed regions.[8] Finally, there is a disproportionate focus on cities in Europe and North America, while other regions remain largely underserved (Froehlich *et al.*, 2019).

A minority of academic studies went beyond experimental stages, being limited to few participants and narrow geographical areas. The 13 crowdsourcing studies published until 2019 in the

[8] *"Nothing about us, without us"*: the motto that calls for persons with disabilities having active involvement in the planning of strategies that affect their lives seems to be appropriate in this context too.

ACM Conference on Accessible Computing (ASSETS) and the ACM CHI Conference on Human Factors in Computing Systems (two of the most important computer science conferences related to accessibility) have a median of 153 participants (Mack *et al.*, 2021, Table 4).

It seems that the most successful platforms are community-driven projects. But these open VGIs also face another challenge: because they allow any person to participate, there is a concern about the reliability of their data (Goodchild and Li, 2012). Within the context of accessibility, persons with disabilities can be suspicious about information coming from persons without disabilities.

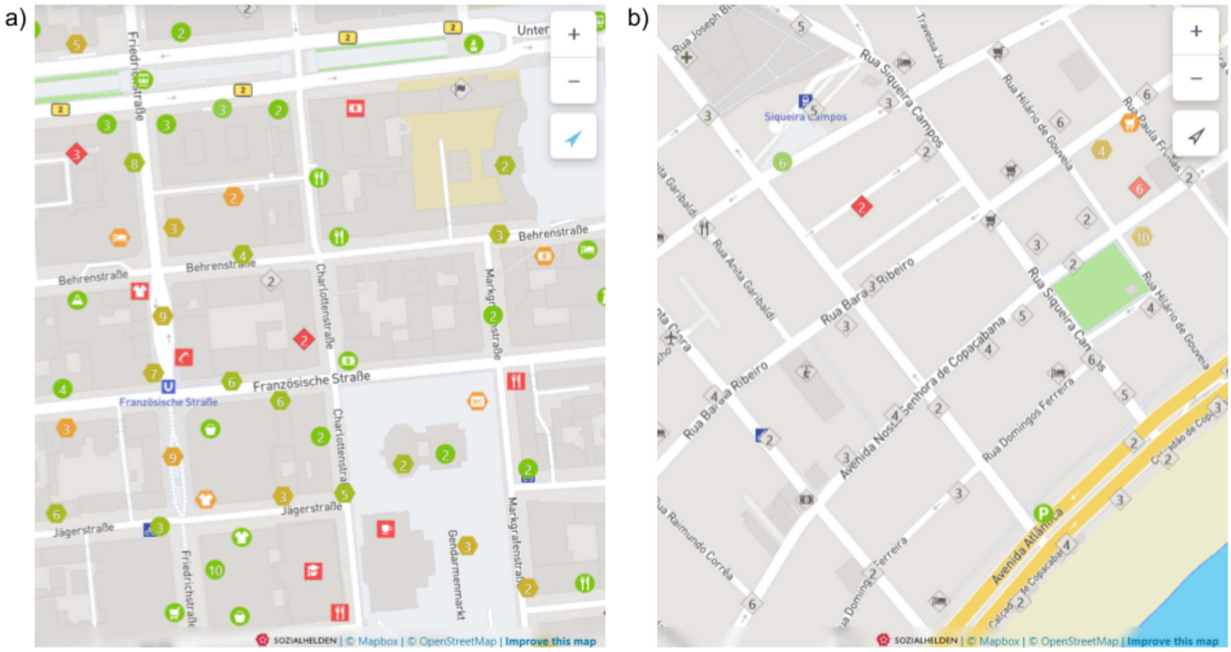


Figure 3.6: *Data sparseness in Wheelmap: spatial density of available data differs when comparing central Berlin (Figure* a*) and Copacabana, Rio de Janeiro (Figure* b*). Both screenshots were taken at the same zoom level. Colored icons denote Points of Interest (POIs) where accessibility information is available, and gray icons represent POIs without data. Numbered shapes mean that neighboring POIs were grouped to minimize visual cluttering.*

Chapter 4

Building a crowdsourcing platform for accessibility

For persons with disabilities, availability of information about accessibility is an important factor in the course of getting out of their homes. When they become travel active, they go through a five-stage process (Yau *et al.*, 2004): (1) coming to terms with their own disability, (2) establishing themselves fully in community life, (3) searching for information to ensure a safe and enjoyable experience, (4) going through the physical journey itself, (5) reflecting about the traveling experience. When choosing hotel accommodations, information about accessibility such as text, floorplans, and photographies is a deciding factor (Darcy, 2010).

Information is fundamental for anticipating whether a place offers the needed degree of support, which depends on the severity of an individual's disability. When it is not available, planning may be hindered, and unpleasant surprises may happen at most inconvenient times.

This not only true for traveling. It also exists in much shorter-term outings, such as going out for shopping, having a meal, or meeting friends. Knowing beforehand whether a place is accessible or not helps avoiding incidents and useless trips, in which after going through the trouble of reaching somewhere, the visit has to be cut short or is rendered impossible due to lack of accessibility.

Crowdsourced accessibility maps are valuable sources of information. However, as previously discussed, many of the existing initiatives are barely more than mere proofs of concepts, with very limited audiences in small geographical areas. Community-driven VGIs generally offer greater amounts of data, but the largest of these platforms also have considerable amerocentric and eurocentric biases, which is in itself a significant problem (Froehlich *et al.*, 2019). Countries in Africa, Asia, and Latin America are mostly destitute of such services.

In what follows, we describe how we built the largest crowdsourced accessibility database in Latin America over the course of 90 months. The approach was based on building a company around an impact-based business model for ensuring financial sustainability. Participation in awards at both national and international level helped by adding an extra layer of credibility to the project, which in turn increased relevance for corporate stakeholders and the public. In order to maximize volunteer participation, the adopted communication strategy was aimed at the broadest audience possible, not limited to persons with disabilities only. Exposure in traditional media and, most importantly, extensive usage of social media were vital for spreading awareness and nurturing a healthy community of contributors, which is the most critical aspect in a crowdsourcing system. Careful design of crowdwork, limiting the scope of action, and simplifying demands helped maximize volunteer participation.

We describe experiments on what affects user engagement. In particular, we investigated the effects of gamification, with and without the addition of extrinsic incentives (Hossain, 2012) such as coupons and gifts. Positive effects were observed even when gamification was performed outside the crowdsourcing platform itself (*e.g.,* in social media).

As with any crowdsourced venture, data quality was also an issue. Obvious vandalism was prevented through content moderation (*e.g.,* filtering comments with blacklisted words), but a more serious concern was commonly voiced by users with disabilities: how can we trust data provided by persons without disabilities, since they "don't know anything" about accessibility? This was addressed through the application of a serious game, detailed in chapter 5, which served a dual-purpose: educating the audience (specially persons without disabilities) and estimating data quality.

4.1 A brief history of Guiaderodas

Guiaderodas is a crowdsourced accessibility guide for persons with disabilities and limited mobility. It was created by a Brazilian social entrepreneur, who, as a wheelchair user himself, was trying to solve his own headache: knowing if a place was accessible or not before leaving home. He was in a constant search for places he could go, and whenever he found an accessible pub or restaurant, he would go there *ad nauseam*, to the point of memorizing their menus. He knew this tool could be life-changing for millions of people who also faced the same challenges.

The vision was to build something that any person could contribute to, even persons without disabilities. Anyone should be able to evaluate any place in the the world, and the tool would aggregate and display all information for interested users to navigate it, for free. It was launched in early 2016 as a mobile app, available for both iOS and Android. GPS coordinates were fed into Foursquare API[1] (later changed to Google Places API[2]) and the platform's own database, and the combined results were displayed as a list containing nearby places. Accessibility levels were represented by colorful pins in a semaphore-like scale: green (accessible), yellow (partially accessible), red (not accessible), or gray (no information available), as seen in Figure 4.1. Users could evaluate places by filling a short questionnaire.

Since launch, the system has grown considerably, reaching tens of thousands of registered users and hundreds of thousands of data points. Between 2019 and early 2020, observed growth was significant. But when the COVID-19 pandemic hit, the platform was hugely impacted. The *"guide that helps you go out"* was suddenly irrelevant. The main audience was part of groups considered to be at highest risk, and advised by authorities to stay home. Nobody knew when (or even if) things would go back to normal, so Guiaderodas had to reinvent itself as a company. Even though the crowdsourcing platform is no longer as central in its strategy as it used to be, it is still available for download and on the web.[3]

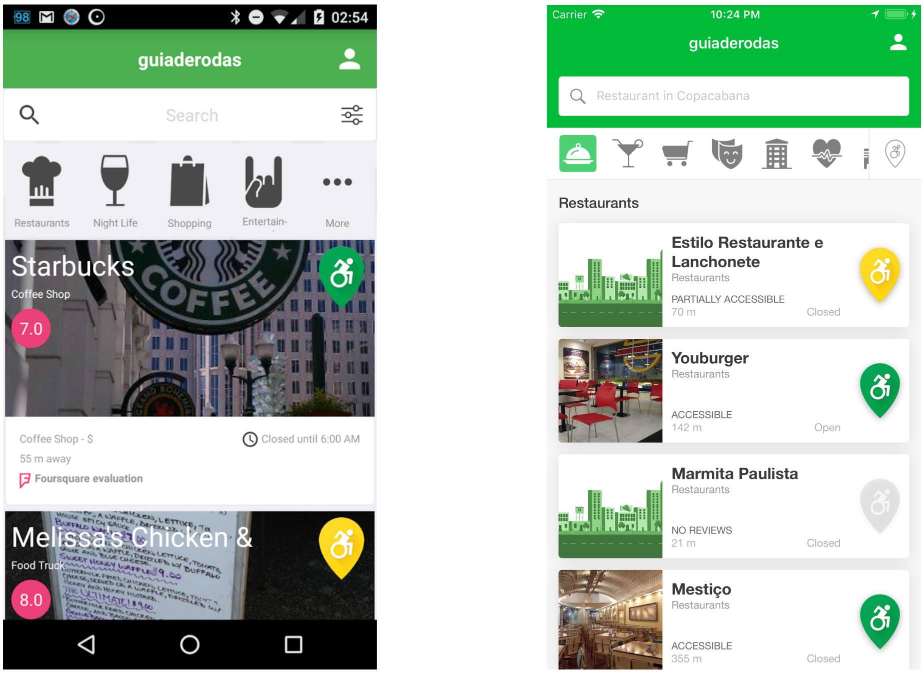
From its conception, Guiaderodas has always been built around a community. It can be understood as an example of crowdcreation (Geiger and Schader, 2014), in which value is derived from the collection of contributions, not individually. Despite being standardized through questionnaires, contributions are non-homogeneous by nature, because volunteers have varying degrees of expertise in accessibility. In fact, persons without disabilities were significant contributors (as discussed in section 4.5), something that has caused questioning about the quality of produced data. This criticism prompted further investigation (chapter 5).

Guiaderodas has served as a complete environment for experiments on crowdsourcing, mainly about user engagement and data quality. It provides the fundaments for this thesis, as events described and data analyzed herein were obtained through the platform between February 2016 and August 2023.

4.2 Scope definition

Given its broad extent and indefinite nature, accessibility is a hard thing to measure. Therefore, it is necessary to establish boundaries as to turn it into a tractable problem.

[1] https://location.foursquare.com/developer/reference/places-api-overview (retrieved Aug. 2023).
[2] https://developers.google.com/maps/documentation/places/web-service (retrieved Aug. 2023).
[3] https://guiaderodas.com/aplicativo-guiaderodas/ (retrieved Aug. 2023).



(a) (b)

Figure 4.1: *The Guiaderodas app. The app's first interface (a), and its redesign, launched in 2019 (b).*

The first choice was to assess physical accessibility for wheelchair users. The fact that Guiaderodas' founder is a wheelchair user himself was undeniably an influence, but there is a much more important reason for it: accessibility for wheelchairs demands physical infrastructure, which can be more direct to evaluate. Because the idea has always been to recruit persons without disabilities as volunteers, verifying the existence of *e.g.*, ramps and elevators is arguably a less daunting task for them than verifying Braille signage and staff's training for serving persons with autism. Even though this reduction of scope means that not all disabilities are covered by the evaluation, it was deemed necessary for the success of the initiative, otherwise evaluation could become too complex, harming user engagement. Nonetheless, the door was not completely closed for adding information about other kinds of disabilities, as discussed in subsection 4.2.2.

Considering the geographical and cultural context Guiaderodas operated in, the choice for focusing on indoor accessibility of venues (restaurants, shops, cafés, etc.) seemed natural. It was based in Brazil, a country in which good sidewalks are the exception, not the norm. This is, in fact, the case of the overwhelming majority of cities in the developing world. So mapping sidewalk obstacles and potholes would arguably result in a rather redundant database devoid of useful information, as users from these regions already expect poor maintenance conditions by default. This is the same approach adopted by some of the most relevant VGIs created by persons with disabilities, such as AXS Map, Euan's Guide, and Wheelmap, which further attests the relevance of this kind of information.

By “indoor accessibility”, the scope includes, in a broad sense, washrooms and baby changing stations, corridors and other internal circulation aspects, counter/tables, etc. Outdoors aspects, however, are not completely ignored. Places’ entrances and parking options are also considered.

In summary, the tool’s purpose can be understood as informing users about what to expect *when* they get to a place – *how* they will get there, whether by car, bus, or walking, is another problem. More formally, the scope is defined as **indoor physical accessibility for wheelchairs**, but without completely ignoring other closely related aspects, such as the direct surroundings and accessibility for other kinds of disabilities.

4.2.1 Embracing subjectivity instead of avoiding it

Whatever the kind of rating scale used, it should ideally be able to take at least some subjectivity into account. One option is to adopt numeric scales, such as grades ranging from 0 to 10 or 5-star rating systems. But these inevitably lead to some sticky questions: if a place is rated “9”, does it mean that it is three times more accessible than a “3”? How much better is a 4-star place, compared to one with 3 stars? And how much worse than a 5-star? Etc.

While supporting comparisons is a useful feature, numeric scales may not be the best alternative. Category-based schemes, such as semaphore-like color-based systems (green, yellow, and red), seem to be more suitable, as they allow comparisons without encouraging mathematical calculations. Variations of this approach have been adopted in some of the biggest accessibility VGIs. They do, however, establish definite criteria between categories. In Wheelmap, users are given specific points to look for, *e.g.,* accessible bathrooms must have *“clear turning space ⩾ 59 inches wide”*. In AXS Map, users respond to a series of yes/no questions, such as *“Does this location have a permanent ramp at the entrance?”*, and the final color category is automatically calculated. Even though the objectiveness is somewhat hidden behind an opaque algorithm, it still exists, as computers are not capable of subjective analysis.

In Guiaderodas’ case, the option was to not shy away from subjectivity at all. Users have three choices: green (accessible), yellow (partially accessible), and red (not accessible). Classification criteria is left entirely for the user to decide. The design assumption is that “accessible” and “not accessible” are self-explanatory – users do not need orientation on what “green” or “red” places look like. The intermediate alternative “partially accessible” leaves some room for wiggle, and implicitly communicates *“when in doubt, choose yellow.”* This, however, creates its own set of challenges that must be addressed by careful design of the work presented to volunteers.

Users can evaluate any place, even ones that are already reviewed. If they do not agree with a place's classification, they can submit their review, and the score is updated. Under the hood, numeric values are assigned to user ratings: 1 (red), 3 (yellow), and 5 (red). A place's rating is the arithmetic mean of its user ratings, and its color is defined as red $\leqslant 2.0 <$ yellow $\leqslant 4.0 <$ green.

This choice is not arbitrary. These specific values produce a desirable behavior for aggregating conflicting reviews. The calculated result gravitates towards either yellow, signaling a controversy, or red, in a conservative approach to preserve user experience (as it is better to be positively surprised with a place's accessibility than the opposite). For instance, one yellow vote and one green vote will result in yellow, whereas one yellow vote and one red vote will result in red. Only a majority of green votes results in green. Table 4.1 gives some examples of this behavior. Considering that the variables are qualitative, other statistical procedures could have been adopted. However, as the goal is to aggregate rather than summarize, this choice was deemed appropriate.

Votes			**Result**	
Red	Yellow	Green	Score	Color
0	1	1	4.0	Yellow
0	1	2	$4.\overline{3}$	Green
1	1	0	2.0	Red
1	2	0	$2.\overline{3}$	Yellow
1	0	1	3.0	Yellow
2	0	1	$2.\overline{3}$	Yellow
3	0	1	2.0	Red
1	0	2	$3.\overline{6}$	Yellow
1	0	3	4.0	Yellow
1	0	4	4.2	Green

Table 4.1: *Examples of vote aggregation in the system.*

4.2.2 Crowdwork design

Complexity of work done by participants and clarity of given instructions directly affect data quality and user participation (Eickhoff, 2014; Kittur *et al.*, 2013). There are strong reasons for keeping tasks and entry barriers to a minimum. Good crowdwork design can be decisive, even in cases when workers have low expertise in the subject. It can also help compensate personal cognitive biases, ever present in crowdsourced systems (Eickhoff, 2018) – and even more in a subjective context like this.

Considering this, any evaluation work that demands longer times (*i.e.,* minutes or hours) or demands difficult analyses (*e.g.,* questions like *"are tables at least 31 inches high?"* may be hard for the average person, who does not usually carry measuring devices) is not appropriate.

The selected approach is based on a questionnaire, which is kept as simple and straightforward as possible, and ideally takes seconds to complete. It is divided in three parts:

1. A single mandatory multiple-choice question *"Do you consider this place as accessible?"* about the whole place in general, in which users must choose between red, yellow, or green.

2. At most 6 optional multiple-choice questions about specific features of the place (*e.g.,* entrance, washrooms);

    - Users can choose between red, yellow, or green.
    - Questions are presented in two pages with at most three questions each.
    - Presented questions depend on the place's category.

3. An optional free-text comment field, for adding relevant information that may not have been covered by previous questions (up to 1000 characters).

The same accessibility features are not relevant in all kinds of places. For example, the existence of tables with appropriate height for wheelchairs is relevant in a restaurant, whereas a a drugstore does not have tables at all. For this reason, every place in the system is assorted into one of 9 possible categories, listed in Table 4.2. They were defined based on the categories existing in services like Foursquare and Google Maps at the time of creation of Guiaderodas.

Because restaurants are a common destination in leisure outings, they were reserved their own category, which also includes bakeries and cafés. Bars, pubs and night clubs comprise the Night Life category, while malls, stores, and markets are included in Shopping. Cinemas, theaters, and museums fall into the Entertainment category. Hotels and motels are listed as Accommodation. Health includes hospitals, pharmacies, etc. Buildings are office buildings, and Public Places include government offices, courthouses, places of worship, etc. Others is an umbrella category for diverse kinds of places, such as parks, schools, libraries, airports, etc.

Each category corresponds to a questionnaire that can be presented to users, comprised by different combinations of the same question pool. Table 4.3 shows all possible questions and the questionnaires corresponding to each category. Figure 4.2 shows an example of what a review looks like for users of the mobile app.

As discussed in sections 3.4 and 3.5, questionnaires have already been used in several different ways for measuring accessibility. Still, they are not the only possible approach for applying crowdsourcing to this goal. Another possibility would be having users vote on photos of places and

Category	Name
1	Restaurants
2	Night Life
3	Shopping
4	Entertainment
5	Accommodation
6	Health
7	Buildings
8	Public Places
9	Others

Table 4.2: *Place categories in the system.*

	Category								
Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Do you consider this place as accessible? *	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is there parking space for handicapped and/or valet parking?	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is the entrance appropriate for persons with limited mobility?	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is internal circulation appropriate?	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is there counter/table with appropriate height for wheelchairs?	✓	✓		✓					
Are there rooms for persons with disabilities?					✓				
Is there an accessible bathroom?	✓	✓	✓	✓	✓	✓	✓	✓	✓
Is there a family bathroom or baby changing room?	✓	✓	✓	✓	✓	✓		✓	✓
Is there an electricity generator?							✓		
Write your evaluation on this location. **	✓	✓	✓	✓	✓	✓	✓	✓	✓

* Mandatory question.
** Free text answer.

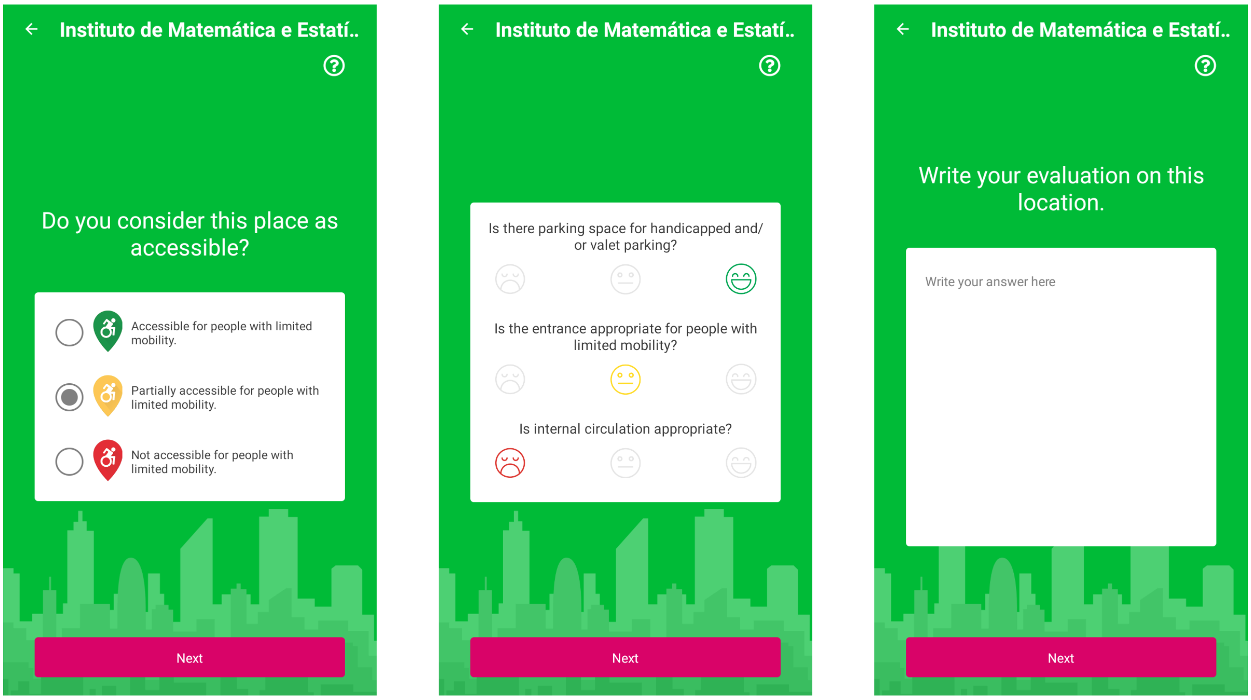
Table 4.3: *Evaluation questionnaire for each place category.*

decide whether they are accessible or not (an adaptation of the method discussed in chapter 5). This, however, could break down and lead to wrong conclusions due to photos being outdated, not adequately representing conditions of a place, etc. Questionnaires implicitly demand that respondents be physically present at the place being evaluated, or at least frequently visit it.

Improvements after user feedback

A common feedback received from persons without disabilities was that they were not sure what an "accessible" environment should look like. To provide some guidance, a "help" button was added to the evaluation screen which showed examples of what reviewers should pay attention to, *e.g., "accessible entrances should have ramps or platform lifts"*, *"accessible restrooms should have handlebars and space for maneuvering a wheelchair"*. These helping texts were carefully chosen to be informative while avoiding the definition of strict guidelines, and therefore keep some room for personal interpretation.

But even after the addition of these help texts, some users still reported they had occasionally



Instituto de Matemática e Estatí..

Do you consider this place as accessible?

○ Accessible for people with limited mobility.

● Partially accessible for people with limited mobility.

○ Not accessible for people with limited mobility.

(a)

Instituto de Matemática e Estatí..

Is there parking space for handicapped and/or valet parking?

Is the entrance appropriate for people with limited mobility?

Is internal circulation appropriate?

(b)

Instituto de Matemática e Estatí..

Write your evaluation on this location.

