



*Mémoire de Maîtrise universitaire en Humanités numériques
et Informatique pour les sciences humaines*

Understanding Digital Citizens: An Analysis of Participation in Sociolinguistic Citizen Science

par

Philippa Payne

sous la direction du Professeur **François Bavaud**

et la direction du Professeure **Anita Auer**

Session de printemps 2025

Understanding Digital Citizens: An Analysis of Participation in Sociolinguistic Citizen Science

Philippa Payne

Contents

1	Acknowledgements	5
2	Abstract	6
3	List of Acronyms	7
4	Introduction	8
5	Theoretical Framework	12
5.1	Demographic Criteria	16
5.1.1	Student or staff member at the University of Lausanne	16
5.1.2	Gender identity	16
5.1.3	Highest level of education completed	18
5.1.4	Region of origin	21
5.1.5	Age	23
5.2	Digital Literacy	27
5.2.1	Years of Internet usage	27
5.2.2	Performance of activities on a computer, tablet or mobile phone .	29
5.2.3	Competency in using digital tools and services	32
5.2.4	Understanding of how algorithms are used on data online	35
5.3	Citizen Science	37

5.3.1	Contribution of data, time or skills to a research project	37
5.3.2	Organisations to which data, time or skills were contributed	39
5.4	Potential Engagement	41
5.4.1	Concerns when sharing data of online communications with research	41
5.4.2	Motivations to share data of online communications with research	44
5.4.3	Interest in using a mobile application to facilitate data analysis .	49
5.5	Existing Knowledge	51
5.5.1	Familiarity with terminology	51
6	Empirical Study	52
6.1	Methodology	53
6.2	Techniques	59
6.2.1	Data cleaning and preparation	59
6.2.2	Variable types	60
6.2.3	Summary statistics	61
6.2.4	Statistical approaches	61
6.2.5	Data visualisation	61
6.3	Results	62
6.3.1	Descriptive statistics	62
6.3.2	Digital use, competency and literacy	69
6.3.3	Categories of activities performed via digital technology	75
6.3.4	Familiarity with sociolinguistics and research	83
6.3.5	Data concerns and algorithmic literacy	84
6.3.6	Motivations for research engagement	88
6.3.7	Interest in a future mobile application	90
6.4	Evaluation	95
6.4.1	Methodological limitations	95
6.4.2	Survey limitations	96
6.4.3	Analytical limitations	97
6.4.4	Future recommendations	98

7	Conclusion	100
7.1	Who is the target audience?	100
7.2	What would help and hinder engagement?	101
7.3	What do people already know?	102
8	Bibliography	104
9	Annex	116
9.1	English survey	116
9.2	French survey	125

1 Acknowledgements

I would like to extend my immense gratitude to the people who have made the completion of this *mémoire* possible. Without the theoretical and methodological suggestions of Professor **Anita Auer** and the statistical and technical support of Professor **François Bavaud**, I would not have achieved such a thorough work. Furthermore, their moral support throughout this project has been exceptional.

It is also of paramount importance to thank Professor **Aris Xanthos**, who made this work possible in so many ways. Since my arrival at the University of Lausanne, he has offered me great kindness. This project would not have been set into motion without our prior discussions.

Sofia Boteva and **Mariana Pereira Alves** have equally offered immeasurable support, not only by contributing to the pilot study and publicising the survey, but by listening to me at challenging moments.

Lastly, to my other friends and family. I can truly state that this work would not have been completed without their belief in my ability during what has undoubtedly been the most testing period of my life.

2 Abstract

In a world where society is organised through digital technology, it should be no surprise that “digital citizenship” has become an inescapable condition for societal agency. With people becoming more conscious of attempts to exploit their data and more reluctant to opt in to data sharing, significant tensions with participation have arisen, particularly for institutions that are more transparent with their data usage and do not attempt to obfuscate their data collection efforts. As such, research institutions encounter challenges in encouraging participation in research projects that involve data collection. It is thus critical for the quality, size and reputation of a project to uphold ethical standards when attracting participants. However, there are many unknowns in terms of what helps and hinders participation in citizen social science. It is essential to understand ways in which participation might therefore be improved. In this study, I offer a framework for identifying the target audience for future CMC-oriented research projects. I identify what helps and hinders such engagement and provide insights to optimise engagement with future research projects, with an eye to the development of a mobile application.

Keywords: citizen social science, computer-mediated communication, data sharing, data collection, digital technology, digital citizenship, digital literacy, mobile application, research participation.

3 List of Acronyms

The following table offers an overview of the acronyms to which I will make reference in this study, for the sake of readability:

Acronym	Meaning
AI	Artificial Intelligence
ANOVA	Analysis of Variance
BCU	Bibliothèque cantonale et universitaire
CA	Cambridge Analytica
CMC	Computer-Mediated Communication
EIL	English as an International Language
EPFL	École polytechnique fédérale de Lausanne
FAIR	Findable, Accessible, Interoperable and Reusable
FLE	Français langue étrangère
GDPR	General Data Protection Regulation
HN	Humanités numériques
IPR	Intellectual Property Rights
IQR	Interquartile Range
PII	Personally Identifiable Information
PWA	Progressive Web Application
SES	Socioeconomic Status
SLI	Sciences du langage et de l'information
UK	United Kingdom
UNIL	University of Lausanne
US	United States

4 Introduction

In 2013, Edward Snowden leaked more than 1.7 million intelligence files to journalists, underscoring the extent to which people are being monitored in civil society (Bakir, 2015). Following on from this leak, the Cambridge Analytica (CA) scandal in 2018 once again revealed the dangers associated with “datafication” (Büchi et al., 2021; van Dijck, 2014). The vast system of surveillance that was uncovered demonstrated the lack of control people had at the hands of organisations collecting data *en masse*, irrespective of their knowledge or consent. Resembling something akin to the “panopticon” of Michel Foucault, from the concept of Jeremy Bentham, this system represented an attempt to maintain power through constant watching (Bentham, 1791; Foucault, 1975). Whilst this controversy eroded people’s trust in both institutions’ and companies’ handling of their data, it only served to encourage organisations to resort to subterfuge instead. As a result, the only way completely to avoid questionable data practices would mean not engaging in digital communication or participating in digital culture at all. Such a move would be tantamount to removing oneself from the 21st century (Anderson, 2015, p. 160). In a world where society is organised through digital technology, it should be no surprise that “digital citizenship” – engaging with digitised spaces – has become an inescapable condition for societal agency, regardless of one’s digital competence (Hintz et al., 2019).

These changing practices were at odds with the “participatory culture” that had been growing since the popularisation of the personal computer in the 1970s, the development of the World Wide Web in the 1990s, and the emergence of Web 2.0 in the 2000s (Jenkins et al., 2016). This so-called participatory culture provided an inclusive community in which people could express themselves, foster connections, and offer valuable contributions (Jenkins & Clinton, 2006). But, with people becoming more conscious of attempts to exploit their data and more reluctant to opt in to data sharing, significant tensions with participation arose, particularly for those institutions that were more transparent with their data usage and did not attempt to obfuscate their data collection efforts.

This post-Snowden, post-CA era goes some way to explain the challenges that research institutions may encounter in encouraging participation in research projects that

involve data collection. However, another element to consider is how actors can offer effective incentives for participation, to offset the risks of data sharing. For many platform-based organisations – such as Facebook, Instagram or TikTok – this “benefits-harms binary” is swiftly handled; the provision of services outweighs the potential risks (Hintz et al., 2019). For other organisations, designing appropriate mechanisms for garnering interest is more perplexing.

Citizen science, whereby volunteers outside academia contribute to or enact scientific research, has had variable success in encouraging participation and has produced a large body of research that considers how best to conceptualise projects in order to motivate engagement (Irwin, 1995; Kullenberg & Kasperowski, 2016). Whilst digital technology and an increasingly “datafied society” have provided unprecedented data collection opportunities, it is critical for the quality, size and reputation of a project to uphold ethical standards when attracting potential participants (Cavalier et al., 2020).

Although citizen science is dominated by the natural sciences, a burgeoning area of citizen science is in humanities and social science disciplines, already constituting 11.0% of European citizen science in 2018 (Hecker, Garbe, & Bonn, 2018; Vohland et al., 2021). Indeed, 4.0% of European projects were sociology based (Hecker, Garbe, & Bonn, 2018). Historically, individuals have long played a passive role as the “objects” of such research, but there has been a noticeable shift towards their playing a more active role as research “subjects” (Kullenberg & Kasperowski, 2016). Since much research into encouraging participation in citizen science has focused on the hard and environmental sciences, there are many unknowns in terms of what helps and hinders participation in citizen social science (Albert et al., 2021). Much of the success of a citizen science project requires the discovery of one’s niche or target audience, as well as improving an audience’s familiarity with its research goals. Even the term “citizen science” itself can be unfamiliar to many (Lewandowski et al., 2017). Alongside this knowledge barrier, financial, administrative and technical difficulties, as well as negative experiences and cultural differences, can stymie project success (Martek et al., 2022). These existing barriers make it all the more essential to understand ways in which participation might be improved. Many of the

barriers that prevent individuals from engaging in digital technology prevent them from engaging in citizen science.

In a Swiss context, one example of sociolinguistic research that successfully elicited data-based contributions from digital citizens is *What's Up, Switzerland?*, conducted under the auspices of the Swiss National Science Foundation (SNSF) (Kucharska, 2021). Between 2014 and 2018, this cross-university project collected more than 600 multilingual chats from the messaging application WhatsApp, in order to build a corpus through which digital communication could be studied (Ueberwasser & Stark, 2017). Building a corpus of linguistic data can greatly facilitate the study of language in authentic contexts (Birner, 2013). Another corpus-building project, *What's New, Switzerland?*, undertaken at the University of Lausanne (UNIL) from 2011 onwards, succeeded in collecting 70 French-language WhatsApp chats (Doudot, 2021). However, the project has experienced severe limitations in terms of data collection. Moreover, the over-representation of the 25 to 34 age group and the fact that a single chat accounted for one third of the corpus further limited its potential (Doudot, 2021). An additional project that merits recognition is the *World Internet Project* survey conducted in 2015. Instead of appealing for data contributions, this project sought to describe different elements of Internet usage in Switzerland and other countries via telephone interview, gathering responses from over 1,000 individuals (Büchi et al., 2021).

From a citizen science perspective, it is critical to find a balance between encouraging the participation of diverse groups and seeking out one's target audience (Paleco et al., 2021). A further challenge to understanding which demographic groups are not represented in these corpora stems from the option of anonymising one's contributions (Ueberwasser & Stark, 2017). Although this practice gives participants control of their personal identities, it stymies attempts to understand limiting factors.

In a primarily "descriptive" discipline such as linguistics, it is all the more crucial to comprehend the significance of diversity and representativeness in gathering data for analysis (Birner, 2013). As forwarded by John J. Gumperz, "ethnographic" observations are strongly interlinked with the concept of linguistic diversity, since our language choices

are fundamentally shaped by our sociocultural contexts (Gumperz, 1982; Gumperz & Cook-Gumperz, 1982). As more diverse contexts of communicating materialise with the expansion of Computer-Mediated Communication (CMC) – whereby humans communicate by means of a computer – a concerted effort must be made to understand the new voices and communities that occupy digital spaces (December, 1997). Regardless of whether the features analysed are part of language or “paralinguistic” (verbal or non-verbal) in nature, it is essential to include all participants in such descriptive research (Luangrath et al., 2017; Wharton, 2003).

In this study, I will bring together understandings of digital citizenship, theories of citizen science and techniques of statistical analysis in order to ascertain how participation in sociolinguistic citizen science might best be encouraged. Firstly, I explicate the theoretical reasoning behind the structuring of a survey, responding to the following research questions:

1. Who is the target audience?
2. What would help and hinder engagement?
3. What do people already know?

These questions will be answered with an eye to the future development of a mobile application to facilitate engagement and automatise data collection (Payne, 2023). Secondly, I outline the methodological process that was followed to achieve results in this study. Thirdly, I analyse and evaluate the results of this survey using the statistical programming language R. I hope that the results of this study offer a first step towards the optimisation of research projects, at the UNIL, in Switzerland, and beyond, assisting institutions in overcoming the existing barriers to data collection and, by extension, restoring people’s faith in data-based research.

5 Theoretical Framework

I consider this study as the first step towards the conceptualisation of a social citizen science framework for the collection of text-based CMC data in a sociolinguistic research context, at the UNIL and beyond.

Erving Goffman states that “social frameworks [...] provide background understanding for events that incorporate the will, aim and controlling effort of an intelligence, a live agency, the chief one being the human being” (Goffman, 1974, p. 22). As such, Goffman’s conceptualisation can be employed both to describe CMC and to shape an enquiry into human engagement in CMC research (Carey, 1980).

On the one hand, CMC, encompassing networked exchanges through “email, computer conferencing, and chat systems” (Tidwell & Walther, 2002, p. 318), as opposed to “face-to-face” communication (Kiesler et al., 1984, p. 110), offers a new frame for organising social experiences (Goffman, 1974). With the previous frame of social interaction broken, considerable adaptations had to be made (Goffman, 1974). In particular, these predominantly text-based communications necessitated the development of linguistic features to “help to regulate interaction between speakers” (Carey, 1980, p. 67). These paralinguistic adaptations, falling “outside the boundaries of [...] analysis”, forced a shift in linguistic analysis to account for an absence in non-verbal cues, such as gestures and facial expressions (Carey, 1980, p. 67). Similarly, interjections, considered by Goffman himself as “non-words” that “can’t quite be called part of a language” (Goffman, 1981, p. 115) and long “neglected” (Ameka, 1992, p. 101) by linguists as akin to gestures, became an engrained part of written communication for the first time (Wharton, 2003).

On the other hand, understanding people’s will for participating in or abstaining from CMC-related research projects also entails the construction of a framework.

Much work has been done, centred on the United States (US), to describe how, how much and how effectively people engage in computer-mediated environments (Kim & Hargittai, 2021; Kimm & Boase, 2021; Redmiles & Buntain, 2021). These studies have focused predominantly on issues pertaining to use, impact and trust in specific contexts, such as mobile devices or the Internet (Kim & Hargittai, 2021; Kimm & Boase, 2021;

Redmiles & Buntain, 2021).

Connected to these issues is a body of literature dedicated to digital citizenship and reflecting on what groups can be deemed digital citizens. Digital citizens are commonly considered to be “those who use the Internet regularly and effectively”, referring currently to accessibility to and competency with digital technology (Mossberger et al., 2008, p. 1). Moreover, with the emergence of a datafied society and increased access to digital products, digital citizenship has been reimagined, rather, as “the informed and knowledgeable use of digital infrastructures” and “the competent creation of digital acts” (Hintz et al., 2019, p. 150). Thus, the bar for digital citizenship has been set high; it entails education, creation and action. However, it is more appropriate to consider digital citizenship as referring to “our roles, positions and activities in a society that is organised through digital technologies” (Hintz et al., 2019, p. 144). Since digital technology has become an inescapable and universal frame in which all citizens must act regardless of competency, I consider a digital citizen to be any individual who engages with digital platforms or tools to any extent.

Instead of dividing individuals into those who are digital citizens and those who are not, it is more valuable to distinguish between “digital natives” and “digital immigrants” (Prensky, 2001, p. 1). This distinction originated from the idea that younger generations are “native speakers of the digital language of computers, video games and the Internet”, whilst those “who were not born into the digital world” have had to “adapt to their environment” as well as they are able (Prensky, 2001, pp. 1–2). In an educational context, digital natives “learn differently compared with past generations”, whereas digital immigrants are slower to adopt new practices.

Even though this distinction is an improvement on an exclusive interpretation of digital citizenship, it has become evident that distinguishing between pre-1980 and post-1980 generations is unsatisfactory in accounting for digital skills (Coşkunserçe & Aydoğdu, 2022). More precisely, differences in age have not reliably predicted digital literacy, “a person’s ability to perform tasks effectively in a digital environment” (Jones-Kavalier & Flannigan, 2006, p. 9). Thus, I will continue to use the terms digital native and

digital immigrant to refer to the generations that were born into the Internet era and those that were not, respectively. These terms do not presuppose intrinsic competence or incompetence.

To identify individuals with digital competence, I will employ the term “digitally literate” to remove any presuppositions of generational relevance. Furthermore, I will employ the more respectful term “digital learner” to reflect individuals with limited technological competencies, regardless of demographic criteria (Bullen & Morgan, 2011; Bullen et al., 2011; Gallardo-Echenique et al., 2015). This gap in digital skills represents the “second-level digital divide” after accessibility (Hargittai, 2002, p. 1). A further category is what I conceptualise as the “digitally naive”; those digital citizens who are uninformed as to the processes of digital tools rather than lacking in their ability to use them.

Aside from attempts to provide a descriptive framework to better understand how people engage with digital technology, much work has been conducted to model the different kinds and phases of citizen science projects. Mostly centred on the natural sciences, such frameworks strive to provide guidance for project design (Phillips et al., 2018; Toft et al., 2017). They outline the key roles, levels of involvement and ethical considerations of citizen science. In particular, the internationally developed *Ten Principles of Citizen Science* highlights the importance of active involvement, knowledge generation, tangible outcomes, collaboration, agency, feedback, openness, acknowledgement and ethical concerns to the success of a research project (Robinson et al., 2018). Although finding a straightforward definition for citizen science is challenging, growing subject- and country-specific literature is working on defining what it means within disparate contexts (Haklay et al., 2021). For the purposes of this study, I consider citizen science in its broadest sense as “public participation in scientific research” (Phillips et al., 2018, p. 1).

Given the nature of the social sciences, in which “there has been a long tradition of engaging closely with citizens as objects of study”, a citizen science approach involving “active participation or contribution” resituates them as “research subjects” (Kullenberg & Kasperowski, 2016, p. 13). The potential for further engagement in sociological research carries significant impact due to this sense of reflexivity. In this case, evoking something

akin to the the *ouroboros*, digital citizens are empowered to reflect on how they could be engaged in research about their own practices, reimagining the hierarchy separating academic from ordinary citizen.

Despite these two existing bodies of research, there has been limited academic output seeking to determine how the nature of an agent determines their willingness to participate. Therefore, this study seeks to understand the nature, motivations and concerns of digital citizens in the context of contributing to such reflexive research projects. I hope that this study offers valuable insights into digital citizens, digital literacy and research participation that can be utilised to optimise future initiatives, particularly within the field of sociolinguistics.

The following sections provide theoretical justifications and discussions for the survey questions that form an integral part of this study. It is critical to then test how accurately these theories translate into a practical context and whether the results of previous studies are replicable. Firstly, I consider the demographic criteria to be captured. Secondly, I consider a reliable way to measure digital literacy. Thirdly, I consider how people's existing familiarity with the research process might alter their experience. Fourthly, I consider the aspects which might help or hinder potential engagement. Lastly, I consider whether existing sociolinguistic knowledge determines engagement.

5.1 Demographic Criteria

5.1.1 Student or staff member at the University of Lausanne

Including markers of socioeconomic status (SES) in this study is a critical component of understanding how individuals act and make decisions within different environments (Park, 2021). In fact, “social demographics [...] serve as a collective body of proxies that represent socialisation” (Park, 2021, p. 287). A “proxy” refers to a variable that can substitute another unmeasurable variable (Upton & Cook, 2014). In this sense, proxy variables offer an insight into how individuals engage in computer-mediated participatory cultures. Moreover, Goffman recognises how one’s “social establishment” is determined by such “fixed barriers to perception” (Goffman, 1956, p. 238).

Whilst demographic criteria are essential for this study, I view this first question as a more practical way of ascertaining whether a respondent is affiliated with the UNIL or not, since I will be appealing to all respondents to diffuse the survey within their own networks. As of the 2023 to 2024 academic year, the university community was constituted of 16,950 students, 3,929 full-time staff and 460 incoming exchange students (UNIL, 2023, 2024a, 2024b). These statistics provide a useful basis on which to understand the composition of the university, which will act as the principal population for the study. Although this categorisation is practical in nature, there is potential to explore the relationship between affiliation with the UNIL and behaviour in the subsequent analyses.

Further SES questions – concerning gender, education, region, and age – will provide more fundamental data to analyse their relationship with behavioural data. In the following breakdown, I review the existing literature pertaining to these SES markers.

5.1.2 Gender identity

Previous studies have demonstrated, in both a US and international context, that gender can play a decisive role in determining how someone uses the Internet, as well as how they assess their digital skills (Kim & Hargittai, 2021). In the 2000s, much research was conducted into the “genderisation of computing”, particularly from the perspective of the United Kingdom and the US (Durnell & Haag, 2002, p. 521). This phenomenon

developed from differentiated educational environments in terms of gender expectations. In fact, research suggested that “females engage with technology less frequently than males” and “generally have less positive attitudes and greater feelings of anxiety towards technology” (Rees & Noyes, 2007, p. 483).

A sense of “technophobia” (Rees & Noyes, 2007, p. 482), however, did not consistently extend to all technologies or, for that matter, all studies, especially over time (Kim & Hargittai, 2021). Whilst results “indicate that there is no longer a gender gap” (Ono & Zavodny, 2003, p. 120) in terms of access to digital technology like the Internet, “the sexes may favour different uses of each technology” (Rees & Noyes, 2007, p. 482). Moreover, although gender no longer appeared to influence competency with key digital activities or the range of features used, it became clear that “gendered perceptions of competency may diverge from actual skill levels” (Hargittai & Shafer, 2006, p. 5). In other words, males often over-exaggerated their digital competencies, whereas females tended to perceive their competency as lower than the skills they truly had. This dimension reflects the stereotyped differences in “agency” between the sexes (Bakan, 1966, p. 1). This separation has become all the more distinct on the issue of privacy, especially among younger generations, as the “confidence gap between men and women magnifies among the younger users, while women’s confidence remains low regardless of age” (Park, 2015, p. 255). Males of all ages are, therefore, more confident in their ability to take privacy measures than their female counterparts, and this gender gap has only widened over time.