(c)

Figure 4.2: *Review questionnaire in the app. The only mandatory question is a generic question about the whole place, which is presented first and on its own page (a). This is followed by two pages, each with up to 3 extra multiple-choice questions about other specific features of the place, depending on its category (b). The last page lets the user write a free-text comment (c). All pages feature a "?" button on the top-right corner, which displays help texts suggesting tips for answering each question.*

given up reviewing a place due to this feeling of insecurity. This particular issue has great relevance because it directly impacts both quantity and quality of collected data. The search for solutions resulted in the creation of a game that was initially designed to encourage users and minimize review abandonment, but evolved into a tool for gauging data quality and detecting user bias. This methodology can be generalized, and is discussed in chapter 5.

4.3 Implementing a social impact business

Social ventures implement business models that function according to market mechanisms, aiming to increase the quality of life for all (Jäger *et al.*, 2020). They differ from traditional companies because their products and services generate positive impacts that can be either societal, economical, environmental, or a combination of all three. The (paying) *customers* provide the necessary revenue for footing the company's bills, but they are not necessarily the same stakeholders as the *beneficiaries*, which are the ones who benefit from the social venture's activities, both directly or indirectly, and possibly at no cost.

As an example impact company, take *Mamut*,[4] from Bolivia, which recycles car tires into rubber materials for construction. It operates in the regular market competing with traditional suppliers, but generating a positive collateral effect: the more its sales grow, the more tires are prevented from going to landfills and oceans. Another example is *Recupera Tu Silla*,[5] from Colombia, which revamps trash-worthy furniture into as-new office furniture sold to companies, and into student desks donated to schools in rural areas. On top of reducing waste, their business supports education in remote regions of the country, and employs rehabilitated ex-guerrilla fighters from the Colombian civil war in their manufacturing process. In both cases, their *customers* are other companies. Their *beneficiaries* include the environment, and in Recupera Tu Silla's case, their employees, school children, and their families.

Financial viability in Guiaderodas is guaranteed through an impact-based business model. The costs associated with keeping the crowdsourced accessibility guide online are covered by offering paid services to companies. Information is provided to the beneficiaries – persons with disabilities or limited mobility, their friends and family, and the general public – for free.

The *Guiaderodas Certification* is a multidisciplinary program that evaluates accessibility throughout companies' operations. It includes architectural assessment of their premises, mandatory attitudinal accessibility training of their staff, and awareness campaigns using the Guiaderodas app. Companies pay to be admitted into the program and evaluated. Upon achieving excellent results in all criteria, they are certified, or else, they have up to 12 months to implement improvements. Among certified companies, are included a few international banks operating in Brazil, some publicly traded companies in global stock markets, among others.

This business model is similar to the *B Corp Certification*,[6] which evaluates governance practices, and societal and environmental performance. Both serve the same purpose: to be exempt certificates of a company's efforts for better Environmental, Social, and Governance (ESG) (Friede *et al.*, 2015), a significant matter for firms dedicated to meeting investors' ever increasing demands for sustainability (Amel-Zadeh and Serafeim, 2018; Gillan *et al.*, 2021).

The Guiaderodas Certification is deemed relevant thanks to two main factors. Firstly, the startup has amassed the largest audience in the segment (a topic further discussed in the following sections). Secondly, it has won several social entrepreneurship prizes and awards throughout the years – *"Best Digital Solution for Inclusion"* by the United Nations (2017), *"Best App for Accessibility"* by the International Telecommunications Union (2018), *"Innovators Under 35"* by

[4] https://grupo-mamut.com (retrieved Aug. 2023).
[5] https://recuperatusilla.com (retrieved Aug. 2023).
[6] https://www.bcorporation.net (retrieved Aug. 2023).

MIT Technology Review (2018), *"Best Companies with Social Impact in Latin America"* by VIVA Schmidheiny (2019), to name a few – which help cement Guiaderodas' authority and reputation. As a result, the certification is a valuable ESG asset for firms of any size.

4.4 Strategies for user engagement

From a user's perspective, the utility of Guiaderodas is strongly correlated to how much relevant information it provides. If there is a large amount of available information, it is more useful; if there is none, it is useless. Due to its crowdsourced nature, information availability depends on volunteer input. Therefore, if there is a large number of contributing volunteers, potential users are more likely to join. Conversely, if there are few volunteers, the probability of conversion is low. The platform's value for any given user increases every time a new volunteer joins. This textbook-like case of *network effect* (Shapiro and Varian, 1998, chapter 7) presents a dilemma: how to acquire users when the database is still empty and utility is virtually zero?

Because this crowdsourced information is geographical, the "kickoff challenge" must be faced multiple times: the first user in any given region will always find no information about their immediate vicinity. When this happens, they may see the platform as useless and quit; but if their pioneering spirit compels them to review a nearby place, the next user is more inclined to do the same. As more local volunteers join, the amount of available information grows, creating a positive feedback loop that accelerates user acquisition in that area. When a critical mass is reached, the regional kickoff challenge is overcome.

Creating incentives that tap into this "pioneering spirit" is key. The approach for boosting engagement can be divided in two pillars:

1. Communication targeted at the broadest audience possible;
2. Gamification appealing to people's sense of community contribution;

Communicaton

A crucial aspect for any crowdsourcing effort is to communicate its own existence. Because no one can participate in something they are not aware of, the upper bound for volunteer count is defined by the number of people who have at least heard about the platform. Moreover, greater awareness and audience engagement positively correlate with greater data completeness (Mobasheri *et al.*, 2018b).

Guiaderodas benefitted from remarkably early media exposure. About a week after launch, it had already been featured on one of the most popular Brazilian morning television shows[7] and many vehicles of national and regional relevance. This was also an important factor for reducing the initial "there's no nearby information" frustration: as stories made clear that the platform was new, users adjusted their own expectations accordingly. This allowed for quickly overcoming the initial engagement barrier in many Brazilian cities.

In the following years, Guiaderodas was continually featured on Brazilian media.[8] An aspect that promoted exposure was the recognition received through national and international prizes, mentioned in section 4.3. Media channels often pick up these stories for publishing,[9] generating buzz with little to no extra effort.

From launch, the app was published in both App Store (Apple/iOS) and Play Store (Android) featuring full localization in three languages: Portuguese, English, and Spanish. This meant that all strings were translated, date format was appropriately changed, etc. But for all languages, the same name *"Guiaderodas"* was adopted. Following advice from Apple's App Store managers, the app was renamed in 2019 to *"Wheelguide"* and *"Guiaderuedas"* for English and Spanish, respectively. Depending on their smartphone language preferences, users see an app that was named differently, but was essentially the same. Logos, marketing material, and privacy policies were updated, and corresponding internet domains were also registered. Figure 4.3 shows different app versions.

Social networks

While media exposure was fundamental for broadcasting the existence of the platform, social networks such as Instagram[10] were channels of major importance for increasing user engagement and educating potential contributors. As previously mentioned, targeting the largest possible audience is worthwhile. This meant reaching out to persons without disabilities nor friends with disabilities, an audience for whom accessibility is often a foreign topic. For this reason, published content focused mainly on teaching about accessibility and showing how accessibility benefits all, even persons with no disabilities.

Having limited resources, the option was to target the Brazilian audience, limiting content production to Portuguese only, despite the interest in growing the audience abroad and the platform's

[7] https://globoplay.globo.com/v/4839474/ (retrieved Aug. 2023).

[8] Examples: https://recordtv.r7.com/sp-no-ar/videos/calcadas-desafiam-cadeirantes-em-sao-paulo-04062022 and https://www.estadao.com.br/pme/vivencia-de-fundadores-e-diferencial-em-startups-de-acessibilidade/ (retrieved Aug. 2023).

[9] Example: https://www.baguete.com.br/noticias/20/11/2018/mit-destaca-empreendedores-brasileiros (retrieved Aug. 2023).

[10] https://www.instagram.com/guiaderodas (retrieved Aug. 2023).

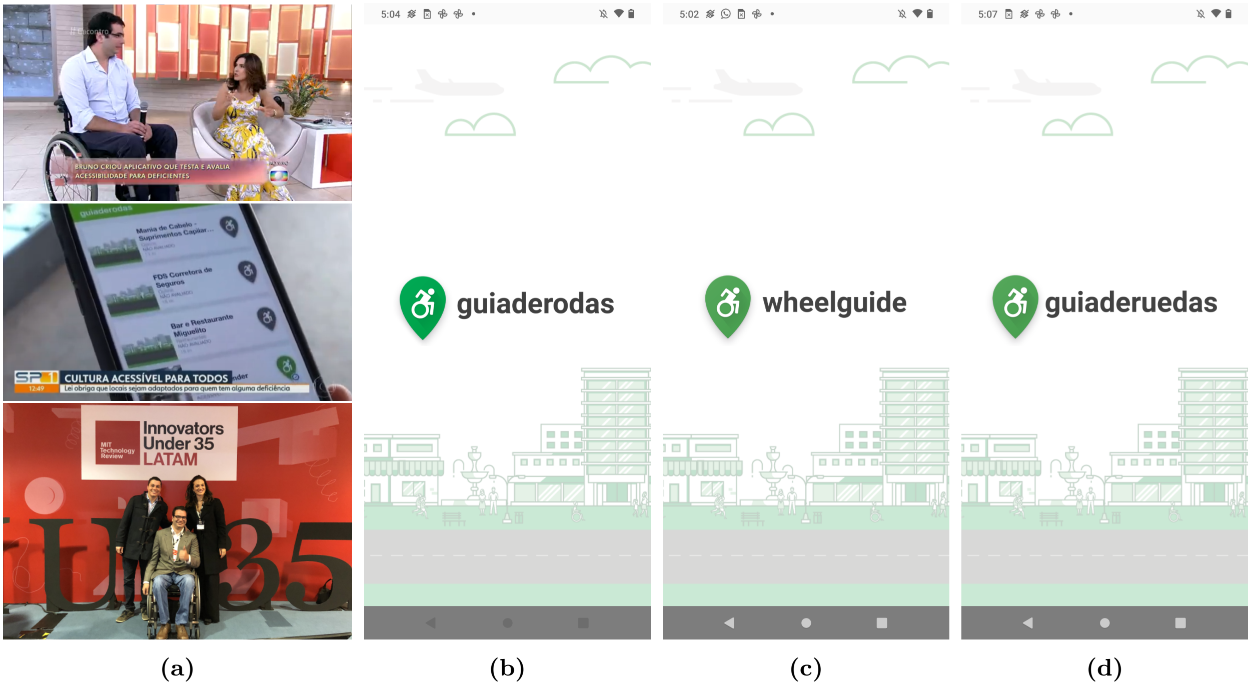

BRUNO CRIOU APLICATIVO QUE TESTA E AVALIA ACESSIBILIDADE PARA DEFICIENTES
CULTURA ACESSÍVEL PARA TODOS
Lei obriga que locais sejam adaptados para quem tem alguma deficiência
Innovators Under 35 LATAM
MIT Technology Review
guiaderodas
wheelguide
guiaderuedas
(a)
(b)
(c)
(d)

Figure 4.3: *Examples of media exposure and localized versions of the app. Guiaderodas was featured numerous times on the press,* e.g., *in talk shows and regional news, and the app was awarded by several international prizes of innovation, which also generated media buzz (a). The app was available in Portuguese (b), English (c), and Spanish (d), under different names. Depending on a user's language settings, they would see either Guiaderodas, Wheelguide, or Guiaderuedas in both App Store (iOS) and Play Store (Android). Apart from that, the app was essentially the same.*

name already being internationalized.

Besides purely educational content, storytelling posts were crafted, showcasing real stories of users that benefitted from the platform and highly participative volunteers. The goal was to encourage intrinsic motivators for participation, such as altruism, charity, identification, etc. (Hossain, 2012), and to thank participants and explain the importance of requested tasks, which also increase motivation (Chandler and Kapelner, 2013).

Paid traffic generation was also a technique applied. Advertisement campaigns were crafted for both user acquisition and engagement. New users were invited to follow Guiaderodas' social media profiles and download the app, while already registered users were prompted to review places in their vicinity. These marketing campaigns ran in Facebook, Instagram, YouTube, and Play Store. Their impact on platform usage is reported in the following sections.

Experiments with gamification

While the previously mentioned communication strategies happened autonomously in relation to this research (and were beneficial to it), they also served as a platform for conducting experiments on the effect of adding incentives on the engagement of crowdsourcing volunteers. In particular, gamification strategies were implemented in a variety of ways, as an exploration of aforementioned intrinsic and extrinsic motivators.

Firstly, gamified content was published throughout several months, in the form of a simple game in which users voted whether the place pictured in a photo seemed to be accessible or not. It served a twofold purpose: educating the audience and reminding them about evaluating places in the crowdsourcing platform. This strategy is further discussed in chapter 5, and it was also the embryo of another experiment on a different topic: data quality.

Another strategy was also experimented with, in a more competition-like fashion and directly linked to the desired crowdwork output: awarding the persons for reviewing places in the Guiaderodas platform. For the duration of the experiment, a monthly competition was carried out, rewarding contributors with the highest numbers of reviewed places. Whenever the defined period of time expired, winners were published, scores were reset, and a new cycle began. Offered prizes included extrinsic incentives of two types (Hossain, 2012): social incentives (*e.g.,* publishing the winner's name and picture), and financial incentives (*e.g.,* coupons and free products).

Both games received positive feedback from users and their creation correlates with an increase in both user registration and place reviews in the crowdsourcing platform. These results are discussed in subsection 4.5.3.

4.5 Results

The results herein presented refer to data collected in the Guiaderodas platform between the app's public launch in February 2016 and July 2023, a period of 90 months.

4.5.1 Crowd profile

During this time, a total of 27,318 accounts were created in the platform. Users could either define an email/password pair, or login with a an external account via Oauth2 (Hardt, 2012). Among the supported services, Facebook was the most common, used by 17,652 (64.6%) users.

Filling demographic information was initially mandatory, but after a redesign of the platform aimed at increasing crowd acquisition it was made optional (more information on this topic later in this chapter). 4,959 (18.14%) users informed their gender. 3,791 (13.9%) users provided data that included their date of birth, whether they had disabilities or not, and if they had friends with disabilities or not. Figure 4.4 shows the demographic profiles of these users. Genders were approximately evenly divided, with a slight majority of males. The same happened for users with

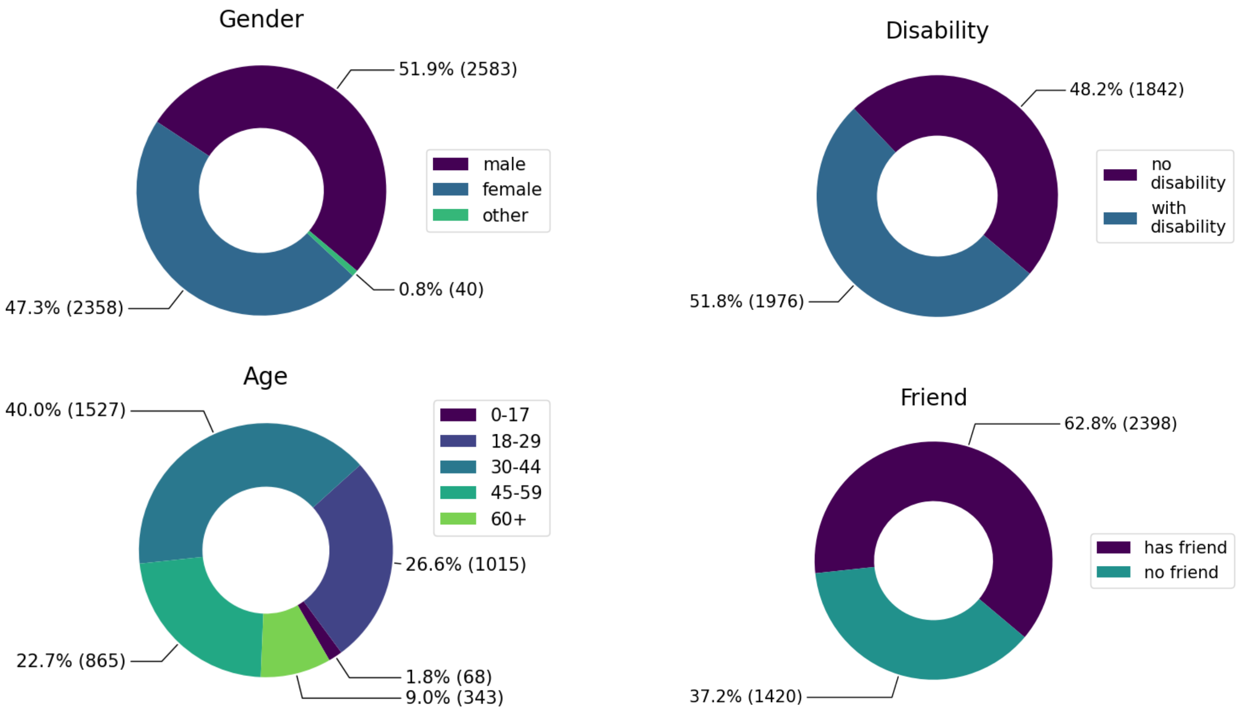


Figure 4.4: *Crowd demographics.*

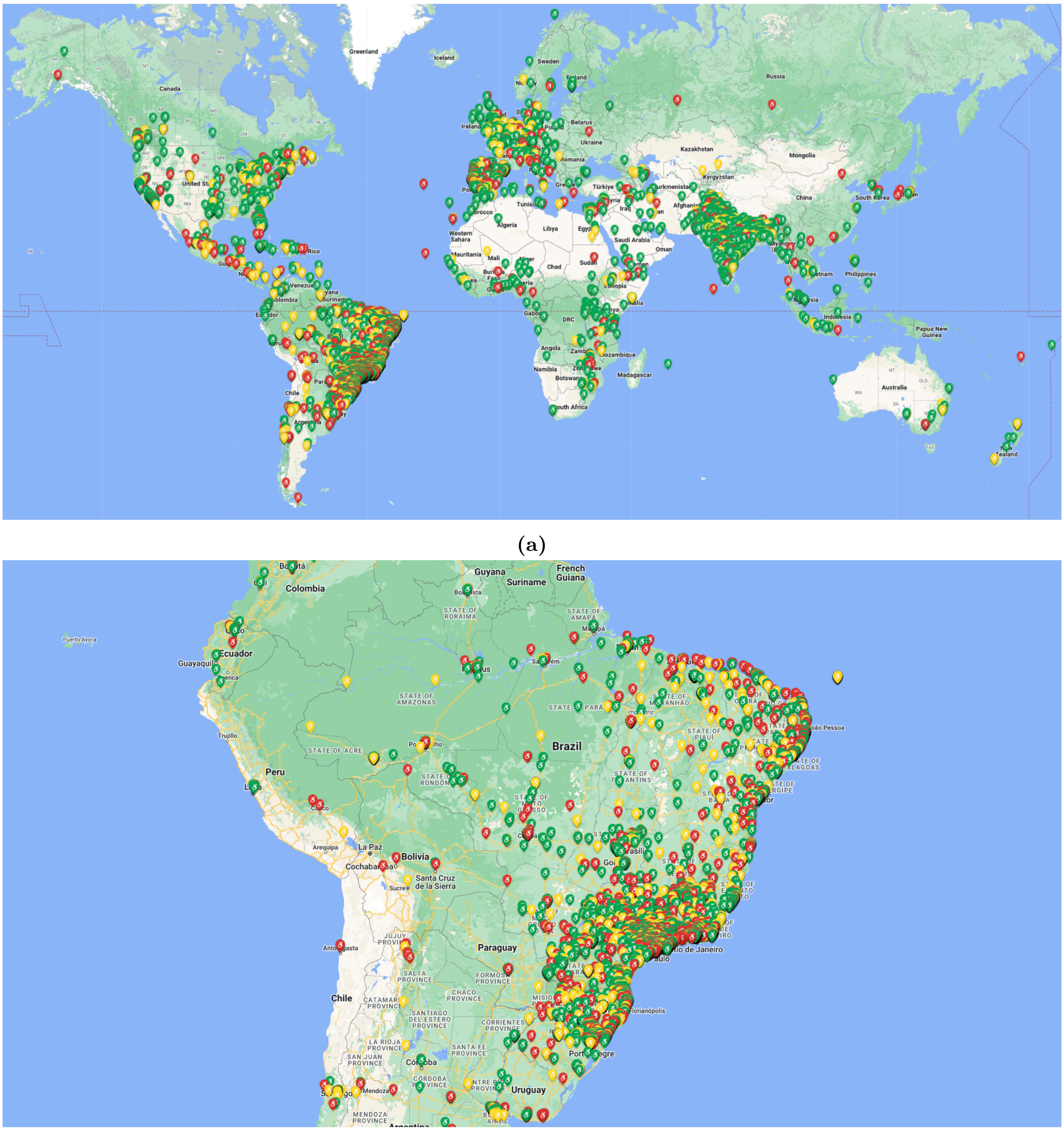
and without disabilities, with a slim advantage for persons with disabilities. The largest age group was between 30 and 44 years old, with two-fifths of the users, followed by users between 18 and 29, and between 45 and 59. 62.8% of the users reported having friends or family with disabilities, and 37.2% did not. Of this group, about half of them did not have disabilities, meaning that 19.3% of the users had no disabilities nor any friends with disabilities. This can be credited to the communication strategy of targeting a broader audience.

4.5.2 Crowdwork output

These users generated a total of 316,527 data points about 35,483 places, including 277,641 answers to multiple-choice questions, and 15,285 free-text comments. The remaining data points (23,601) are related to the Guiaderodas certification program and were not included in this analysis.

The distribution of reviewed places in the world can be seen in Figure 4.5. There were reviewed places in 139 countries, with considerable geographical distribution in India, Europe, and the East Coast of the USA. The country with the largest density of reviews was the Cayman Islands, a small nation in the Caribbean Sea, with 1.08 reviewed places per 1000 inhabitants.

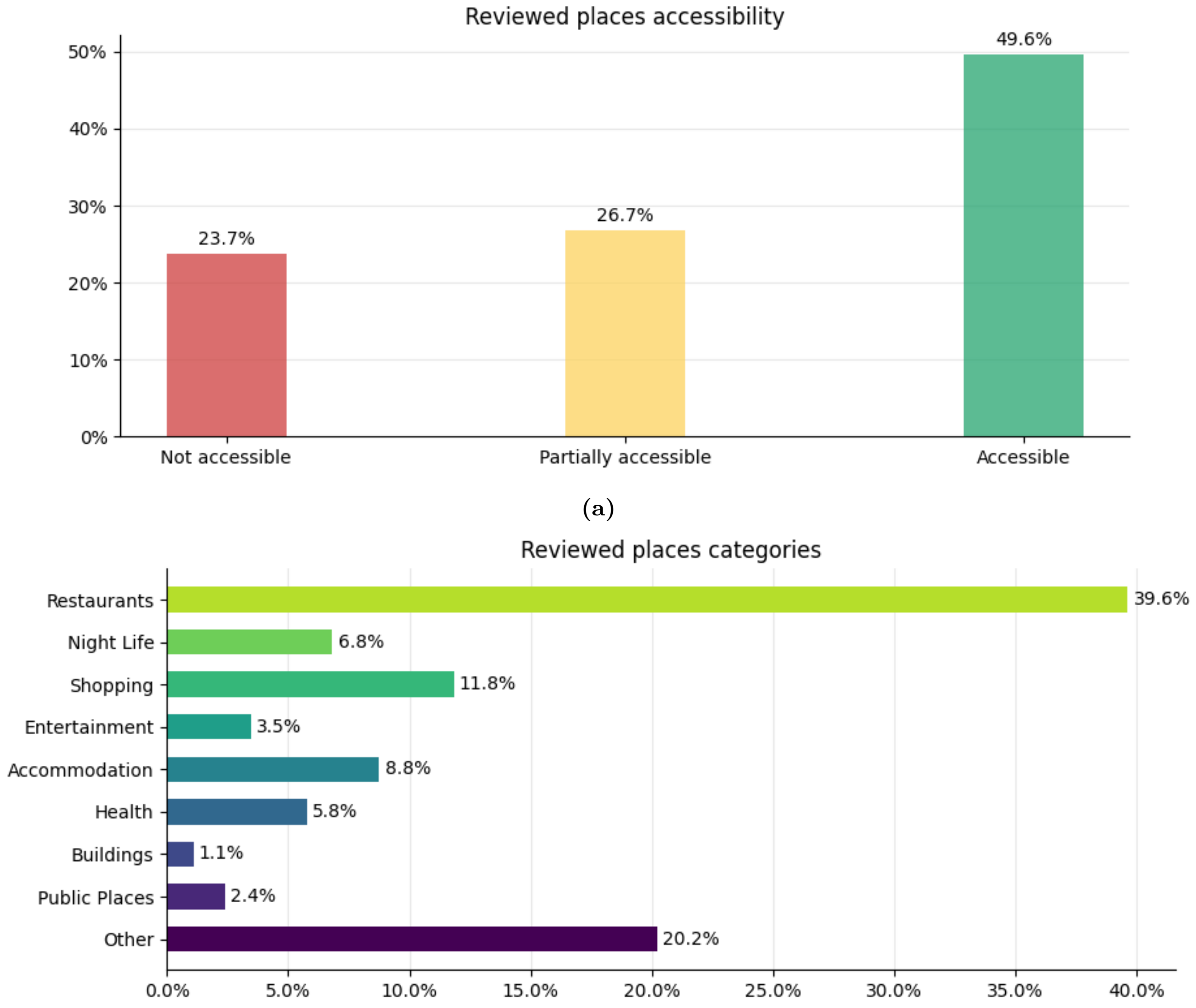
Nonetheless, Brazil concentrates more than 85% of reviewed places, another direct consequence of communication approach. Southern and Southeastern states account for almost 85% of reviewed places in Brazil, totaling almost 72% of all places reviewed in the platform.



(a)

(b)

Figure 4.5: *Map of places reviewed in the platform.*



(a)

(b)

Figure 4.6: *Reviewed places accessibility and categories.*

Figure 4.6 shows a breakdown of how the places were evaluated and their categories. Almost half of reviewed places were rated as “accessible” (Figure 4.6a). Empirically, this does not correspond to the perceived reality, specially in Brazilian cities, where lack of accessibility is rampant. But this proportion shows an important behavioral bias of volunteers: they tended to review more accessible places than inaccessible ones.

Another interesting usage pattern arises in Figure 4.6b, which shows the categories of reviewed places, which are the same listed in Table 4.2. Over three-fifths (61.7%) of them are places usually visited in leisure activities, such as restaurants (“Restaurants”), bars and pubs (“Night Life”), malls, stores, and markets (“Shopping”), and entertainment venues like cinemas, museums, and stadiums (“Entertainment”). This result is not unexpected, considering that the tool had always been marketed as a guide for “helping users to go out”. Furthermore, it also indicates that volunteers often reviewed places when they were “going out” themselves.

4.5.3 Crowd acquisition and engagement

Increasing the size of the crowd (crowd acquisition), and having them perform the desired work (crowd engagement) are two of the biggest challenges in crowdsourcing. In this context, the following metrics are most relevant:

1. User signups (*i.e.,* creation of new user accounts), which directly measures crowd acquisition;

2. Number of active crowdworkers (*i.e.,* users that generated valid output in a given time frame), which indicates the number of volunteers contributing to the platform;

3. Number of reviewed places, a direct measure of the amount of produced output.

Although crowd acquisition is always fundamental, increasing the number of volunteers does not necessarily imply an increase of the crowd’s throughput. Not all crowd members create equal value. The Pareto rule is a heuristic suggesting that 20% of contributors often create 80% of value, and crowdsourced systems commonly provide empirical evidence to support it (Täuscher, 2017). As Figure 4.7 shows, indeed a small group of volunteers was responsible for most contributions, while almost two-thirds (65%) of registered users did not perform any work whatsoever. By contrast, the group of more engaged users who reviewed at least 3 places comprises only 9.8% of the population, but generated more than three-quarters (76.9%) of the output.

Identifying and nurturing healthy relationships with highly engaged workers is desirable. For this reason, crowd engagement is measured as a combination metrics number 2 and 3. An increase

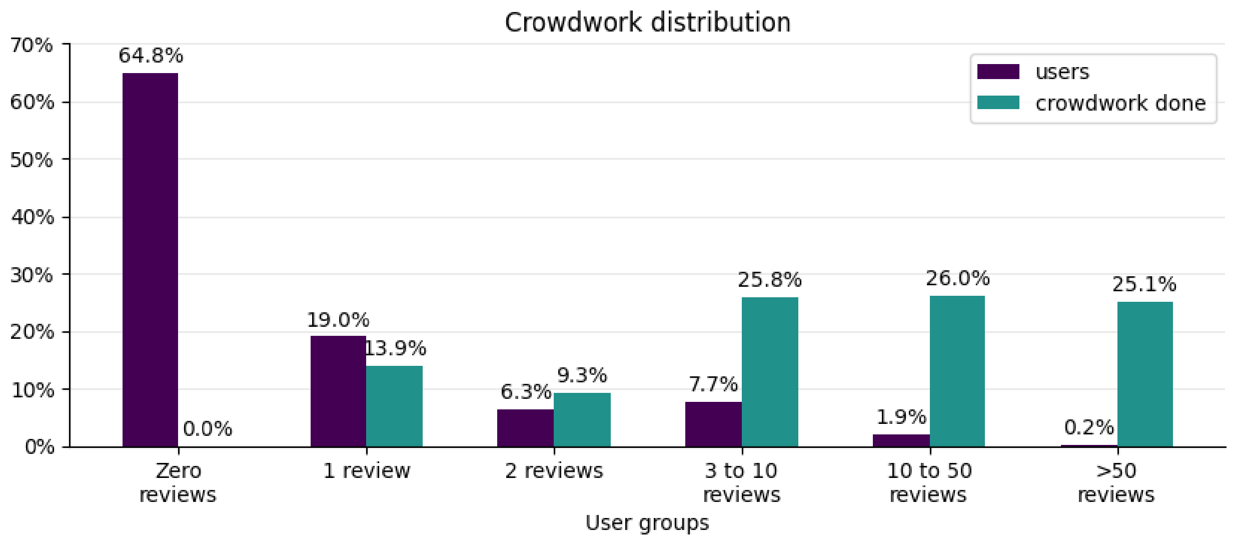


Figure 4.7: *Crowdwork output by engagement level.*

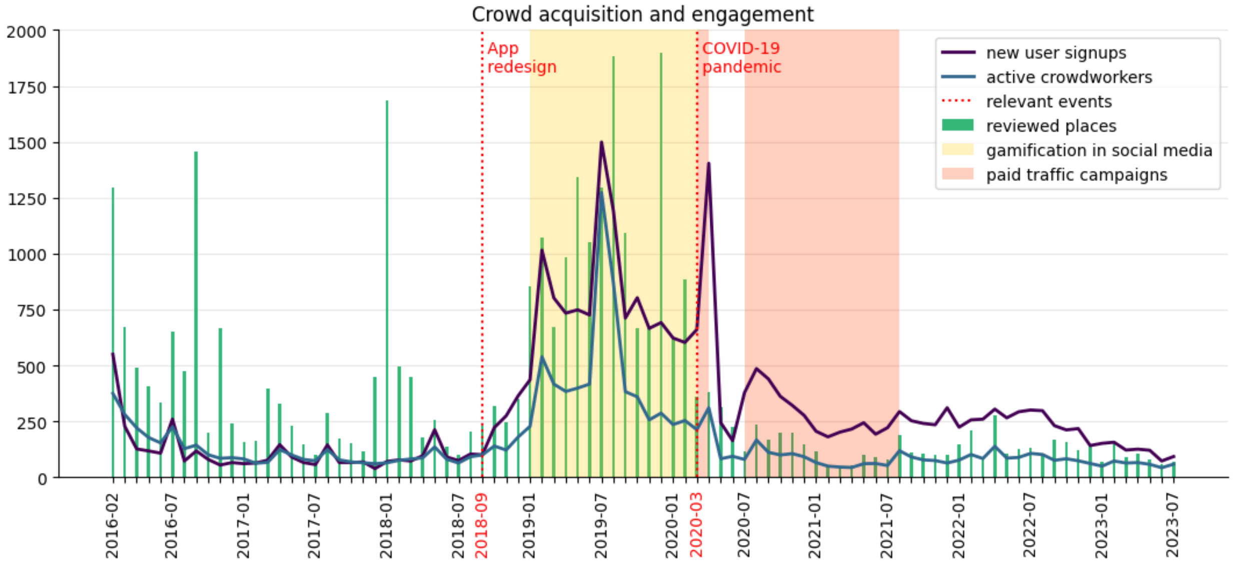


Figure 4.8: *Crowd acquisition and engagement over time.*

in the number of reviewed places without a proportional increase in the number of reviewers means that at least one (potentially high-value) user has reviewed multiple places.

Figure 4.8 displays the three key metrics over time. Media exposure in early 2016 generated an influx of new users, and because it was clear for most users that the platform was new, many contributors were pioneers that reviewed numerous places as a way to help reaching critical mass and gaining traction. Over time, the effect of this initial burst was attenuated. During 2017 and most of 2018, numbers of monthly new users and active crowdworkers stayed essentially stable and close to each other, indicating that users created accounts exclusively when they wanted to review a place. Occasional bursts of the number of reviewed places can be observed. They were caused by highly motivated individuals, but these stints usually did not last for long.

In September 2018, a redesigned version of the app was launched, featuring two major changes. The first one was motivated by analytics data showing that an alarmingly high proportion of would-be volunteers clicked on the button to review a place, but gave up when immediately presented with a screen demanding them to either log in or create an account. The solution was to move this screen to the end of the reviewing process. The rationale was that after spending some effort in the review, users would be less inclined to abandon it. Additionally, a mandatory demographic survey was made optional. The second change was to offer users the option to create an account when they opened the app for the first time. Even though this was not mandatory, the assumption was that many would accept (specially because with Oauth2, accounts can be created in a few clicks), and by collecting e-mail addresses, further engagement strategies were made possible. After the launch of this new version of the app, the number of newly registered users started to increase, but even though the number of active crowdworkers grew, the change is not so obvious that it can be directly attributed to the redesign alone.

In early 2019, experiments with gamification in social media channels were started. The hypothesis was that by creating compelling content in other platforms, users would be inspired to participate in the crowdsourcing process. The first method implemented was a poll-based game in which users voted whether pictures represented accessible places or not (this topic is more thoroughly discussed in chapter 5). In May 2019, a competition that awarded crowdworkers with most reviews in each month was introduced. Initially, awards included publishing the winner's name and picture, but other incentives such as coupons and t-shirts were added in July 2019. The peak seen in the graph corresponds exactly to this event. These prizes were distributed for the remainder of the experiment, but the observed boost on user engagement wore off gradually in the following months.

For the duration of the experiment, all relevant metrics reached higher averages. When compared to the preceding time period of same length using the Kruskal-Wallis test (Kruskal and Wallis, 1952), this increase is shown to be statistically significant for new user signups (p-value $= 6.70 \times 10^{-6}$), number of active crowdworkers (p-value $= 6.67 \times 10^{-6}$), and number of reviewed places (p-value $= 7.76 \times 10^{-5}$).

March 2020 was the start of the COVID-19 pandemic, a point of time marked by general uncertainty. The public was advised by sanitary authorities to stay home if possible, and persons with disabilities and the elderly, in particular, were deemed to be at higher risk. All gamification experiments were suspended. As a consequence, usage of the crowdsourcing platform dove into all-time lows. However, this period coincides with an anomalous increase in Figure 4.8, due to another factor: paid advertising campaigns, which were also started in March 2020. Unrelated to the pandemic, a big marketing initiative focused on user acquisition was deployed. The ads invited potential users to download the app on Facebook and Google display networks, which include several channels (Facebook and Google proper, but also Instagram, YouTube, Play Store, banners on third-party websites, etc.). It ran for about five weeks before being interrupted for a few months, and later resumed. While this strategy is outside the scope of this study, it explains why the second peak observed in April 2020 affects new user signups, but not the other metrics, as the ads were focused on user acquisition only. The subsequent drop during the interruption of the campaigns between April and July 2020 is also noticeable in the graph.

From August 2021 onwards, neither gamification in social media nor paid traffic campaigns were operating. In this period, the number of active crowdworkers fell to the same levels that preceded the app's redesign, but user signups stabilized around higher averages. This can be credited to the app's experience that favors creation of new accounts, and statistical significance was again checked using the Kruskal-Wallis test, which yielded a p-value $= 0.28$ for active crowdworkers and p-value $= 1.28 \times 10^{-7}$ for new user signups. For the number of reviewed places, the p-value $= 4.24 \times 10^{-3}$ indicates significant difference, which can be credited to the lack of occasional burts of reviewed places.

4.5.4 Crowd behavior

Another investigation path was to check if demographic characteristics affected user behavior. Firstly, engagement levels were analyzed, and they were not equal among all groups. For each demographic trait, groups were compared using the Kruskal-Wallis test (Kruskal and Wallis, 1952). Results are summarized in Table 4.4. The resulting *p-values* indicate the probability that these results would be observed assuming that the null hypothesis is true, *i.e.,* the probability that they are due to chance if there are no differences between the groups. As a common rule of thumb, differences are considered to be statistically significant if the resulting p-value ≤ 0.05. Lower values (*e.g.,* 0.01, 0.001) are also frequent in some stricter scenarios. Considering that nowadays these calculations are done by computers, it seems natural to fully publish p-values and let any decisions about significance to be done by readers themselves.

Kruskal-Wallis was chosen because it is a nonparametric method of analysis of variance. One-way ANOVA was initially considered, but it depends on the assumption of data normality, which is unjustified in this case. Nonetheless, an extra cautionary step for validating these results was also performed: one-way ANOVA was calculated over the rank transformation dataset (*i.e.,* re-

	Group	**Average reviews**	**Standard deviation**	**p-value (Kruskal-Wallis)**
Gender	male	3.63	19.38	0.73
	female	4.48	30.67	
	other	1.95	3.64	
Age	0-17	2.73	9.83	3.26×10^{-4}
	18-29	3.44	20.43	
	30-44	6.05	39.34	
	45-59	3.24	13.50	
	60+	4.78	23.32	
Disab.	yes	3.83	19.99	0.67
	no	5.31	35.72	
Friend	yes	5.48	31.80	2.08×10^{-15}
	no	2.98	22.40	
All	no data	0.86	3.11	1.39×10^{-109}
	has data	4.54	28.69	
	all users	1.37	11.13	–

Table 4.4: *User group engagement comparison.*

placing data by their ranks), a procedure recommended for cases in which normality cannot be assumed (Montgomery, 2017, subsection 3.1.12). These calculations generated similar p-values to the ones presented in Table 4.4.

The most determining factor for predicting volunteer engagement seemed to be whether users had supplied their demographic data or not. This should not be surprising: users that did not bother to fill their profiles are, in average, less engaged than ones that did. As previously mentioned, the overwhelming majority of users (86.1%) did not provide demographic data. The remaining 13.9% of volunteers, despite being the minority, accounted for 48.0% of all reviews.

Gender differences did not affect user engagement, and neither did having a disability or not. On the other hand, age difference was a demographic trait that influenced engagement, and having a friend with disabilities or not was the characteristic that had the most significant difference in average reviews produced by users.

Besides engagement, another possible analysis is about behavioral differences groups might have. Data shown in Table 4.5 comes from the investigation about whether demographic attributes influence the probability of users evaluating places as accessible and of including comments in their reviews. The ratio of votes for each accessibility level and of the addition of comments in their reviews was determined based on the data points produced by users of each group. P-values were generated by Pearson's chi-squared test (χ^2) (Pearson, 1900), chosen because analyzed data was categorical (the three accessibility levels, and a boolean indicating whether a review has a free-text comment or not). Because only a small number of individuals declared their gender as "other", this

		Votes				Comments		
	Group	Not accessible (%)	Partially accessible (%)	Accessible (%)	p-value (χ^2)	With (%)	Without (%)	p-value (χ^2)
Gender	male	30.7	22.0	47.2	0.90	36.3	63.7	0.39
	female	30.9	24.5	44.6		29.6	70.4	
	other	30.8	27.8	41.5	–	51.9	48.1	–
Age	0-17	24.4	26.9	48.7	0.68	34.4	65.6	2.58×10^{-4}
	18-29	38.3	20.5	41.2		22.3	77.7	
	30-44	27.9	26.1	45.9		26.3	73.7	
	45-59	32.7	21.1	46.2		46.1	53.9	
	60+	31.9	22.0	46.2		46.3	53.7	
Disab.	yes	31.0	25.2	43.8	0.91	39.2	60.8	0.03
	no	31.3	22.6	46.1		24.0	76.0	
Friend	yes	29.1	23.7	47.2	0.38	31.6	68.4	0.64
	no	37.4	24.1	38.4		27.6	72.4	
All	no data	29.5	22.0	48.5	0.89	43.7	56.3	0.08
	has data	31.2	23.8	45.1		30.7	69.3	
	all users	30.3	22.9	46.8	–	37.4	62.6	–

Table 4.5: *User group review behavior comparison.*

group was not considered.

The proportion of votes for each category stayed generally stable in all groups. Observed p-values indicate that volunteers tend to behave approximately the same way, regardless of their demographic group. For all groups except one, the proportion of votes for “not accessible” was greater than “partially accessible”. This may seem contradictory compared to Figure 4.6a, in which the number of places evaluated as not accessible is less than the number of partially accessible. But it can be explained by the fact that Figure 4.6a shows places’ overall grades, whereas Table 4.5 was built considering all data points. It is possible that some aspects of a place are rated as not accessible, but the place as a whole is considered partially accessible.

Both gender and having friends with disabilities did not seem to affect the probability of leaving comments. Age, however, had higher influence. Volunteers aged 45 or more included comments considerably more frequently than younger ones, specially ones between 18 and 29. Similarly, persons with disabilities also left comments more frequently, though number differences were not as stark.

4.6 Chapter conclusion

We discussed the history of Guiaderodas and its implementation. The process for designing the application and the tasks for capturing data from volunteers, including all decisions behind them, were specified. Developing and deploying an impact-based business model was fundamental for ensuring the necessary financial sustainability for supporting data collection over an extended period of time. Finally, strategies for increasing volunteer acquisition and engagement were distilled, and results were presented.

What was herein presented is, to the best of our knowledge, the largest crowdsourced accessibility platform in Brazil, perhaps in Latin America and the entire developing world. It is also one of the longest-running experiments reported in this field, going way beyond mere proofs of concept in limited academical settings. Furthermore, it has also served as a platform for experimentation on what owners of crowdsourcing initiatives can do for increasing their chances of success.

Finally, we peered into the question of whether demographic traits influence user behavior. This is an important matter, as it is an indicator of the quality of produced data. Results show that these attributes generate hardly any significant differences that may threaten the reliability of volunteers. In particular, persons with disabilities had very similar reviewing behavior to their counterparts without disabilities, an indicator that the commonly received criticism that *"persons without disabilities are not reliable evaluators"* is not grounded in reality.

There are, however, two possible arguments against these conclusions. Firstly, that looking at groups' averages may not be the best path, as averages can be skewed by a few outliers. Also, that analyzing review behavior in this context may not be the best approach, as users were evaluating different places. Ideally, behavior differences should be probed by having volunteers evaluate the same place. These claims are, indeed, the background motivation for what is laid out in chapter 5.

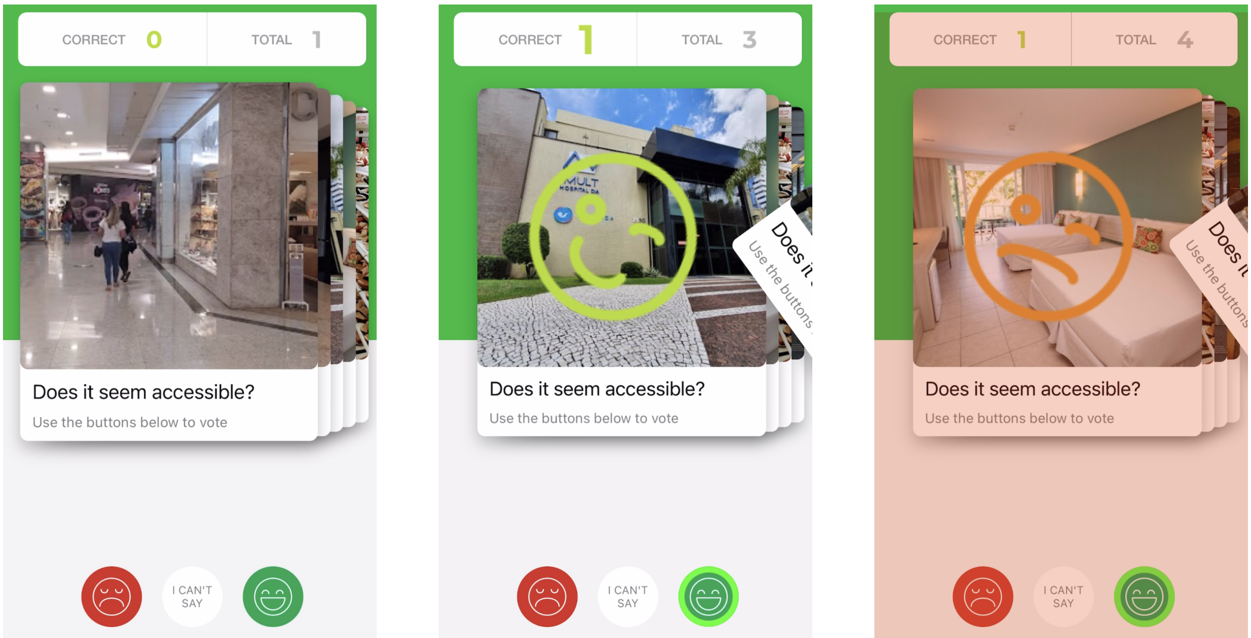
Chapter 5

A serious game for assessing data quality

Data quality is an everlasting concern in crowdsourced databases, as discussed in chapter 2. In the particular case of Guiaderodas, a common criticism from persons with disabilities was that allowing anyone to evaluate places would produce questionable results. The basis for this argument is that persons without disabilities have no experience with accessibility, and therefore cannot be relied upon for this kind of review. But does this line of thinking really hold water?

Indeed, even persons without disabilities themselves may often feel insecure about their own skills for the task. This was another commonly received feedback, and, as mentioned in subsection 4.2.2, the crowdsourcing platform was adapted to accommodate this concern. The addition of some guidance to the assessment interface was a first push for educating the audience, and the search for other alternatives resulted in the concept of a game. A quick-and-dirty pilot was put together using social media platforms. The idea was simple: letting players vote whether places pictured in series of photos appeared to be accessible or not. Only two options were available: "yes" or "no", with no middle ground. After voting, users could see the percentage of votes for each option, and thus had near real-time feedback for their answers. User engagement throughout the first months of testing indicated that the game was appealing to the audience, so it was adapted and implemented within the main crowdsourcing tool, as shown in Figure 5.1.

More importantly, the game enabled a new approach for studying the influence of demographic traits in user behavior. This matter was explored in section 4.5, but in that case, users were evaluating a large number of unique locations, which meant that most of the time they were looking at different things. In the game, players are presented with subsets of the same pool of



(a) **(b)** **(c)**

Figure 5.1: *The game interface. Users were presented with series of photos of real places, alongside the question "Does it seem accessible?" and three possible answers: "no", "I can't say", and "yes" (a). Immediately after voting, visual and sound feedback signaled whether the user answered correctly (b) or not (c), while the photo at the top of the stack was removed and user scores were updated accordingly. Answers were counted as correct when they agreed with the benchmark previously defined by the experts committee.*

photos. This serves as a new platform for testing if various user groups behave disparately when judging the same thing. In this way, it is possible to verify whether distinct groups have different biases in comparison with each other or not.

The practice of having diverse crowds work on the same data has already been applied in crowdsourcing, *e.g.,* for cross-validation, in which a contributing crowd generates data that is later validated by another crowd (Luo and Zeynalvand, 2017; Luo *et al.*, 2019). This indicates that the method herein described can be adapted for a similar purpose – it is possible to set up a system to assess accessibility remotely, based on photos of places, in a fashion that benefits from cross-validation and/or the wisdom of the crowds. Here, however, the aim is to uncover hidden biases in population groups, in particular between persons with and without disabilities, but also investigating other demographic characteristics, such as gender or age.