As David Bakan outlined, “agency manifests itself in self-protection, self-assertion, and self-expansion” (Bakan, 1966, pp. 14–15). In this sense, the male expectation within a given environment is to display a high level of self-efficacy, explaining this disparity in perceived digital competence (Durndell & Haag, 2002; Eagly, 1987). Opposing Bakan’s notion of “agency” is that of “communion”, the “participation of the individual in some larger organism” (Bakan, 1966, pp. 14–15). Given that social stereotypes expect a higher level of communion from the female populous, it might be expected that women display a higher interest in participation and, by extension, in citizen science. However, depending on the project discipline, citizen science often attracts either a higher proportion of male

participants or gender parity (Paleco et al., 2021). Even so, these two modalities mark a critical tension in this study, which seeks in part to understand the relationship between digital competency (agency) and research participation (communion).

This tension manifests itself in differences in terms of digital media use styles (Kim & Hargittai, 2021). Among the male demographic, it is more common to use technology “as a means to an end”, whereas it is more common among the female demographic “as an end in itself” (Kelan, 2007, p. 9). Thus, the masculine style is likened to a “hammer” (a tool), whereas the feminine style is likened to a “harpichord” (a mode of expression) (Kirkup, 1992, p. 269). As such, males and females are more likely to employ technology for technical and social purposes, respectively (Lai & Katz, 2012). This dynamic parallels the concepts of agency and communion (Bakan, 1966).

It is essential to note that non-binary gender identities will and should also be represented within the data set. Historically, those identifying as not male or female were integral to the early development of participatory online communities (Jenkins et al., 2016). Understanding more about this demographic would be enlightening and bring greater acknowledgement to non-binary individuals (Correa et al., 2021). However, it is not always possible to draw reliable conclusions about this group, especially if the proportion of non-binary respondents is low. Therefore, some studies have chosen to absorb the non-binary category within a binary variable (i.e., by including these individuals within the male, or female, group) (Park, 2021). Others have chosen to maintain a male-female binary, omitting data in order to not make spurious judgments about this demographic (Correa et al., 2021). I will offer two alternative responses here: “Other” (in which case the respondent will be prompted to specify in their own words) and “Prefer not to say” for those who do not wish to share this information.

5.1.3 Highest level of education completed

Social demographics do not solely act as proxies for socialisation (Park, 2021). An individual’s level of education can also be used to measure SES itself (Kim & Hargittai, 2021). In fact, research has proposed that those with higher SES “use the Internet more,

show better skills and get higher benefits from it”, primarily by using the Internet in an instrumental way, for personal development and information retrieval (Gui & Gerosa, 2021, p. 133).

Since reliable access to digital technology does not pose such a significant problem in developed countries, economic status is no longer as relevant to affording digital products (Hintz et al., 2019; Mossberger et al., 2008). In research relating to young adults – the generation considered “digital natives” – it is parental education that is used as a measure of SES (Kim & Hargittai, 2021). Given that this study seeks to capture a range of age groups, rather than focusing on young people’s digital usage in particular, I have chosen to establish the respondent’s own level of education instead.

Prior research has demonstrated that “those with higher education [and] higher income [...] tended to display higher confidence in dealing with [...] personal data” (Park, 2021, p. 290). In turn, “confidence had a positive effect on the level of digital participation” (Park, 2021, p. 290). The relationship between education, income, confidence and participation highlights the importance of considering digital inequalities and how they inhibit engagement in citizen science.

Not only are more educated individuals more confident in using digital technology, but research suggests that they are also more competent users, both with regards to specific skills and overall (van Deursen & van Dijk, 2014; van Deursen et al., 2014). From another perspective, research has established that less educated individuals spend more time surfing the Internet and scrolling their smartphones than more educated individuals (Gui & Gerosa, 2021; van Deursen & van Dijk, 2014). They spend a higher amount of time engaging in gaming and social interaction in comparison, marking once again the boundary between instrumental and expressive uses of technology, between agency and communion (Bakan, 1966; Kim & Hargittai, 2021). In this sense, the concept of a digital divide no longer relates to the “classification of haves and have nots”, but rather to the nature of technology use (van Deursen & van Dijk, 2014, p. 520).

In terms of benefits, this discrepancy represents a “third-level digital divide”, whereby “internet use and online activities will confer greater benefits to internet users in life realms

where the user already has significant resources” (van Deursen & Helsper, 2015, p. 31).

In other words, greater socioeconomic benefits offline result in greater benefits online.

The issue of under-representation of certain groups in citizen science has largely been broached in a US-centric context. In terms of age, gender, race and affluence, “participation in citizen science does not reflect the demographics of the US” (Pandya, 2012, p. 314). For instance, younger, less educated, less affluent individuals participate less than those who are older, more educated and affluent, although there is some variation in terms of subject matter (Paleco et al., 2021; Trumbull & Bonney, 2000). Indeed, studies have often “revealed that participants tended to be older and better educated than the general population, and that they also were interested in science” (Trumbull & Bonney, 2000, p. 268).

The understanding that having an existing interest in science determines a potential participant’s involvement reinforces the importance of measuring educational level, beyond its use as a proxy for SES. Individuals who already hold “positive beliefs and good knowledge about science” are more likely to participate in research projects (Trumbull & Bonney, 2000, p. 268). One of the key factors as to why those from less educated backgrounds do not engage in academic research (and academic environments in general) is because of their perceptions (Gloria et al., 2001). When individuals from outside academia become involved in it, many find that their cultural values are at odds with those upheld within the environment. As such, many less educated people experience a lack of “cultural congruity”; a sense of isolation (Gloria et al., 2001, p. 545). It stands, then, that individuals with a higher educational level may be expected to have greater interest in engaging with a future citizen science project. Individuals without a prior academic experience may be hesitant in engaging.

Whilst knowledge of science is one aspect that contributes to participation, knowledge of citizen science itself is another aspect. Indeed, “familiarity with citizen science and interest in [it] are likely to affect recruitment and retention” (Lewandowski et al., 2017, p. 4). Furthermore, previous research has suggested a link between familiarity with and confidence in citizen science (Lewandowski et al., 2017). Thus, existing knowledge of

both science and citizen science can contribute to public perceptions of a given project.

Lastly, with the understanding that familiarity with science in general and citizen science in particular are both of importance for encouraging interest in a project, it must be noted that familiarity with the subject of research may relate to willingness to participate. Although this question does not account for existing expertise in the field of research, the question of familiarity with sociolinguistic terminology will be considered later in the survey.

5.1.4 Region of origin

Many studies into digital divides and under-representation within citizen science consider the question of racial inequalities (Kim & Hargittai, 2021; Paleco et al., 2021; Pandya, 2012; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021). As I expect to garner responses from within the UNIL community and those connected to it, I have adapted this question to suit the context of my study. At the UNIL, the Swiss nationality accounts for 74.17% of the student community and 56.68% of the staff, whilst Europeans account for 19.97% of the student community (UNIL, 2023, 2024a). A high percentage of Swiss nationals constitutes the UNIL community, followed by which Europeans are well represented.

Thus, I have chosen a granularity of response options that reflects the makeup of the academic community whilst preserving the opportunity to analyse regional differences. I have chosen to break down the response options into (1) individuals of Swiss origin, organised by canton, (2) non-Swiss Europeans, and (3) non-Swiss non-Europeans. This breakdown effectively segments respondents to this survey enough to explore the role that regional differences play in digital literacy and research participation, without losing focus on the university population's makeup. Ideally, this categorisation allows me to study both regional differences within Switzerland and to compare Switzerland as a whole to the global picture.

Results from the *European Citizen Science Survey* highlighted that Switzerland was the fifth highest producer of citizen science projects within Europe as of 2017. Switzerland found itself well above average in its contributions to citizen science projects, although

it lagged behind Germany, the UK, Austria and Spain. From 2014 to 2019, citizen science within a European context has been growing rapidly, with an increasing number of proposals every year (Warin & Delaney, 2020). However, an additional challenge for European-based citizen science projects is in communication, since “unlike the citizen science landscape in the United States or Australia”, they struggle with “interoperability” due to their being carried out in the native language of a given country (Hecker, Garbe, & Bonn, 2018, pp. 198–199). To remediate this issue of interoperability in my own study, I have chosen to prepare this survey in two language formats: French and English. With 90.54% of students emanating from French-speaking countries and a minimum B2 French language requirement as per the Common European Framework of Reference for Languages (CEFR), it is to be expected that the vast majority of students will engage with the French-language version of the survey (UNIL, 2024a). However, the UNIL community is constituted of a high proportion of incoming exchange students (69.57%) and staff of international, non-francophone backgrounds, so it is appropriate to offer this survey in what is considered an international language (UNIL, 2023, 2024b). Indeed, English is a “language of international, and therefore intercultural, communication”, arguably due to the fact that digital technology is “dominated by English-medium programming” (Sharifian, 2009, p. 2). English as an International Language (EIL) has enabled the development of participatory cultures on the Internet, by serving “as a contact language with people from throughout the world” (Sharifian, 2009, p. 73).

On the one hand, this relatively new dominance of anglophone communication is another reason why studying the development of CMC is of great purport to researchers in sociolinguistics. On the other hand, this hegemony can perpetuate communicative inequalities for those individuals who do not speak this shared language. Arguably, this paradigm has necessitated the development of non-standard, universal forms of communication outside the normal bounds of language. Due to the challenges of text-based communication, “without physical and social cues or immediate feedback”, it can be more challenging, not only to reach a “consensus” in terms of a decision but to “create a solidarity through developing interpretive consensus” (Baym, 1995; Kiesler & Sproull, 1992;

Wellman & Gulia, 1999). This challenge leads communities to engage in “performances” – such as the use of emoticons, emojis, and paralanguage – as “artful use[s] of language that stand apart” and provide “a frame that invites critical reflection on communicative processes” (Bauman & Briggs, 1990; Hine, 2000, p. 60).

In recognition of how linguistic limitations can, therefore, interfere with operability, I offer a French- and an English-language version of the survey. The dual-language provision of this survey will reduce the cognitive load in responding, for those unfamiliar with either French or English, whilst making it more accessible to the community. I hope that these considerations will increase the number and diversity of respondents. It would equally be valuable to keep in mind how such potential communication difficulties between regions might affect participation, in both digital environments and citizen science.

5.1.5 Age

As with gender, previous studies have given inconsistent results as to the relationship between age and aspects of digital technology (Correa et al., 2021). The most well-evidenced relationship is that between age and Internet usage, demonstrating the “strong negative effect of age on Internet use, be it time online, frequency, or variety of usage”, with young people engaging with “all usage types” more frequently (Büchi & Latzer, 2016, p. 8). It has been found that younger people, those known as digital natives, engage in activities such as social interaction and entertainment more readily and frequently than older individuals (Büchi & Latzer, 2016). A respondent’s age strongly relates to “different uses of the Web” (Redmiles & Buntain, 2021, p. 314). Although the “gender gap” between male and female Internet users appears to be at best narrowing and at worst inconclusive, age still maintains its significance when it comes to specific use cases, such as social media (Bimber, 2000; Redmiles & Buntain, 2021). Younger people tend to use social media more, and more frequently, than older people (Redmiles & Buntain, 2021). This result is likely to be related to the fact that the “vast majority of young adults [...] own a smartphone” (Kim & Hargittai, 2021, p. 114). However, recent studies into the type of Internet access used (“mobile-only, PC-only, and hybrid users”) has not demonstrated a

significant difference relative to age (Correa et al., 2021, p. 67). Age, therefore, has a significant effect on digital participation (Park, 2021). In fact, “the amount, frequency and diversity of Internet uses are relevant aspects of the digital-inclusion process”, thus, usage of digital technology reflects one’s inclusion in society (Correa et al., 2021, p. 64).

This increased participation may reflect a sense of confidence among young people in engaging with digital technology (Park, 2021). This notion brings the discussion back to Bakan’s concept of “agency”, questioning whether a sense of self-assertiveness is indeed at odds with a desire to participate (Bakan, 1966). One theory that might evidence this point is that of Albert Bandura, who states that “self-efficacy”, or an individual’s personal effectiveness in completing given cognitive tasks, can differ from their self-assessment of their abilities (Bandura, 1977, p. 191). Bandura suggests that “the stronger the perceived self-efficacy, the more active the efforts”, a cyclic process in which participation improves perceived self-efficacy, which itself improves actual self-efficacy (Bandura, 1977, p. 194). In short, asserting one’s own ability psychologically prepares oneself to undertake a task well; undertaking a task well improves one’s future undertaking of that task. Self-assertion, then, may make one more willing to participate in a given environment.

As older generations who preceded the mainstream Internet tend to use digital technologies less than so-called digital natives, it is no surprise that this group’s self-assessed level of confidence in using digital technologies and using them well is also lower (Park, 2021; Reisdorf & Blank, 2021). This point is particularly pertinent in the context of perceptions of privacy and trust in digital technology (Büchi et al., 2021; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021). Firstly, age is “significantly and negatively related to digital privacy-protection skills and to the conduct of digital activities” (Büchi et al., 2021, p. 299). It is also significant when it comes to the understanding of algorithms, a component of digital literacy which can be termed “algorithmic literacy” (Reisdorf & Blank, 2021, p. 341). Whilst younger people use digital technologies more frequently and in more diverse ways, it seems that experience has taught them to protect themselves in the face of privacy threats. On the contrary, older people are more “vulnerable” and may engage in behaviours that “compromise privacy” (Büchi et al., 2021, p. 299). Older

demographics also tend to have a higher level of concern and a lower sense of control in the context of disclosing personal information via digital technology (Büchi et al., 2021; Redmiles & Buntain, 2021).

Oddly, studies have also demonstrated that younger people have a higher level of “trust” in the way that their personal information is handled, reflecting, perhaps, a certain level of digital naivety, especially when it comes to sharing information via social media (Redmiles & Buntain, 2021, p. 311). In many of these studies, self-assessment of privacy-related skills is employed as a proxy for the measurement of real competency (Büchi et al., 2021; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021). Thus, it is critical to recognise that self-assessment can vary significantly from the reality; young people may well overestimate their control in online environments.

Self-efficacy is integral to engaging individuals in citizen science projects. In a citizen science context, self-efficacy can also refer to “the extent to which a learner has confidence in his or her ability to participate in a science or environmental activity” (Phillips et al., 2018, p. 8). It has already been mentioned that knowledge of the subject of research, knowledge of science as a whole, and knowledge of citizen science itself all contribute to enhanced and continued participation (Lewandowski et al., 2017; Paleco et al., 2021; Trumbull & Bonney, 2000). Studies in citizen science have demonstrated that one of the key outcomes of collaborative research projects is gains in knowledge and perceived self-efficacy, especially among less educated participants. Citizen science that succeed in increasing participants’ perceptions of self-efficacy tend to retain participants for longer and attract new participants more easily (Tillotson-Chavez & Weber, 2024).

For instance, pre- and post-project surveys of knowledge gains in citizen science projects have “exhibited positive and significant rank increase between surveys”, especially among groups who do not have scientific backgrounds (Tillotson-Chavez & Weber, 2024, p. 10). Moreover, if a citizen science project serves to empower individuals who feel themselves othered from academia, it has a key role to play in empowering individuals in their interactions with and via digital technology. Thus, a successful citizen science project design would not only study participants’ use of computer-mediated environments

but improve their relationship with them.

In addendum, considering whether there is a relationship between perceived competence in using digital technology and willingness to participate in academic research would give an insight into whether individuals experience a similar or different sense of self-efficacy between these two domains. However, it must be noted that “academic performance” can serve as “a better predictor of economic and social well-being than the mere quantity of years spent in lower or higher education” (Gui & Gerosa, 2021, p. 132). Even so, I have chosen to retain the more traditional indicator, since lack of knowledge or willingness to self-disclose past or present academic achievement may introduce resistance into response giving.

5.2 Digital Literacy

5.2.1 Years of Internet usage

Before taking up the discussion regarding years of Internet usage, it is necessary to contextualise the concept of digital literacy within the realm of citizen science. The development of the Internet and digital technology have “broadened the range of tasks that amateurs can tackle” (Grey, 2011, p. 41). This diversification has not only meant the provision of a plethora of tools and environments, but has also opened the door for the expansion of citizen science. Amateur science projects, undertaken outside an academic context, have proliferated, and previously distributed communities have united on online platforms. The proliferation of citizen science, then, is another effect of the Internet’s fostering of participatory cultures, those cultures “with relatively low barriers to [...] civic engagement”, where “members believe their contributions matter” (Jenkins & Clinton, 2006, p. 3).

For people wishing to participate in this new frame of citizen science, effective participation was not only facilitated by some level of familiarity with scientific concepts and the scientific process – “scientific literacy” – but familiarity with digital technology (Trumbull & Bonney, 2000, p. 266). In such circumstances, “digital literacy was one of the skills necessary for participating in citizen science projects online” (Aristeidou & Herodotu, 2020, p. 6). An increased level of participation in an online project would mean greater exposure to its “software and hardware” and, by extension, increase one’s digital competencies (Aristeidou & Herodotu, 2020, p. 6). In fact, research also demonstrates that “participation in the social components” of an online citizen science project also improves gains in scientific literacy (Jennett et al., 2016, p. 5). If someone has the digital competency to participate more actively, it can result in further knowledge gains. Again, it appears that Bakan’s communion has the capacity of reaffirming agency and independence (Bakan, 1966).

However, as the complexity of the performable tasks increased and ethical issues began to arise, the level of pre-existing competency required for effective and informed participation also increased. For instance, the question of data sharing entailed the

creation of consent processes, requiring “a high level of information literacy” to understand (Eleta et al., 2019, p. 5). As suggested in research into algorithmic literacy, the ability to understand information processes and make judgments about them is strongly related to the current paradigm of digital literacy (Reisdorf & Blank, 2021). Thus, participation in citizen science projects is improved by scientific literacy, necessitates a certain level of digital literacy, and also has the capacity to improve both scientific and digital literacy (Aristeidou & Herodotu, 2020).

In previous studies, years of Internet usage has been measured in different ways (Correa et al., 2021; Kim & Hargittai, 2021). One possibility is to ask at what “stage in their academic career [...] the respondent first became an Internet user”, treating the question as an event-based measurement (Kim & Hargittai, 2021, p. 119). Another possibility is to ask more generally about the “first time they used the Web” (Correa et al., 2021, p. 66). I have chosen a more duration-based framing of the question, referring to the length of time in years that a respondent has used. Although this question is self-reported and may therefore lack a degree of reliability, it would be no different if the question were asked in an event-based manner. In fact, it might be argued that first use may find itself as an outlier that significantly precedes general use. A self-reported estimate of experience with the Internet is therefore as, if not more, appropriate as asking about first use.

Moreover, I ask this question as well as the age of participants to account for any divides in terms of accessibility to the Internet, which will otherwise not be focused on in the scope of this study, since recent research has demonstrated this question to be moot in the context of developed countries, in which this survey will be undertaken (van Deursen & Helsper, 2015). Indeed, “internet access is near-universal” (van Deursen & Helsper, 2015, p. 31). Alongside the demographic question of age, this question will offer an insight into the group considered digital learners and those considered digital natives.

5.2.2 Performance of activities on a computer, tablet or mobile phone

Essentially, many of the variables measured as part of the survey are proxies, since measuring the actual indicator would be too time consuming and complex for the scope of this project. Ideally, there are more foolproof ways to measure SES and literacy. For instance, more complex calculations and estimations of salary, parental salary or parental education can be integrated into a survey to understand a respondent's SES (Gui & Gerosa, 2021; Kim & Hargittai, 2021). Observational studies may create a series of tasks to complete to demonstrate digital literacy, or a knowledge-based assessment could be made to ascertain someone's pre-existing competencies (Kim & Hargittai, 2021). However, such tasks would require extended participation and would increase the obligation of involvement, potentially reducing engagement. I considered it critical to remove as many barriers to participation as possible in order to garner as many responses as possible, especially given that encouraging adequate survey participation can in any case be challenging.

Taking impetus from existing studies into the use of digital technologies, such as the Internet or mobile devices, the survey was constructed in such a way as to calculate a score for data usage in general. A Likert scale was constructed in order to ascertain the frequency with which a respondent performed a given task. A higher diversity and frequency of a set of tasks thus corresponds to a higher level of digital literacy. Previous studies have suggested a plethora of activities that might be included in such an analysis; however, it was critical to update and reshape these lists to reflect both the digital zeitgeist and the diversification of the kinds of tasks one can perform using digital technology. Some studies chose to group the activities in a pre- or post-survey context into categories. For example, Teresa Correa et al. measured 16 different activities within five categories: "social media use", "recreation", "email exchange", "information seeking", and "e-banking and e-commerce" (Correa et al., 2021, pp. 66–67). The activities included were:

- "use social media
- share content on social media
- online videogaming

- download and/or watch and listen to music, movies, games
- email exchange
- information seeking for goods and services
- information seeking for work opportunities
- information seeking for professional life
- information seeking for education
- information seeking for health
- download and fill out forms
- bank transactions
- buy or pay online
- contact clients/suppliers online
- apply for jobs online
- job/professional online training courses” (Correa et al., 2021, pp. 66–67).

Another example, from Bianca C. Reisdorf and Grant Blank, measured 12 activities:

- “buy a product online
- get information about local events
- order groceries
- send email to a list
- make or receive phone calls
- watch TV
- use email

- check a fact
- use instant messaging
- look for news
- investigate topics of personal interest
- watch videos or movies” (Reisdorf & Blank, 2021, p. 346).

Information retrieval, or search skills, were measured via a different variable (Reisdorf & Blank, 2021).

A final example comes from Alexander J. A. M. van Deursen and Ellen J. Helsper, who divided digital uses into four categories: economic, social, educational, and institutional (van Deursen & Helsper, 2015, p. 37). These categories were comprised of 12 activities:

- “trading goods
- booking holidays
- buying products
- job searching
- meeting people
- social interaction
- online dating
- searching educational information
- political participation
- online voting
- contacting the government
- searching medical information” (van Deursen & Helsper, 2015, p. 37).