For better emulating the mechanics of the game implemented in social media platforms, it was desirable to provide users with real-time or near real-time feedback about their answers. For this end, an expert committee was formed for pre-evaluating photos and setting expected answers for calculating user scores. Benchmarks generated by specialists have also been previously employed in crowdsourcing, but for different reasons, such as gauging the quality level of individual contributors (Sorokin and Forsyth, 2008), and in situations in which clear "correct" and "wrong" answers

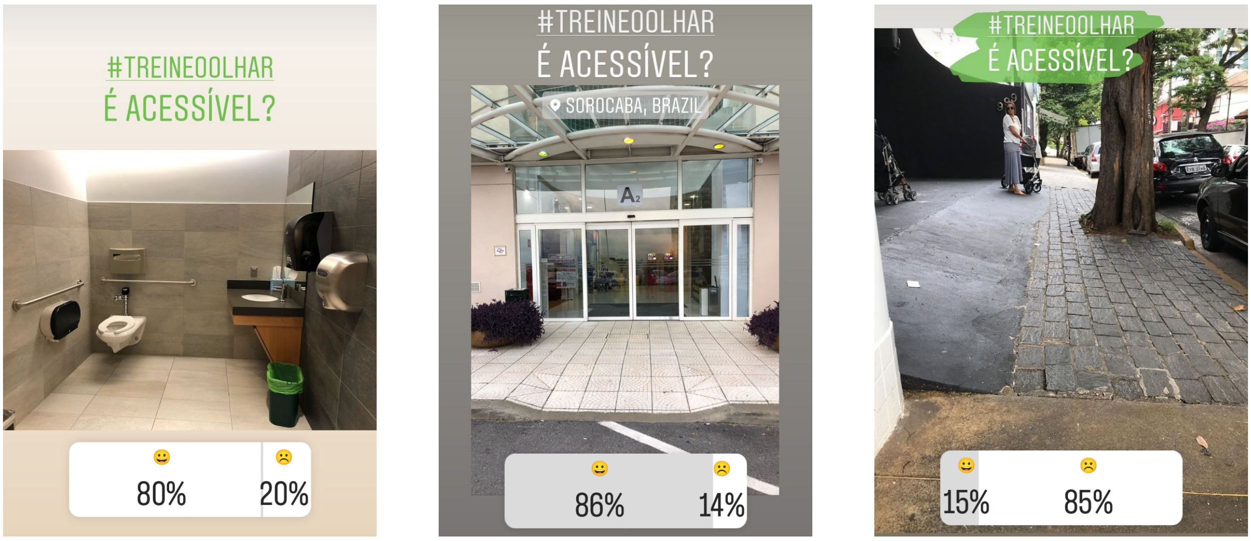


Figure 5.2: *The game pilot implemented in social media.*

existed (Wang *et al.*, 2020, 2019). In this case, the context is arguably interpretative, and the intention was to provide a common basis for comparison, not decide whether individual contributions should be corrected or outright discarded.

The remainder of this chapter details our methodology for designing and implementing the game, as well as observed results.

5.1 Game implementation

5.1.1 The game pilot

To test whether the envisioned game would appeal to the target audience, a first simplified version was implemented using native mechanics of the Instagram platform, called *polls*. Polls can be created within *user stories*, a feature that allows users to post volatile content that disappears after 24 hours. This pilot involved posting series of photos of real places, featuring both good and bad examples of several aspects relevant for accessibility (washrooms, entrances, parking spaces, etc.). Photos were presented with the question *"Is it accessible?"*, and for each of them, users could vote in a poll with only two alternatives, "yes" and "no", represented by emojis. Figure 5.2 shows some examples.

All used photos were from real places. Most of them were taken *in loco* by the author, and others were shared by the audience. There were no "correct" answers, but after voting, users were able to see the percentage of votes received by each alternative, so they could know if their vote agreed with the majority's opinion or not. This default mechanics of the chosen platform provided near real-time feedback to users.

	Viewers	Votes	Engagement ratio
Minimum	337	48	9.41%
Maximum	1436	334	45.94%
Mean	594.69	173.37	29.64%
Median	557.5	162.5	29.59%
σ	191.69	54.00	5.98%

Table 5.1

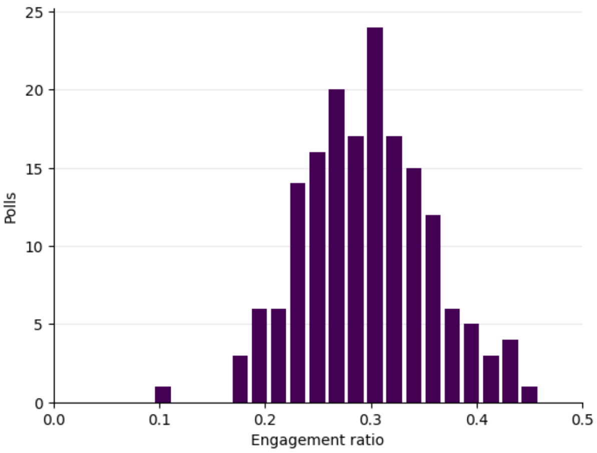


Figure 5.3

Table 5.1 and Figure 5.3: *Descriptive analysis and engagement ratio of the game pilot (n = 170).*

In total, 170 polls were created as part of the experiment, which lasted for about 7 months. Table 5.1 and Figure 5.3 summarize the engagement ratio observed, calculated as the amount of votes the poll received divided by the number of people who viewed it. In addition to quantitative measurements of engagement, qualitative feedback was also collected from users. Besides mostly positive impressions regarding the game mechanics, users also expressed the desire for some form of "official" correct answer for each poll. Users also reportedly wanted to discuss more about details of some photos, highlighting concerns and sharing personal opinions. This happened specially in polls that had tighter voting (*i.e.,* with results close to 50%), an indication that deciding whether these places were accessible or not was a debatable topic. While these points were not addressed in the pilot, they can also be listed as relevant outputs from the experiment.

The main learning points from the pilot included:

1. The overall game mechanics engaged the audience;
2. The absence of middle-ground vote alternatives forced users to make decisions even when they had doubts, but controversial cases could create discussion;
3. While (near) real-time feedback was appreciated, users still wanted to decisively know whether they answered correctly or not, as if participating in a quiz.

5.1.2 The real game

The first challenge in implementing a fully functional version of the game is to amass a large, ideally infinite, number of relevant photos of real places. They should also be previously unseen by users, for minimizing the effect of hidden cognitive biases (Eickhoff, 2018) of the player base.

Because the game and social media audiences intersected each other, the photos from the game pilot were not an option. Moreover, manually curating and/or taking pictures is a laborious task: during the 7 months of the pilot, a mere 170 publishable photos were collected, an indication of potential scalability issues this alternative might present. The photo set should ideally overlap with the geographical and cultural environments that players are accustomed to. Because the largest portion of the audience was spread throughout Brazil, selecting photos from all parts of the world was not optimal. The best alternative was to depict places from that particular country.

Because giving feedback about users' answers was an essential part of game dynamics, simply building a large photo set was not enough. The second implementation challenge was to ensure that all photos were labeled as either "accessible" or "not accessible". This allowed immediate visual and sound feedback to users upon checking their answers. Furthermore, scores could also be calculated, adding a valuable supplementary gamification feature. While a clear solution would be to manually label each photo (a job that could be accelerated via crowdsourcing), the inherent difficulty of deciding what an accessible place looks like, discussed in chapter 3, turned this hard issue into a disheartening problem.

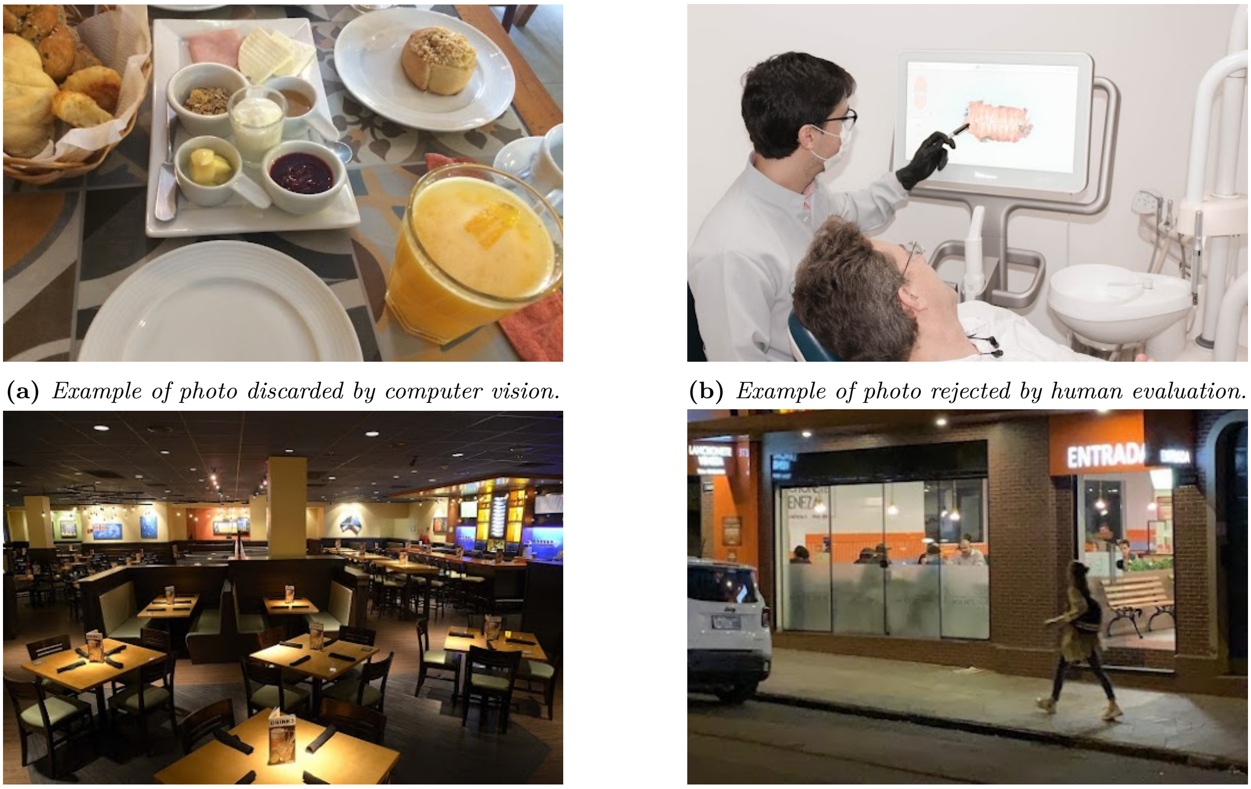
It was necessary to set up a process for importing, filtering, and labeling large amounts of photos. The devised solution started with importing photos from Google Places API,[1] a database that is itself also crowdsourced from individual contributors of the Google Maps platform. A total of 11,307 photos were imported, linked to 1,222 places in Brazil that had been reviewed in Guiaderodas.

Using Google Places as a source had its own complications: due to its crowdsourced origin, photos had high variability in both quality and the things they depicted. Many of them were not related to accessibility (*e.g.,* plates of food), lacked quality (*e.g.,* badly framed, out of focus), or were otherwise unfit for usage (*e.g.,* featuring children faces). Hence, the first filtering step was to apply computer vision algorithms to automatically discard photos that were obviously out of scope. The tagging feature of Microsoft Azure AI Vision API[2] was used for detecting food, drinks, text, and human faces. At this stage, 4,606 (40.7%) photos such as Figure 5.4a were discarded.

The remaining phases were executed by a small group of volunteers, recruited from top users of the Guiaderodas platform. All of them had either reviewed over 100 places or were active contributors during the game pilot stage. The group was comprised of persons with different kinds of disabilities, and persons in close proximity to persons with disabilities (*e.g.,* immediate family). This specific body of experts was selected due to their empirical knowledge about accessibility.

[1] https://developers.google.com/maps/documentation/places/web-service (retrieved Aug. 2023).

[2] https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/ (retrieved Aug. 2023).



(a) *Example of photo discarded by computer vision.*

(b) *Example of photo rejected by human evaluation.*

(c) *Example of photo rejected for lack of consensus.*

(d) *Example of photo included in the game.*

Figure 5.4: *Examples of rejected and approved images in the vetting process.*

Their profiles are summarized in Table 5.2. Additionally, two individuals that had no disabilities nor proximity to persons with disabilities (participants #8 and #9) worked on discarding photos that were not related to accessibility but still passed through computer vision filters, such as Figure 5.4b, eliminating further 3,167 (28.0%) images. Other 2,327 (20.6%) images did not get to be manually reviewed within the time frame defined for preparation of the game.

The remaining photos were evaluated by different members of the image vetting committee. They used the same gameplay interface of Figure 5.1 for labeling each photo as "accessible" or

#	**Gender**	**Age range**	**Has disability?**	**Family member with disability?**
1	female	60+	Yes	–
2	female	45-59	No	Yes
3	female	30-44	Yes	–
4	male	30-44	Yes	–
5	female	30-44	No	Yes
6	female	18-29	Yes	–
7	female	18-29	No	Yes
8*	female	45-59	No	No
9*	male	30-44	No	No

Table 5.2: *Members of the image vetting committee.*

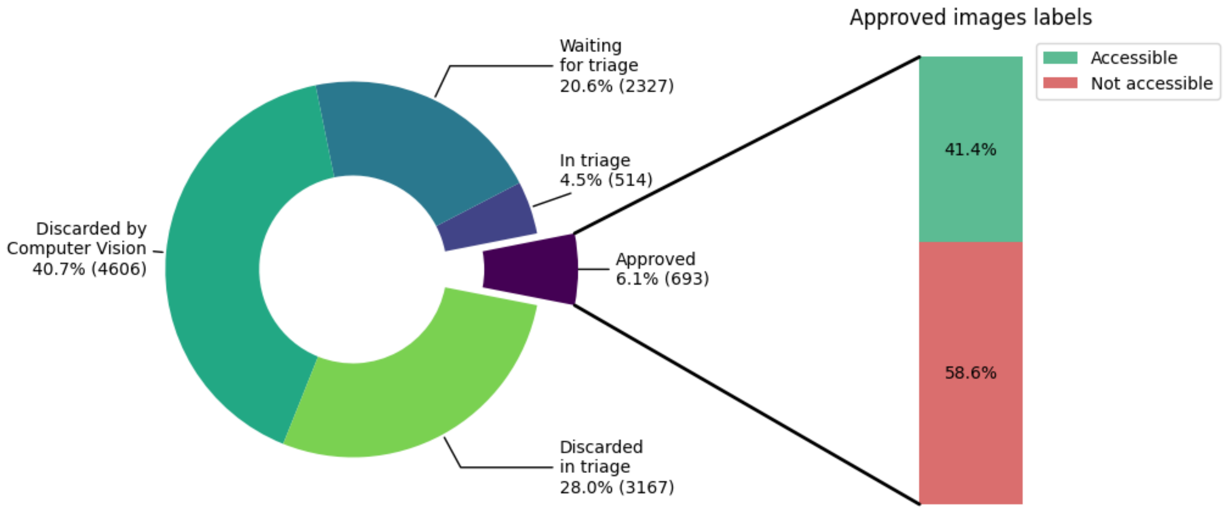


Figure 5.5: *Image vetting process results.*

"not accessible". Experts saw each photo at most once, and were preferably presented with ones that had already been evaluated by others. This practice of "funneling" votes was a measure to maximize the number of images approved, as a photo was approved when:

1. it received three votes, and all of them were equal, **OR**

2. it received five votes, four of were them equal and only one differed.

Photos that received two votes for each label were deemed undecidable and discarded, such as Figure 5.4c. In the end, 514 (4.5%) images did not receive the minimum amount of votes and were not included in the game. Of the 693 (6.1%) approved images, 287 (41.4%) of them were labeled as "accessible", and 406 (58.6%) as "not accessible". Figure 5.4d is an example of an approved photo labeled as "not accessible".

Figure 5.5 summarizes the outcomes of this process, which was performed over the course of approximately four weeks. The resulting dataset was used in a different manner when compared to previously mentioned works, in which a gold standard was used for deciding whether individual contributions should be compensated or discarded. Here, expected answers were used for gamification purposes, such as providing feedback to players and calculating scores, and for analyzing behavioral tendencies, discussed in the remainder of this chapter.

Being a crowdsourcing system itself, this method's scalability is proportional to the size of the working crowd. Photos were labeled by a rather small group of experts, but a larger crowd would provide higher work throughput, and therefore generate a bigger dataset in the same time period. An adapted version of this method could potentially be applied for labeling a larger amount of photos and serve as a platform for remote assessment of accessibility.

5.2 Results

The goal is to generate some kind of measurement for the quality of work produced by different groups of volunteers. Since the surrounding context is highly subjective and experts themselves may be biased, the dataset of expected answers is used for measuring whether different demographic groups are *relatively* biased in comparison with each other.

Results herein presented are based on data collected between August 2021 and August 2023.

5.2.1 Crowd profile

A total of 1059 individuals played the game. This crowd is a subset of the one described in the previous chapter, and had a somewhat similar profile, shown in Figure 5.6. Again, filling demographic information was not mandatory. 520 (49.1%) users informed their date of birth and whether they had disabilities and friends with disabilities. Of these, 486 (45.9% of the total) also informed their gender.

Gender, disability, and friend proportions were essentially the same as in the previous case, with only slight differences. In age, a smaller proportion of individuals aged between 45 and 59 played the game, and even less for ones aged over 60 or below 18. Users aged between 30 and 44 were the largest group, followed by the group between 18 and 29.

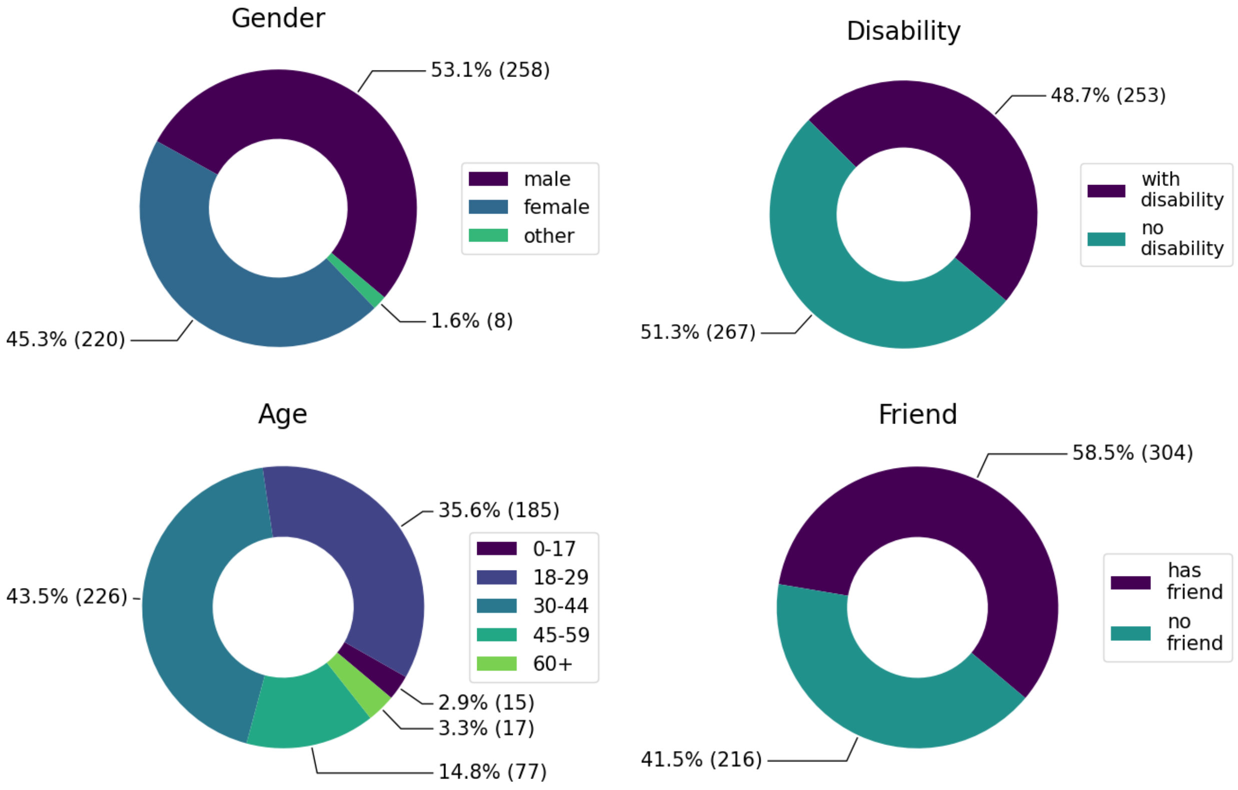


Figure 5.6: *Game crowd demographics.*

5.2.2 Crowd behavior

Player votes

A player's total number of votes is calculated as the sum of votes for *"not accessible"* (*"red"*), *"accessible"* (*"green"*), and skips:

$$V_{total} = V_{red} + V_{green} + V_{skip}$$

Figure 5.7 shows the distribution of players' total votes, and Table 5.3 contains the corresponding descriptive analysis. 272 (52.31%) players voted on 50 images or less, 90 (17.31%) of them cast between 51 and 100 votes, 92 (17.69%) between 101 and 200 votes, and 66 (12.69%) voted on 201 or more images. 14 (2.69%) individuals voted for all images available in the game at their time of playing. 127 (24.42%) cast exactly 17 votes due to an implementation detail: to minimize network overload, votes were transmitted together with requests for more images. Because this event was triggered when three or less images were locally available and images were fetched 20 at a time, 17 was the minimum possible value of votes that were transmitted. This also means that users that voted on less than 17 images were not taken into account.

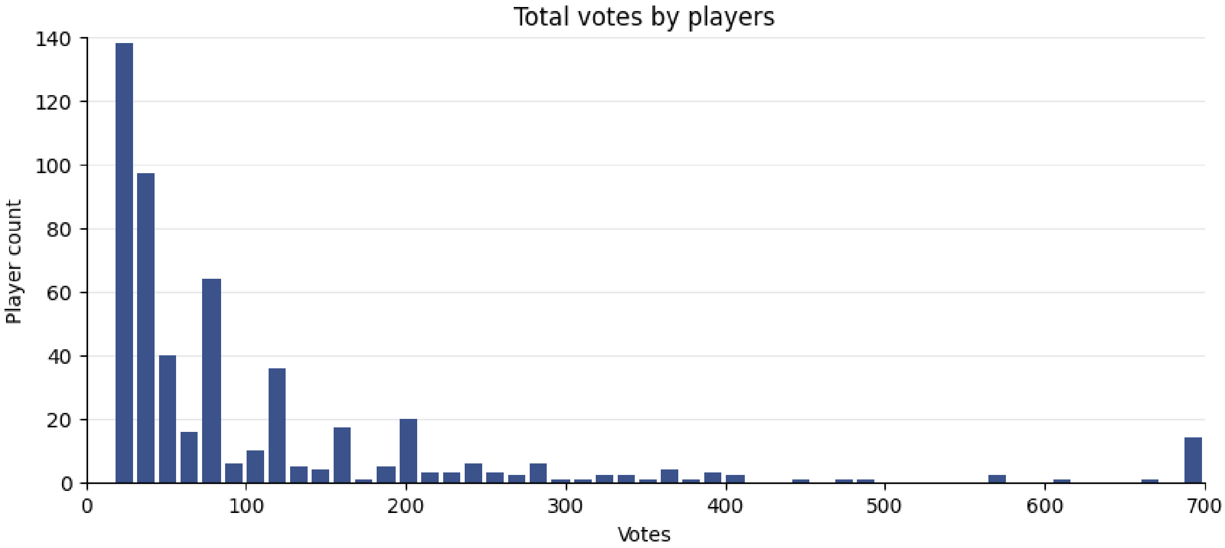


Figure 5.7: *Distribution of total votes cast by players.*

min	max	mean	med	σ
17	693	105.148	45.000	137.272

Table 5.3: *Descriptive analysis of total votes cast by players.*

Votes received

Figure 5.8 shows the distribution of votes received per image, and Table 5.4 the corresponding descriptive analysis. Images shown to each player were selected randomly from the pool of labeled photos. The chi-squared test (Pearson, 1900) performed over the frequency of votes received by each image yields p-value = 0.999, indicating that the probability of selection was uniformly distributed. This means that every image was equally likely to be selected at any given time, for any given user. This was the choice for implementation because privileging a set of images could potentially introduce unknown biases to the analysis.

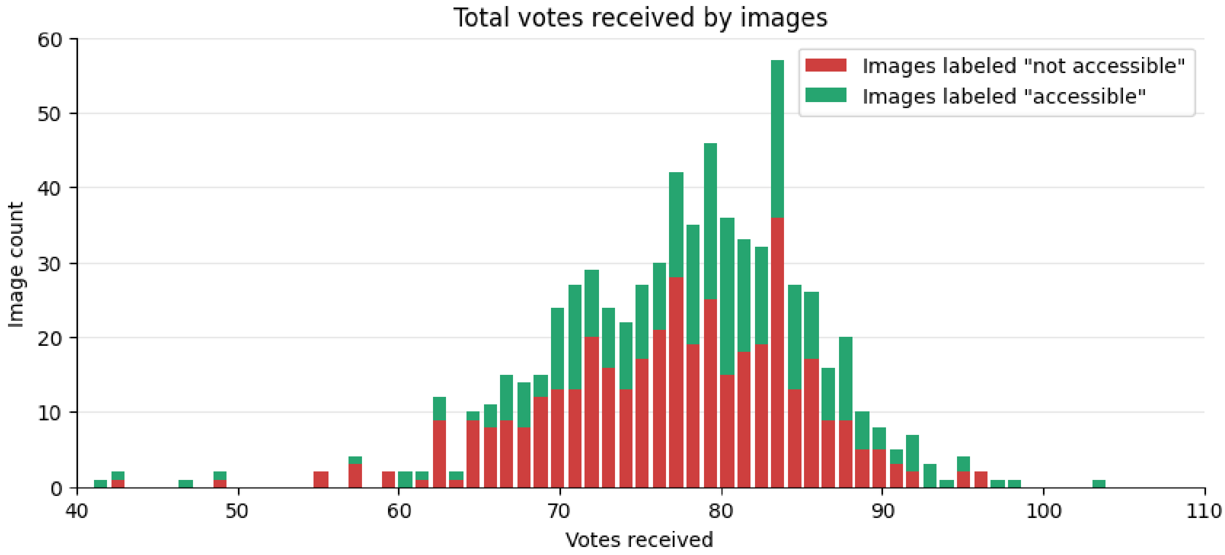


Figure 5.8: *Distribution of votes received by images.*

Image label	min	max	mean	med	σ
"not accessible"	43	96	76.983	77.000	7.749
"accessible"	41	104	78.422	79.000	8.407

Table 5.4: *Descriptive analysis of votes received by images.*

Ratio of correct answers

Because all images had been previously labeled, player votes are considered as *correct answers* when they agree with the image's label. By defining $N(V_x \mid L_y)$ as the number of votes for "x" on images labeled as "y", with $x \in \{$*"red"*, *"green"*, *"skip"*$\}$ and $y \in \{$*"red"*, *"green"*$\}$, we calculate the ratio of correct answers for images labeled as *"not accessible"* (*"red"*) and *"accessible"* (*"green"*):

$$C_{red} = \frac{N(V_{red} \mid L_{red})}{N(V_{red} \mid L_{red}) + N(V_{green} \mid L_{red})}$$

$$C_{green} = \frac{N(V_{green} \mid L_{green})}{N(V_{red} \mid L_{green}) + N(V_{green} \mid L_{green})}$$

Figure 5.9 shows the distribution of players' ratio of correct answers for images labeled as *"not accessible"* and *"accessible"*. Table 5.5 contains the corresponding descriptive analysis. Both datasets have negative skew, *i.e.,* longer left tails and data concentration to the right of the figure. The greater skewness observed for *"accessible"* indicates that players tend to get higher ratios for images labeled as such. In comparison to *"not accessible"*, both the mean and median were higher for *"accessible"*, and standard deviation (σ) was lower. 367 (70.58%) players had $C_{green} > C_{red}$, and 77 (14.82%) achieved a perfect score for *"accessible"* (*i.e.,* $C_{green} = 1.0$), while only 8 (1.54%) had a perfect score for *"not accessible"* ($C_{red} = 1.0$). None had a perfect score for both.

This data does not follow a normal distribution. This can be verified by using the Shapiro-Wilk test for normality (Shapiro and Wilk, 1965), which tests the null hypothesis that data was sampled from a normal distribution. Resulting p-values of 4.38×10^{-17} and 1.15×10^{-19} for the distribution of correct answers for images labeled as *"not accessible"* and *"accessible"*, respectively, are a strong indication of the non-normality of data.

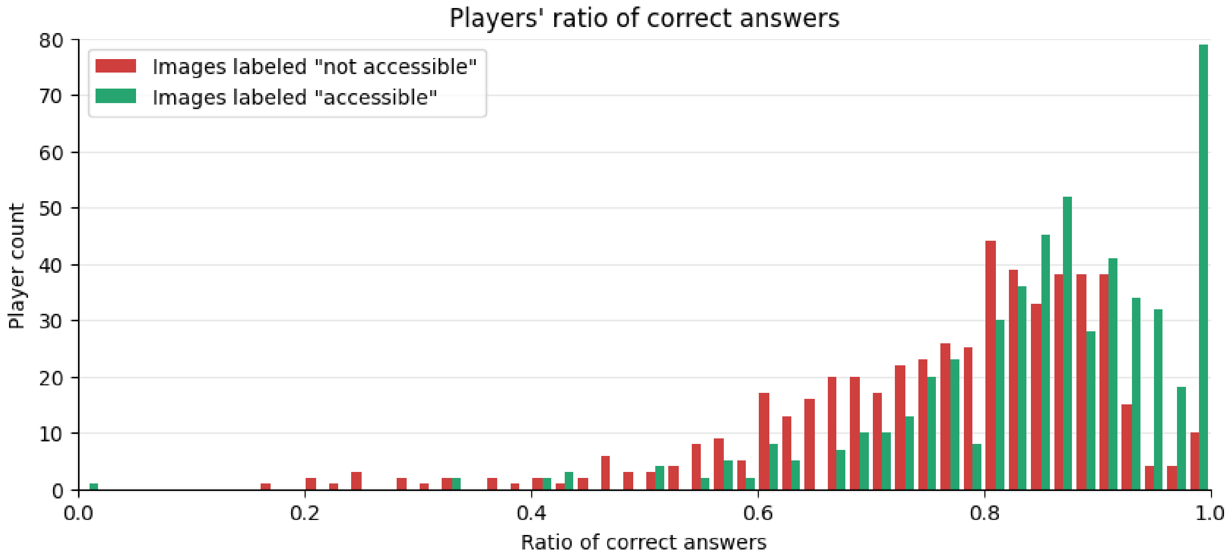


Figure 5.9: *Distribution of players' ratio of correct answers.*

Image label	min	max	mean	med	σ
"not accessible"	0.169	1.000	0.763	0.800	0.145
"accessible"	0.000	1.000	0.852	0.871	0.125

Table 5.5: *Descriptive analysis of players' ratio of correct answers.*

Ratio of skipped votes

The ratio of skipped votes for images labeled as *“not accessible”* and *“accessible”* is the number of times a player skipped images divided by the total number of images seen for each label:

$$S_{red} = \frac{N(V_{skip} \mid L_{red})}{N(V_{red} \mid L_{red}) + N(V_{green} \mid L_{red}) + N(V_{skip} \mid L_{red})}$$

$$S_{green} = \frac{N(V_{skip} \mid L_{green})}{N(V_{red} \mid L_{green}) + N(V_{green} \mid L_{green}) + N(V_{skip} \mid L_{green})}$$

Figure 5.10 shows the distribution of players’ ratio of skipped votes for both image labels, and Table 5.6 the corresponding descriptive analysis. Again, data is not normally distributed. Shapiro-Wilk yields p-values of 1.084×10^{-32} for *“not accessible”* and 7.662×10^{-36} for *“accessible”*. Datasets are positively skewed, with long right tails.

Players did not skip votes very often. The median is zero for both labels, as 263 (50.57%) users did not skip any images at all. However, when players skipped images, it was more frequently for *“not accessible”* images than for *“accessible”* ones, resulting in the observed higher mean and standard deviation (σ) of the former, in comparison with the latter. A large majority of 387 (74.42%) players did not skip any *“accessible”* images, and 302 (58.08%) did not skip any *“not accessible”* images.

These numbers indicate an interesting behavior pattern: players tended to be more sure of their answers when looking at *“accessible”* images, while *“not accessible”* ones incited more doubts. Moreover, this also explains why *“accessible”* images had a slightly higher mean of votes received, as seen in Table 5.4.

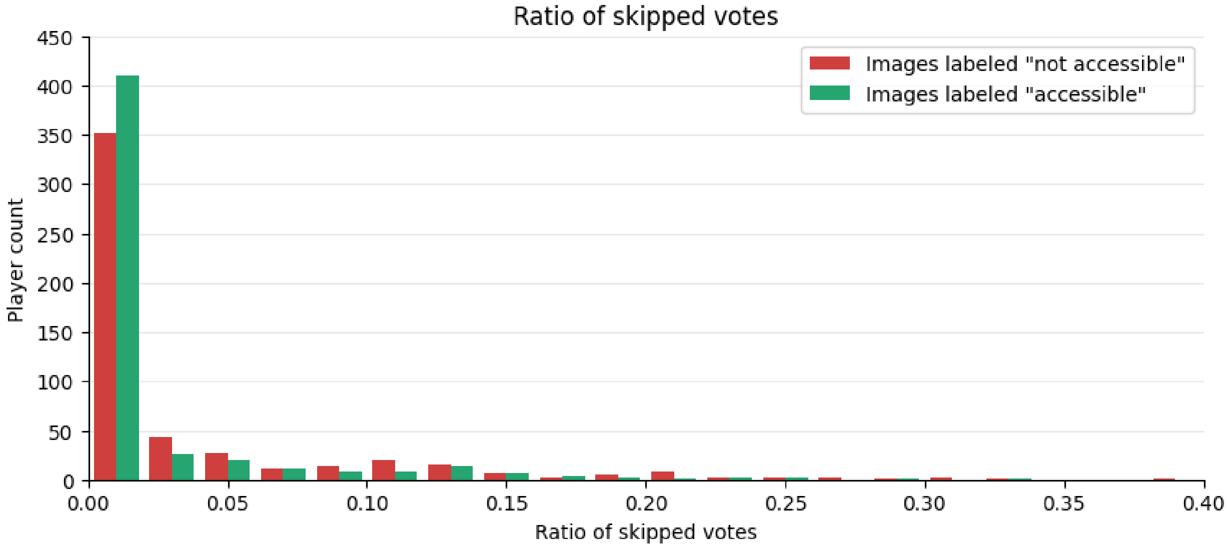


Figure 5.10: *Distribution of players’ ratio of skipped votes.*

Image label	min	max	mean	med	σ
"not accessible"	0.000	0.400	0.033	0.000	0.061
"accessible"	0.000	0.333	0.020	0.000	0.048

Table 5.6: *Descriptive analysis of players' ratio of skipped votes.*

Influence of image labels

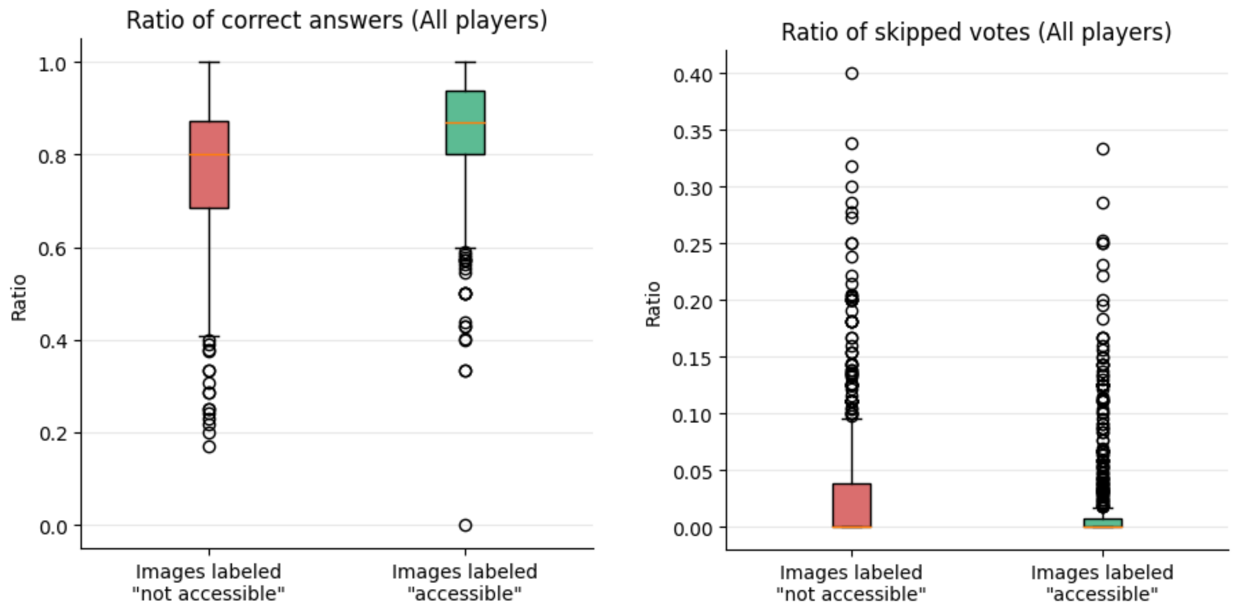
The box-and-whisker plots (Tukey *et al.*, 1977) seen in Figure 5.11 are another way to visualize the same data from Figures 5.9 and 5.10 and Tables 5.5 and 5.6. Medians are represented by red lines. The lower and upper limits of boxes are respectively the first and third quartile (Q1 and Q3), *i.e.,* the medians of the lower and upper halves of the dataset. Whiskers extend from the box to the data points closest to 1.5 times the interquartile range (IQR), where $IQR = Q_3 - Q_1$. Outliers beyond whiskers are plotted as individual points.

Inspecting Figure 5.11a, it is noticeable that the boxes have a limited overlap, an indication that there might be significant difference between the distributions. In fact, the Kruskal-Wallis test (Kruskal and Wallis, 1952) results in a p-value $= 2.498 \times 10^{-28}$, showing that, indeed, players behave significantly differently for *"not accessible"* and *"accessible"* images. Kruskal-Wallis was chosen for being appropriate for datasets that are not normally distributed (Montgomery, 2017).

Applying the test to the ratio of skipped votes yields p-value $= 2.235 \times 10^{-7}$, proving that players really tend to skip more often for *"not accessible"* images. With pronounced skewness and relatively large excess kurtosis (Zwillinger and Kokoska, 1999) (6.86 for *"not accessible"*, and 10.40 for *"accessible"*, where a normal distribution would yield 0.00), the observed distributions produce large quantities of outliers, a fact that can be observed by the large number of individual points plotted in Figure 5.11b.

Considering that the notion of *"correct"* answers actually denotes an agreement with an image's label, we can conclude from the observed behavior of the crowd that players tend to be more confident on their answers when looking at *"accessible"* images, or that images labeled as *"not accessible"* are more prone to generate doubt.

While there is no clear, set in stone definition for what an accessible place is, players appeared to somehow reach a consensus more easily for positively evaluated images. On the other hand, they seemed less sure on where to draw the line for *"not accessible"*, meaning that different individuals gave less or more leeway when voting.



(a) **(b)**

Figure 5.11: *Ratio of correct answers and skipped votes by image label.*

Influence of demographic traits

After establishing some existing behavior patterns of the crowd, the next question that naturally arises is whether demographic traits also affect behavior. To investigate this, we compared the ratios of correct answers and skipped votes for each demographic group.

The four traits shown in Figure 5.6 (gender, age, disability, and having friends with disabilities) were examined separately. Some groups had too small sample sizes and were disregarded, namely: gender *"other"* (8 players), age *"0-17"* (15 players), and age *"60+"* (17 players). Individuals in these groups were ignored when analyzing these specific traits, but where included for others. This means that the dataset was comprised of 478 players for gender, 488 for age, and 520 for disabilities and friends with disabilities.

For each demographic trait and image label, groups were compared using the Kruskal-Wallis test. Figure 5.12 displays data for the ratio of correct answers, and Table 5.7 shows the corresponding descriptive analysis and the resulting p-values. For all groups, ratios for the two labels differ from each other, akin to what was already observed for the whole population (Figure 5.11). An overlap of green boxes can be observed in all four cases, and large p-values show that the we cannot reject the null hypothesis that demographic traits do not affect correct answer ratios for *"accessible"* images. Similarly, red boxes also overlap for gender, age, and friend, and corresponding p-values were also large. However, the overlap was smaller for disability, and a p-value $\approx 10^{-7}$ strongly rejects the null hypothesis.

In summary, demographic traits did not affect correct answers ratio for *"accessible"* images, but the difference between players with and without disabilities was statistically significant for *"not accessible"* images.

Figure 5.13 and Table 5.8 show data for the ratio of skipped votes per demographic group. All groups also present a behavior similar to the whole population, with higher ratios of skips for *"not accessible"* images. Comparisons between groups of each trait yielded high p-values for all but two cases: players who did not have friends with disabilities skipped *"not accessible"* images more than ones who had (p-value = 0.009), and women skipped *"accessible"* images more frequently than men (p-value = 0.017). But even in these cases, however, half of the users did not skip any votes, a fact denoted by the median being zero.

Again, results were also verified by performing the same comparisons using one-way ANOVA over the rank transformation dataset (Montgomery, 2017, subsection 3.1.12), which resulted in approximately the same p-values.

After analyzing demographic traits separately, there was still the question about whether their combinations influenced behavior. All possible group permutations were generated, *e.g., "men aged between 30 and 44 with disabilities and no friend"*, *"women of any age without disabilities with friends"*, etc., resulting in 84 of such sub-groups that had at least 20 players. Because samples had different sizes, the Tukey-Kramer variant of Tukey's honestly significant difference (HSD) test (Kramer, 1956; Tukey, 1949) was used for testing all possible pairwise combinations and finding ones that had divergent means.

18 out of 3,486 combinations produced p-values ≤ 0.05 for correct answer ratios of images labeled *"not accessible"*. No pairs produced such p-values for *"accessible"* images, nor for ratios of skipped votes for any label. However, all of these 18 pairs had the same composition: they consisted of a subset of the group of players with disabilities compared to a subset of players without, *e.g., "persons with disabilities aged between 30 and 44 of any gender"* vs. *"persons without disabilities of any age who are female and have a friend"*. This is another evidence that having a disability or not is the deciding factor in this particular case.

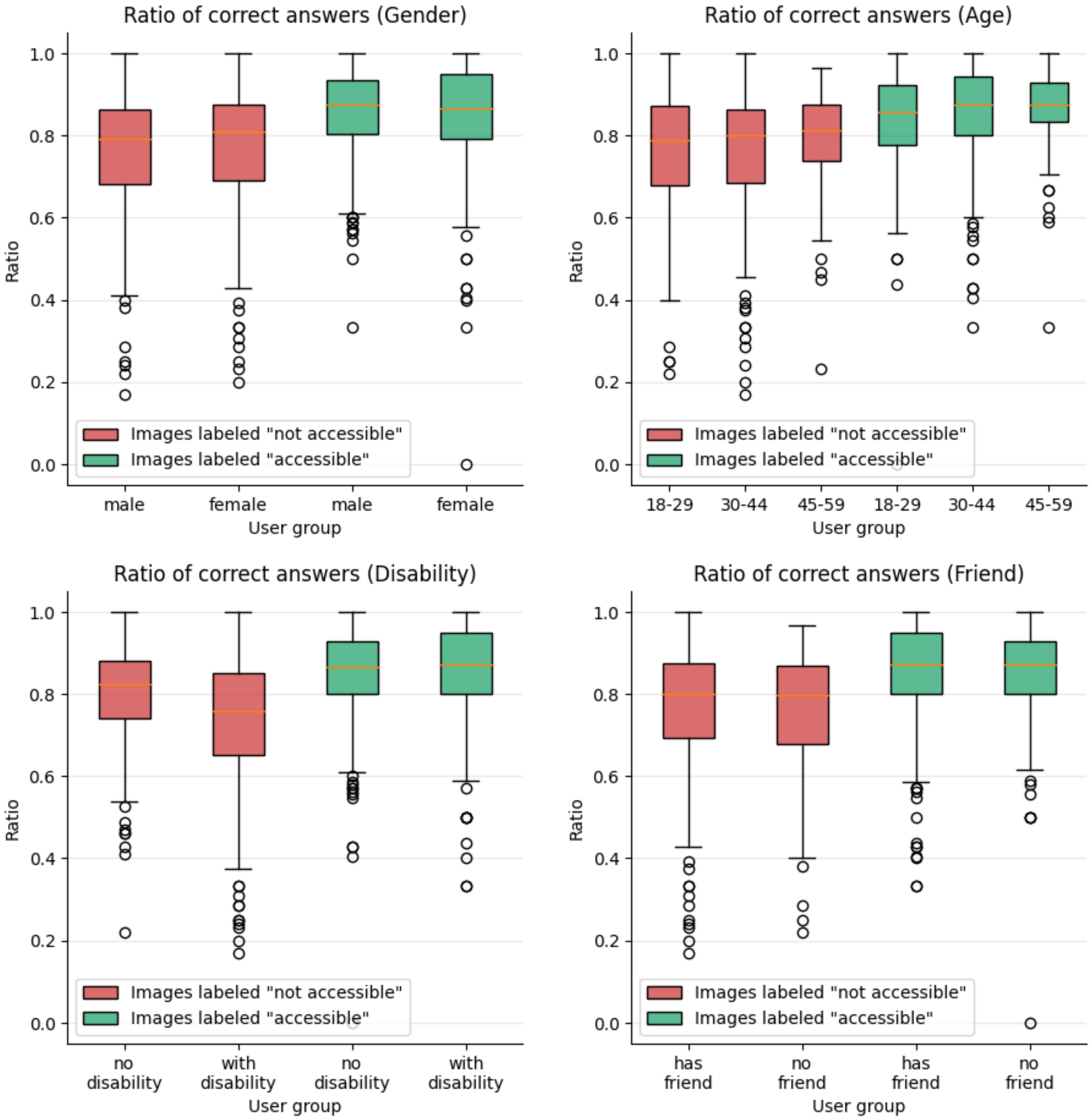


Figure 5.12: *Ratio of correct answers by user group.*

Image label	Group		min	max	mean	med	σ	p-value
“not accessible”	Gend.	male	0.169	1.000	0.760	0.792	0.142	0.295
		female	0.200	1.000	0.768	0.810	0.151	
	Age	18-29	0.219	1.000	0.762	0.789	0.141	0.553
		30-44	0.169	1.000	0.760	0.800	0.153	
		45-59	0.231	0.966	0.779	0.813	0.133	
	Disab.	yes	0.169	1.000	0.729	0.759	0.163	4.781×10^{-7}
		no	0.219	1.000	0.796	0.823	0.117	
	Friend	yes	0.169	1.000	0.764	0.800	0.151	0.675
		no	0.219	0.966	0.762	0.798	0.136	
“accessible”	Gend.	male	0.333	1.000	0.858	0.875	0.109	0.849
		female	0.000	1.000	0.848	0.867	0.141	
	Age	18-29	0.000	1.000	0.839	0.857	0.130	0.095
		30-44	0.333	1.000	0.858	0.875	0.123	
		45-59	0.333	1.000	0.865	0.874	0.116	
	Disab.	yes	0.333	1.000	0.854	0.872	0.131	0.361
		no	0.000	1.000	0.851	0.867	0.120	
	Friend	yes	0.333	1.000	0.853	0.870	0.130	0.552
		no	0.000	1.000	0.851	0.871	0.119	

Table 5.7: *Descriptive analysis of players’ ratio of correct answers by demographic group.*