In this study, I have synthesised, altered and extended the lists of activities included in the survey with the objective of comprehensiveness. As many previous studies have focused on a large but partial area of digital technology, such as the kinds of tasks one can perform on a browser, or the kinds of task one can perform on a mobile device, I have tried to provide a more inclusive list (Correa et al., 2021; Gui & Gerosa, 2021; Kim & Hargittai, 2021; Kimm & Boase, 2021; Redmiles & Buntain, 2021). This list of 30 items may not be exhaustive but includes key activities irrespective of access type and whether the activity can be conducted online or offline. However, this divide is becoming more obscure given almost-constant access to the Web in developed countries, standards for compatibility across multiple devices, and the development of Progressive Web Applications (PWAs) – “websites that you can install on your device” which run both online and offline (Boamah, 2024; Hilchenbach, 2023).

With the proliferation of activities one can perform with digital tools and platforms, I thought it critical to capture both the instrumental and expressive tasks one can perform (Kirkup, 1992; Lai & Katz, 2012; van Deursen & van Dijk, 2014). A comparable divide could be introduced between agency-based and communion-based activities (Bakan, 1966). These tasks will later be divided into categories that represent different use cases, use styles and levels of participation, in order to explore the different personalities of digital citizens. These personalities will offer an insight into the relationship between different profiles of digital citizen and willingness to engage in research projects. Categorisation might allow the evaluation of theories pertaining to societally embedded profiles of use.

5.2.3 Competency in using digital tools and services

Practical studies of an individual’s competency are time intensive and thus fall outside the scope of this study (Kim & Hargittai, 2021). An alternative to such skills assessments, other “survey measures of people’s online skills [are] derived from measures about the actual online skills of users assessed through performance tests” (Hargittai, 2005, p. 372). Eszter Hargittai has carried out extensive research into Internet skills in order to propose an aggregated measure based on 27 “Internet-related terms” (Hargittai, 2005, p. 372),

whereby respondents are asked to suggest their “level of understanding” with them, via a Likert scale (Hargittai, 2009; Kim & Hargittai, 2021, p. 119). This measure of online skills has been demonstrated to reflect accurately the actual skills of respondents (Hargittai, 2005, 2009).

It has already been discussed how people’s actual competency can differ from their perceived competency, with research demonstrating that demographic differences in gender, age, income and education influencing one’s level of confidence with digital technology (Park, 2015, 2021; van Deursen & van Dijk, 2014; van Deursen et al., 2014). Indeed, “perceived skills and actual skills are not always the same thing and vary by gender, with women perceiving their skills lower than men” (Reisdorf & Blank, 2021, p. 353). Confidence in using digital technology is often linked to trust in the digital technology itself, especially when considering the question of data privacy (Büchi et al., 2021; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021). Interestingly, whilst “men tended to have higher technical privacy skills and have greater confidence in their own privacy protection behaviour” (Büchi et al., 2021, p. 300), they “are less likely to trust social-media providers to protect their personal data” (Park, 2015; Redmiles & Buntain, 2021, p. 318). Thus, whilst the ideas of confidence and trust are conceptually linked, they may not share a positive relationship; confidence in one’s competency does not entail trust in the environment in which you have that competency. In fact, increased confidence may instead be linked to decreased trust. Arguably, differences in the use of digital technologies could be related to differences in confidence and trust, if men do indeed have a propensity towards agency (independence) and women towards communion (participation).

Aside from gender, “low-income people are less likely to have confidence in and use privacy settings” and, hence, “are especially vulnerable to discriminatory uses of big data by employers” (Madden et al., 2017, p. 82). Not only do they not take the appropriate measures or demonstrate the appropriate skills to protect their data, but they also lack confidence in their efforts to do so. The reason that this fact is so concerning is that it entails that vulnerable groups, those groups with lower confidence in their abilities to navigate digital environments, may find themselves falling prey more readily to undue

surveillance and data harvesting (Büchi et al., 2021).

Even so, decreased levels of confidence have become the norm regardless of demographic criteria; both people's confidence and trust have been dampened by the datafication of society:

In the age of big data, however, the confidence level associated with privacy prognostication has decreased considerably, even when conscientious people exhibit due diligence (Hartzog & Selinger, 2013a).

In a post-Snowden, post-CA era, then, the knowledge of constant surveillance serves not so much to erode trust in the system of surveillance itself as to erode one's trust in one's own actions within that system. This notion echoes Foucault's theory of surveillance, whereby:

He who is subjected to a field of visibility, and who knows it, assumes responsibility for the constraints of power; he makes them play spontaneously upon himself [...] he becomes the principle of his own subjection (Foucault, 1977, pp. 202–203).

In other words, many digital citizens acknowledge that they are being watched when participating (out of necessity) in digital environments and, as a result, alter their own behaviour accordingly. In some, these behavioural changes manifest themselves as taking greater protective measures; in others, as withdrawing from full participation in the digital space. Lack of trust results in assuming responsibility for surveillance practices, whilst lack of confidence in one's own abilities, due to the inescapable nature of surveillance, maintains order. Either way, individuals do not perceive themselves as having the ability to change the system itself, and resort to “obscurity” or making data “hard to obtain or understand” in order to maintain “safety” (Hartzog & Selinger, 2013b).

However, enough time has passed since the Snowden and CA scandals made these surveillance practices visible – time enough for individuals to become less vigilant and for institutions to obscure such practices once again. Indeed, this obfuscation runs contrary to Foucault's theory in that surveillance has once again been rendered invisible and coded. In

a sense, institutions also employ “obscurity” to mask their data collection efforts, making it impossible for individuals to unveil the extent of their surveillance whilst giving them the illusion of autonomy. This approach works since “less committed folks [...] experience great effort as a deterrent” (Hartzog & Selinger, 2013b). Even so, this illusion of choice also mirrors Foucault’s theory, as a digital citizen remains the “the principle of his own subjection” (Foucault, 1977, pp. 202–203). For instance, in rejecting or consenting to cookies, a digital citizen is given an artificial choice which internalises a feeling of control, despite hidden data collection processes.

Individuals who limit themselves as a reaction to surveillance practices also limit their capacity to participate, and participate effectively, in digital environments and research projects alike. Those who either lack confidence or lack trust in using digital spaces cannot fully instrumentalise or express themselves via the technologies available. Those who have taken personal responsibility for data manipulation by institutions may be more suspicious of academic projects with legitimate scientific aims of furthering knowledge by consensually gathering their data.

For the purposes of this study, I have chosen to take a simple self-assessment of competency, since it is one’s perceived skill, rather than actual skill, that encourages or discourages one to participate (Bandura, 1977; Phillips et al., 2018). Since this study does not seek to comment on digital inequality itself, but rather on how confidence in one’s digital skills impacts engagement with academic research, this question is sufficient. Any limitations pertaining to the gap between perceived and actual skill will be considered in my evaluation of the study.

5.2.4 Understanding of how algorithms are used on data online

As individuals’ confidence in their own digital skills drop, it could be stated that the actual skills required to participate effectively in digital society are mounting. Thus, to be considered a skilled user, individuals must know how to perform a breadth of activities via digital technology, at significant depth. Not only are digital citizens expected to have a vast array of increasingly nuanced practical skills, but they must equally possess

knowledge of the internal workings of the system in order to exercise any power. In particular, “increasingly powerful and often secretive [...] algorithms [...] are eroding” people’s agency and autonomy to the benefit of those who control such algorithms (Hartzog & Selinger, 2013a). As such, digital literacy no longer simply refers to what activities one performs, how frequently, and how well, but should account for understandings of such algorithms too (Reisdorf & Blank, 2021).

Understanding of algorithms is strongly related to questions of trust and amount of use (Reisdorf & Blank, 2021). As a result, Reisdorf and Blank have posited that algorithmic literacy be “included in general Internet skills measures of quantitative studies” (Reisdorf & Blank, 2021, pp. 341–342). Similarly, Dimitar Christozov and Stefka Toleva-Stoimenova conceptualise “Big Data literacy” as the ability both to learn from vast amounts of data and to understand how that data is manipulated in order to reach a new understanding (Christozov & Toleva-Stoimenova, 2015, p. 156).

I have decided to include a question pertaining to algorithmic literacy in this study as it links well to current theories of citizen science which suggest that concerns over how data will be treated in the scope of a research project may limit potential engagement (Bowser et al., 2020; Rudnicka et al., 2022; Tauginienė et al., 2021; Toft et al., 2017). Furthermore, algorithmic literacy is strongly associated with digital literacy, so this question provides both a comparative measure of digital literacy in a datafied society and an element to be combined with the previous question on digital uses to reflect an individual’s overall digital literacy (Reisdorf & Blank, 2021). Since algorithmic literacy affects both “trust in online platforms and amount of use of these platforms”, I consider it a useful point of exploration (Reisdorf & Blank, 2021, p. 342).

5.3 Citizen Science

5.3.1 Contribution of data, time or skills to a research project

Prior experience with citizen science is likely to predispose an individual to future participation (Lewandowski et al., 2017). Individuals who are familiar with citizen science and have “interest in participating in it are likely to affect recruitment and retention” (Lewandowski et al., 2017, p. 4). Moreover, individuals who have already participated in citizen science are equally more confident in its results (Lewandowski et al., 2017). However, definitions of what constitutes citizen science vary depending on “national and cultural differences in interpreting” it (Haklay et al., 2021, p. 25). This difficulty is one reason why I have chosen not to explicitly use the term citizen science in this question, since I want to pinpoint the activities undertaken as part of this study’s interpretation of citizen science, rather than stimulating preconceptions that vary from country to country. Another reason for not explicitly mentioning the term citizen science is that research has suggested that not all individuals (and particularly those outside an academic environment) are familiar with this term, and so I do not want to be exclusionary in the framing of this question (Lewandowski et al., 2017). It is crucial to maintain an accessible and inclusive “communication strategy” (Land-Zandstra et al., 2021, p. 252) when constructing questions and presenting information (de Vries et al., 2019; Rüfenacht et al., 2021).

Several studies have sought to identify different types of citizen science. Early on in debates regarding citizen participation, Sherry R. Arnstein constructed an 8-rung ladder of citizen participation, spanning from “non-participation” to “degrees of tokenism” to “degrees of citizen power” (Arnstein, 1969, p. 216). The first level, non-participation, referred to those projects that were created by someone else. The second level, degrees of tokenism, referred to those projects that allowed “the have-nots to [...] have a voice” (Arnstein, 1969, p. 216). The third level, degrees of citizen power, allowed “have-nots [...] the [...] right to decide” (Arnstein, 1969, p. 216). More recent models of participation have reinterpreted this model of citizen participation. This updated model starts with “contributory” projects (data collection), then “participatory” projects (collaboration), followed by “co-creation” projects (initiation) (Eleta et al., 2019; Hecker, Garbe, & Bonn,

2018, p. 2). In this model, the first layer involves citizens contributing data passively to a research study, the second layer involves actively participating in research tasks, and the third layer involves citizens designing and initiating a project independently.

In existing projects like *What's up, Switzerland?* and *What's new, Switzerland?*, the level of participants' integration in the project can be considered as "contributory" in that they contributed chats following a call for the collection of CMC data (Doudot, 2021; Kucharska, 2021; Ueberwasser & Stark, 2017). Alternatively, this level of participation could be considered as a low level of "tokenism" under Arnstein's model, since participants provided valuable data but had no influence or decision-making power; their voices were, in a sense, "heard" through the CMC data that they proffered, but were unable to ordain the use of that data (Arnstein, 1969). This approach echoes the traditional paradigm in social science research. In this paradigm, participants are allowed to contribute at an early stage of a research project, usually through surveys or data, but are excluded from determining the project's development, even though they are usually the ones being studied (Kullenberg & Kasperowski, 2016).

However, there is a growing trend towards increased participation in "citizen social science" (Albert et al., 2021, p. 119). Whilst "participatory methods [...] have a long legacy in the social sciences", social actors can be further engaged in "some or all research processes" (Albert et al., 2021, pp. 119–120). Although it may only be one rung further on Arnstein's ladder, I hope that this study serves to involve citizens to a greater extent in various stages of future projects, such as "ideation", "data collection", "dissemination" and "impact" (Albert et al., 2021; Arnstein, 1969, p. 120). By understanding the target audience for sociolinguistic research better, future projects can be founded on their needs and tailored to their expectations. This shift moves future research projects from humdrum contributory initiatives towards participatory ones. In this sense, "members of the public contribute data but also help to refine project design" (Hecker, Garbe, & Bonn, 2018, p. 194).

Another helpful way of categorising projects grounds itself in legal and governmental definitions. The kind of research expected to be undertaken under the guidance of this

study is “institutional” in nature, since it is likely to be organised by members of a university like the UNIL (Cooper et al., 2019, p. 1). The subject matter, relating to CMC, pertains to “humans” and any “Personally Identifiable Information” (PII), such as the data contributions themselves, should be kept private (Cooper et al., 2019, p. 2). The reason for the anonymity of contributions is that personal language use (as evidenced in CMC data) represents a sort of “cognitive biometric”, much like a fingerprint, that allows an individual to be identified even without the inclusion of their name and other data. Cognitive biometrics is “the process of identifying an individual through extracting and matching unique signature”; this signature is often determined by demographic criteria (Pokhriyal et al., 2015, p. 69). As such, researchers must continue to be careful in the handling of such PII, limiting a heightened level of citizen involvement.

In this question, I include contributions of “data”, “time” and “skills” in order to reflect all levels of engagement in citizen science.

5.3.2 Organisations to which data, time or skills were contributed

Other than level of participant involvement, another way of categorising projects is through their political, scientific or societal contexts, which entail different “descriptive, instrumental, and normative aspects” (Haklay et al., 2021, p. 21). Political contexts, in particular, carry more instrumental weight in seeking to advance policies. In terms of the descriptive aspect, referring to the kinds of activities involved, societal contexts might seek to further “public understanding” of a given issue, whereas scientific contexts might seek to further “scientific knowledge” (Haklay et al., 2021, p. 21). Whilst a great number of citizen science projects, regardless of context, follow norms such as the *Ten Principles of Citizen Science*, the norm for level of participation may vary depending on the given context. Here, the distinction could once again be drawn between “institutional” and “non-institutional” contexts, or even “public” and “private” organisations, since such contexts would have a bearing on the norms adopted, as well as the instrumentalising of research output (Cooper et al., 2019; Haklay et al., 2021; Nuessle et al., 2020).

In Europe, around 45% of citizen science projects are “coordinated by a scientific

organisation”, 14% by an educational organisation, and 11% by a non-governmental organisation, as of 2018 (Hecker, Garbe, & Bonn, 2018, p. 191). These contexts may provide added nuance as to the kind of digital citizen who is interested in sociolinguistic citizen science.

This dimension could allow for the exploration of whether one’s involvement with particular types of citizen science determines one’s interest in sociolinguistic citizen science. As such, I have expanded the categorisations presented in the literature, accounting for the following spheres:

- educational institutions
- medical research centres
- government organisations
- non-governmental organisations
- commercial companies.

5.4 Potential Engagement

5.4.1 Concerns when sharing data of online communications with research

Since the inception of scientific study, predating the arrival of citizen science per se, choices have had to be made regarding not only “who participates” but how scientific output is handled such that it respects “norms, access, sharing, privacy, and ownership” (Lynn et al., 2019, p. 1). The proliferation of digital technology and the Internet have both facilitated scientific enquiry and brought to bear “new challenges to information management and associated privacy” (Bowser et al., 2020; Lynn et al., 2019, p. 2). As such, there tends to be a lack of “ethical literacy” among project coordinators, particularly because privacy regulations are constantly updated and digital technology is constantly changing (Tauginienė et al., 2021, p. 411).

The reality is that issues relating to data sharing “may not be at the forefront of citizen science project concerns at the time of a project’s design” (Lynn et al., 2019, p. 3). This reality is not always a problem, given that citizen science projects, especially those focused on data collection, follow an “iterative” process in which citizens’ input is often mediated by academics to produce some form of output which informs the next input phase (Straub, 2016, p. 2). This circular process allows a project to be scaled up over time (Straub, 2016). Even so, it is crucial to make ethical and appropriate choices during the design phase of a project, regardless of whether these choices will change over time (Lynn et al., 2019; Tauginienė et al., 2021). Moreover, it is equally crucial to avoid “consultation fatigue amongst research participants” due to incessant requests for feedback (Tauginienė et al., 2021, p. 410).

At the same time, there is significant friction in citizen science between openness and privacy (Tauginienė et al., 2021). Indeed, “there is a clear tension between the ideals of openness and accessibility [...] and [...] data protection” (Suman & Pierce, 2018; Tauginienė et al., 2021, p. 410). On the one hand, the *Ten Principles of Citizen Science* highlight the value of making data “publicly available” and “open access” (Robinson et al., 2018, p. 29). In parallel, the advancement of FAIR (Findable, Accessible, Interoperable and Reusable) practices in academic research has further advocated accessibility (Bowser

et al., 2020). On the other hand, project leaders must “take into consideration legal and ethical issues” regarding the so-called “state of the data” (Bowser et al., 2020; Robinson et al., 2018, p. 39). This tension has been exacerbated by regulatory developments, such as the *General Data Protection Regulation* (GDPR), *Intellectual Property Rights* (IPR) and copyright licences (Lynn et al., 2019; Tauginienė et al., 2021).

However, citizen science encompasses a vast range of projects and it is not always realistic or ethical to maintain open access to data. This statement applies, in particular, to medical research, but equally to any project that collects data “that could compromise the privacy, safety, or security” of participants (Lynn et al., 2019, p. 9). Hence, the “goal should not always be to become more open” (Lynn et al., 2019, p. 7). As has been established, examples of an individual’s language production, such as their CMC, are equivalent to biometric data and must, therefore, be treated accordingly (Pokhriyal et al., 2015).

In this study, I recentre the agency of digital citizens by employing a more bottom-up approach to project design in terms of data concerns, “giving governance over those decisions” to potential participants (Lynn et al., 2019, p. 9). This approach demonstrates “respect” for those who will contribute their data (Tauginienė et al., 2021, p. 412).

Sadly, it is all too common for citizen science project leaders to exhibit a lack of ethical literacy, as well as a “lack of familiarity” as to how data is stored (Bowser et al., 2020; Tauginienė et al., 2021, p. 8). Many project leaders are not able to “articulate specifics around data security approaches” and some are not even able to identify “whether their project [has] a licence” that stipulates data access (Bowser et al., 2020, pp. 8–9).

This deplorable state of affairs has the potential both to undermine the perceived quality of contributed data and to put vulnerable groups at risk. Certain demographics do not have the ability or confidence to protect themselves effectively from the mishandling of data (Redmiles & Buntain, 2021). As a result, it will be valuable to assess the relationship, if any, between level of concern regarding data and demographic criteria.

One of the *Ten Principles of Citizen Science* is that “citizen scientists benefit from taking part” (Robinson et al., 2018, p. 29). I consider this question, then, a valuable

opportunity to encourage potential participants' reflections on data practices; how they do and do not want their data to be used. It is a moment of empowerment that allows them both to contribute more effectively to future project design and to inspire their education and confidence regarding data use, something lacking due to the simultaneous visibility and invisibility of data manipulation. Future research projects might use the results of this study to inform attempts to present data processes to participants, by responding to their principal concerns in communication. In addition, it raises the bar for project leaders to maintain best practices.

To inform my preparation of the concerns listed in this question, I referred to the concept of the “data lifecycle” to understand further the points in data collection that can be of particular concern. I ascertained that “data infrastructure and storage”, “data security and protection”, “data governance and ownership” and “data privacy and access” are particular points of concern when it comes to data collection for research (Bowser et al., 2020; Lynn et al., 2019, p. 2). Thus:

Data protection is about securing data against unauthorised use, whereas data privacy and access focus on who has data, who defines it, and who uses it (Lynn et al., 2019, p. 6).

Considering data protection and data privacy measures is of great import when it comes to avoiding the exposure of one's identity or PII (Cooper et al., 2019). Moreover, data protection regulations, such as the GDPR, mean that “data can no longer be stored without a clear purpose for an unlimited period of time” (Tauginienė et al., 2021, p. 409). Participants have the right under GDPR to request that data be removed at any time (Tauginienė et al., 2021).

If these issues are central to digital citizens' concerns over data use, a possible solution might be “dynamic informed consent”, whereby participants' consent is reassessed on a regular basis (Tauginienė et al., 2021, p. 408). A secondary solution would be to offer participants choices in terms of openness at the point of “data acquisition” (Bowser et al., 2020, p. 2). These solutions could be considered if the results of this study demonstrate the significance of these issues.

The problem of “clear purpose” can be difficult to account for in citizen science, especially in long-running projects that seek to build corpora or datasets. Although there is a clear research topic in mind – sociolinguistic research into CMC and paralanguage – there is not a clear “scientific hypothesis” (Elliott & Rosenberg, 2019, p. 2). Not all academic research is “hypothesis-driven”; instead, scientific enquiries might rather be “exploratory” in nature (Elliott & Rosenberg, 2019, p. 2). As a result, data might be reused over a length of time (Bowser et al., 2020). Therefore, whilst it is possible to describe the field of research that this study will inform, it is difficult to reason how such corpora-building efforts, exploring different aspects of this field, are compatible with the GDPR. It would be critical for future projects to explicate the goals and length of a project at the point of data acquisition.

Outside “unauthorised use” and “who has data”, data governance refers to who it is that determines the nature of data protection and access (Lynn et al., 2019, p. 6). If someone else governs one’s data, it is crucial for them to respect the participant’s right to “transparency” (Land-Zandstra et al., 2021, p. 255). In this study, I hope to encourage a bottom-up approach to data governance that empowers participants to educate themselves and determine the nature of data use.

5.4.2 Motivations to share data of online communications with research

In terms of the “benefits-harms” binary, the previous question about the concerns of sharing data with research represents the potential “harms” that might inhibit participation in contribution-based citizen science (Hintz et al., 2019, p. 110). This question seeks to ascertain the “benefits” that might encourage participation instead.

There is a considerable body of literature relating to the question of what motivates individuals to engage in citizen science projects. One considerable factor is pre-existing interest in the subject matter; a factor that cannot easily be manipulated to encourage engagement without extensive education first (Paleco et al., 2021). Since extensive education would necessitate a high level of investment (read engagement) in a project, it is not realistic to expect that an individual’s interest can be altered in a pre-project context.

Pre-existing interest does not encourage a diverse audience to participate in a project, since the “typical citizen science participant” is not only “Caucasian, older [and] highly educated” but “already demonstrates a high interest in science when joining a project” (Phillips et al., 2018, p. 12). This notion is reinforced by most research, demonstrating that “well-educated Western males with a pre-existing interest in science and technology” constitute the majority of participants (Kloetzer et al., 2021, p. 295). Although this group may appear more clear cut within the hard and environmental sciences, citizen social science also has a responsibility to “attract a larger and more diverse audience”, especially when it comes to projects that collect data online (Kloetzer et al., 2021, p. 295).

It is crucial to evaluate the “target group alignment” of a project, principally in terms of “motivation and engagement” (Kieslinger et al., 2018, p. 86). Whilst organisers of future sociolinguistic research projects benefit from the data-based contributions of participants, participants can be motivated to contribute by unlocking some “benefit” of taking part (Land-Zandstra et al., 2021). As per the *Ten Principles of Citizen Science*:

Benefits may include the publication of research outputs, personal enjoyment, social benefits, satisfaction through contributing to scientific evidence [and] the potential to influence policy (Robinson et al., 2018, p. 29).

Apart from pre-existing interest, I discount three further unfeasible motivations from this study. Firstly, “the potential to influence policy” is not a realistic aim in the context of this type of project (Robinson et al., 2018, p. 29). Usually, sociolinguistic research is: “inherently descriptive [...] rather than prescriptive” in nature (Birner, 2013, p. 6). Therefore, it is inappropriate to present political change as a project outcome, since sociolinguistics studies how language is used, not how it should be used (Birner, 2013). Secondly, “behaviour change” (Land-Zandstra et al., 2021, p. 256), relating to altered behaviour as an outcome of the project, will not be easy to define in this context either; descriptive research into CMC seeks rather to observe “language in actual use under natural conditions” than to alter it (Birner, 2013, p. 7). Thirdly, it has already been established that “openness” is not a realistic prospect in this field, given that linguistic

data is commensurate with PII, acting as cognitive biometric data (Cooper et al., 2019; de Vries et al., 2019; Pokhriyal et al., 2015).

A distinction should be made between what “attracts” potential participants to a research project, and what “sustains” them in the long term (Iacovides et al., 2013, p. 1). Often research has demonstrated that what attracts participants does not always sustain participation and vice versa (Land-Zandstra et al., 2021). For instance, studies have found that “volunteers’ initial interest in citizen science projects stemmed from [...] egoism”, but their interest in “community involvement” grew over time (Rotman et al., 2012, p. 5). Since this study is concerned predominantly with encouraging initial engagement (i.e. attracting new participants), rather than sustained recruitment, I expect to observe more egotistical motivations for sharing data with research.

The possibility of making a meaningful contribution to “real” scientific research is a key motivation in existing studies (de Vries et al., 2019; Land-Zandstra et al., 2021). Interestingly, this motivation parallels understandings of the participatory cultures facilitated by the Internet, communities in which contributions are perceived to be valuable (Jenkins & Clinton, 2006). Individuals who are motivated by contributing to science might equally exhibit “intellectual curiosity” or familiarity with the subject (Bowser et al., 2020, p. 12).

In order to account for those who have a desire to contribute to scientific research, I have prepared the following motivation: “having a more active role in the project”. Since this project requires a minimum of a few minutes to send one’s CMC data to research, those who desire to make a more meaningful and sustained participation can select this statement. Furthermore, this statement ascertains whether a more involved type of citizen science project, higher up the ladder of participation, would be appreciated by potential participants (Arnstein, 1969; Land-Zandstra et al., 2021).

However, those who join a research project on the basis of desiring to contribute to science must also be convinced that “they are contributing to something important” (Land-Zandstra et al., 2021, p. 254). A critical way of evidencing the importance of a project is by communicating information about its “scientific output” (de Vries et al.,

2019, p. 1) effectively (Land-Zandstra et al., 2021; Rüfenacht et al., 2021). In fact, studies have suggested that participants who are motivated by contributing to science also assess “scientific publications” resulting from collected data to be very important (Alender, 2016; de Vries et al., 2019, p. 7).

Effective communication strategies often involve the publication of results, especially in scientific journals, where possible (de Vries et al., 2019). Research has suggested that acknowledging participants in this scientific output is also a way to motivate and sustain engagement (Alender, 2016; de Vries et al., 2019).

In particular, studies have suggested that “name recognition” via citations in academic output and project communications is meaningful to participants (Alender, 2016; Ganzevoort et al., 2017, p. 13). However, this preference for name recognition is not always universal, especially in terms of academic publications (Alender, 2016). Such acknowledgements appeal mainly to younger demographics, who actively seek to improve their “reputation and career” (Alender, 2016; de Vries et al., 2019, p. 8). Furthermore, the possibility of co-authorship – referring to a participant’s inclusion in the list of authors for an academic publication – has also been highlighted as a way of recognising citizens’ contributions (Curtis, 2015). This recognition can be achieved through acknowledging the participant group as a whole, or by acknowledging individual participants who have made major contributions to a project (Curtis, 2015; de Vries et al., 2019). As such, the survey offers two options to account for those motivated by “acknowledgement in citations” and those motivated by “co-authorship in publications”.

In terms of “reputation and career”, I have also added the option of “networking opportunities” which is not covered so extensively in the literature. Given that career advancement is so significant for younger demographics, the possibility to gain access to new networks may be of import. Although this motivation might appear highly “egocentric” since it seeks to maintain “face”, this option is also related to more social motivations for engaging in research, fostering a sense of “belongingness” and interaction (Land-Zandstra et al., 2021; Rotman et al., 2012, p. 249). Another such motivation is “the possibility to share with friends and family”, which again relates to “one’s feeling of being secure,

accepted, included, valued, and respected” (Land-Zandstra et al., 2021, p. 249). In both these cases, respondents would demonstrate their motivation to “become part of a community of like-minded people”, either in a professional or personal context (Land-Zandstra et al., 2021, p. 248). Given the future intention of creating a mobile application to collect, anonymise and analyse CMC data, one possible feature would be immediately to allow the sharing of basic data analyses via social media upon uploading one’s CMC data. Thus, this motivation also accounts for whether sharing results within one’s existing network motivates engagement.

Also related to the concepts of reputation and status, recent studies have focused on the field of “gamification” and its potential for online citizen science (de Vries et al., 2019, p. 9). An aspect of “gamification” would be feasible in the context of future projects at the UNIL, with the prospect of developing a mobile application with competitive aspects, such as rankings and badges. Gamification can offer many advantages; it offers immediate “feedback to participants” (de Vries et al., 2019, p. 9) and “serves as a motivating factor” to contribute more data than one’s peers (Land-Zandstra et al., 2021, p. 248). However, research suggests that “game mechanics” are more effective ‘in helping to sustain volunteer involvement’ than in attracting new volunteers (Iacovides et al., 2013, pp. 4–6).

Even so, gamification can also act as a key tool to encourage learning within a project (Kloetzer et al., 2013). In fact, learning itself is a key motivation for research participation; many participate “because they want to learn something new” (de Vries et al., 2019, p. 6). In terms of attracting and sustaining participation, “many of those who start citizen science projects are motivated primarily by [...] educational goals” (Bowser et al., 2020, p. 11). Not only does the prospect of learning attract participants, but it is also “linked to sustained participation” (Kloetzer et al., 2021, p. 293). As has been discussed, in the realm of citizen science increased participation (communion) can even bolster self-efficacy (agency) (Kloetzer et al., 2021; Land-Zandstra et al., 2021; Phillips et al., 2018).

However, learning can take diverse forms, and individuals who would be motivated by one form of learning may not be motivated by others. I have identified three main forms

of learning that align with research into learning outcomes of citizen science (Kloetzer et al., 2021; Phillips et al., 2018). Firstly, scientific literacy is a critical part of learning within a citizen science context (Kloetzer et al., 2021; Land-Zandstra et al., 2021; Rudnicka et al., 2022). Learning about sociolinguistic research might be a valuable gain for both participants and researchers. Secondly, increasing one’s skills in a specific area, such as one’s practical or technological skills, is another key area (Land-Zandstra et al., 2021). It might be expected that young people, seeking to advance their careers, might find this area of particular interest. Thirdly, “the opportunity to self-learn” is a final aspect of learning that will be considered (Rudnicka et al., 2022, p. 5). Often encouraged by data-harvesting online quizzes, learning about oneself might also harness the initially “egotistical” focus of potential participants (Rotman et al., 2012, p. 3). Learning about oneself might also encourage attitudinal change by demonstrating how much can be learned from one’s data (Land-Zandstra et al., 2021). In a potential mobile application, data visualisations of one’s CMC might instigate both learning about oneself and learning about science.

Therefore, with regards to learning, the three options for potential motivations to sharing data (and thus engaging) with citizen science are “learning about science or research”, “learning about myself” and “learning a skill”.

Although citizen science projects (especially those of a non-institutional or grassroots nature) often do not have the funds to offer financial rewards, I have included a final option, “financial compensation (money or vouchers)”, to account for the fact that this option may nonetheless motivate potential participants (West & Pateman, 2016).

5.4.3 Interest in using a mobile application to facilitate data analysis

With the near-universal access to smartphones in developed countries, mobile applications present unprecedented opportunities for citizen science projects (Lemmens et al., 2021; Mazumdar et al., 2018). There are great benefits to the development of a mobile application to facilitate the collection and analysis of CMC data. One notable benefit is that using “third-party” platforms to facilitate data collection does not allow project

organisers sufficient governance over data infrastructure, security and privacy (Bowser et al., 2020; Lynn et al., 2019, p. 11). Given the diversity of definitions for what citizen science is, no single platform can meet the specific demands of a project. Indeed, such platforms do not offer sufficient “control of the technical infrastructure to impose [...] field-specific or project-specific preferences” (Bowser et al., 2020, p. 11).

Alternatively, mobile applications “provide a new way to steer the data gathering process” (Lemmens et al., 2021, p. 462). They have “enabled direct participation”, “real time” contributions, “interactive features” and data visualisation (Lemmens et al., 2021; Mazumdar et al., 2018, pp. 462–463). Moreover, as with the participatory cultures facilitated by the Internet, “mobile phones may facilitate access by larger, more diverse populations” (Mazumdar et al., 2018, pp. 309–312). Even so, it is still essential to employ appropriate “safeguard mechanisms” and data encryption to protect “sensitive data”, corresponding to both legal stipulations and the concerns of participants (Lemmens et al., 2021, p. 465). In addition, mobile applications have to be maintained with regulatory and technical changes over time (Mazumdar et al., 2018).

Paralleling social media, the networked capabilities of mobile applications allow for greater sharing and interaction opportunities, encouraging greater communion, even at a low level of participation in citizen science (Arnstein, 1969; Bakan, 1966). This networking furthers “public awareness” of a project and gives autonomy to the mobile user (Lemmens et al., 2021, p. 463). Thus, developing mobile applications for citizen science can “harness the power of social networking”, to access a more diverse audience and help to tailor their functionality to the participants themselves (Lemmens et al., 2021; Mazumdar et al., 2018, p. 320).

Mobile applications thus facilitate learning, gamification and socialisation of citizen science projects (Aristeidou & Herodotu, 2020). In this study, it would be particularly valuable to explore whether there is any relationship between extensive use of social media and interest in a mobile application for sociolinguistic research. Moreover, a mobile application might be more popular among digital natives (Lemmens et al., 2021).

5.5 Existing Knowledge

5.5.1 Familiarity with terminology

This question sought to assess pre-existing familiarity with and interest in both citizen science and sociolinguistics. I drew inspiration from survey measures relating to digital literacy, which utilised key Internet-related terms to understand a respondent’s competency (Hargittai, 2002, 2005, 2009). This question will allow me to explore the relationship between existing familiarity and interest – what might be referred to as “intellectual curiosity” relating to the subject matter – and attitudes towards research participation (Bowser et al., 2020). On a more fundamental level, I could equally explore whether a higher level of education is associated with a higher level of familiarity with these advanced theoretical concepts, or whether familiarity with sociolinguistics is associated with familiarity with citizen science.

In terms of project planning, I consider it essential to understand more fully what digital citizens’ existing knowledge basis is in terms of sociolinguistic research and citizen science. Understanding existing familiarity will inform future projects on how best to communicate with participants, underlining the value of inclusivity (de Vries et al., 2019; Rüfenacht et al., 2021). Since projects that engage citizens in science have a responsibility to encourage learning outcomes from their projects, I wish to provide descriptive data regarding respondents’ level of familiarity. Moreover, it would be valuable to ascertain whether familiarity with citizen science is linked to previous participation in research.

Considering whether engagement in certain digital activities, particularly those activities involving CMC, is connected to familiarity with key CMC vocabulary, would also be insightful. Lastly, it would be valuable to consider whether there is a relationship between familiarity and motivation. For instance, it might be suggested that a higher level of familiarity relates to a lower level of interest in learning as part of engagement, and vice versa.

6 Empirical Study

In this section, I will describe the key aspects of the study that I undertook and evaluate its results in five parts. Firstly, I will outline my methodology in preparing the survey that I used to gather data regarding digital literacy and research engagement. Secondly, I will outline the statistical techniques with which I analysed the resultant data, to inform the preparation of the R script. Thirdly, I will detail the results I garnered from the collected data, responding to each of the questions raised by the theory, as well as the three overarching questions I seek to answer in this study:

1. Who is the target audience?
2. What would help and hinder engagement?
3. What do people already know?

Fourthly, I will evaluate the study and its results, based upon my own feedback and the feedback of others. I will address the strengths and weaknesses of the research question, as well as the survey design and methodology. I will consider whether the study produced valuable results that contribute to existing research in the fields of digital literacy and citizen science.

Lastly, I will offer recommendations both for future studies and for research projects, based on these strengths and weaknesses. These recommendations will include further exploring any significant results produced by the study and resolving any issues that arose from the survey design and methodology. Most importantly, it will give recommendations as to how future sociolinguistic research projects into CMC can employ newfound understandings of digital citizens and citizen science approaches to maximise and optimise data collection.

6.1 Methodology

The first step in the preparation for this study was to identify key literature to ascertain which areas had and had not been explored, and which areas needed to be explored again. This research would then inform the creation of the survey that would constitute an integral part of this study. As part of this research, I searched for several key words in the *Renouvaud* library network, with access to physical and online resources, via the *Bibliothèque cantonale et universitaire* (BCU) website (Lausanne, 2024). These key words would be searched for alongside three principal themes, as follows:

Citizen science	<i>participation, scientific literacy, self-efficacy, data collection, data privacy, data protection, online, inclusion, diversity, motivation, concern, gamification, mobile application, gender, Europe, engagement, big data</i>
Digital citizen	<i>digital literacy, self-efficacy, surveillance, data privacy, data protection, digital inequality, digital divide, digital native, digital learner, digital immigrant, big data, inclusion, skill, participatory culture, gender, trust</i>
Sociolinguistics	<i>interjection, part of speech, computer-mediated communication, paralinguage</i>

Figure 1: Key words searched for.

In particular, three key texts were identified from this search: *Citizen Science: Innovation in Open Science, Society and Policy*, *The Science of Citizen Science* and *Handbook of Digital Inequality* (Hargittai, 2021; Hecker, Haklay, et al., 2018). In addition, one academic journal, *Citizen Science: Theory and Practice* was a valuable discovery. From contributions within these four key sources, the search process was then reiterated to uncover further literature.

Following the research process, I drafted questions to constitute a survey that I would

then seek to publicise. The questions of the survey were shaped by my research into existing studies, both in the field of digital citizenship and that of citizen science. Upon receiving initial feedback from my supervisors as to the wording of the questions, as well as the content and structure of the survey, I made a number of alterations:

- An introductory text was prepared explaining the purpose of the survey and assuring respondents that their data would be both confidential and anonymous.
- For the question “What is the highest level of education you have completed?”, the response “Did not complete high school” was changed to “Completed compulsory education” in order not to discredit this stage of education.
- The question “What is your regional background?” was changed to “What is your region of origin?” to be clearer as to that which background refers.
- The multiple choice options for the question “What is your age?” were changed to a simple number input field, since age grouping can be determined at a later stage.
- The list of activities under the question “How much do you perform the following activities on a computer, tablet or mobile phone?” were extended.
- The question “Considering the above activities, how competent are you in using digital tools and services?” was changed to “As a whole, how competent do you consider yourself in using digital tools and services?” This change was made to reflect that this question represents a self-assessment. Further clarification was given as to “using digital tools and services” by giving reference to the previous question about activities.
- The questions “Have you ever contributed data, time or skills to a research project?” and “If you have contributed to a research project, to which of the following organisations have you contributed?” were added to ascertain prior involvement with citizen science and placed in a separate section of the survey.

Originally, the choice was made to prepare this survey using the software LimeSurvey. However, since gaining access to this software via the UNIL would take around 3 weeks

and the study was time sensitive, Google Forms was instead chosen to produce the survey. The survey was prepared in both English and French to represent both the francophone and international communities at the UNIL (Sharifian, 2009).

Following the preparation of the survey, a pilot study was conducted for both the English and the French versions of the survey. Four individuals were chosen, to try to cover different demographics and survey access types:

	UNIL?	Education	Digital Native?	Access	Time Taken
Person A	No	High school	No	Tablet	12 minutes
Person B	No	High school	No	Laptop	10 minutes
Person C	Yes	Bachelor's degree	Yes	Laptop	7 minutes
Person D	Yes	Bachelor's degree	Yes	Mobile	5 minutes

Figure 2: *Pilot study participants.*

As part of the pilot study, participants were required to time how long the survey took to complete, in order to create an assessment for when the full survey was publicised. The participants recorded the above completion times (see *Time Taken*).

After completing the survey, a Zoom call was scheduled with each participant in order to discuss any feedback they might have regarding the survey. Participants also had the opportunity to give further written feedback after the call via email. Critical feedback was obtained from participants which shaped the survey significantly:

- The approximate time (5-12 minutes) for survey completion was added to the descriptive text at the top of the survey.
- A question was added at the top of the survey to determine whether a respondent was a member of the UNIL community (either as a member of staff, a current student or alumnus) or not.

- Numbers were added to each question so that the survey was easier to follow.
- Any minor spelling errors identified by participants were corrected.
- E-readers were explicitly included as an access type in the question “How much do you perform the following activities on a computer, tablet or mobile phone?”
- Another level of frequency, “Less than a year” was added to the same question concerning performing various activities, to account for situations in which a respondent rarely performs a given activity.
- For the same question, the option “Using AI tools (chatbots, writing assistants, translators)” was added to the list of activities, in order to account for burgeoning uses of digital technology.
- The question “How concerned are you about the following when sharing data of online communications with research?” was made clearer in the French version of the survey, since it was unclear whether “online” referred to the nature of communications or the mode of sharing.

One final piece of feedback regarding the question “How much do you perform the following activities on a computer, tablet or mobile phone?” was that its length did not facilitate responding. It was suggested that the frequencies be “pinned” to the top of the question so that the frequency categories were still visible as one scrolled through the extensive list of activities. Since this format was not easily achievable with Google Forms, highlighting the constraints of a third-party platform, the question was left as before (Bowser et al., 2020; Lynn et al., 2019).

Once these alterations had been made to the survey, the aim was set to gather at least 100 responses from within the university community, as well as the networks of those within it. Publicising the survey throughout the university community via email would have been too extensive (with 16,950 students, 3,929 full-time staff and 460 incoming exchange students) and was in any case impossible due to the moderation of the principal

university mailing lists (UNIL, 2023, 2024a, 2024b). Instead, solicitations to publicise the survey were made through the three sections, within the Faculty of Letters:

- Sciences du langage et de l'information (SLI)
- Humanités numériques (HN)
- Français langue étrangère (FLE)

At the same time, 16 posters with QR codes to both the English and French versions of the survey were displayed in the Anthropole building at the UNIL. The email and poster advertisements were released on 11th November 2024 and an informal deadline set for 25th November 2024.

However, these three emails and 16 posters only encouraged 5 respondents to the English version of the survey and 29 respondents to the French version in this time. It was clear that further efforts would be required to encourage responses.

As such, on 25th November 2024, personal emails were made to 12 professors who had taught me during my studies, requesting that they both answer the survey and mention it to their current students. A digital copy of the poster was included in this email as visual endorsement for my study. In addition, a message with the two links to the survey was sent via the WhatsApp group for *dhelta*, “the association of UNIL and EPFL digital humanities students” (EPFL, 2020). More posters were printed and left on tables within the Anthropole and Geopolis buildings. Requests were also made to the administrative staff within the three sections to resend the survey, with only one section (SLI) facilitating this reminder. An extension of two weeks, with a final deadline of 9th December 2024 was set for data collection. Taking into account this extension, 79 responses were gathered in total from the UNIL community.

In parallel to these rather underwhelming data collection results, copies were made of the two surveys, as a secondary dataset to supplement the UNIL datasets if there was a lack of responsivity. These copies were sent out via my personal networks and social media, with requests to share the survey within respondents' own networks. In total, these efforts produced 128 responses. Although the survey is not as appropriate for

respondents outside the UNIL and Switzerland, this supplementary data allows further comparative analysis of different demographic subgroups.

At the end of the data collection stage, there is a total number of 207 responses in the combined datasets, of which 79 were collected from inside and 128 from outside the UNIL community. Since the 79 responses do not meet the aim of 100 set at the start of the study, it is clear that engaging individuals within the UNIL community, or perhaps within the Faculty of Letters, can be challenging. Although the additional 128 responses from outside the UNIL community may introduce further biases into the dataset, I do not intend to make suppositions about the UNIL community in particular through this study, and so this data still adds a valuable dimension to the study. Further reflections will be made on this subject as part of the evaluation section.

6.2 Techniques

6.2.1 Data cleaning and preparation

The decision was made to programming the statistical analysis of my data with R. Upon collecting the data for analysis, I began by cleaning and improving the data I had received to improve its usability.

Firstly, I opened, translated and combined the 4 comma-separated files. I then altered the names of the columns so that the variables could be referred to with greater ease in the following analyses.