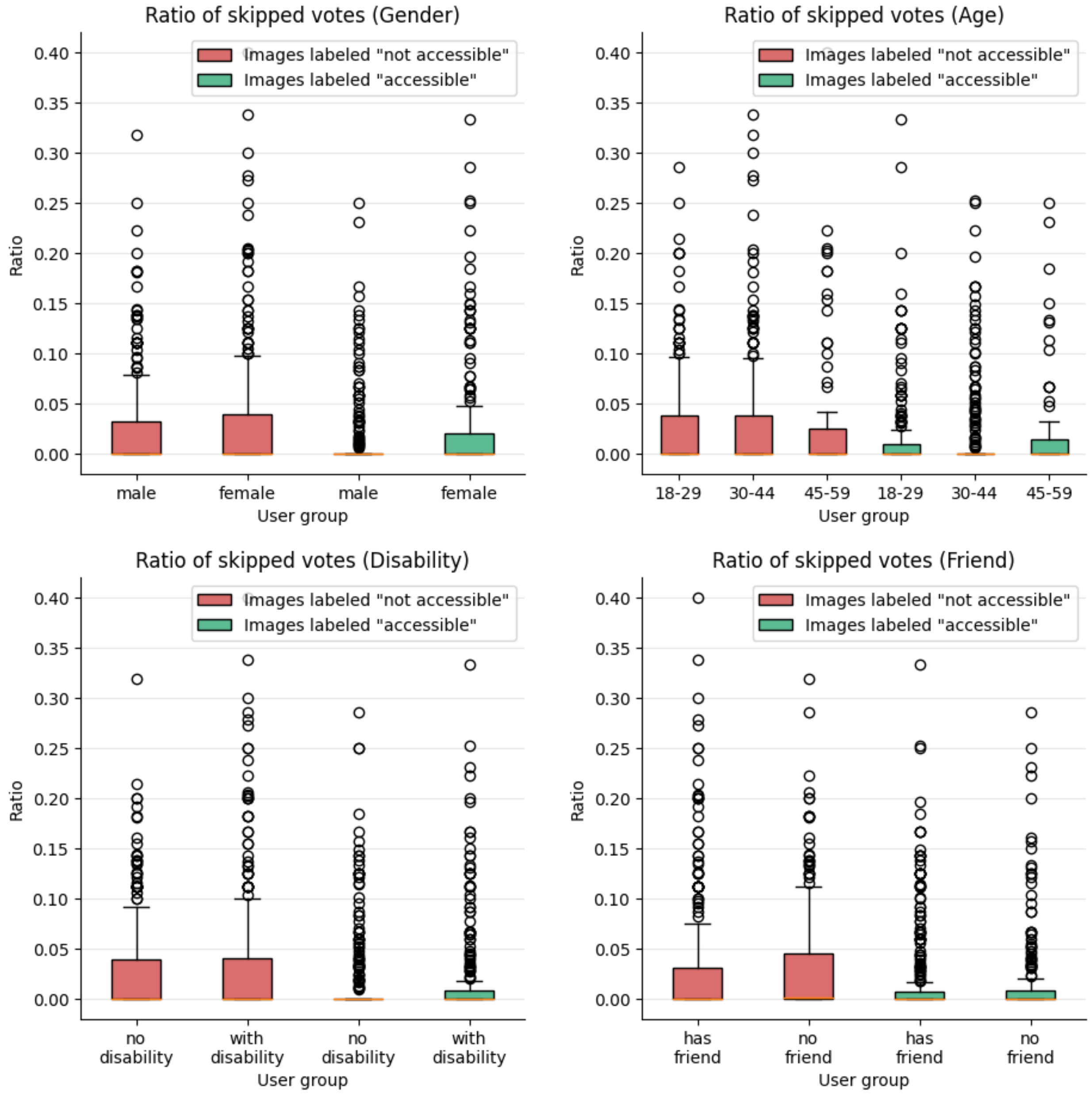


Figure 5.13: *Ratio of skipped votes by user group.*

Image label	Group		min	max	mean	med	σ	p-value
“not accessible”	Gend.	male	0.000	0.318	0.026	0.000	0.049	0.412
		female	0.000	0.400	0.037	0.000	0.070	
	Age	18-29	0.000	0.286	0.030	0.000	0.054	0.808
		30-44	0.000	0.338	0.033	0.000	0.062	
		45-59	0.000	0.400	0.038	0.000	0.074	
	Disab.	yes	0.000	0.400	0.038	0.000	0.071	0.839
		no	0.000	0.318	0.028	0.000	0.050	
	Friend	yes	0.000	0.400	0.031	0.000	0.064	0.009
		no	0.000	0.318	0.035	0.001	0.058	
“accessible”	Gend.	male	0.000	0.250	0.014	0.000	0.037	0.017
		female	0.000	0.333	0.026	0.000	0.056	
	Age	18-29	0.000	0.333	0.020	0.000	0.048	0.872
		30-44	0.000	0.253	0.020	0.000	0.047	
		45-59	0.000	0.250	0.023	0.000	0.053	
	Disab.	yes	0.000	0.333	0.022	0.000	0.050	0.626
		no	0.000	0.286	0.019	0.000	0.045	
	Friend	yes	0.000	0.333	0.021	0.000	0.048	0.930
		no	0.000	0.286	0.019	0.000	0.047	

Table 5.8: *Descriptive analysis of players’ ratio of skipped votes by demographic group.*

5.3 Chapter conclusion

In this chapter, we discussed a method for tackling the inherent problem of assessing data quality in crowdsourcing systems. Moreover, the context in which we operated was the same subjective topic of chapter 4: accessibility.

The method relied on a serious game in which players voted on images sampled from a pool of pre-labeled photos of real places. This ensured that crowdworkers were evaluating the same thing, so as to allow comparative analyses, but at the same time adding enough randomness to avoid introducing unexpected implicit biases. The preparation process for labeling images was done by recruiting a smaller crowd of experts, and serves a dual purpose of enabling gamification features such as scores, and also defining a ground zero for studying groups’ behavior. The method is applicable even in contexts where there are no clear definitions of what “correct” or “wrong” look like, such as ours, and can be executed in a timely manner.

Game data was collected for 25 months, and post-hoc analysis shows that there are no statistically significant differences in behavior between different demographic groups when evaluating good accessibility. In this case, persons without disabilities are, therefore, as good judges as (or at least as *bad* as) persons with disabilities, and women are as good as men, younger persons are as good as older ones, and so on.

When evaluating bad accessibility, tests indicate that having disabilities affects behavior, *i.e.,* persons with and without disabilities behave differently. Data also shows that these negative cases also generate more doubt, and reviewers in general tend to be less sure about their votes. Here, the subjectivity of the underlying context seems to come into play: as empirical experts who face daily challenges due to lack of accessibility, persons with disabilities have different personal criteria for evaluation, resulting in a more heterogeneous behavior when analyzed as a group. This hypothesis is also supported by the fact that the group of persons with disabilities had greater variance in comparison to their counterparts without disabilities.

Chapter 6

Discussion

Some of the most important open problems in crowdsourcing include how to recruit and engage crowdworkers (Dahlander and Piezunka, 2020; Kaartemo, 2017; Täuscher, 2017), and how to ensure quality of the produced output (Bouguettaya *et al.*, 2017; Kittur *et al.*, 2013; Lease, 2011). We have presented the techniques used for building a large-scale crowdsourced accessibility platform, running continuously for more than 7.5 years and in which more than 27,000 volunteers generated over 300,000 data points. We reported on our experiments about crowd acquisition and engagement. We developed a new method for gauging crowd behavior biases as a proxies for data quality, and analyzed results from its application in a real-world setting.

Literature about crowdsourcing suffers, on the one hand, from immense success bias in reporting, and on the other, from a lack of field-testing. Most published studies are purely theoretical (Lenart-Gansiniec, 2018) and do not focus on the kernels and essentials of crowdsourcing (Zhao and Zhu, 2014). Nonetheless, the challenges of forming a crowd should not be taken for granted, as that is precisely the aspect where most initiatives fail (Dahlander and Piezunka, 2020). To illustrate, the 13 crowdsourcing studies on accessibility analyzed by Mack *et al.* (2021) have a median of only 153 participants. Merely 7 out of 137 databases reviewed by Kamikubo *et al.* (2021) included more than 1000 individuals, and they typically involved passive data collection, not necessarily active participation from a crowd. In contrast, our results reached the magnitude of tens of thousands of volunteers, comparable to some of the most successful community-led accessibility maps such as Wheelmap in Germany and AXS Map in the United States. Additionally, Guiaderodas is, to the best of our knowledge, the largest crowdsourced accessibility platform in the developing world, helping to fill an important geographical gap. It is also one of the longest experiments reported in this field, running continuously since early 2016.

Several attempts to measure accessibility are, in fact, measuring other objective features and using them as proxies to infer accessibility (or lack thereof). This is, for instance, the case of systems for reporting problems in sidewalks like potholes and steps (Cardonha *et al.*, 2013; Koch *et al.*, 2012; Mirri *et al.*, 2016; Prandi *et al.*, 2016, 2017b; Rice *et al.*, 2012; Saha *et al.*, 2019; Shigeno *et al.*, 2013; Weld *et al.*, 2019). In our case, the approach was to embrace subjectivity into the evaluation by adopting a semaphore-like scale and leaving space for personal interpretation by volunteers.

Games have been applied in crowdsourcing to make monotonous tasks more engaging, increase work quality and motivate users (Choi *et al.*, 2014; Kavaliova *et al.*, 2016). This has been done through *gamification*, which combines crowdwork with game design elements such as points and badges (Morschheuser *et al.*, 2016; Prandi *et al.*, 2016, 2017b), and by creating games that generate useful data without users even realizing they are doing so (Breazeal *et al.*, 2013; Law and Von Ahn, 2009; Law *et al.*, 2007). In our case, we applied games in two separate ways, for two distinct goals. The first was to increase volunteer participation: simple games were implemented in social media (*e.g.,* Instagram), and together with design changes in the crowdsourcing platform itself, monthly means of user signups and active reviewers increased by around 10x and 6x, respectively. We also found that the most engaged users generate disproportionately high value. The top 10^{th} percentile of most engaged volunteers accounted for more than three-quarters of total output. We also investigated whether demographic traits influenced engagement, and discovered that the most significant characteristic is not demographic, but behavioral: the 13.9% of users who bothered to fill their profiles generated 48.0% of all data. Because filling a survey form is, in itself, an indicator of greater commitment, any crowdsourcing initiative could use them strategically for identifying potentially high-value contributors.

The second application of games in our studies was for measuring quality of data. Some approaches for this issue include variations of majority voting (which may or may not weight individual contributions according to some measurement of workers' expertise) and cross-validation (*i.e.,* splitting the crowd into groups that validate each other's output), *e.g.,* Burnap *et al.* (2015); Ipeirotis *et al.* (2010); Luo and Zeynalvand (2017); Luo *et al.* (2019); Sheng *et al.* (2008); Snow *et al.* (2008); Sorokin and Forsyth (2008); Wang *et al.* (2020). These methods are built on the premise that workers having low expertise breaks down quality (Eickhoff, 2014; Kittur *et al.*, 2013), depending on the existence of a definite ground truth. For this reason, they are unsuitable for subjective tasks (Haralabopoulos *et al.*, 2020; Kairam and Heer, 2016), which may benefit from applying the notion of *crowd truth* (Aroyo and Welty, 2015) and take into account multiple perspectives and interpretations. Our approach was to create a serious game based on a surrogate truth generated by a small group of experts, but instead of using this benchmark to evaluate the quality of individual workers, we used it to probe the existence of behavioral biases between demographic groups. We found strong evidence that persons without disabilities can be equally reliable sources for accessibility data as persons with disabilities, as they did not have worse performance.

The *network effect* (Shapiro and Varian, 1998, chapter 7) is determinant for many crowdsourcing platforms, which are initially of only marginal importance to users, but become increasingly more valuable as the number of participants grows. For them, acquiring users is hard until reaching a critical mass of contributors. Location-sensitive platforms are even harder, as they must overcome this initial traction challenge at least once for every served region. Going forward, a possible path of investigation is finding what is the turning point in these cases. Guiaderodas benefitted from early media exposure, but this was not enough for identifying a replicable factor that can be used for accelerating growth.

While the implemented serious game was used for studying user behavior, it can also be applied for investigating whether initially training crowdworkers affects their performance and motivation. Additionally, it can also be adapted to serve as a tool for remote evaluation of accessibility, as long as a minimum number of appropriate images is somehow guaranteed for each location. The threshold amount of such images for ensuring fair evaluation is, however, still unknown.

In summary, considering each research question addressed in this thesis, we have that our experiments on platform design and gamification in a real-world, large-scale crowdsourcing platform show that both can positively affect crowd acquisition and engagement. And because traditional methods for ensuring data quality in crowdsourcing largely depend of the definition of a ground truth, we also conceived a method for testing biases in subjective contexts, and tested it by applying it in our platform through a serious game.

One of the most relevant limitations of this work is that while some findings are generic enough for application in other crowdsourcing contexts, such as that the most decisive predictor of volunteer engagement being behavioral, many results are specific to the field of crowdsourced accessibility databases. Furthermore, most data collected is about places located in Latin America, mainly in Brazil. While this helps close an important geographic gap, many regions are still left uncovered, such as Africa and Southeastern Asia.

Some possible avenues for continuation of this work include investigating factors that can be generalized to help accelerate growth of all kinds of crowdsourcing platforms. Potentially, these results may also be generalized to social computing in general. Other possibilities are to explore other applications of the implemented serious game, in particular whether if it can be used for remote evaluation of accessibility. More broadly, it can also be combined with experiments that involve *in loco* evaluation and test if initially training workers with a serious game can improve performance and motivation.

Machine learning systems make use of crowdsourcing for input generation, and techniques which recursively re-label data can largely benefit from methods for assembling surrogate truths, specially when working with subjective contexts. Examples of such applications include language processing, a field that has recently received quite a lot of publicity thanks to systems like ChatGPT, based on complex recurrent or convolutional neural networks (Vaswani *et al.*, 2017) and semi-supervised learning (Van Engelen and Hoos, 2020; Zhu, 2005). Moreover, most of the existing mathematical work on social computing focuses on a single application (Chen *et al.*, 2016). By detailing our approach and isolating the core components of our strategy, we also contribute to the still open challenge of laying out foundations of a framework that can generalize to a large number of social computing applications.

Bibliography

108th Congress(1990) 108th Congress. Americans With Disabilities Act of 1990, Public Law 101-336, 1990. Cited in page 30

ABNT(1985) Associação Brasileira de Normas Técnicas ABNT. *NBR 9050: Adequação das edificações e do mobiliário urbano à pessoa deficiente - Procedimento.* ABNT, 1st edition. Cited in page 28

ABNT(1994) Associação Brasileira de Normas Técnicas ABNT. *NBR 9050: Accessibility of the handicapped to buildings and the urban environment - Procedure.* ABNT, 1st revised edition. Cited in page 28

ABNT(2004) Associação Brasileira de Normas Técnicas ABNT. *NBR 9050: Accessibility to buildings, equipment and the urban environment.* ABNT, 2nd edition. Cited in page 28, 29, 30

ABNT(2015) Associação Brasileira de Normas Técnicas ABNT. *NBR 9050: Accessibility to buildings, equipment and the urban environment.* ABNT, 3rd edition. Cited in page 28, 29, 30

ABNT(2020) Associação Brasileira de Normas Técnicas ABNT. *NBR 9050: Accessibility to buildings, equipment and the urban environment.* ABNT, 4th edition. Cited in page 25, 28, 29, 31

Ahmad *et al.***(2018)** Waqas Ahmad, Shengling Wang, Ata Ullah, Muhammad Yasir Shabir *et al.* Reputation-aware recruitment and credible reporting for platform utility in mobile crowd sensing with smart devices in IoT. *Sensors*, 18(10):3305. Cited in page 19

Ahmetovic *et al.***(2017)** Dragan Ahmetovic, Roberto Manduchi, James M Coughlan and Sergio Mascetti. Mind your crossings: Mining GIS imagery for crosswalk localization. *ACM Transactions on Accessible Computing (TACCESS)*, 9(4):1–25. Cited in page 40

Akrout *et al.***(2019)** Ismail Akrout, Amal Feriani and Mohamed Akrout. Hacking google reCAPTCHA v3 using reinforcement learning. *arXiv preprint arXiv:1903.01003*. Cited in page 5

Amel-Zadeh and Serafeim(2018) Amir Amel-Zadeh and George Serafeim. Why and how investors use ESG information: Evidence from a global survey. *Financial analysts journal*, 74 (3):87–103. Cited in page 52

Arditte *et al.***(2016)** Kimberly A Arditte, Demet Çek, Ashley M Shaw and Kiara R Timpano. The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological assessment*, 28(6):684. Cited in page 5

Aroyo and Welty(2015) Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24. Cited in page 19, 88

Aroyo *et al.***(2019)** Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield and Rachel Rosen. Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions. In *Companion Proceedings of the 2019 World Wide Web Conference*, pages 1100–1105. Cited in page 19

ATBCB(2010) Architectural and Transportation Barriers Compliance Board ATBCB. *Americans with Disabilities Act: Accessibility Guidelines for Buildings and Facilities (ADAAG).* US Access Board. Cited in page 30, 31

Bakillah *et al.***(2013)** Mohamed Bakillah, Steve HL Liang, Amin Mobasheri and Alexander Zipf. Towards an efficient routing web processing service through capturing real-time road conditions from big data. In *2013 5th computer science and electronic engineering conference (CEEC)*, pages 152–155. IEEE. Cited in page 37

Bakillah *et al.***(2014)** Mohamed Bakillah, Johannes Lauer, Steve HL Liang, Alexander Zipf, J Jokar Arsanjani, Amin Mobasheri and Lukas Loos. Exploiting big vgi to improve routing and navigation services. *Big data techniques and technologies in geoinformatics*, pages 177–192. Cited in page 37

Barowy *et al.***(2012)** Daniel W Barowy, Charlie Curtsinger, Emery D Berger and Andrew McGregor. AutoMan: A platform for integrating human-based and digital computation. In *Proceedings of the ACM international conference on Object oriented programming systems languages and applications*, pages 639–654. Cited in page 7, 18, 19

Barroso *et al.***(2020)** João Barroso, Lourdes Moreno Lopez, Hugo Paredes, Franz Puehretmair and Tania Rocha. Special issue on accessibility and software design for all, 2020. Cited in page 31

Beale *et al.***(2006)** Linda Beale, Kenneth Field, David Briggs, Phil Picton and Hugh Matthews. Mapping for wheelchair users: Route navigation in urban spaces. *The Cartographic Journal*, 43 (1):68–81. Cited in page 38

Behrend *et al.***(2011)** Tara S Behrend, David J Sharek, Adam W Meade and Eric N Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800–813. Cited in page 5

Berriel *et al.***(2017)** Rodrigo F Berriel, Andre Teixeira Lopes, Alberto F De Souza and Thiago Oliveira-Santos. Deep learning-based large-scale automatic satellite crosswalk classification. *IEEE Geoscience and Remote Sensing Letters*, 14(9):1513–1517. Cited in page 40

Bi(2006) Yuhua Bi. Accessibility and attitudinal barriers encountered by travelers with physical disabilities in China. Master's Thesis, University of Missouri–Columbia. Cited in page 27

Bi *et al.***(2007)** Yuhua Bi, Jaclyn A Card and Shu T Cole. Accessibility and attitudinal barriers encountered by Chinese travellers with physical disabilities. *International Journal of Tourism Research*, 9(3):205–216. Cited in page 27

Biagi *et al.***(2017)** Ludovico Biagi, Sara Comai, Raffaella Mangiarotti, Matteo Matteucci, Marco Negretti and Secil Ugur Yavuz. Enriching geographic maps with accessible paths derived from implicit mobile device data collection. In *Enriching Urban Spaces with Ambient Computing, the Internet of Things, and Smart City Design*, pages 89–113. IGI Global. Cited in page 39

Biagi *et al.***(2020)** Ludovico Biagi, Maria Antonia Brovelli and Lorenzo Stucchi. Mapping the accessibility in OpenStreetMap: A comparison of different techniques. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:B4–2020. Cited in page 39

Bikhchandani *et al.***(1992)** Sushil Bikhchandani, David Hirshleifer and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026. Cited in page 9

Biljecki *et al.***(2013)** Filip Biljecki, Hugo Ledoux and Peter Van Oosterom. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2):385–407. Cited in page 37

Blanco *et al.***(2011)** Roi Blanco, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, Henry S Thompson and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 923–932. Cited in page 19

Bless and Fiedler(2014) Herbert Bless and Klaus Fiedler. *Social cognition: How individuals construct social reality.* Psychology Press. Cited in page 9, 20

Borgo *et al.***(2018)** Rita Borgo, Luana Micallef, Benjamin Bach, Fintan McGee and Bongshin Lee. Information visualization evaluation using crowdsourcing. In *Computer Graphics Forum*, volume 37, pages 573–595. Wiley Online Library. Cited in page 4

Bos and Von Weizsacker(1989) Dieter Bos and Robert K Von Weizsacker. Economic consequences of an aging population. *European economic review*, 33(2-3):345–354. Cited in page 21

Boudreau and Jeppesen(2015) Kevin J Boudreau and Lars B Jeppesen. Unpaid crowd complementors: The platform network effect mirage. *Strategic Management Journal*, 36(12):1761–1777. Cited in page 3

Bouguettaya *et al.***(2017)** Athman Bouguettaya, Munindar Singh, Michael Huhns, Quan Z Sheng, Hai Dong, Qi Yu, Azadeh Ghari Neiat, Sajib Mistry, Boualem Benatallah, Brahim Medjahed *et al.* A service computing manifesto: the next 10 years. *Communications of the ACM*, 60(4): 64–72. Cited in page 9, 87

Brabham(2008) Daren C Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90. Cited in page 2

Brabham *et al.***(2014)** Daren C Brabham, Kurt M Ribisl, Thomas R Kirchner and Jay M Bernhardt. Crowdsourcing applications for public health. *American journal of preventive medicine*, 46(2):179–187. Cited in page 4

Brasil(2015) Brasil. Lei Federal nº 13.146: Lei Brasileira de Inclusão da Pessoa com Deficiência, 2015. Cited in page 30

Breazeal *et al.***(2013)** Cynthia Breazeal, Nick DePalma, Jeff Orkin, Sonia Chernova and Malte Jung. Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1):82–111. Cited in page 16, 88

Bringa *et al.***(2010)** Olav Rand Bringa, Einar Lund and Kristi Ringard. Norway's planning approach to implement universal design. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Browne(2021) Ryan Browne. OnlyFans says it will no longer ban porn in stunning U-turn after user backlash. *CNBC*. Cited in page 4

Buhrmester *et al.***(2018)** Michael D Buhrmester, Sanaz Talaifar and Samuel D Gosling. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2):149–154. Cited in page 5

Buolamwini and Gebru(2018) Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR. Cited in page 7

Burnap *et al.***(2015)** Alex Burnap, Yi Ren, Richard Gerth, Giannis Papazoglou, Richard Gonzalez and Panos Y Papalambros. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design*, 137(3):031101. Cited in page 9, 18, 88

Butlewski and Jabłońska(2014) M Butlewski and J Jabłońska. Ergonomic model of hotel service quality for the elderly and people with disabilities. In *Occupational Safety and Hygiene II-Selected Extended and Revised Contributions from the International Symposium Occupational Safety and Hygiene, SHO*, pages 633–638. Cited in page 27, 36, 37

Callison-Burch and Dredze(2010) Chris Callison-Burch and Mark Dredze. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 1–12. Cited in page 19

Campbell *et al.***(2014)** Megan Campbell, Cynthia Bennett, Caitlin Bonnar and Alan Borning. Where's my bus stop? Supporting independence of blind transit riders with StopInfo. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*, pages 11–18. Cited in page 38

Cappa *et al.***(2019)** Francesco Cappa, Raffaele Oriani, Michele Pinelli and Alfredo De Massis. When does crowdsourcing benefit firm stock market performance? *Research Policy*, 48(9):103825. Cited in page 4

Card *et al.***(2006)** Jaclyn A Card, Shu T Cole and Amanda H Humphrey. A comparison of the accessibility and attitudinal barriers model: Travel providers and travelers with physical disabilities. *Asia Pacific Journal of Tourism Research*, 11(2):161–175. Cited in page 27

Cardonha *et al.***(2013)** Carlos Cardonha, Diego Gallo, Priscilla Avegliano, Ricardo Herrmann, Fernando Koch and Sergio Borger. A crowdsourcing platform for the construction of accessibility maps. In *Proceedings of the 10th international cross-disciplinary conference on web accessibility*, pages 1–4. Cited in page 38, 39, 88

Casadesus-Masanell and Hałaburda(2014) Ramon Casadesus-Masanell and Hanna Hałaburda. When does a platform create value by limiting choice? *Journal of Economics & Management Strategy*, 23(2):259–293. Cited in page 3

CEN-CENELEC(2014) CEN-CENELEC. *CEN-CENELEC GUIDE 6: Guide for addressing accessibility in standards.* International Organization for Standardization. Cited in page 30

Cesari *et al.***(2018)** Ugo Cesari, Giuseppe De Pietro, Elio Marciano, Ciro Niri, Giovanna Sannino and Laura Verde. A new database of healthy and pathological voices. *Computers & Electrical Engineering*, 68:310–321. Cited in page 37

Chandler and Kapelner(2013) Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133. Cited in page 9, 18, 55

Chandler *et al.***(2013)** Jesse Chandler, Gabriele Paolacci and Pam Mueller. Risks and rewards of crowdsourcing marketplaces. In *Handbook of human computation*, pages 377–392. Springer. Cited in page 10, 18