Secondly, I removed any individuals with missing data values. Since age was the only optional question in the survey, I made the decision to remove any individuals that did not provide their age. This decision was made because age was a significant aspect of the study. In addition, I removed the individuals that did not list their gender identity as Male or Female, since I did not feel comfortable generalising on the basis of so few respondents. This procedure led to a total loss of around 7.25% from the dataset, bringing the number of individuals from $N = 207$ to $N = 192$.

Thirdly, I converted any binary variables to values of 0 (Female; No) and 1 (Male; Yes). I converted the Likert scale questions to ordinal scales. I provided extra columns in the dataset for regional origin and created some preliminary scores, combining various values to produce new variables, such as amount of use, level of concern, level of familiarity and various categories of activities. Also, I converted the categorical variable representing approximate years of Internet usage into a discretised 7-option numerical variable using the range midpoints.

Lastly, I converted certain variables to factors or numerical values in order to further facilitate the data analysis.

6.2.2 Variable types

The following table gives an overview of the types of variables employed in the analysis:

Variable	Type
Member of the UNIL	Binary categorical
Gender identity	Binary categorical
Level of education completed	Categorical
Region of origin	Categorical
Age	Continuous numerical
Years of Internet usage	Discrete numerical
Activities performed	Ordinal; Likert scale (1 – 7)
Digital competency	Ordinal; Likert scale (1 – 5)
Level of diversity in activities	Numerical; 0 – 1
Algorithmic literacy	Ordinal; Likert scale (1 – 5)
Contribution to a research project	Binary categorical
Organisations contributed to	Binary dummy
Concerns when sharing data	Binary dummy
Overall level of concern	Numerical; 0 – 1
Motivations for sharing data	Binary dummy
Interest in mobile application	Binary categorical
Research interest	Numerical; 0 – 1
Familiarity with terms	Ordinal; Likert scale (1 – 3)

Figure 3: *Types of variables.*

6.2.3 Summary statistics

I first displayed the summary statistics for the dataset. These summary statistics included minimum and maximum values, the 2nd and 3rd quartiles, the median and the mean.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

From these statistics, I manually calculated range and interquartile range (IQR) for added clarity. In addition, I outputted the standard deviation for key variables.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

In this study, I have chosen to display any figures to 2 decimal places. All statistical tests will be conducted at a significance level of $\alpha = 0.05$.

6.2.4 Statistical approaches

I used chi-square to produce contingency tables and conduct tests to understand the association between pairs of categorical variables, using the defaults provided by R, which included Yates' correction for continuity for all 2x2 tables.

$$\chi^2_{Yates} = \sum \frac{|O_i - E_i| - 0.5}{E_i}$$

I used the default correlation calculation with R, which is Pearson correlation, for numerical variables. I then tested the significance of outstanding correlation coefficients through a t-test.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

I used Analysis of Variance (ANOVA) to compare categorical variables, such as region of origin and level of education, with numerical variables. Additionally, I employed logistic regression in specific cases where I was handling a binary variable.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

6.2.5 Data visualisation

Using the *ggplot2* package for R, I produced several box-plots and scatter plots (with the line of regression marked) to reinforce my findings.

6.3 Results

6.3.1 Descriptive statistics

Out of the $N = 192$ individuals in the final dataset, 36.98% were members of the UNIL community and 63.02% were not, reflecting the combined datasets that were gathered both internally and externally.

We find a relatively balanced distribution in terms of gender identity, with 55.21% identifying as Female and 44.79% identifying as Male. Although this distribution differs from the expectations relevant to research participation in a citizen science context, where we see those identifying as Male being more involved in projects, it is in line with studies into survey responsiveness by gender (Correa et al., 2021; Paleco et al., 2021). Such studies demonstrate that those identifying as Female disproportionately contribute to surveys. Even so, the distribution is balanced enough to proceed with gender-based analyses.

Expectedly, there is a high proportion of respondents with further education. This result is expected, given the sampling context; the survey was publicised both within a university context and among a university graduate’s circles. This sampling choice is one of the most perplexing limitations of this study, since convenience – nay, necessity – determined the population groups and sampling method chosen. As such, the most represented educational group in this study is those with a bachelor’s degree, accounting for 45.31% of respondents. Following this group, the next best represented is those with a master’s degree, accounting for 30.21% of the dataset. Furthermore, given that people with doctorates account for less than 2% of the world’s population, respondents with a doctoral degree were vastly overrepresented in the dataset, at 9.38% (Stapleton, 2024). Again, the data collection strategy accounts for these discrepancies in representation. Furthermore, there is a very low proportion of respondents with only compulsory education. Whilst the surveying strategy might partly account for this under-representation, it is possible that the categories “High school, or equivalent” and “Compulsory education”, particularly in the English-language version of the survey, do not account for regional differences in terms of education. For instance, high school might be synonymous with compulsory education in some countries, whilst in others it may refer to 16-18 educa-

tion. This aspect is something to consider in future research, especially as understanding differences in educational level would be of great value in this field.

Due to the necessity to request responses from both inside and outside the UNIL community, it became clear following the data collection process that more reliable results would be obtained by comparing three regional groupings: “Swiss”, “Non-Swiss, European” and “Non-Swiss, non-European”. The majority of respondents, 59.38%, identified their region of origin as “Non-Swiss, European”, reflecting both the international community at UNIL and my largely Europe-dominated network. Those of a Swiss origin were the next largest group, representing 23.44% of the dataset, followed by those of a non-European origin, with 17.19%. Had I appreciated the difficulty of encouraging responses from within the UNIL community in advance, I would have also asked for the current domicile of respondents to further ascertain whether region has a significant impact on response. However, the addition of the first question – whether someone is part of the UNIL community or not – does well in accounting for this difference.

The youngest respondent was 18 years old and the oldest was 79, giving a range of 61 years. The standard deviation is $s_A = 16.37$ years and the IQR is 31, suggesting that the dataset is varied in terms of age. The mean of $\bar{x}_A = 40$ highlights the dominance of external respondents in the dataset and reflects the most common age demographic in research participation (Phillips et al., 2018). The discrete variable representing the years that a respondent has used the Internet and acting as an estimation of access to digital technology highlights that respondents tend to have had access to the Internet for a mean of $\bar{x}_{YI} = 19.67$ years. The minimum years of Internet usage was 8.0, highlighting the high level of access respondents have had to digital technology since the arrival of the Internet around 35 years ago.

As a first appraisal of the activities that respondents performed, frequently or infrequently on the Internet, we find general consistency in terms of the summary statistics. However, 2 categories stand out in particular: both “Searching information” and “Browsing websites” have minimum values of 2.0 in the dataset. This value demonstrates that all respondents perform these two activities at least “Less than once a year”; none of the

respondents have never performed these activities.

The commonality of performing these two activities, “Searching information” and “Browsing websites”, is bolstered by their means, at $\bar{x}_{SI} = 6.08$ and $\bar{x}_{BW} = 6.0$, respectively, highlighting that the average respondent reports a daily performance of these two activities. In addition, “Checking the time” ($\bar{x}_{CTT} = 6.12$), “Using instant messaging applications” ($\bar{x}_{UIMM} = 5.95$), “Sending text or multimedia messages” ($\bar{x}_{STOMM} = 5.87$) and “Emailing” ($\bar{x}_E = 5.77$) all reflected a near-daily average. Since 3 of these high-frequency activities (messaging, texting and emailing) offer the potential for the production of textual paralanguage, this overview was valuable both in terms of demand and supply when it comes to future research in this area (Baym, 1995; Carey, 1980; December, 1997; Kiesler et al., 1984; Luangrath et al., 2017).

The two activities with the lowest means were “Live streaming” and “Making GIFs or memes” ($\bar{x}_{LS} = 2.44$ and $\bar{x}_{MGOM} = 2.30$), reflecting perhaps the novelty of these 3 use cases for digital technology (Dean, 2018). Both of these activities might require greater investments of time and effort than other activities considered, or might equally appeal to niche subsets of digital citizens. More research into these newer use cases of digital technologies is necessary, since existing research has not focused on these relatively new developments. However, since the focus of this study is on textual CMC and research participation, this question will not be broached extensively.

From the sum of these 30 activities performed via digital technology, a total score was created to represent each respondent’s amount of use. The minimum possible score was 30 and the maximum possible score was 210. This dataset tended towards the higher end of this range, with a minimum of 60 and a maximum of 194, suggesting a higher amount of digital technology use. Moreover, the mean, at $\bar{x}_{DU} = 146.9$, reflects respondents’ high level of use. However, we must exercise caution here since the score does not distinguish between the breadth (diversity) and depth (frequency) of uses, so that these two qualities get lost in a single variable. Even so, both of these qualities are essential as part of an assessment of digital literacy (Correa et al., 2021; Grey, 2011). In addition, this variable was then divided by 30 to achieve a value that parallels the Likert

scale used for each individual activity. A combined score representing digital literacy was then prepared, by normalising the score for amount of use and the value for digital competency and multiplying them together, so that amount of use and competency are equally represented in this new variable, with values between 0 and 1. As such, the mean for digital literacy was $\bar{x}_{DL} = 0.57$.

To allow for the study of the diversity of activities performed via digital technology, a new variable was created, with 1 added to the value if the activity was performed “Yearly” or more. This value was then divided by 30 (the number of activities listed), to produce a value between 0 and 1. We find, then, that the digital citizens included in the dataset have a very high diversity of use, at $\bar{x}_{DS} = 0.89$, and very low standard deviation, at $s_{DS} = 0.12$. Thus, respondents’ consistently engaged in a diversity of digital activities.

Prior to analysing the dataset in detail, several categories of use were prepared to establish further the most common uses of digital technology. In particular, respondents engaged frequently in 2 categories, “Self-development” (with a mean of $\bar{x}_{SD} = 5.51$) and “Written communication” ($\bar{x}_{WC} = 5.78$). The former category reflects a more instrumental, self-centred use of digital technology, whereas the latter re-establishes the importance of activities like messaging and emailing to this dataset (Kelan, 2007; Kim & Hargittai, 2021; Kirkup, 1992; Lai & Katz, 2012; Rotman et al., 2012; van Deursen & van Dijk, 2014). Out of these pre-prepared categories, “Creativity” had the lowest mean, at $\bar{x}_C = 3.92$. This result is valuable, since it demonstrates that respondents engage more frequently or more broadly with instrumental rather than expressive uses (Kelan, 2007; Kim & Hargittai, 2021; Kirkup, 1992; Lai & Katz, 2012; van Deursen & van Dijk, 2014).

In fact, 2 further categories were prepared to include instrumental and expressive use cases, respectively. The descriptive results from these two variables also reflects a slight dominance of instrumental over expressive use cases. The instrumental use category has both a higher minimum value (2.0 over 1.78) and a higher mean ($\bar{x}_{IU} = 5.04$ over $\bar{x}_{EU} = 4.62$) than the expressive category. Of course, what is considered instrumental and expressive is highly subjective; however, these categories were based on theoretical insights (Kelan, 2007; Kim & Hargittai, 2021; Kirkup, 1992; Lai & Katz, 2012; van

Deursen & van Dijk, 2014).

More self-centred or more collective use cases, echoing the divide between agency and communion, were also taken into account with 2 more categories representing egotistical and collective uses. Paralleling the distinction between instrumental and expressive uses, we find that, overall, respondents engage more frequently in egotistical uses (with a mean of $\bar{x}_{EU} = 0.24$) than collective uses ($\bar{x}_{CU} = 0.17$). Again, further studies need to be conducted on more expansive and representative databases to understand whether this result is replicable.

The mean self-assessment of respondents in terms of competency in using digital technology, as per the given activities, was $\bar{x}_{DC} = 4.03$, indicating an assessment of high competency (“Very competent”) in using digital technology overall, something reinforced by the relatively low standard deviation of $s_{DC} = 0.84$, suggesting a level of consistency in respondents’ assessments. However, there is an astonishing difference when it comes to respondents’ self-assessed understanding of how algorithms work upon their data online; the mean is $\bar{x}_{AL} = 3.06$ (“Moderately understand”) and the standard deviation of $s_{AL} = 0.98$ indicates a greater variety of responses to this question. This result endorses including algorithmic literacy in assessments of digital literacy (Reisdorf & Blank, 2021).

In terms of prior research contributions, we find that the majority have already contributed to a research project of some description. Indeed, 61.98% report having to have already contributed data, time or skills to a research project. Since studies have demonstrated that prior contributions to research relate to future interest in research participation, this result is also valuable in nature.

As for the type of research project to which respondents have contributed, we find that 35.94% indicated that “I have not contributed to a research project”. This result represents a 2.08% difference from the previous question, asking whether respondents had already contributed data, skills, or time to research. There could be many reasons for this minimal difference. Firstly, respondents might not have remembered their contributions on responding to the previous question and this question served to remind them of this contribution. Secondly, respondents might not consider activities such as answering a

survey as contributing data, since this kind of research does not clearly indicate that it is gathering data (Rudnicka et al., 2022). Lastly, contributing to a for-profit organisation might not come to mind when one considers research.

In any case, the most common organisation to which respondents have contributed is “A university or educational institution”. It is expected that this result emanates once again from the over-representation of the university-educated demographic in the dataset, and further research must be done to ascertain whether those who are not university educated have also engaged in academic research.

Furthermore, a variable combining the value for previous research contributions and that for present interest in an application was created, with both given equal weight to produce a value between 0 and 1. This variable sought to represent interest in research in a more general sense. With a mean of $\bar{x}_{RI} = 0.37$ and a standard deviation of $s_{RI} = 0.29$, we find that there is relatively low interest in research as a whole and limited variability in this interest level.

Moving on to respondents’ primary concerns when it comes to offering data to research projects, 3 particular concerns manifest themselves in the dataset: “My personal information will be exposed” (with a mean of $\bar{x}_{MPIWBE} = 3.12$), “My data will be misused” ($\bar{x}_{MDWBM} = 3.15$, and “My identity will not be protected” ($\bar{x}_{MIWNBP} = 3.22$). The level of concern rests at “Moderately concerned”, demonstrating that respondents are not overly concerned with data use. This moderation could be ascribed to digital citizens’ resignation, after Foucault, to surveillance practices and data mishandling, or it could be related to the framing of the question. Respondents, especially from an educated background or university community, may place higher trust in research; another reason why it is critical in future studies to ascertain academic outsiders’ relationship with research practices and data sharing (Gloria et al., 2001). Of least concern to respondents was “My data will be used for a long time”, a valuable insight in the context of corpora-building research projects in sociolinguistics, which rely on continued access to data over time (Birner, 2013).

From the list of concerns, an overall concern score was created by summing the

values for each concern and dividing it by the number of concerns. In general, this score reinforced people’s middling concern with data processes in regards to data, with a mean of $\bar{x}_{CS} = 3.0$ (“Moderately concerned”).

In considering motivations, it must first be noted that 31.77% of respondents reported that “None of the above” options would encourage their engagement in research. In future research, it would be valuable to conduct a more interview-based study or comment-based survey to ascertain whether there are other motivations, beyond those listed, that would encourage research participation and data sharing. Perhaps a further option “Nothing would motivate me to do so” would have avoided the ambiguity of this response. Converting this question to a Likert scale, on par with the other questions, would also have provided more valuable data.

In line with previous research, however, we find that learning – in general – is the most motivational of the available options (Bowser et al., 2020; de Vries et al., 2019; Kloetzer et al., 2021; Land-Zandstra et al., 2021; Phillips et al., 2018; Rudnicka et al., 2022). Again, further research into those who do not belong or who have not belonged to an academic community would be valuable to ascertain whether this result is caused by a rather educationally biased dataset. This need is especially evidenced by the fact that “Learning about science or research” was the most listed motivation among respondents, with 41.15% listing this possibility as motivating. Unlike previous studies comparing “Learning about myself” and “Learning a skill”, we find that both are on par – motivating 37.5% of respondents. Further research should be conducted to understand in which contexts these two motivations would be appropriate (Rudnicka et al., 2022). The least motivating option among respondents was “Competitive aspects”, interesting only 4.09% of respondents. This result is also substantiated by research, which highlights that gamification sustains rather than attracts participation in research. Future studies should separate these questions in similar surveys, asking respondents instead to assess, firstly, what would motivate initial participation and, secondly, what would sustain it.

A further category, “Acknowledgement”, combining motivations that involved some form of acknowledgement, was created in order to consider what groups might be moti-

vated by this form of motivation the most. However, on first perusal, interest in acknowledgement appeared to be low.

Overall, interest in the future development of an application to encourage data collection of CMC, one of the underlying bases for this study, was tepid, with only 52.08% reporting interest in such an application. Statistical analysis will determine whether this interest is related to any of the other variables measured in this study.

In regards to familiarity with terminology pertaining to sociolinguistic research, we find that overall people “Recognise but do not understand” key terms, as the means for 3 of the terms – “Computer-Mediated Communication”, “Interjections” and “Parts of Speech” – all tend towards $\bar{x} = 2.0$, with “Parts of Speech” being the best understood ($\bar{x}_{POS} = 2.08$). However, “Paralanguage”, with a mean of $\bar{x}_P = 1.58$, is clearly not a widely understood concept. These results indicate that researchers in this field must invest themselves in educating future participants. Regardless of whether educational gains motivate participants, education would be a valuable outcome, moving participants from recognition to understanding (Bowser et al., 2020).

Familiarity with the term “Citizen Science” also appeared rather mediocre, with a mean of $\bar{x}_{CS} = 1.84$, suggesting a lack of understanding as well as a lack of recognition among some respondents. However, the standard deviation is $s_{CS} = 0.83$, indicating that there is a variety of responses to this 3-part Likert scale question. Both the mean and standard deviation could reflect the diversity of definitions of what citizen science is considered to be, leading to some degree of confusion as to whether respondents truly understand this term (Haklay et al., 2021; Kullenberg & Kasperowski, 2016).

6.3.2 Digital use, competency and literacy

In the following sections, I will outline the key results from my study, with the provision of appropriate data visualisations. Although the dataset can be used to study a wide range of research questions, I hope that the selection I have made offers an insight into the central questions of this study.

The first question I considered was which groups use digital technology the most

frequency. Of course, it has already been mentioned that the score for the amount of use of digital technology does not discriminate a moderate yet diverse use and an extensive yet focused use of digital technology. However, I wish to consider the amount of use in general whether diverse or not.

I found a slight negative correlation ($r = -0.25$) between age and amount of use, which reinforces results of previous studies which demonstrate that as age increases, the amount of use of digital technology decreases (Bimber, 2000; Büchi & Latzer, 2016; Correa et al., 2021; Kim & Hargittai, 2021; Redmiles & Buntain, 2021). This result emphasises the difference between digital natives and digital learners. Although this result was significant, with a p-value of $p < 0.01$, it must be recognised that this result does not necessarily entail that digital learners engage in fewer use cases, or even that they engage in all use cases less.

The ANOVA results demonstrated that use differences in terms of gender identity and level of education were not significant. However, region of origin demonstrated a clear difference in terms of use of digital technology. Indeed, this result was significant, offering a p-value of $p < 0.01$, suggesting that individuals from different regions of origin engage in digital technology differently. However, caution must again be exercised here, since the rather imbalanced dataset, caused by engaging several distinct population groups, might easily have skewed results. Even so, the following box-plot (*see Figure 3*) demonstrates that, whilst non-Swiss Europeans and Swiss respondents use digital technology to a similar extent, non-Europeans tend towards a higher amount of digital use, suggesting that those from a non-European background either use digital technology more frequently or more diversely.

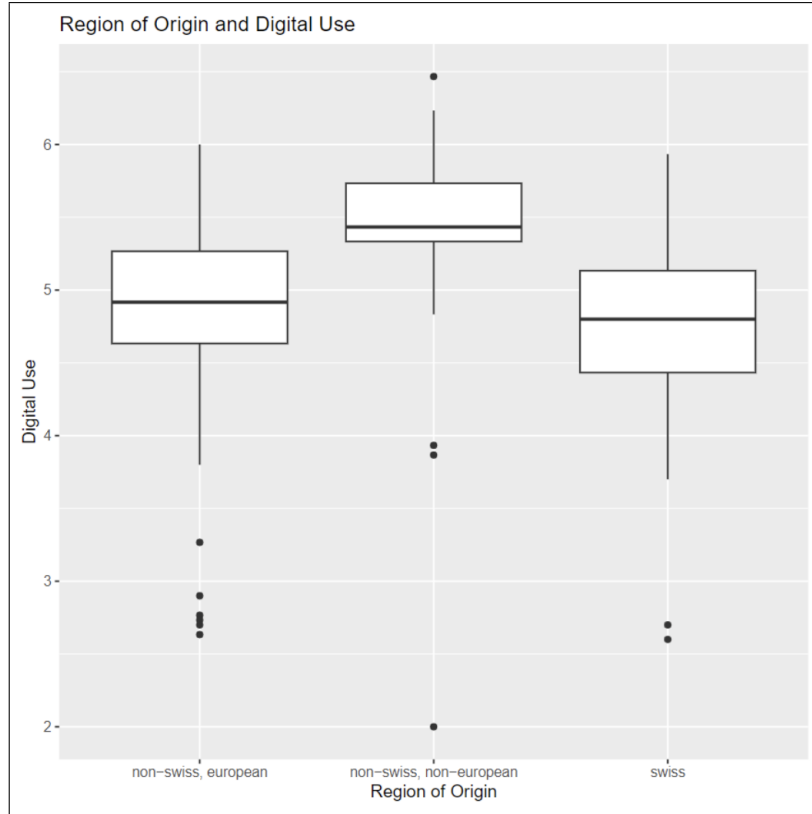


Figure 4: *Region of origin and digital use.*

There is equally a slight negative correlation between age and self-assessment of competency in using digital technology, suggesting that older respondents assessed themselves as less competent in using digital technology. The correlation coefficient of $r = -0.18$ was thus of significance, with a p-value of $p = 0.01$. Interestingly, although there was a slight positive correlation between years of using the Internet and digital competency, suggesting that those individuals who had used the Internet for longer had greater digital confidence and self-assessed skills, this result was not significant. Even so, the age variable often mirrored the results of the variable representing years of Internet usage, so this difference in correlation was intriguing.

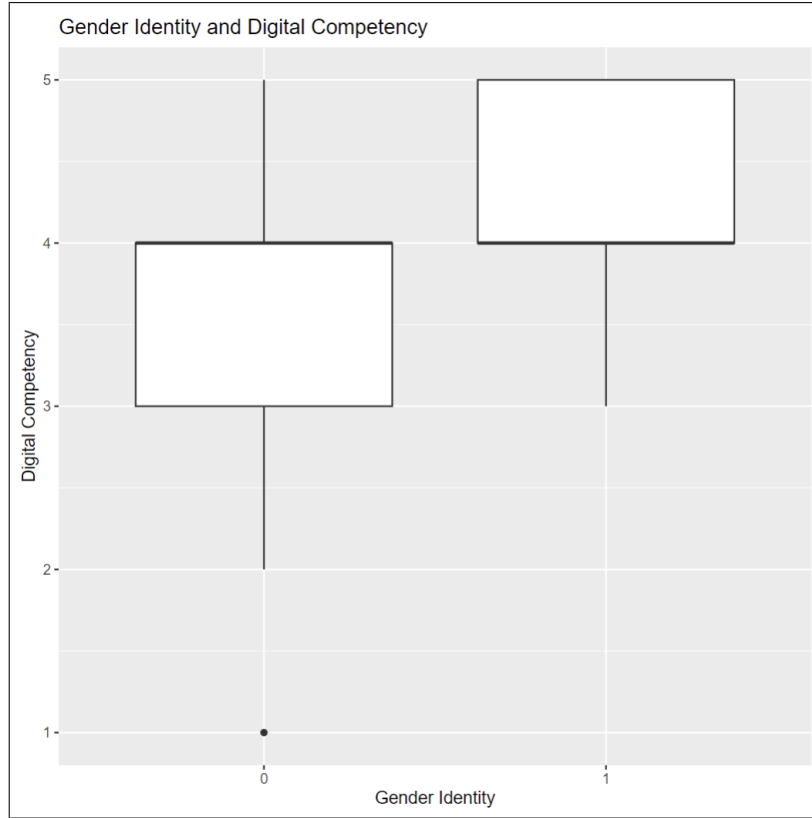


Figure 5: *Gender identity and digital competency.*

As suggested by previous research, gender identity had a major impact on self-assessment of digital competence (Büchi et al., 2021; Hargittai & Shafer, 2006; Rees & Noyes, 2007; Reisdorf & Blank, 2021). The p-value from the ANOVA result was $p < 0.01$, reflecting the fact that gender identity had a significant impact on the digital competency variable. A future study should include a more practical assessment of actual competence to understand if there remains a discrepancy between perceived and actual competence. In any case, this result demonstrates the male respondents' greater degree of agency in digital environments. The above box-plot (see Figure 4) further articulates this continued gender gap. Although male and female respondents have a similar median of 4 on the Likert scale, representing "Very competent", the IQR spans upwards for males and downwards for females. Level of education and region of origin were not significant.

I created a variable to represent digital literacy overall, which combined the variables representing amount of use of digital technology and self-assessment of digital competency (with equal weighting). This variable thus reflected diversity, frequency, and competency,

3 key aspects of digital literacy according to previous research. Once again, I found that there was a slight negative correlation of $r = -0.26$ between age and digital literacy, suggesting perhaps that those born before the development of the Internet lack digital skills to a greater extent than those born into the Internet age. This correlation was significant, with a p-value of $p < 0.01$.

As for the ANOVA results, gender identity and region of origin remained significant in terms of digital literacy, with p-values of $p = 0.03$ and $p < 0.01$, respectively. That region of origin potentially entailed greater digital literacy disputes prior research (especially that dating from access-based digital divides) that Europeans tend to be more digitally literate. Interestingly, the dataset suggests (in spite of the potentially distorted educational level of Swiss respondents, emanating from within the UNIL community) that Swiss respondents might have a lower level of digital literacy. Further comparative studies should be conducted to ascertain the difference between the Swiss perspective and that of other countries.

The effect of educational level on digital literacy was not significant; a more well-distributed dataset with a higher number of respondents without a bachelor's degree would be valuable. Another measure of education, such as an assessment of logic skills or learning abilities, might be more fruitful than determining educational ability from degree title alone.

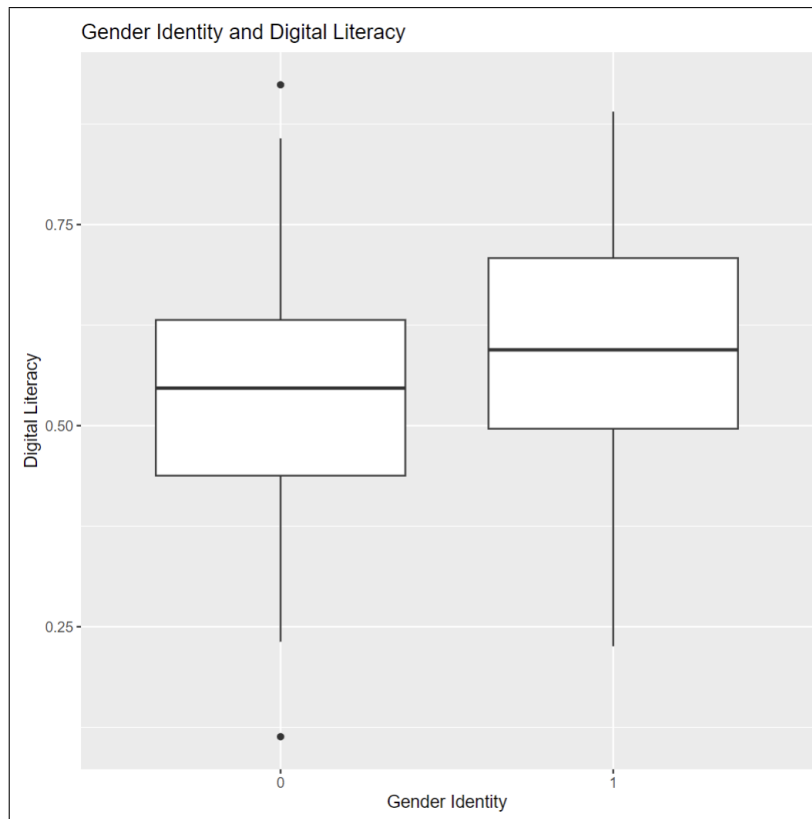


Figure 6: *Gender identity and digital literacy.*

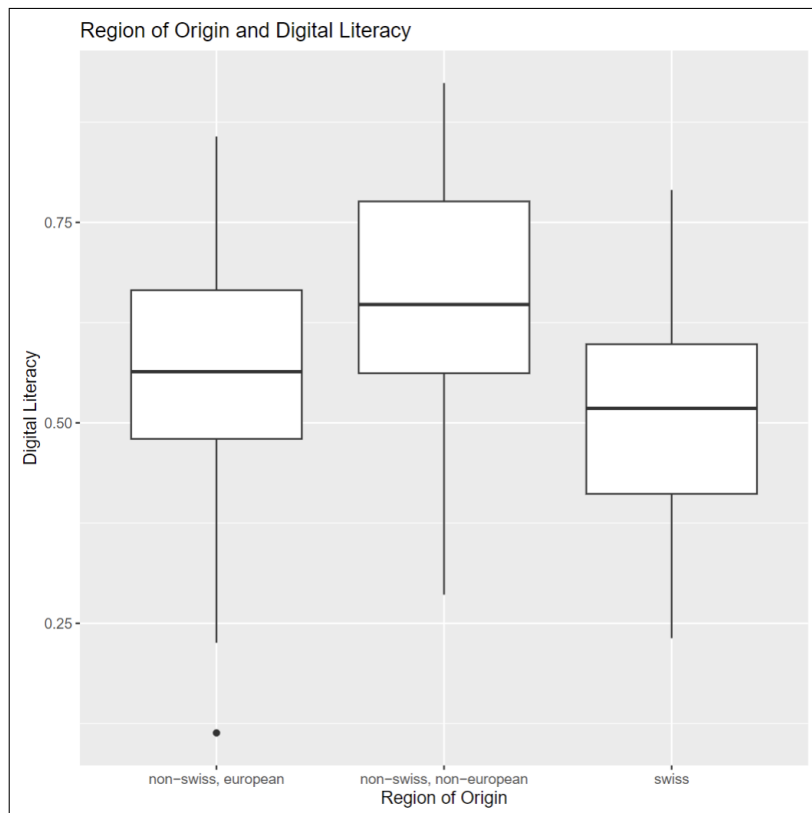


Figure 7: *Region of origin and digital literacy.*

In terms of the relationship between digital competency and digital technology usage, it is expected to observe that as the amount of use of digital technology increases, so too does digital competency (Aristeidou & Herodotu, 2020; Hargittai, 2002). Even though this survey measured only a self-assessment of digital competency, the results of linear regression demonstrate a clear positive relationship between the two variables. On the one hand, the relationship is significant, with a p-value of $p = 0.02$. On the other hand, the R-squared value is $R^2 = 0.03$, suggesting that use of digital technology only explains around 3.0% of the variance in digital competency. This result suggests that there are other factors beyond amount of use that determine digital competency.

To remove the effect of frequency on the performance of activities, I created a diversity score, whereby an addition of 1 was made to the value each time an activity was performed “Yearly” or more. This variable was created in order to account for the fact that diversity and frequency got lost in the amount of usage score.

The ANOVA results demonstrated that gender identity, level of education and region of origin did not significantly impact the diversity of activities performed online. This result was valuable since it demonstrated that, whilst non-Europeans performed activities more frequently, they did not perform activities more diversely.

Age and years of Internet usage were both negatively correlated (-0.31 and -0.17 , respectively) with diversity of use, and significantly so (the p-values were $p < 0.01$ and $p = 0.02$, respectively). This result bolsters the theory that digital natives tend to employ digital technology in more diverse ways than digital learners.

6.3.3 Categories of activities performed via digital technology

One of the key facets of this study is to ascertain who uses paralinguage and, thus, which groups should be targetted when attempting data collection efforts in the field of CMC. Although there is a slight negative correlation ($r = -0.13$) between age and uses of digital technology that potentially produce paralinguistic output, this result is not significant. This result suggests that it is therefore inappropriate that previous studies, such as *What’s New, Switzerland?*, have an over-representation of younger age demographics

in their corpora; all ages have the potential to produce paralinguistic and should therefore constitute part of a corpus (Doudot, 2021). More effort should be made to attract the contributions of diverse age groups within such research projects.

According to the ANOVA results, neither gender identity nor level of education are significant in determining potential paralinguistic output. Of course, it must be recognised here that the potential to produce paralinguistic does not necessarily entail that an individual will produce paralinguistic when they engage with such use cases.

It should be noted that a lack of use of paralinguistic in CMC is of equal value to such research, to better understand how much paralinguistic is employed more globally and among specific demographics. Even so, in this study, region of origin is significant, with a p-value of $p < 0.01$ in determining potential for paralinguistic output. In the following box-plot (see Figure 7), it appears that non-Europeans have the highest potential for paralinguistic output, followed by non-Swiss Europeans. Swiss respondents have the greatest IQR in terms of paralinguistic use cases and the median suggests that they perform these activities less than the other regional groups.

However, it would be valuable to conduct a more extensive study with a larger and more balanced dataset, to ascertain whether such results are reproducible. Moreover, introducing a variable that measures how much each respondent actually uses paralinguistic would help better understand who uses paralinguistic when given the opportunity. For instance, a survey question asking respondents whether they have previously seen or personally use popular examples of paralinguistic and interjections would be insightful.

Even so, part of what makes the study of paralinguistic so critical is its diversity depending on one's language; such a question in a cross-regional study would take extensive research and even collaboration and would thus have been beyond the scope of this study.

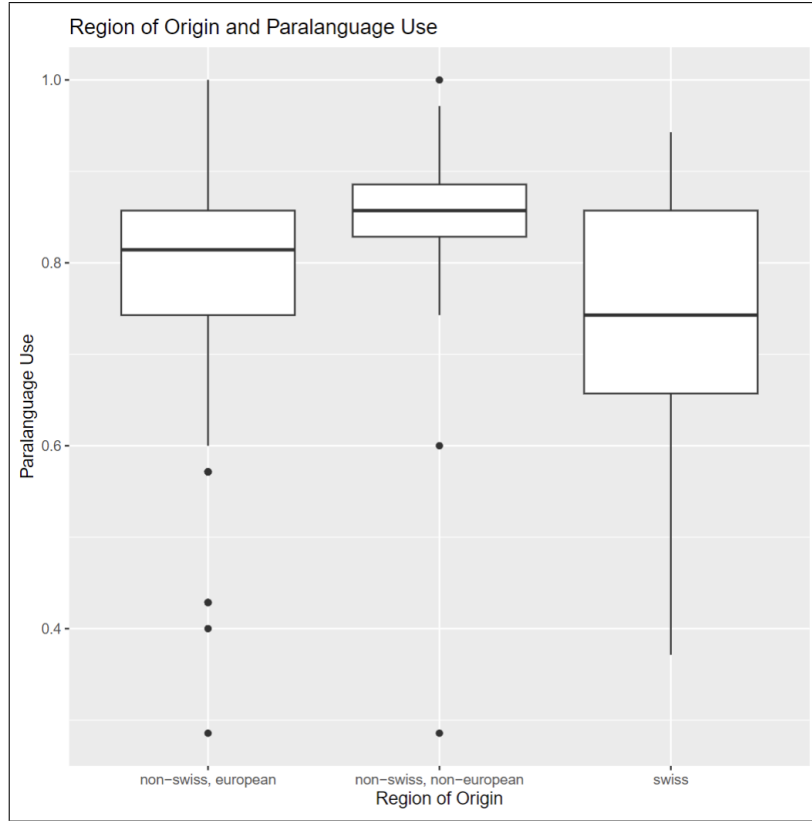


Figure 8: *Region of origin and paralanguage use.*

Contrastingly, when I analysed the performance of more general communication activities (including written, audio and visual modes), there was a significant relationship between age and usage. The correlation coefficient for age and communication output was $r = -0.26$, with a minute p-value of $p < 0.01$, suggesting perhaps that there are differences in the chosen method of communication via digital technology, depending on age. This study is focused on the potential for paralanguage output via computer-mediated environments, and therefore will not consider this question extensively. However, further research could be undertaken to identify whether younger demographics employ, for instance, non-written forms of communication more than older demographics.

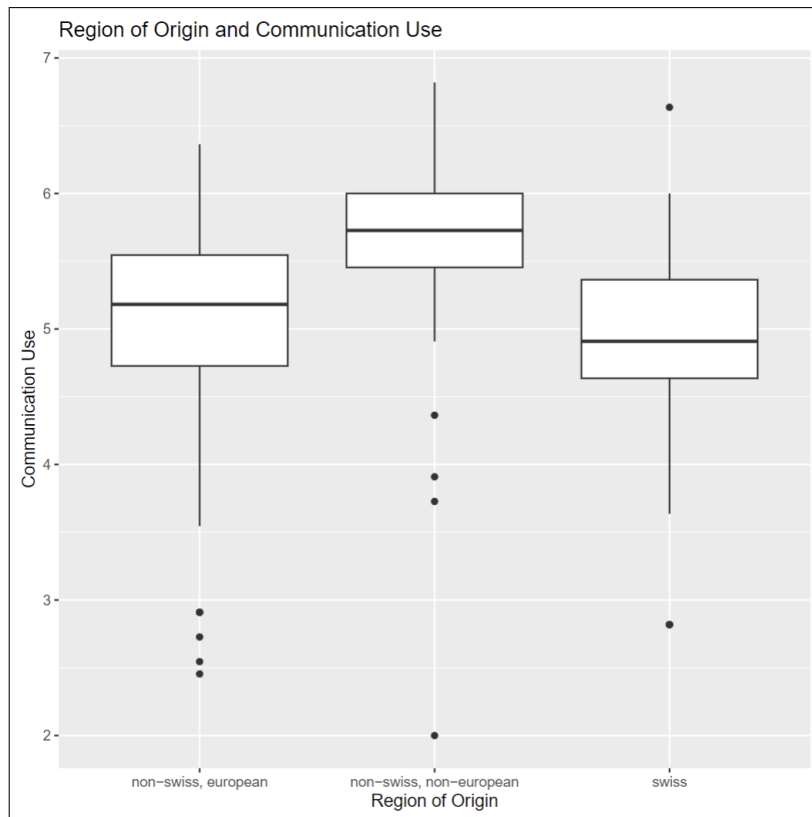


Figure 9: *Region of origin and communication use.*

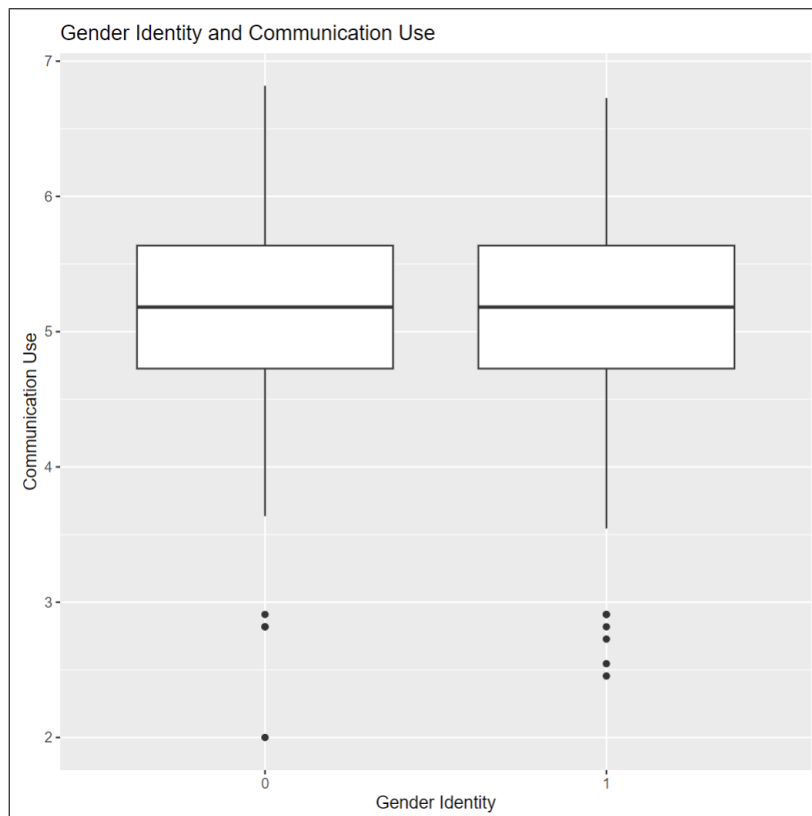


Figure 10: *Gender identity and communication use.*

The box-plot for communication use cases (*see Figure 8*) resembles that of par-
alanguage use cases, with non-Europeans engaging in communicative activities more.
However, there appears to be more similarity between non-Swiss Europeans and Swiss
respondents here.

Gender identity and level of education do not have a significant effect on the perfor-
mance of communicative activities. In fact, the box-plot for gender identity (*see Figure*
9) demonstrates a high level of similarity in terms of practices. Both the median (av-
erage) and IQR (distribution) of male and female respondents were similar, centred and
concentrated slightly above a “Weekly” usage, as per the original Likert scale.

The next stage was to consider whether different demographic groups engaged in
instrumental and expressive uses of digital technology to differing extents, since previous
theoretical work would suggest a potential gender gap when considering these categories
(Gui & Gerosa, 2021; Kelan, 2007; Kim & Hargittai, 2021; Kirkup, 1992; Lai & Katz,
2012; van Deursen & van Dijk, 2014).

Gender identity did not have a significant relationship with instrumental use. Level
of education was not significantly related with instrumental use either, countering the
expectations set up by previous studies that suggested that males and those with a
higher educational level would engage in instrumental uses of digital technology more.

However, I found a negative correlation of $r = -0.20$ between age and instrumental
use; a significant relationship given its associated p-value of $p = 0.01$, suggesting that
younger demographics engage in higher instrumental than older demographics.

The ANOVA result for region of origin was again significant, offering a p-value of
 $p < 0.01$. As such, regional groups in the dataset exhibited different levels of instrumental
use. Again, non-Europeans tended to use digital technology instrumentally more.

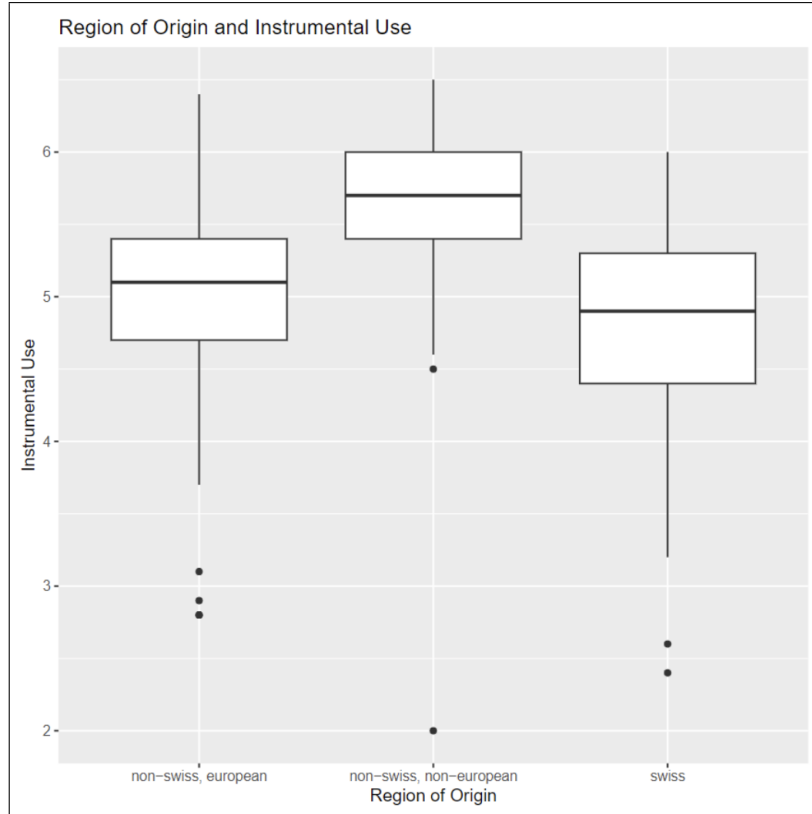


Figure 11: *Region of origin and instrumental use.*

The second aspect of this question was whether certain groups engage in expressive forms of digital activities more than others, with the theory suggesting that female respondents might display a higher level of expressive use than male respondents.