Chandler *et al.***(2014)** Jesse Chandler, Pam Mueller and Gabriele Paolacci. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1):112–130. Cited in page 5

Chen *et al.***(2016)** Yiling Chen, Arpita Ghosh, Michael Kearns, Tim Roughgarden and Jennifer Wortman Vaughan. Mathematical foundations for social computing. *Communications of the ACM*, 59(12):102–108. Cited in page 4, 8, 90

Choi *et al.***(2014)** Joohee Choi, Heejin Choi, Woonsub So, Jaeki Lee and JongJun You. A study about designing reward for gamified crowdsourcing system. In *International Conference of Design, User Experience, and Usability*, pages 678–687. Springer. Cited in page 17, 88

Chung *et al.***(2014)** Michael Jae-Yoon Chung, Maxwell Forbes, Maya Cakmak and Rajesh PN Rao. Accelerating imitation learning through crowdsourcing. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4777–4784. IEEE. Cited in page 5

Church and Marston(2003) Richard L. Church and James R. Marston. Measuring Accessibility for People with a Disability. *Geographical Analysis*, 35(1):83–96. ISSN 1538-4632. doi: 10.1111/j.1538-4632.2003.tb01102.x. URL http://dx.doi.org/10.1111/j.1538-4632.2003.tb01102.x. Cited in page 35, 37

Cooper *et al.***(2010)** Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović *et al.* Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760. Cited in page 15

Correia *et al.***(2020)** António Correia, Daniel Schneider, Shoaib Jameel, Hugo Paredes and Benjamim Fonseca. Empirical investigation of the factors influencing researchers' adoption of crowdsourcing and machine learning. In *International Conference on Intelligent Systems Design and Applications*, pages 1257–1270. Springer. Cited in page 9, 18

Coughlan and Shen(2013) James M Coughlan and Huiying Shen. Crosswatch: a system for providing guidance to visually impaired travelers at traffic intersection. *Journal of assistive technologies*. Cited in page 40

Council *et al.***(1990)** National Research Council *et al. Human factors research needs for an aging population*. National Academies Press. Cited in page 21

Dahlander and Piezunka(2020) Linus Dahlander and Henning Piezunka. Why crowdsourcing fails. *Journal of Organization Design*, 9(1):1–9. Cited in page 8, 87

D'Anastasio(2021) Cecilia D'Anastasio. Twitch and reddit protests may be only the beginning. *Wired magazine*. Cited in page 3

Darcy(2010) Simon Darcy. Inherent complexity: Disability, accessible tourism and accommodation information preferences. *Tourism Management*, 31(6):816–826. Cited in page 43

Davtyan *et al.***(2015)** Martin Davtyan, Carsten Eickhoff and Thomas Hofmann. Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 783–790. Cited in page 19

de Albuquerque *et al.***(2016)** João Porto de Albuquerque, Melanie Eckle, Benjamin Herfort and Alexander Zipf. Crowdsourcing geographic information for disaster management and improving urban resilience: an overview of recent developments and lessons learned. *European handbook of crowdsourced geographic information*, pages 309–321. Cited in page 37

de Alfaro and Shavlovsky(2013) Luca de Alfaro and Michael Shavlovsky. Crowdgrader: Crowdsourcing the evaluation of homework assignments. *arXiv preprint arXiv:1308.5273*. Cited in page 4

de Assis and Toledo(2016) Diva Carolina A de Assis and Alexandre Márcio Toledo. Concepção de calçadas à luz da NBR 9050–Interpretações equivocadas das recomendações da norma. *Blucher Engineering Proceedings*, 3(3):67–78. Cited in page 28

Deca(2018) Deca. Deca - louças e metais para banheiros, cozinhas e áreas de serviço. https://www.deca.com.br, 2018. Retrieved May 2018. Cited in page 29

Deterding *et al.***(2011)** Sebastian Deterding, Dan Dixon, Rilla Khaled and Lennart Nacke. From game design elements to gamefulness: defining "Gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pages 9–15. Cited in page 3, 16

Difallah *et al.***(2013)** Djellel Eddine Difallah, Gianluca Demartini and Philippe Cudré-Mauroux. Pick-a-crowd: tell me what you like, and I'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web*, pages 367–374. Cited in page 19

Ding *et al.***(2014)** Chaohai Ding, Mike Wald and Gary Wills. A survey of open accessibility data. In *proceedings of the 11th web for all conference*, pages 1–4. Cited in page 41

Ding *et al.***(2011)** Dan Ding, Shivayogi Hiremath, Younghyun Chung and Rory Cooper. Detection of wheelchair user activities using wearable sensors. In *International Conference on Universal Access in Human-Computer Interaction*, pages 145–152. Springer. Cited in page 39

Djaouti *et al.***(2011a)** Damien Djaouti, Julian Alvarez and Jean-Pierre Jessel. Classifying serious games: the G/P/S model. In *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches*, pages 118–136. IGI Global. Cited in page 15

Djaouti *et al.***(2011b)** Damien Djaouti, Julian Alvarez, Jean-Pierre Jessel and Olivier Rampnoux. Origins of serious games. In *Serious games and edutainment applications*, pages 25–43. Springer. Cited in page 15

Duan *et al.***(2019)** Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pages 255–260. Cited in page 37

D'Innocenzo and Morini(2010) Assunta D'Innocenzo and Annalisa Morini. Accessible design in italy. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Easley and Ghosh(2016) David Easley and Arpita Ghosh. Incentives, gamification, and game theory: an economic approach to badge design. *ACM Transactions on Economics and Computation (TEAC)*, 4(3):1–26. Cited in page 16

Edinger *et al.***(2019)** Janick Edinger, Alexandra Hofmann, Anton Wachner, Christian Becker, Vaskar Raychoudhury and Christian Krupitzer. Wheelshare: Crowd-sensed surface classification for accessible routing. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 584–589. IEEE. Cited in page 39

Eichhorn and Buhalis(2011) Victoria Eichhorn and Dimitrios Buhalis. Accessibility: A key objective for the tourism industry. *Accessible tourism: Concepts and issues*, pages 46–61. Cited in page 27

Eickhoff(2014) Carsten Eickhoff. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 53–56. Cited in page 9, 18, 48, 88

Eickhoff(2018) Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170. Cited in page 9, 20, 48, 70

Eickhoff and de Vries(2013) Carsten Eickhoff and Arjen P de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2):121–137. Cited in page 18

Faltings *et al.***(2014)** Boi Faltings, Radu Jurca, Pearl Pu and Bao Duy Tran. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*. Cited in page 9, 19

Fänge and Iwarsson(2003) Agneta Fänge and Susanne Iwarsson. Accessibility and usability in housing: construct validity and implications for research and practice. *Disability and Rehabilitation*, 25(23):1316–1325. Cited in page 25

Feng *et al.***(2018)** Yuanyue Feng, Hua Jonathan Ye, Ying Yu, Congcong Yang and Tingru Cui. Gamification artifacts and crowdsourcing participation: Examining the mediating role of intrinsic motivations. *Computers in Human Behavior*, 81:124–136. Cited in page 3

Ferguson(2011) Kitty Ferguson. *Stephen Hawking: His life and work.* Random House. Cited in page 22

Flores and Manduchi(2018) German H Flores and Roberto Manduchi. WeAllWalk: An annotated dataset of inertial sensor time series from blind walkers. *ACM Transactions on Accessible Computing (TACCESS)*, 11(1):1–28. Cited in page 39

Franklin *et al.***(2011)** Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh and Reynold Xin. CrowdDB: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 61–72. Cited in page 4, 7, 8

French *et al.***(2008)** Brian French, Divya Tyamagundlu, Daniel P Siewiorek, Asim Smailagic and Dan Ding. Towards a virtual coach for manual wheelchair users. In *2008 12th IEEE international symposium on wearable computers*, pages 77–80. IEEE. Cited in page 39

Friede *et al.***(2015)** Gunnar Friede, Timo Busch and Alexander Bassen. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of sustainable finance & investment*, 5(4):210–233. Cited in page 52

Froehlich *et al.***(2019)** Jon E Froehlich, Anke M Brock, Anat Caspi, João Guerreiro, Kotaro Hara, Reuben Kirkham, Johannes Schöning and Benjamin Tannert. Grand challenges in accessible maps. *interactions*, 26(2):78–81. Cited in page 27, 41, 43

Gadiraju *et al.***(2017)** Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(4):1–26. Cited in page 18, 20

Geiger and Schader(2014) David Geiger and Martin Schader. Personalized task recommendation in crowdsourcing information systems – Current state of the art. *Decision Support Systems*, 65:3–16. Cited in page 8, 13, 14, 45

Ghosh *et al.***(2015)** Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden and Antonio Reyes. SemEval-2015 Task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 470–478. Cited in page 19

Ghosh and Hummel(2014) Arpita Ghosh and Patrick Hummel. A game-theoretic analysis of rank-order mechanisms for user-generated content. *Journal of Economic Theory*, 154:349–374. Cited in page 17

Ghosh and McAfee(2011) Arpita Ghosh and Preston McAfee. Incentivizing high-quality user-generated content. In *Proceedings of the 20th international conference on World wide web*, pages 137–146. Cited in page 18

Gilart-Iglesias *et al.***(2015)** Virgilio Gilart-Iglesias, Higinio Mora, Raquel Pérez-delHoyo and Clara García-Mayor. A computational method based on radio frequency technologies for the analysis of accessibility of disabled people in sustainable cities. *Sustainability*, 7(11):14935–14963. Cited in page 39

Giles(2012) Jim Giles. Computational social science: Making the links. *Nature News*, 488(7412): 448. Cited in page 4

Gillan *et al.*(2021) Stuart L Gillan, Andrew Koch and Laura T Starks. Firms and social responsibility: A review of ESG and CSR research in corporate finance. *Journal of Corporate Finance*, 66:101889. Cited in page 52

Goodchild(2007) Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221. Cited in page 37

Goodchild and Li(2012) Michael F Goodchild and Linna Li. Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120. Cited in page 42

Google(2021) Google. Accessibility scanner. https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.auditor, 2021. Retrieved September 2021. Cited in page 32

Graser *et al.*(2015) Anita Graser, Markus Straub and Melitta Dragaschnig. Is OSM good enough for vehicle routing? a study comparing street networks in Vienna. In *Progress in Location-Based Services 2014*, pages 3–17. Springer. Cited in page 37

Grosbois(2010) Louis-Pierre Grosbois. The evolution of design for all in public buildings and-transportation in france. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Guimarães(2010) Marcelo Pinto Guimarães. Writing poetry rather than structuring grammar: Notes for the development of universal design in brazil. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 41

Guy and Truong(2012) Richard Guy and Khai Truong. CrossingGuard: exploring information content in navigation aids for visually impaired pedestrians. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 405–414. Cited in page 40

Hachmann *et al.*(2018) Samyra Hachmann, Jamal Jokar Arsanjani and Eric Vaz. Spatial data for slum upgrading: Volunteered Geographic Information and the role of citizen science. *Habitat international*, 72:18–26. Cited in page 37

Haider *et al.*(2019) Md. Masud Haider, Mohammad Rokibul Hoque, Md. Khaliluzzaman and Mohammad Mahadi Hassan. Zebra crosswalk region detection and localization based on deep convolutional neural network. In *2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON)*, pages 93–97. IEEE. Cited in page 40

Hamari(2013) Juho Hamari. Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. *Electronic commerce research and applications*, 12(4):236–245. Cited in page 17

Handy and Niemeier(1997) S L Handy and D A Niemeier. Measuring Accessibility: An Exploration of Issues and Alternatives. *Environment and Planning A: Economy and Space*, 29(7): 1175–1194. doi: 10.1068/a291175. URL https://doi.org/10.1068/a291175. Cited in page 35

Handy and Clifton(2001) Susan L Handy and Kelly J Clifton. Evaluating neighborhood accessibility: Possibilities and practicalities. *Journal of transportation and statistics*, 4(2/3):67–78. Cited in page 35

Hara *et al.*(2015) Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H Ng and Jon E Froehlich. Improving public transit accessibility for blind riders by crowdsourcing bus stop landmark locations with google street view: An extended analysis. *ACM Transactions on Accessible Computing (TACCESS)*, 6(2):1–23. Cited in page 38

Hara *et al.***(2019)** Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham and Chris Callison-Burch. Worker demographics and earnings on Amazon Mechanical Turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6. Cited in page 5

Haralabopoulos *et al.***(2020)** Giannis Haralabopoulos, Myron Tsikandilakis, Mercedes Torres Torres and Derek McAuley. Objective assessment of subjective tasks in crowdsourcing applications. In *Proceedings of the LREC 2020 Workshop on" Citizen Linguistics in Language Resource Development"*, pages 15–25. Cited in page 9, 19, 88

Hardt(2012) Dick Hardt. The OAuth 2.0 Authorization Framework. RFC 6749, 2012. URL https://www.rfc-editor.org/info/rfc6749. Cited in page 56

Haselton *et al.***(2015)** Martie G Haselton, Daniel Nettle and Damian R Murray. The evolution of cognitive bias. *The handbook of evolutionary psychology*, pages 1–20. Cited in page 9, 20

Hassan *et al.***(2020)** Saad Hassan, Larwan Berke, Elahe Vahdani, Longlong Jing, Yingli Tian and Matt Huenerfauth. An isolated-signing RGBD dataset of 100 American Sign Language signs produced by fluent ASL signers. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 89–94. Cited in page 37

Haworth and Bruce(2015) Billy Haworth and Eleanor Bruce. A review of volunteered geographic information for disaster management. *Geography Compass*, 9(5):237–250. Cited in page 37

Heaven(2020) Will Douglas Heaven. Predictive policing algorithms are racist. they need to be dismantled. *MIT ZTechnology Review*, 17:2020. Cited in page 7

Hochmair *et al.***(2013)** Hartwig H Hochmair, Dennis Zielstra, Pascal Neis, PN Hartwig, H Hochmair and D Zielstra. Assessing the completeness of bicycle trails and designated lane features in openstreetmap for the united states and europe. In *Transportation Research Board Annual Meeting*. Cited in page 37

Hornby *et al.***(2000)** Albert Sydney Hornby, Michael Ashby and Sally Wehmeier. *Oxford advanced learner's dictionary of current English*. Oxford University Press. Cited in page 23

Hossain(2012) Mokter Hossain. Users' motivation to participate in online crowdsourcing platforms. In *2012 International Conference on Innovation Management and Technology Research*, pages 310–315. IEEE. Cited in page 44, 55, 56

Hosseini *et al.***(2012)** Mehdi Hosseini, Ingemar J Cox, Nataša Milić-Frayling, Gabriella Kazai and Vishwa Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *European Conference on Information Retrieval*, pages 182–194. Springer. Cited in page 19

Hossen *et al.***(2020)** Md Imran Hossen, Yazhou Tu, Md Fazle Rabby, Md Nazmul Islam, Hui Cao and Xiali Hei. An object detection based solver for Google's image reCAPTCHA v2. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2020)*, pages 269–284. Cited in page 5

Howe(2006) Jeff Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4. Cited in page 1

Howe(2008) Jeff Howe. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House. Cited in page 1

Hube *et al.*(2019) Christoph Hube, Besnik Fetahu and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Cited in page 20

Huenerfauth and Kacorri(2014) Matt Huenerfauth and Hernisa Kacorri. Release of experimental stimuli and questions for evaluating facial expressions in animations of american sign language. In *Proceedings of the 6th workshop on the representation and processing of sign languages: beyond the manual channel, the 9th international conference on language resources and evaluation (LREC 2014), Reykjavik, Iceland*. Cited in page 37

IBGE(2010a) Instituto Brasileiro de Geografia e Estatística IBGE. Censo 2010. https://censo2010.ibge.gov.br/, 2010a. Retrieved May 2018. Cited in page 21

IBGE(2010b) Instituto Brasileiro de Geografia e Estatística IBGE. Censo demográfico, amostra - pessoas com deficiência. https://cidades.ibge.gov.br/brasil/sp/sao-paulo/pesquisa/23/23612/, 2010b. Retrieved February 2018. Cited in page 21

Ipeirotis *et al.*(2010) Panagiotis G Ipeirotis, Foster Provost and Jing Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. Cited in page 9, 18, 88

ISO(2014) International Organization for Standardization ISO. *ISO/IEC Guide 71: Guide for addressing accessibility in standards*. International Organization for Standardization. Cited in page 30

ISO(2011) International Organization for Standardization ISO. *ISO 21542: Construction–Accessibility and Usability of the Built Environment*. International Organization for Standardization. Cited in page 30

ISO(2006) International Organization for Standardization ISO. *ISO 20282-1: Ease of operation of everyday products–Part 1. Design requirements for context of use and user characteristics*. International Organization for Standardization. Cited in page 30

ISO(2013) International Organization for Standardization ISO. *ISO/TS 20282-2: Usability of consumer products and products for public use*. International Organization for Standardization. Cited in page 30

Iwarsson and Ståhl(2003) Susanne Iwarsson and Agnetha Ståhl. Accessibility, usability and universal design—positioning and definition of concepts describing person-environment relationships. *Disability and rehabilitation*, 25(2):57–66. Cited in page 25, 34

Jäger *et al.*(2020) Urs Jäger, Felipe Symmes and Guillermo Cardoza. *Scaling Strategies for Social Entrepreneurs*. Springer. Cited in page 51

Jeonnong-dong and Dongdaemun-gu(2008) Jeonnong-dong and Dongdaemun-gu. Indoor spatial analysis using space syntax. *ISPRS Silk Road for Information from Imagery*, pages 1065–1069. Cited in page 36

Kaartemo(2017) Valtteri Kaartemo. The elements of a successful crowdfunding campaign: A systematic literature review of crowdfunding performance. *International Review of Entrepreneurship*, 15(3):291–318. Cited in page 9, 87

Kacorri *et al.*(2016) Hernisa Kacorri, Sergio Mascetti, Andrea Gerino, Dragan Ahmetovic, Hironobu Takagi and Chieko Asakawa. Supporting orientation of people with visual impairment: Analysis of large scale usage data. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 151–159. Cited in page 37

Kairam and Heer(2016) Sanjay Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 1637–1648. Cited in page 9, 19, 88

Kamikubo *et al.*(2021) Rie Kamikubo, Utkarsh Dwivedi and Hernisa Kacorri. Sharing practices for datasets related to accessibility and aging. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–16. Cited in page 37, 87

Kanasi *et al.*(2016) Eleni Kanasi, Srinivas Ayilavarapu and Judith Jones. The aging population: demographics and the biology of aging. *Periodontology 2000*, 72(1):13–18. Cited in page 21

Kanefsky *et al.*(2001) Bob Kanefsky, Nadine G Barlow and Virginia C Gulick. Can distributed volunteers accomplish massive data analysis tasks. *Lunar and Planetary Science*, 1:32. Cited in page 14

Karger *et al.*(2014) David R Karger, Sewoong Oh and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24. Cited in page 19

Katz *et al.*(2017) Daniel Martin Katz, Michael James Bommarito and Josh Blackman. Crowdsourcing accurately and robustly predicts Supreme Court decisions. *Available at SSRN 3085710.* Cited in page 4

Kaufmann *et al.*(2011) Nicolas Kaufmann, Thimo Schulze and Daniel Veit. More than fun and money: Worker motivation in crowdsourcing-a study on Mechanical Turk. In *Proceedings of the 17th Americas Conference on Information Systems - AMCIS*, pages 1–11. Cited in page 15

Kavaliova *et al.*(2016) Maya Kavaliova, Farzad Virjee, Natalia Maehle and Ingeborg Astrid Kleppe. Crowdsourcing innovation and product development: Gamification as a motivational driver. *Cogent Business & Management*, 3(1):1128132. Cited in page 3, 15, 88

Kettunen *et al.*(2016) Juhani Kettunen, Jari Silander, Matti Lindholm, Maiju Lehtiniemi, Outi Setälä and Seppo Kaitala. Changing role of citizens in national environmental monitoring. *European Handbook of Crowdsourced Geographic Information*, page 257. Cited in page 37

Kim(2005) JongBae Kim. *Development and Effectiveness Evaluation of a Virtualized Reality Telerehabilitation System for Accessibility Analysis of Built Environment.* PhD Thesis, University of Pittsburgh. Cited in page 36

Kirkham *et al.*(2017) Reuben Kirkham, Romeo Ebassa, Kyle Montague, Kellie Morrissey, Vasilis Vlachokyriakos, Sebastian Weise and Patrick Olivier. WheelieMap: an exploratory system for qualitative reports of inaccessibility in the built environment. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12. Cited in page 39

Kittur *et al.*(2013) Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease and John Horton. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318. Cited in page 8, 9, 18, 48, 87, 88

Koch *et al.*(2012) Fernando Koch, Carlos Cardonha, Jan Marcel Gentil and Sergio Borger. A platform for citizen sensing in sentient cities. In *International Workshop on Citizen in Sensor Networks*, pages 57–66. Springer. Cited in page 38, 88

Koester *et al.*(2016) Daniel Koester, Björn Lunt and Rainer Stiefelhagen. Zebra crossing detection from aerial imagery across countries. In *International Conference on Computers Helping People with Special Needs*, pages 27–34. Springer. Cited in page 40

Kohler(2015) Thomas Kohler. Crowdsourcing-based business models: how to create and capture value. *California management review*, 57(4):63–84. Cited in page 2, 3

Kolmonen(2017) Lauri Kolmonen. Business Opportunities in Crowdsourced Stock Market Analysis. Master's thesis, Aalto University. School of Business. URL http://urn.fi/URN:NBN:fi:aalto-201711137563. Cited in page 4

Kose(2010) Satoshi Kose. The impact of aging on japanese accessibility standards. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 21, 34

Kovashka *et al.***(2016)** Adriana Kovashka, Olga Russakovsky, Li Fei-Fei and Kristen Grauman. Crowdsourcing in computer vision. *arXiv preprint arXiv:1611.02145*. Cited in page 5

Kraft(2016) Amy Kraft. Microsoft shuts down AI chatbot after it turned into a Nazi. *CBS News*. Cited in page 8

Kramer(1956) Clyde Young Kramer. Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, 12(3):307–310. Cited in page 81

Krauss(2010) Ingrid Krauss. Manifestations of universal design in germany. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Krempi(2004) Ana Paula Krempi. Exploring spatial statistics tools for an accessibility analysis in the city of Bauru. Master's Thesis, Escola de Engenharia de São Carlos da Universidade de Sâo Paulo. Cited in page 36

Kruger and Dunning(1999) Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121. Cited in page 9, 19

Kruskal and Wallis(1952) William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621. Cited in page 62, 63, 79

Kubásek and Hřebíček(2013) Miroslav Kubásek and Jiří Hřebíček. Crowdsource approach for mapping of illegal dumps in the Czech Republic. *International Journal of Spatial Data Infrastructures Research*, 8:144–157. Cited in page 37

Kutchukian and Kutchukian(2017) Lúcia Kutchukian and Eric Kutchukian. Análise das condições de acessibilidade no ambiente urbano da área central de Guarapuava. *Revista Técnico-Científica*, 1(7). Cited in page 28

Law and Von Ahn(2009) Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. Cited in page 18, 88

Law *et al.***(2007)** Edith LM Law, Luis Von Ahn, Roger B Dannenberg and Mike Crawford. TagATune: A Game for Music and Sound Annotation. In *ISMIR*, volume 3, page 2. Cited in page 18, 88

Lazar *et al.***(2015)** Jonathan Lazar, Daniel F Goldstein and Anne Taylor. *Ensuring digital accessibility through process and policy*. Morgan kaufmann. Cited in page 31

Lazer *et al.***(2009)** David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann *et al.* Computational social science. *Science (New York, NY)*, 323(5915):721–723. Cited in page 4

Lease(2011) Matthew Lease. On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Cited in page 9, 18, 87

Lee and Kacorri(2019) Kyungjun Lee and Hernisa Kacorri. Hands holding clues for object recognition in teachable machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Cited in page 37

Leightley *et al.*(2015) Daniel Leightley, Moi Hoon Yap, Jessica Coulson, Yoann Barnouin and Jamie S McPhee. Benchmarking human motion analysis using kinect one: An open source dataset. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–7. IEEE. Cited in page 37

Lenart-Gansiniec(2018) Regina Lenart-Gansiniec. Methodological challenges of research on crowdsourcing. *Journal of Entrepreneurship, Management and Innovation*, 14(4):107–126. Cited in page 8, 87

Lerner and Tirole(2002) Josh Lerner and Jean Tirole. Some simple economics of open source. *The journal of industrial economics*, 50(2):197–234. Cited in page 1