Again, there is no significant difference in terms of gender identity or level of education when it comes to engaging more or less extensively with more expressive digital activities.

Similarly to previous results, age is negatively correlated ($r = -0.23$) with a more extensive performance of expressive activities. Moreover, years of Internet usage also produces a negative correlation of $r = -0.17$. Both results are significant, with age and years of Internet usage producing p-values of $p < 0.01$ and $p = 0.02$, respectively. These results suggest more concretely that age relates to the amount of use of digital technology in general, since these variables reflect both frequency and diversity of use. Thus, those who are not digital natives engage in instrumental and expressive uses of digital technology less.

In future studies, a score could be calculated from the performed activities variables,

thus removing the frequency of usage and instead focusing on the number of different uses within these 2 categories.

Again, regional origin does have a significant relationship (with a p-value of $p < 0.01$) with the extent to which expressive activities are performed, with non-Europeans performing expressive activities more than Europeans.

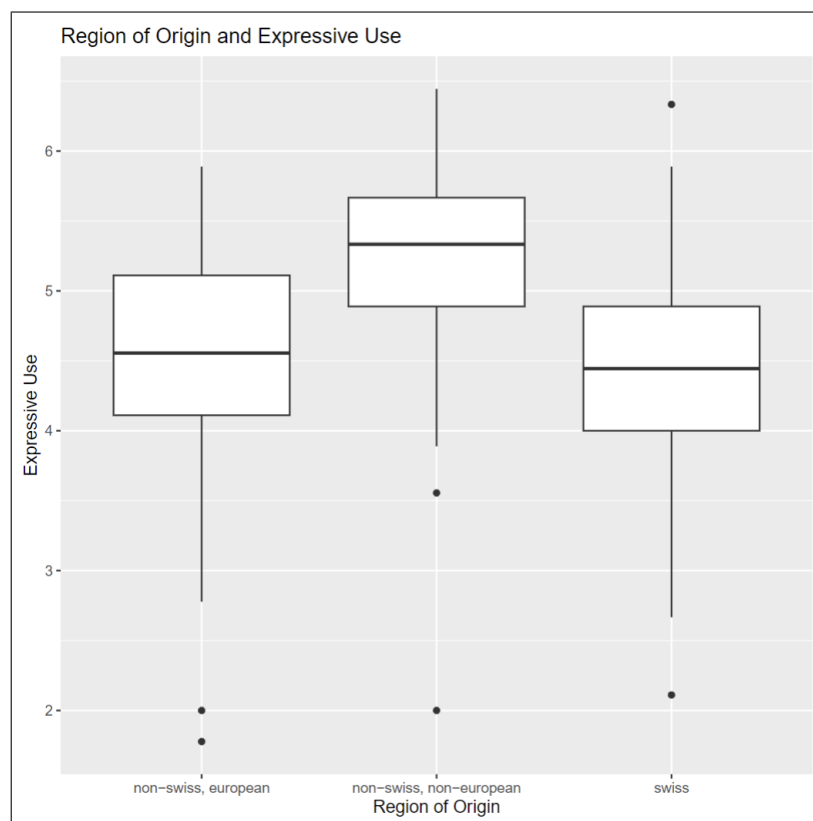


Figure 12: *Region of origin and expressive use.*

One specific activity performed using digital technology in which paralanguage is employed extensively is via social media platforms. Previous research has suggested that different demographic groups engage in such networked environments more or less frequently (Bimber, 2000; Redmiles & Buntain, 2021). In this sample, gender identity did not have a significant effect on frequency of engagement with social media; gender did not entail greater or lesser communion through computer-mediated environments. In addition, region did not significantly affect engagement with social media.

In terms of age and years of Internet usage, significant results were produced. Both age and years of Internet usage exhibited negative correlations ($r = -0.24$ and $r = -0.20$)

of a significant nature in terms of their p-values ($p < 0.01$ and $p = 0.01$). This result suggests that younger demographics, as per previous studies, engage in social media more frequently than older demographics (Bimber, 2000; Redmiles & Buntain, 2021).

ANOVA also highlighted the significance of level of education in terms of social media usage, with a p-value of $p = 0.02$. However, on inspection of the following box-plot (see Figure 14), it should be noted that this relationship might not be linear in nature; both compulsorily educated and highly educated individuals exhibit a tendency to use social media less frequently than moderately educated individuals, even though the median is similar for most groups, at 6 – or “Daily” usage – and without much variability at all.

It might be conjectured that both highly and less educated respondents may not have the luxury of time to dedicate towards extensive social media usage. Further studies should consider this question in greater detail.

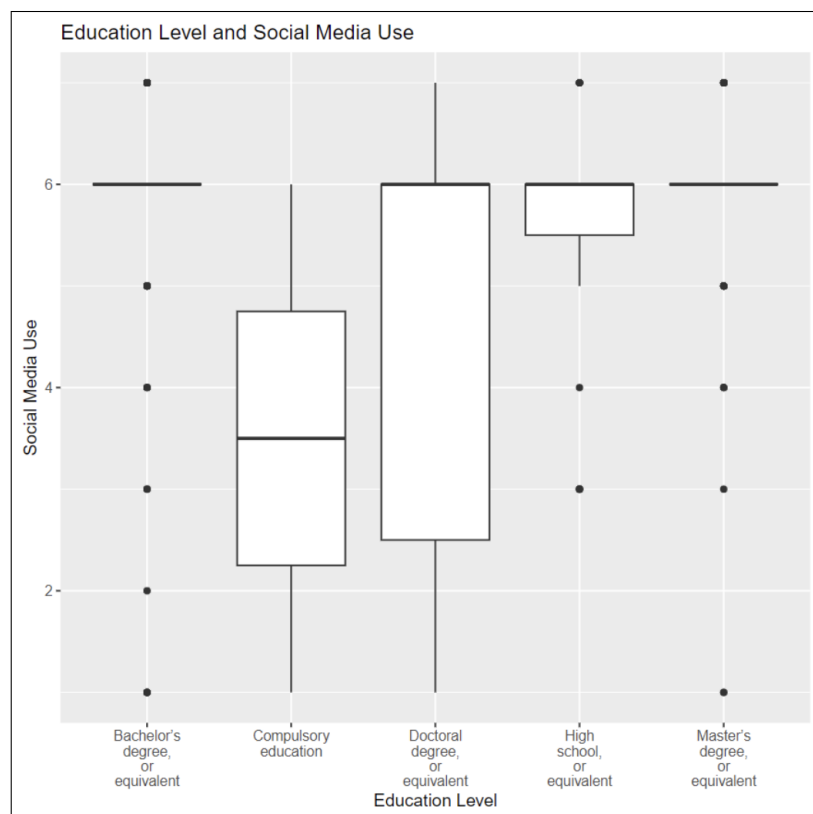


Figure 14: *Education level and social media use.*

There is no significant difference between different demographics in terms of performing activities for self-development via digital technology, despite research suggesting that

Males and highly educated people might engage more frequently in such activities (Gui & Gerosa, 2021; Kim & Hargittai, 2021).

Although previous research has suggested that lower educated, younger individuals were more likely to perform entertainment-related activities via digital technology, this study found no relationship between education and digital entertainment activities (Gui & Gerosa, 2021; van Deursen & van Dijk, 2014). It did, however, find a relationship between age and such usage.

Firstly, there was a coefficient of $r = -0.39$ and an associated p-value of $p < 0.01$ between age and entertainment usage, suggesting that age is significantly and negatively correlated with using digital technology for entertainment purposes. In other words, the older one is, the less one uses digital technology for entertainment. Similarly, years of Internet usage was significantly (a p-value of $p = 0.01$) and negatively (a coefficient of $r = -0.18$) correlated with using digital technology for entertainment; those who have used the Internet for longer, use it for entertainment less. This result underscores the differences in use between digital natives and those born before the Internet.

Secondly, when considering the performance of two activities – browsing websites and playing games – age demonstrated significant, negative correlation, with a coefficient of $r = -0.30$ and a p-value of $p < 0.01$, reinforcing this finding. No other demographic factors were significant here in determining entertainment-related activities.

6.3.4 Familiarity with sociolinguistics and research

To be able to better judge who would be more receptive to contributing to research, it is valuable to consider what groups have contributed to research already. On the one hand, age, years of Internet usage, region of origin and gender identity do not yield significant results when it comes to differences in research contributions. On the other hand, level of education is revealed through a chi-square test to be significantly (with a p-value of $p = 0.01$) associated with previous research contributions, be it in terms of data, time or skill. This finding is in line with research into citizen science, which suggests that more highly educated individuals are more likely to contribute to research projects – a

particular problem in terms of engaging diverse participants (Phillips et al., 2018).

However, it is also reasonable that increased engagement in research, particularly at master's or doctoral degree level, is more likely due to the fact that both of these educational stages require the conduct of and participation in research as part and parcel of a programme. Of course, then, it is to be expected that research contributions increase at these levels of education.

I considered it valuable to ascertain whether there was a relationship between familiarity with sociolinguistics and familiarity with citizen science, since scientific literacy was so central to engagement with research. Indeed, a linear regression approach demonstrated that there was a strong positive relationship between these two variables, with a p-value of $p < 0.01$ and an R-squared value that indicates that familiarity with citizen science explains 23.59% of the variance in sociolinguistic terminology. This result indicates that knowledge of the field of sociolinguistics is significantly related to knowledge of citizen science practices.

Moreover, familiarity with citizen science is related to previous participation in research. Linear regression highlights that familiarity with citizen science has a significant ($p = 0.02$ positive relationship with previous participation in research. This result could suggest the success of learning outcomes in terms of participation in research; educational gains do arise from participating in research, as indicated by research (Kloetzer et al., 2021; Phillips et al., 2018).

6.3.5 Data concerns and algorithmic literacy

Previous studies have suggested that demographic criteria, especially in terms of gender, age, and SES, have an effect on an individual's trust in data processes, as well as their confidence in taking privacy measures online (Büchi et al., 2021; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021).

I did not find any significant relationships between age, years of Internet usage, or gender identity and the respondent's level of concern in terms of data practices. However, level of education did demonstrate its importance in this context, with a significant p-

value of $p = 0.04$. This result suggests that level of education is related to one's level of concern with how one's data is handled when participating in research. Whilst this insight is valuable, the dataset is rather imbalanced and does not represent well those who have just completed compulsory education; more research is needed to ascertain whether this result is replicable in more representative datasets.

As per previous studies, region of origin is also significant (with a p-value of $p = 0.03$) in terms of concern with the treatment of data (Büchi et al., 2021; Redmiles & Buntain, 2021; Reisdorf & Blank, 2021). This difference might be related to the fact that regulations for data practices, such as the GDPR, have been more extensively developed within a European context (Lynn et al., 2019; Tauginienė et al., 2021).

The following box-plots (*see Figures 12 and 13*) highlight how lower education and a non-European background might relate to higher concern for one's data, and higher education might relate to higher levels of trust.

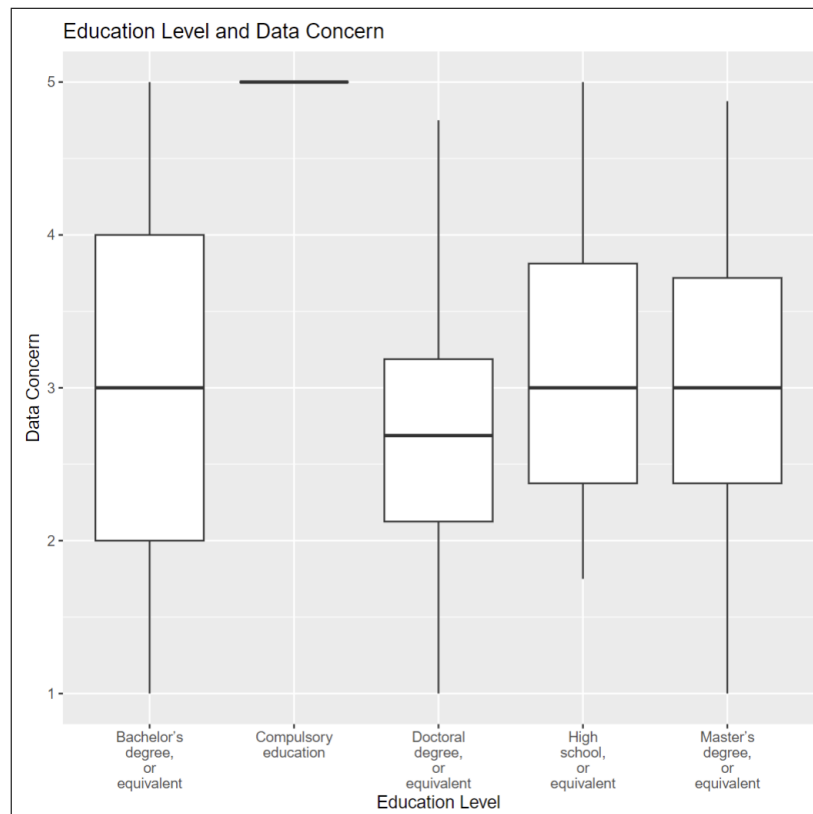


Figure 13: *Education level and data concern.*

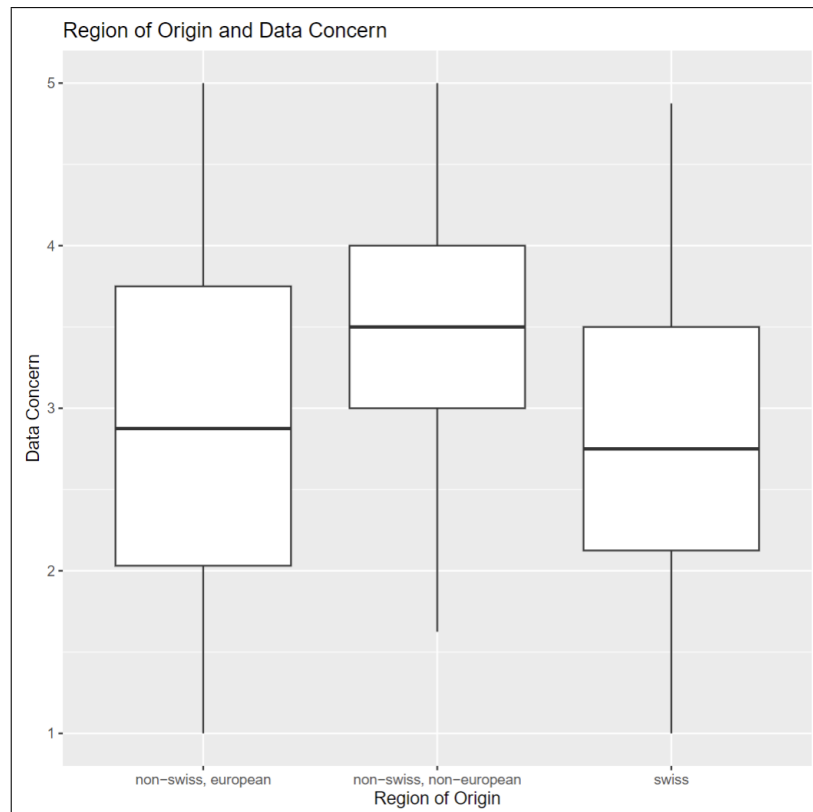


Figure 15: *Region of origin and data concern.*

I then explored the relationship between both frequency and diversity of digital technology usage and concern with the handling of data in a research context, to understand whether the characteristics of a digital citizen impact their perceived risks when contributing data to research. Indeed, previous studies suggested that engaging more in digital environments might entail a greater level of naivety in terms of data mishandling (Redmiles & Buntain, 2021).

Using linear regression, there was no significant result in either the case of frequency of use or that of diversity. In fact, there was almost no relationship between frequency of use and level of concern, suggesting that engaging in digital environments more had no impact at all on one's level of concern. This result might reflect Foucault's paradigm, whereby the visibility of surveillance practices has become so normalised that people's increased concern about them is perceived as futile (Foucault, 1975). There is a slight negative relationship between diversity of use and level of concern, suggesting that increased diversity might be related to decreased concern. However, the results were insignificant and, thus, inconclusive.

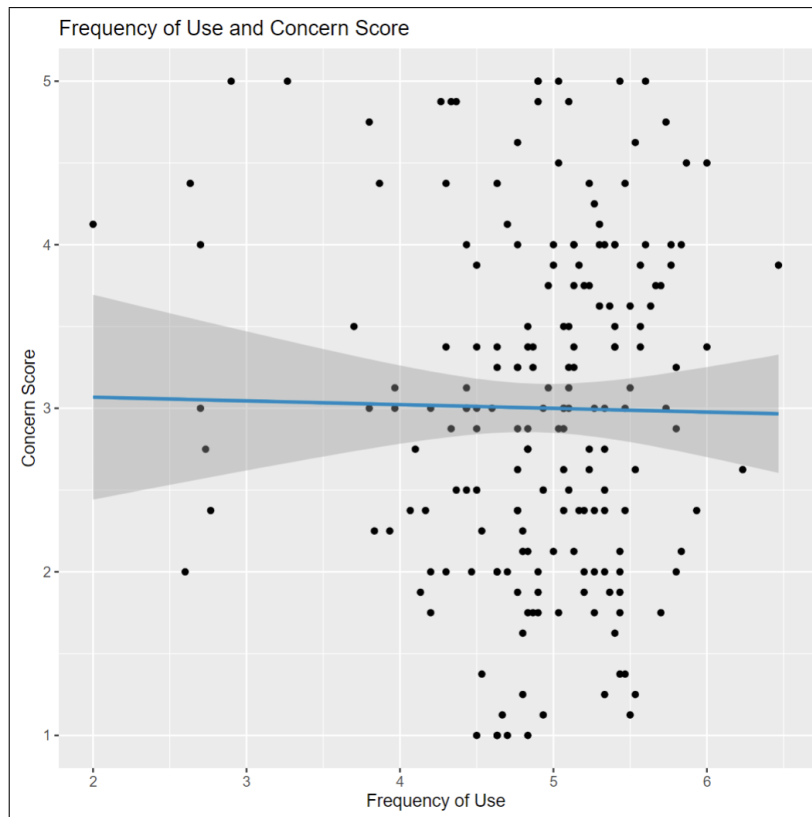


Figure 16: *Frequency of use and concern score.*

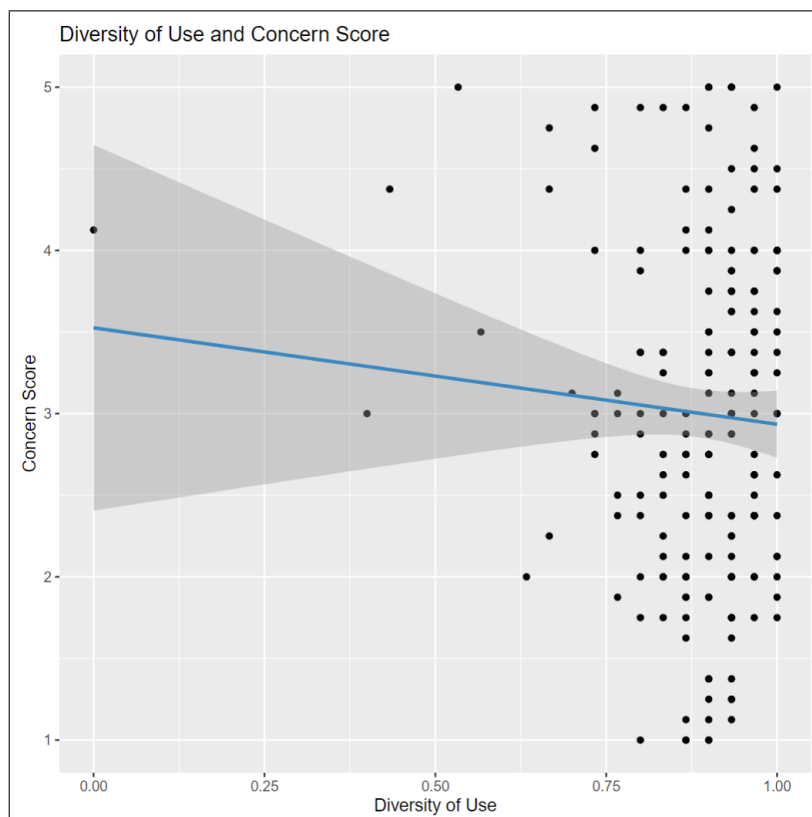


Figure 17: *Diversity of use and concern score.*

Algorithmic literacy has been introduced as a key facet of digital literacy in recent studies, which have highlighted that increased competency is related to an increased understanding of how algorithms work on one’s data (Reisdorf & Blank, 2021). This finding was replicated in this study, which convincingly demonstrated that high algorithmic literacy is positively related to high (self-assessed) competency. Linear regression produced a p-value of $p < 0.01$. Moreover, the R-squared value highlighted how 27.03% of the variance in digital competency is explained by the algorithmic literacy variable. It is thus critical to include algorithmic literacy in future studies of digital literacy. Future studies should explore alternative methods of measuring actual digital competency and actual understanding of algorithms, to reinforce the significance of this result.

6.3.6 Motivations for research engagement

One of the key motivations for contributing to research that was represented in prior studies was an interest in acknowledgement, either in citations or as co-authorship in publications (Alender, 2016; Curtis, 2015; de Vries et al., 2019). This incentive was particularly popular among younger demographics, who were expected to seek more career-advancement than older demographics (Alender, 2016; de Vries et al., 2019). This study did not produce any significant results in this regard. None of the demographic criteria analysed betrayed a relationship of any kind with this particular incentive, either as acknowledgement in citations on its own, or as acknowledgement in general. Gains in terms of reputation did not appear to be important to this sample.