Lévy(1997) Pierre Lévy. Collective intelligence, 1997. Cited in page 4

Lewthwaite and James(2020) Sarah Lewthwaite and Abi James. Accessible at last?: what do new European digital accessibility laws mean for disabled people in the UK? *Disability & Society*, 35(8):1360–1365. Cited in page 31

Luo and Zeynalvand(2017) Tie Luo and Leonit Zeynalvand. Reshaping mobile crowd sensing using cross validation to improve data credibility. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–7. IEEE. Cited in page 19, 68, 88

Luo *et al.*(2019) Tie Luo, Jianwei Huang, Salil S Kanhere, Jie Zhang and Sajal K Das. Improving IoT data quality in mobile crowd sensing: A cross validation approach. *IEEE Internet of Things Journal*, 6(3):5651–5664. Cited in page 19, 68, 88

Mack *et al.*(2021) Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich and Leah Findlater. What do we mean by "Accessibility Research"? a literature survey of accessibility papers in CHI and ASSETS from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18. Cited in page 10, 42, 87

Mackett *et al.*(2008) Roger L Mackett, Kamalasudhan Achuthan and Helena Titheridge. AMELIA: A tool to make transport policies more socially inclusive. *Transport Policy*, 15(6): 372–378. Cited in page 36

Mann(2004) William C Mann. The aging population and its needs. *IEEE Pervasive Computing*, 3(2):12–14. Cited in page 21

Marcus *et al.*(2011) Adam Marcus, Eugene Wu, David Karger, Samuel Madden and Robert Miller. Human-powered sorts and joins. *arXiv preprint arXiv:1109.6881*. Cited in page 4, 9

Martins(2016) Laura Martins. Vaso sanitário com abertura frontal oferece risco e não pode ser instalado! http://cadeiravoadora.com.br/vaso-sanitario-com-abertura-frontal-nao-e-adequado/, 2016. Retrieved May 2018. Cited in page 30

Mason and Suri(2012) Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, 44(1):1–23. Cited in page 5, 9

Matthews *et al.*(2003) Hugh Matthews, Linda Beale, Phil Picton and David Briggs. Modelling Access with GIS in Urban Systems (MAGUS): capturing the experiences of wheelchair users. *Area*, 35(1):34–45. Cited in page 38

Maynard and Bontcheva(2016) DG Maynard and Kalina Bontcheva. Challenges of evaluating sentiment analysis tools on social media. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1142–1148. LREC. Cited in page 19

Michael and Chen(2005) David R Michael and Sandra L Chen. *Serious games: Games that educate, train, and inform.* Muska & Lipman/Premier-Trade. Cited in page 15

Ministry of the Environment(2009) Ministry of the Environment. Universal design as a municipal strategy. http://www.universell-utforming.miljo.no/file_upload/bautaengkorr4.pdf, 2009. Retrieved May 2018. Cited in page 34

Mirri *et al.*(2016) Silvia Mirri, Catia Prandi and Paola Salomoni. Personalizing pedestrian accessible way-finding with mPASS. In *2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1119–1124. IEEE. Cited in page 39, 88

Mobasheri *et al.*(2017) Amin Mobasheri, Jonas Deister and Holger Dieterich. Wheelmap: the wheelchair accessibility crowdsourcing platform. *Open Geospatial Data, Software and Standards*, 2(1):1–7. Cited in page 10, 38

Mobasheri *et al.*(2018a) Amin Mobasheri, Haosheng Huang, Lívia Castro Degrossi and Alexander Zipf. Enrichment of OpenStreetMap data completeness with sidewalk geometries using data mining techniques. *Sensors*, 18(2):509. Cited in page 39, 40

Mobasheri *et al.*(2018b) Amin Mobasheri, Alexander Zipf and Louise Francis. OpenStreetMap data quality enrichment through awareness raising and collective action tools–experiences from a European project. *Geo-spatial Information Science*, 21(3):234–246. Cited in page 39, 40, 53

Moghaddam *et al.*(2011) Athena K Moghaddam, Joelle Pineau, Jordan Frank, Philippe Archambault, François Routhier, Thérèse Audet, Jan Polgar, François Michaud and Patrick Boissy. Mobility profile and wheelchair driving skills of powered wheelchair users: Sensor-based event recognition using a support vector machine classifier. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7336–7339. IEEE. Cited in page 39

Montgomery(2017) Douglas C Montgomery. *Design and analysis of experiments.* John wiley & sons. Cited in page 64, 79, 81

Mora *et al.*(2017) Higinio Mora, Virgilio Gilart-Iglesias, Raquel Pérez-del Hoyo and María Dolores Andújar-Montoya. A comprehensive system for monitoring urban accessibility in smart cities. *Sensors*, 17(8):1834. Cited in page 39

Moraes(2007) Miguel Correia de Moraes. Acessibilidade no Brasil: análise da NBR 9050. Master's Thesis, Universidade Federal de Santa Catarina, Florianópolis, SC. Cited in page 30

Morschheuser *et al.*(2016) Benedikt Morschheuser, Juho Hamari and Jonna Koivisto. Gamification in crowdsourcing: a review. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 4375–4384. IEEE. Cited in page 3, 8, 9, 16, 17, 88

Moss(2018) Richard C. Moss. “The least-worst idea we had” – The creation of the Age of Empires empire. https://arstechnica.com/gaming/2018/01/the-age-of-age-of-empires-as-told-by-the-devs-who-built-it/, 2018. Retrieved September 2021. Cited in page 1

Mourcou *et al.*(2013) Quentin Mourcou, Anthony Fleury, Pascal Dupuy, Bruno Diot, Celine Franco and Nicolas Vuillerme. Wegoto: A smartphone-based approach to assess and improve accessibility for wheelchair users. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1194–1197. IEEE. Cited in page 39

Moysidou and Hausberg(2020) Krystallia Moysidou and J Piet Hausberg. In crowdfunding we trust: A trust-building model in lending crowdfunding. *Journal of Small Business Management*, 58(3):511–543. Cited in page 4

Naroditskiy *et al.***(2014)** Victor Naroditskiy, Nicholas R Jennings, Pascal Van Hentenryck and Manuel Cebrian. Crowdsourcing contest dilemma. *Journal of The Royal Society Interface*, 11 (99):20140532. Cited in page 8

Neidle *et al.***(2012)** Carol Neidle, Ashwin Thangali and Stan Sclaroff. Challenges in development of the american sign language lexicon video dataset (ASLLVD) corpus. In *5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC*. Citeseer. Cited in page 37

Nicolau *et al.***(2011)** Marcos Antonio Nicolau, Cândida Nobre and Ana Cirne Paes de Barros. Fiat mio: um carro para chamar de seu? reflexões sobre comunicação e hábitos de consumo na sociedade em rede. *Revista Comunicação Midiática*, 6(1):58–79. Cited in page 2

Obermeyer *et al.***(2019)** Ziad Obermeyer, Brian Powers, Christine Vogeli and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453. Cited in page 7

Oncins(2020) Estella Oncins. Mapping the european digital accessibility field: The IMPACT project. In *9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, pages 33–37. Cited in page 31

Ostroff(2010) Elaine Ostroff. Universal design: an evolving paradigm. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Otmani *et al.***(2009)** R Otmani, A Moussaoui and A Pruski. A new approach to indoor accessibility. *International Journal of Smart Home*, 3(4):1–14. Cited in page 36

Palazzi *et al.***(2010)** Claudio E Palazzi, Lorenzo Teodori and Marco Roccetti. Path 2.0: A participatory system for the generation of accessible routes. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1707–1711. IEEE. Cited in page 39

Paolacci and Chandler(2014) Gabriele Paolacci and Jesse Chandler. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science*, 23 (3):184–188. Cited in page 5, 10, 19

Pasaogullari and Doratli(2004) Nil Pasaogullari and Naciye Doratli. Measuring accessibility and utilization of public spaces in Famagusta. *Cities*, 21(3):225 – 232. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2004.03.003. URL http://www.sciencedirect.com/science/article/pii/S0264275104000290. Cited in page 35

Pearson(1900) Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175. Cited in page 64, 76

Pérez-delHoyo *et al.***(2016)** Raquel Pérez-delHoyo, Clara García-Mayor, Higinio Mora-Mora, Virgilio Gilart-Iglesias and María Dolores Andújar-Montoya. Making smart and accessible cities: An urban model based on the design of intelligent environments. In *2016 5th International Conference on Smart Cities and Green ICT Systems (SMARTGREENS)*, pages 1–8. IEEE. Cited in page 39

Pirie(1979) G H Pirie. Measuring Accessibility: A Review and Proposal. *Environment and Planning A: Economy and Space*, 11(3):299–312. doi: 10.1068/a110299. URL https://doi.org/10.1068/a110299. Cited in page 27, 35

Prandi *et al.***(2016)** Catia Prandi, Paola Salomoni, Marco Roccetti, Valentina Nisi and Nuno Jardim Nunes. Walking with Geo-Zombie: A pervasive game to engage people in urban crowdsourcing. In *2016 International Conference on Computing, Networking and Communications (ICNC)*, pages 1–5. IEEE. Cited in page 16, 39, 88

Prandi *et al.***(2017a)** Catia Prandi, Silvia Mirri, Stefano Ferretti and Paola Salomoni. On the need of trustworthy sensing and crowdsourcing for urban accessibility in smart city. *ACM Transactions on Internet Technology (TOIT)*, 18(1):1–21. Cited in page 40

Prandi *et al.***(2017b)** Catia Prandi, Marco Roccetti, Paola Salomoni, Valentina Nisi and Nuno Jardim Nunes. Fighting exclusion: a multimedia mobile app with zombies and maps as a medium for civic engagement and design. *Multimedia Tools and Applications*, 76(4):4951–4979. Cited in page 16, 17, 39, 88

Prelec *et al.***(2017)** Dražen Prelec, H Sebastian Seung and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535. Cited in page 18

Raji and Buolamwini(2019) Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435. Cited in page 7

Ramot *et al.***(2010)** Avi Ramot, Yael Danieli-Lahav and Judith Bendel. Planning accessibility in the old city of jerusalem. In Wolfgang Preiser and Korydon Smith, editors, *Universal design handbook*. McGraw-Hill Professional, New York, 2nd edition. Cited in page 34

Rello *et al.***(2014)** Luz Rello, Joaquim Llisterri and Ricardo Baeza-Yates. DysList: An annotated resource of dyslexic errors. Cited in page 37

Rice *et al.***(2012)** Matthew T Rice, Ahmad O Aburizaiza, R Daniel Jacobson, Brandon M Shore and Fabiana I Paez. Supporting accessibility for blind and vision-impaired people with a localized gazetteer and open source geotechnology. *Transactions in GIS*, 16(2):177–190. Cited in page 38, 88

Robson *et al.***(2015)** Karen Robson, Kirk Plangger, Jan H Kietzmann, Ian McCarthy and Leyland Pitt. Is it all a game? Understanding the principles of gamification. *Business horizons*, 58(4): 411–420. Cited in page 3

Rochet and Tirole(2003) Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the european economic association*, 1(4):990–1029. Cited in page 3

Rzeszotarski and Kittur(2012) Jeffrey Rzeszotarski and Aniket Kittur. CrowdScape: interactively visualizing user behavior and output. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 55–62. Cited in page 19

Rzeszotarski and Kittur(2011) Jeffrey M Rzeszotarski and Aniket Kittur. Instrumenting the crowd: using implicit behavioral measures to predict task performance. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 13–22. Cited in page 19

Saab *et al.***(2019)** Farah Saab, Imad H Elhajj, Ayman Kayssi and Ali Chehab. Modelling cognitive bias in crowdsourcing systems. *Cognitive Systems Research*, 58:1–18. Cited in page 9, 20

Safire(2009) William Safire. Fat tail. *The New York Times*. Cited in page 1

Saha *et al.***(2019)** Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara *et al.* Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Cited in page 10, 40, 88

Saha *et al.***(2021)** Manaswi Saha, Devanshi Chauhan, Siddhant Patil, Rachel Kangas, Jeffrey Heer and Jon E Froehlich. Urban accessibility as a socio-political problem: A multi-stakeholder analysis. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–26. Cited in page 35

Saidi Sief *et al.***(2016)** A Saidi Sief, Alain Pruski and Abdelhak Bennia. A new approach for handling element accessibility problems faced by persons with a wheelchair. *Journal of Automation Mobile Robotics and Intelligent Systems*, 10. Cited in page 36

Saldanha *et al.***(2014)** Fabio Prado Saldanha, Patrick Cohendet and Marlei Pozzebon. Challenging the stage-gate model in crowdsourcing: The case of fiat mio in brazil. *Technology Innovation Management Review*, 4(9). Cited in page 2

Schall *et al.***(2011)** Daniel Schall, Florian Skopik and Schahram Dustdar. Expert discovery and interactions in mixed service-oriented systems. *IEEE Transactions on services computing*, 5(2): 233–245. Cited in page 19

Schonfeld(2008) Erick Schonfeld. When Crowdsourcing Fails: Cambrian House Headed to the Deadpool. *Techcrunch.* Cited in page 8

Se and Jasiobedzki(2006) Stephen Se and Piotr Jasiobedzki. Photo-realistic 3d model reconstruction. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 3076–3082. IEEE. Cited in page 36

Senaratne *et al.***(2017)** Hansi Senaratne, Amin Mobasheri, Ahmed Loai Ali, Cristina Capineri and Mordechai Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1):139–167. Cited in page 40

Shankar *et al.***(2017)** Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World.* Cited in page 6, 40

Shapiro and Varian(1998) Carl Shapiro and Hal R. Varian. *Information rules: a strategic guide to the network economy.* Harvard Business Press. Cited in page 3, 53, 89

Shapiro and Wilk(1965) Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611. Cited in page 77

Sheehan(2018) Kim Bartel Sheehan. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs*, 85(1):140–156. Cited in page 5, 9

Sheng *et al.***(2008)** Victor S Sheng, Foster Provost and Panagiotis G Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. Cited in page 9, 18, 88

Shigeno *et al.***(2013)** Kelly Shigeno, Sergio Borger, Diego Gallo, Ricardo Herrmann, Mateus Molinaro, Carlos Cardonha, Fernando Koch and Priscilla Avegliano. Citizen sensing for collaborative construction of accessibility maps. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–2. Cited in page 38, 88

Siegel(2019) Rachel Siegel. Tumblr once sold for $1.1 billion. The owner of WordPress just bought the site for a fraction of that. *The Washington Post*. Cited in page 4

Sivakorn *et al.*(2016) Suphannee Sivakorn, Jason Polakis and Angelos D Keromytis. I'm not a human: Breaking the Google reCAPTCHA. *Black Hat*, pages 1–12. Cited in page 5

Snow *et al.*(2008) Rion Snow, Brendan O'connor, Dan Jurafsky and Andrew Y Ng. Cheap and fast–but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263. Cited in page 5, 9, 18, 88

Sorokin and Forsyth(2008) Alexander Sorokin and David Forsyth. Utility data annotation with Amazon Mechanical Turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE. Cited in page 9, 19, 68, 88

Souza and Thomé(2008) Luiz Alberto Souza and Anderson Viera Thomé. Análise das condições de acessibilidade no ambiente urbano da Área central de Blumenau. *Seminário Internacional NUTAU/USP - Núcleo de Pesquisa em Tecnologia da Arquitetura e Urbanismo da Universidade de São Paulo*, 7. Cited in page 30

Statcounter(2021) Statcounter. Mobile operating system market share worldwide. https://gs.statcounter.com/os-market-share/mobile/worldwide, 2021. Retrieved September 2021. Cited in page 33

Stewart *et al.*(2015) Neil Stewart, Christoph Ungemach, Adam JL Harris, Daniel M Bartels, Ben R Newell, Gabriele Paolacci and Jesse Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, 10(5):479–491. Cited in page 5

Story(2001) Molly Follette Story. Principles of universal design. *Universal design handbook*. Cited in page 34, 35

Sun and Mobasheri(2017) Yeran Sun and Amin Mobasheri. Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using strava data. *International journal of environmental research and public health*, 14(3):274. Cited in page 37

Sun *et al.*(2015) Yeran Sun, Hongchao Fan, Mohamed Bakillah and Alexander Zipf. Road-based travel recommendation using geo-tagged images. *Computers, Environment and Urban Systems*, 53:110–122. Cited in page 37

Surowiecki(2004) James Surowiecki. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations*, 296(5). Cited in page 14

Susarla *et al.*(2012) Anjana Susarla, Jeong-Ha Oh and Yong Tan. Social networks and the diffusion of user-generated content: Evidence from YouTube. *Information systems research*, 23 (1):23–41. Cited in page 2

Tang *et al.*(2015) Duyu Tang, Bing Qin and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432. Cited in page 19

Tanimoto *et al.*(2013) Yoshio Tanimoto, Kuniharu Nanba, Kazunari Furusawa, Hideki Yamamoto, Akihiro Tokuhiro and Hiroyuki Ukida. Small device for counting the number of manual wheelchair strokes. In *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1755–1760. IEEE. Cited in page 39

Täuscher(2017) Karl Täuscher. Leveraging collective intelligence: How to design and manage crowd-based business models. *Business Horizons*, 60(2):237–245. Cited in page 2, 3, 4, 8, 60, 87

Thill *et al.***(2011)** Jean-Claude Thill, Thi Hong Diep Dao and Yuhong Zhou. Traveling in the three-dimensional city: applications in route planning, accessibility assessment, location analysis and beyond. *Journal of Transport Geography*, 19(3):405–421. Cited in page 36

Tong *et al.***(2020)** Yongxin Tong, Zimu Zhou, Yuxiang Zeng, Lei Chen and Cyrus Shahabi. Spatial crowdsourcing: a survey. *The VLDB Journal*, 29(1):217–250. Cited in page 3

Torvalds(1992) Linus Torvalds. LINUX's History. https://www.cs.cmu.edu/~awb/linux.history.html, 1992. Retrieved September 2021. Cited in page 1

Tukey(1949) John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114. Cited in page 81

Tukey *et al.***(1977)** John W Tukey *et al. Exploratory data analysis*, volume 2. Reading, MA. Cited in page 79

Turner(2006) Andrew Turner. *Introduction to neogeography.* " O'Reilly Media, Inc.". Cited in page 37

UNFPA(2017) United Nations Population Fund UNFPA. World population dashboard. https://www.unfpa.org/data/world-population-dashboard, 2017. Retrieved June 2018. Cited in page 21

United Nations(2018) United Nations. World population prospects. https://esa.un.org/unpd/wpp/dataquery/, 2018. Retrieved June 2018. Cited in page 21

United Nations(2006) United Nations. *Convention on the Rights of Persons with Disabilities (CRPD).* United Nations. Cited in page 24, 34

Van Engelen and Hoos(2020) Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440. Cited in page 90

Vanclooster *et al.***(2012)** Ann Vanclooster, Tijs Neutens, Veerle Fack, Nico Van de Weghe and Philippe De Maeyer. Measuring the exitability of buildings: A new perspective on indoor accessibility. *Applied Geography*, 34:507–518. Cited in page 36

Vaswani *et al.***(2017)** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. Attention is all you need, 2017. Cited in page 90

Vatavu and Ungurean(2019) Radu-Daniel Vatavu and Ovidiu-Ciprian Ungurean. Stroke-gesture input for people with motor impairments: Empirical results & research roadmap. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14. Cited in page 37

Von Ahn(2006) Luis Von Ahn. Games with a purpose. *Computer*, 39(6):92–94. Cited in page 15

Von Ahn and Dabbish(2004) Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. Cited in page 15, 18

Von Ahn and Dabbish(2008) Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67. Cited in page 9, 15, 16

Von Ahn *et al.***(2003)** Luis Von Ahn, Manuel Blum, Nicholas J Hopper and John Langford. CAPTCHA: Using hard AI problems for security. In *International conference on the theory and applications of cryptographic techniques*, pages 294–311. Springer. Cited in page 5

Von Ahn *et al.***(2004)** Luis Von Ahn, Manuel Blum and John Langford. Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60. Cited in page 5

Von Ahn *et al.***(2008)** Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468. Cited in page 5, 6, 19

Von Hippel(2006) Eric Von Hippel. *Democratizing innovation.* the MIT Press. Cited in page 1

Vuurens *et al.***(2011)** Jeroen Vuurens, Arjen P de Vries and Carsten Eickhoff. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)*, pages 21–26. Cited in page 19

W3C(2021) World Wide Web Consortium W3C. Introduction to Web Accessibility. https://www.w3.org/WAI/fundamentals/accessibility-intro/, 2021. Retrieved September 2021. Cited in page 32

Wang *et al.***(2020)** Cong Wang, Mingming Zhao, Qinyue Wang and Min Li. A sentinel-based peer assessment mechanism for collaborative learning. *CMC-COMPUTERS MATERIALS & CONTINUA*, 65(3):2309–2319. Cited in page 4, 19, 69, 88

Wang *et al.***(2016)** Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng and Ben Y Zhao. Defending against sybil devices in crowdsourced mapping services. In *Proceedings of the 14th annual international conference on mobile systems, applications, and services*, pages 179–191. Cited in page 3, 9, 19

Wang *et al.***(2013)** Tianyi Wang, Gang Wang, Xing Li, Haitao Zheng and Ben Y Zhao. Characterizing and detecting malicious crowdsourcing. In *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*, pages 537–538. Cited in page 8, 9, 19

Wang *et al.***(2019)** Yufeng Wang, Hui Fang, Qun Jin and Jianhua Ma. SSPA: An effective semi-supervised peer assessment method for large scale MOOCs. *Interactive Learning Environments*, pages 1–19. Cited in page 4, 9, 19, 69

WebAIM(2021) Web Accessibility in Mind WebAIM. WAVE Evaluation Tool. https://chrome.google.com/webstore/detail/wave-evaluation-tool/jbbplnpkjmmeebjpijfedlgcdilocofh, 2021. Retrieved September 2021. Cited in page 32

Weld *et al.***(2019)** Galen Weld, Esther Jang, Anthony Li, Aileen Zeng, Kurtis Heimerl and Jon E Froehlich. Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 196–209. Cited in page 10, 40, 88

Wellcome(2020) Wellcome. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak, 2020. Retrieved September 2021. Cited in page 4

WHO(2018) World Health Organization WHO. Disability and health: fact sheet. http://www.who.int/mediacentre/factsheets/fs352/en/, 2018. Retrieved April 2018. Cited in page 21

Wikipedia(2023) Wikipedia. Wikipedia:about. https://en.wikipedia.org/wiki/Wikipedia:About, 2023. Retrieved August 2023. Cited in page 2

Wikipedia(2021a) Wikipedia. Biblioteca Mário de Andrade. https://pt.wikipedia.org/wiki/Biblioteca_M%C3%A1rio_de_Andrade, 2021a. Retrieved August 2021. Cited in page 23

Wikipedia(2021b) Wikipedia. Coletor eletrônico de voto. https://pt.wikipedia.org/wiki/Coletor_eletr%C3%B4nico_de_voto, 2021b. Retrieved September 2021. Cited in page 32

Xi and Hamari(2019) Nannan Xi and Juho Hamari. Does gamification satisfy needs? A study on the relationship between gamification features and intrinsic need satisfaction. *International Journal of Information Management*, 46:210–221. Cited in page 17

Yau *et al.***(2004)** Matthew Kwai-sang Yau, Bob McKercher and Tanya L Packer. Traveling with a disability: More than an access issue. *Annals of Tourism Research*, 31(4):946–960. Cited in page 43

Zadrija *et al.***(2018)** Valentina Zadrija, Josip Krapac, Siniša Šegvić and Jakob Verbeek. Sparse weakly supervised models for object localization in road environment. *Computer Vision and Image Understanding*, 176:9–21. Cited in page 40

Zhang and Ai(2015) Xiang Zhang and Tinghua Ai. How to model roads in OpenStreetMap? a method for evaluating the fitness-for-use of the network for navigation. In *Advances in Spatial Data Handling and Analysis*, pages 143–162. Springer. Cited in page 37

Zhao and Zhu(2014) Yuxiang Zhao and Qinghua Zhu. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*, 16(3):417–434. Cited in page 8, 15, 87

Zhou *et al.***(2018)** Yuan Zhou, Zesun Yang, Chenxu Wang and Matthew Boutell. Breaking Google reCAPTCHA v2. *Journal of Computing Sciences in Colleges*, 34(1):126–136. Cited in page 5, 6

Zhu(2005) Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Cited in page 90

Zwillinger and Kokoska(1999) Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press. Cited in page 79