Contrastingly, “having a more active role in the project” was a significant motivator in terms of interest in a mobile application for data collection. A chi-square test produced highly valuable results, with a p-value of $p = 0.01$. This result suggests that the possibility of greater involvement is highly motivating when it comes to engagement with a future data collection application.

This finding might be related to the niche subgroups that constitute the sample; responses were elicited from two populations (SLI and HN) who might be expected to exhibit a have greater interest, both in UNIL-based projects and in the field of program-

ming (similarly, respondents from within FLE may have a greater awareness of linguistic concepts than the general population). As such, contributing to such a project might indeed be a considerable form of career advancement for these subgroups, since it could form part of their portfolio. This point should be recognised if the development of such an application is undertaken in the future. Collaborative application development could both contribute to educational goals within the UNIL and support data collection efforts.

Regardless, this finding supports that of previous studies in citizen science, which suggest that opportunities for greater collaboration are highly motivational (Arnstein, 1969; Land-Zandstra et al., 2021).

Another key motivation for engaging in research as stated in previous studies is learning a skill. In the context of this study, there were no significant differences in terms of who was interested in learning a skill as a motivation for contributing data to research.

In terms of more egotistical and more collective-based motivations for engaging in research, education level and region of origin did not play significant roles in determining these categories. Age and years of Internet usage were significantly and negatively correlated with both egotistical ($r = -0.31$ with a p-value of $p < 0.01$; $r = -0.18$ with a p-value of $p = 0.02$) and collective-based ($r = -0.30$ with a p-value of $p < 0.01$; $r = -0.26$ with a p-value of $p < 0.01$). This result is particularly interesting, since older demographics who have used the Internet for longer are even less interested in collective-based motivations than egotistical ones, perhaps suggesting how the development of Web 2.0 has contributed to the blossoming of participatory cultures (Jenkins et al., 2016).

Given that gender identity has played such a minor role in determining the variables under consideration in this study, it was fascinating to observe that gender identity does indeed enjoy a significant relationship ($p = 0.03$) with collective-based motivations for participating in research. Although there is no evidence to support the duality between agency on the one hand and communion on the other, it is clear that female respondents were more motivated by communion than their male counterparts, continuing to reflect their increased interest in participatory activities (Bakan, 1966).

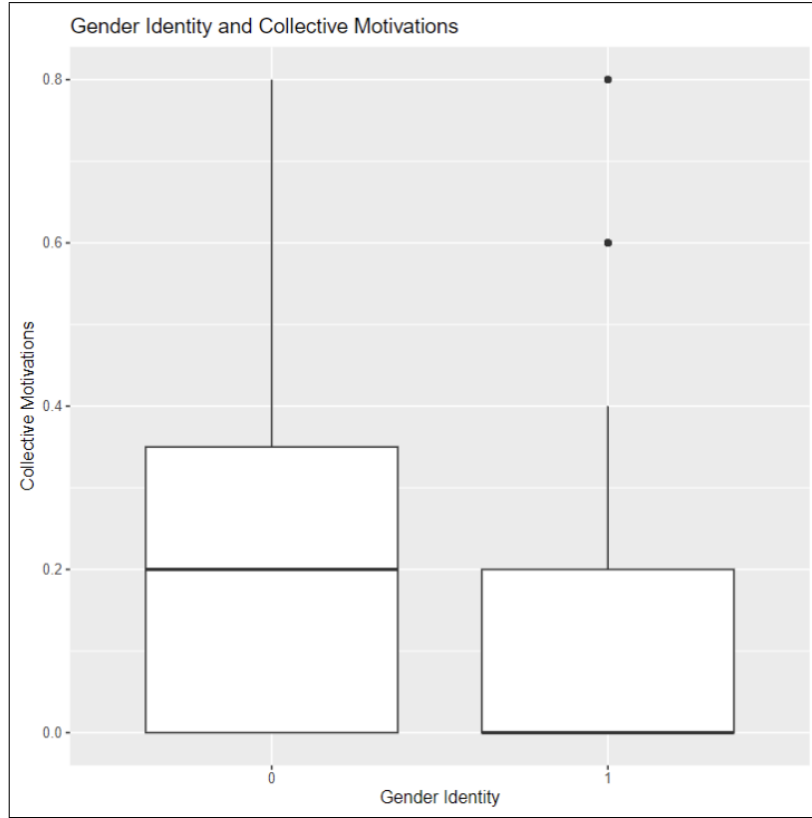


Figure 18: *Gender identity and collective motivations.*

Using logistic regression to trace the relationship between the binary variable representing motivation by competitive aspects and the continuous variable representing (past and present) research interest, I discovered that research interest is related to this motivation. A p-value of $p < 0.05$ exhibits the significance of this relationship. In other words, greater research interest is related to greater motivation by gamification aspects. It could be argued, then, that this result bolsters previous studies which suggest that gamification does more to sustain than to attract participants; those who are already interested in research are more motivated by the prospect of competition (Aristeidou & Herodotu, 2020; de Vries et al., 2019; Iacovides et al., 2013; Kloetzer et al., 2013).

6.3.7 Interest in a future mobile application

In terms of uncovering the demographic who would be most interested in a mobile application, the results of this study were largely inconclusive, with no significant relationship demonstrated between application interest and the variables gender identity, level of ed-

education, age, and years of Internet usage. However, I discovered a significant relationship between region of origin and interest in a mobile application using a chi-square test. As such, non-Swiss Europeans were less interested in such a data-collecting application, Swiss respondents were ambivalent, and non-Europeans displayed far greater interest. This finding could be useful to inform future studies into CMC that intend to gather data; non-Europeans might be more eager to engage in such research, particularly in an application format. The p-value for this test was $p = 0.01$, underscoring its significance.

There was no significant relationship between use of social media and interest in an application, something prior research suggested, since experience with social media can entail experience with mobile applications as a whole (Aristeidou & Herodotu, 2020; Arnstein, 1969; Lemmens et al., 2021; Mazumdar et al., 2018).

Based on previous studies, it is expected that scientific literacy and pre-existing knowledge of the subject matter of a research project are related to increased interest in research participation (Jennett et al., 2016; Kloetzer et al., 2021; Paleco et al., 2021; Phillips et al., 2018; Trumbull & Bonney, 2000). This suggestion was substantiated in my study; linear regression indicated a positive relationship between familiarity with sociolinguistics and interest in a mobile application. This result was significant, with a p-value of $p = 0.01$. Hence, familiarity with the subject of research does encourage interest in participation. However, the R-squared value highlights that familiarity with sociolinguistics only explains 3.53% of the variance in interest.

Similarly, previous research findings indicate that familiarity with citizen science affects interest in research participation. Again, such findings are only reinforced by this study; linear regression demonstrates a positive relationship between interest in a mobile application and familiarity with the term citizen science. Again, this familiarity is self-assessed and it would be valuable to critique how much respondents actually know (i.e., their real level of familiarity). Even so, this relationship is also significant, with a p-value of $p < 0.01$, and a higher R-squared value than the previous result. Indeed, interest in a mobile application explains 5.51% of the variance in familiarity with citizen science.

Similarly, the results of a chi-square test produce significant results (a p-value of

$p = 0.03$) in terms of an association between previous research participation and interest in an application to collect data for research. Respondents who had previously engaged in research are more interested in engaging with such a mobile application. This result again demonstrates the association between existing familiarity with research processes (i.e., scientific literacy) and possible future engagement.

According to scholars, hesitancy in terms of data handling practices degrade individuals' trust in data collection and, thus, their interest in engaging in such activities for the purposes of research (Bowser et al., 2020; Hartzog & Selinger, 2013a, 2013b; Kim & Hargittai, 2021; Kimm & Boase, 2021; Redmiles & Buntain, 2021; Tauginienė et al., 2021). In this study, no significant relationship was uncovered between interest in a data collection application, for research purposes, and concern, suggesting that if all concerns were met as promised, this hesitancy would not continue to affect individual's use of such an application. Moreover, research interest on the whole (based on both past research contributions and current application interest, with equal weighting) is equally unaffected by one's level of concern.

Although competitive aspects on the whole did not appear to motivate respondents' initial contribution to a research project, the result of a chi-square test between the possibility of gamification and interest in a mobile application for research-focused data collection yielded borderline results. At a significance level of $\alpha < 0.05$, this motivation was judged insignificant, since its p-value was $p = 0.05$. Regardless of this result, it is important to note that there is a slight association between interest in gamification aspects and interest in a future application. This question should be broached further in order to prepare for the correct design and development of a future application.

It was necessary to engage in further analysis in order to build a profile of the type of digital citizen who would be interested in contributing to a mobile application, since this basis was a large component of this study. With this necessity in mind, I outputted the correlation matrix between interest in an application and the list of activities performed online. From this matrix, I manually judged the correlations of the highest magnitude, creating the following longlist:

- Using calculators or currency converters ($r = 0.18$)
- Using AI tools ($r = 0.17$)
- Sending voice messages or making voice calls ($r = 0.16$)
- Watching or making short reel videos ($r = 0.15$)
- Making video or conference calls ($r = 0.14$)
- Using calendars or reminders ($r = 0.13$)
- Making lists ($r = 0.12$)
- Taking pictures or videos ($r = 0.11$).

I then performed logistic regression between application interest and these 8 activities, to create a shortlist of 4 key activities that had significant ($p < 0.05$) positive relationships with interest in a mobile application:

- Using calculators or currency converters ($p = 0.02$)
- Using AI tools ($p = 0.02$)
- Sending voice messages or making voice calls ($p = 0.03$)
- Watching or making short reel videos ($p = 0.03$).

It appears that high use of existing applications (calculators, currency converters, and AI tools, as well as short video services such as TikTok and Instagram) is related to interest in a research-based application. These results were echoed by general research interest. This result parallels previous research which suggests that existing experience with applications is associated with interest in digital citizen science (Lemmens et al., 2021; Mazumdar et al., 2018). A high textual paralinguage output is not associated with interest in an application.

Thus, future research projects in this area could consider ways of engaging individuals with the above profile. For instance, more practical features could be added, artificial

intelligence could be integrated into the application, and short explicatory videos (rather than written descriptions) could be made to further advance learning outcomes from the project. Advertisement of the project could focus on audiovisual communication to optimise engagement. Opportunities for participants to create their own audiovisual output could also serve both to engage them in the project and to promote its research.

Lastly, I considered how the various motivations for contributing data to research corresponded with interest in a mobile application, using several iterations of chi-square tests, since I had converted each motivation into a binary variable, coded 0 by default and 1 if the respondent ticked the box for this motivation. Most of these motivations contributed significantly ($p < 0.05$) to interest in an application, with “learning a skill”, “learning about science or research”, and “learning about myself” being the most significant. As it should, ticking the box for “none of the above” motivations was also strongly associated with a lack of interest in an application, with a p-value of $p < 0.01$. The only 3 motivations which did not yield significant results were “networking opportunities”, “the possibility to share with friends and family” and “competitive aspects”, although these motivations were also bordering on significance.

6.4 Evaluation

Whilst, overall, I am satisfied with the results obtained from this study, there are several points that I would improve in future research. I offer these insights as both reflections on my own study and opportunities for others to build upon what I have achieved. These reflections take three parts. Firstly, I will consider the methodological constraints of my study. Secondly, I will consider the analytical weaknesses that I encountered. Lastly, I will give some suggestions for adapting my study for further research.

6.4.1 Methodological limitations

The largest methodological limitation that I experienced was due to the populations from which I drew responses. I encountered a distinct lack of responsivity from these populations. Moreover, the populations I chose, from within specific subsets of the UNIL community and from within my own personal network introduced considerable bias into this study. Even though this method was employed due to necessity, future studies should attempt to garner a more representative and balanced dataset, especially in terms of educational and subject background. In particular, there was a lack of respondents who had followed only a compulsory education. Additionally, it is evident that the concepts of compulsory and higher education might overlap in an international context; as such, taking more care to phrase this question would have alleviated any confusion. It would be preferable to be more explicit by referring to pre-university studies.

Also, a more proactive strategy might be employed in approaching individuals directly within the UNIL campus.

Had I been aware in advance that I would encounter such difficulty in obtaining responses from the UNIL community, I would have framed the survey differently, to make it more internationally oriented, since some of the questions focused on a Swiss context which might not have been accessible for a non-Swiss respondent.

A further practical limitation of this survey was the format it took. Google Forms was utilised out of convenience, given that access to my chosen software, LimeSurvey would take a minimum of three weeks to obtain. In fact, I still have not obtained the

said access, 2 months on.

An alternative software may have circumvented issues identified already during the pilot study stage, regarding the way that lengthy Likert-type questions would displayed to respondents. It was difficult for some individuals to follow, especially in the question regarding motivations. On reflection, I should have split this question into parts for the sake of readability.

6.4.2 Survey limitations

In terms of the content of the survey, there are several aspects that I would improve upon reiterating the study. A discussion with Sonia Petrini, a PhD candidate at the UNIL, offered some particular insights regarding the survey.

Firstly, there was no feedback section to the survey. Although this aspect did not become clear in the pilot study, citizen science research has highlighted how integral it is to maintain two-way communication in the preparatory stages of a project (de Vries et al., 2019; Rüfenacht et al., 2021).

Secondly, the question regarding what would motivate respondents to share their data with research, was not explicit as to whether this data sharing with voluntary. In the context of this question, I intended to refer to informed, voluntary and consensual data sharing. However, Petrini was totally justified in suggesting that this question might be perceived as including more exploitative data sharing, such as cookies.

The question regarding whether or not respondents had already contributed data, time or skills to research was redundant, given the following question regarding the organisations to which respondents had contributed. Whether or not respondents had indeed contributed could be gathered from this second question instead, streamlining the survey.

I would alter the 3-point Likert scale for familiarity with certain terminology to a 5-point Likert scale, ranging from “Not at all familiar” to “Extremely familiar”, offering more valuable data in terms of prior knowledge (Hargittai, 2002, 2005, 2009)

Similarly, I would change the multiple-choice question on data-related concerns to

another 5-point Likert scale, so that more valuable insights could be achieved during data analysis.

Furthermore, the question on what would motivate contributions to research needed fine-tuning. For instance, there is a considerable difference between not being interested in any of the motivations offered and not being motivated at all. As such, I would divide the negative option “None of the above” into two options: “Nothing would motivate me” and “None of the above options motivate me”. I would also frame the question pertaining to motivations through the model of categorisation provided by previous research into motivation in citizen science in order to be more thorough in the kinds of motivation for which I account (Land-Zandstra et al., 2021; Rotman et al., 2012).

In particular, I did not account directly for the motivation of making a contribution to science in my survey; a great flaw considering how important this motivation has been demonstrated to be in previous research (Bowser et al., 2020; Land-Zandstra et al., 2021; Robinson et al., 2018). Additionally, I would add more positive motivations for engaging in citizen science, such as pure recreational enjoyment (Land-Zandstra et al., 2021; Rotman et al., 2012). Also, research has outlined more extensive ways that participants can be acknowledged for their contributions (Alender, 2016; Curtis, 2015; de Vries et al., 2019).

It is critical in the context of citizen science to understand what would motivate participation beyond data contributions; as such, I would alter the question about motivations to refer instead to contribution to research projects more broadly.

A final addition that I would make to the survey would be to include job searching in the list of activities performed via digital technology; an item that was omitted despite its inclusion in previous research (Correa et al., 2021; van Deursen & Helsper, 2015).

6.4.3 Analytical limitations

Two further limitations were revealed during the data analysis stage. The first limitation was that the sum of performed activities did not differentiate well between the sheer amount of use of digital technologies and the diversity of uses. Whilst I circumvented

this issue by creating my own diversity score, I was still dissatisfied with the results I could garner in terms of frequency. This issue was echoed in the categorisations I created for different kinds of use profiles. The second limitation is that there is a distinct difference between confidence in engaging with digital environments (trust in oneself) and confidence in the digital environments themselves (trust in the environments). It is critical to consider this distinction in future research, in order to make valuable interpretations of both of these phenomena.

6.4.4 Future recommendations

It is impossible to include a conclusive list of recommendations for further research in this area, since my study broached many valuable questions.

A more extensive study could engage in a more rigorous digital skills assessment, after the measures successfully developed by Hargittai, to remove a large part of the bias caused by self-assessment (Hargittai, 2002, 2005, 2009). It could also engage individuals in a more interview- or comment-based survey, to gather more extensive feedback pertaining to concerns, motivations and familiarity.

Similarly, a more practical assessment of educational level could be prepared, such as a measurement of logical or comprehension abilities. A more thorough study could even create more reliable indicators of SES, such as employment type.

Moreover, a larger and more collaborative study could provide a question that asks respondents to identify whether they recognise and employ certain popular examples of paralinguage or interjection in CMC.

In terms of research questions that were beyond the scope of this study, I have identified several that would be valuable to consider:

- How does knowledge of surveillance affect participation in digital environments?
- Who considered themselves to still be learning how to use digital technologies?
- Do individuals consider it empowering to play a more active role in research?
- What do individuals expect from a citizen science project?
- Do individuals consider themselves to be benefit from digital technologies?
- What are the differences between attracting and sustaining participation in sociolinguistic research?
- Is there a non-linear relationship between the amount of use of digital technology and level of education?

7 Conclusion

To conclude, I intend to respond to the three research questions outlined at the outset of this study:

1. Who is the target audience?
2. What would help and hinder engagement?
3. What do people already know?

7.1 Who is the target audience?

Overall, this study emphasised a high frequency of use in terms of instant messaging, sending text messages and emailing, three categories of critical importance to research into textual paralinguistics in CMC. Furthermore, a high frequency of written communication was exhibited overall.

It found that people from different regions of origin engage in digital technology very differently. As such, non-Europeans tend to use digital technology more frequently than non-Swiss Europeans and Swiss respondents. However, that does not mean to say that non-Europeans use technology more diversely.

Interestingly, Swiss respondents display a lower level of digital literacy overall - in regards to frequency, diversity and competency of use, suggesting that they might not be the most optimal group to engage in such research, since they might produce less paralinguistics output.

This study indicated that age, gender identity and educational background did not relate to the potential production of paralinguistics output. Region of origin, on the other hand, did relate to the production of paralinguistics; non-Europeans engaged more frequently with use cases that could involve paralinguistics. Swiss respondents engaged the least with paralinguistics-related uses of digital technology. This finding suggests that conducting Swiss-centric studies into CMC may not be the most productive choice.

As for engagement in social media, I found that younger demographics did tend to

engage more. However, engagement in social media may not always mean the creation of textual paralinguage, since many platforms now follow a more audio-visual model.

In terms of the audience who is most interested in contributing to research, this study demonstrated that a higher level of education relates to a higher interest in research. Moreover, individuals with more knowledge of sociolinguistics and citizen science are more interested in contributing to research.

When it comes to interest in a future mobile application, I discovered that non-Europeans are the most interested. Equally, individuals who had previously engaged with research are also more interested in the prospect of a research-based application.

As for the profile of digital citizen who is most interested in such an application, it seems that those who use such tools as calculators, currency converters and AI tools, who use voice features, or who engage with short entertainment videos, are the most interested in an application. These findings provide a kind of framework to inspire the design of such a project in the future, by integrating related aspects into its conception.

7.2 What would help and hinder engagement?

Overall, there appeared to be a moderate understanding and concern for data processes.

This finding might not necessarily hinder engagement in future research projects exactly, but it should nonetheless be taken into account, especially since CMC research involves the handling of sensitive data.

The highest concerns that respondents had concerning data were that their personal information would be exposed, their data misused and their identity left unprotected. These concerns should be taken seriously in future research projects; textual communication is easily identifiable. Furthermore, concerns over data did not affect interest in research or interest in an application.

Even though respondents were only moderately concerned with data practices, this finding should not deter researchers from pursuing best practices.

There was a relative lack of concern over using one's data over a long time. This finding is particularly encouraging in the context of corpus-building projects that require

long-term access to data.

The study reflected a low level of interest in research participation overall and the potential motivations included did not incentivise a third of the sample.

The most attractive motivations related to learning overall, and especially learning about science or research. Competitive aspects, acknowledgement in citations, co-authorship, networking opportunities and the possibility to share with one's existing network were not as motivational in terms of encouraging engagement with research or an application. Furthermore, having a more active role in a project was highly attractive to respondents, suggesting that the possibility of greater collaboration would be valuable.

Interestingly, female respondents appeared to be more motivated by collective-based motivations for participating than male respondents.

Although interest in an application was tepid overall, region of origin and previous participation in research were both critical in determining its audience. **For non-Europeans, who had higher interest in an application and higher concern for their data, pursuing rigorous data practices would be essential for helping engagement.**

For those who have already engaged in research, gamification is found to be slightly more attractive, suggesting perhaps that it is more effective in sustaining participation rather than encouraging initial participation. Moreover, people who are more interested in an application are also more interested in gamification aspects, so it is important not to discredit competitive aspects to motivating engagement entirely.

7.3 What do people already know?

On average, respondents recognised but did not understand key terminology relative to sociolinguistics. To aid engagement, then, it would be equally important to consider this familiarity threshold as a basis for attracting participants. In particular, paralanguage was the least understood of these terms. Since it is also the most central to the research projects to which this study seeks to contribute, this insight is valuable.

Familiarity with the term citizen science was also moderate, with individuals recognising but not understanding it. As a result, it is necessary to integrate understandings

of citizen science into future outreach as further contextualisation, particularly given that playing a more active role was so motivating.

In conclusion, I have offered a valuable framework for identifying the target audience for future CMC-oriented research projects (non-Europeans; previous research participants). Furthermore, I have identified what helps and hinders such engagement (learning; having an active role) and what people already know (they recognise but do not understand sociolinguistic terminology). I hope that, in responding to these three key research questions, I provide some useful insights to optimise engagement with future research projects, especially with an eye to the development of a mobile application.

8 Bibliography

References

- Albert, A., Balázs, B., Butkevičienė, E., Mayer, K., & Perelló, J. (2021). Citizen social science: New and established approaches to participation in social research. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Alender, B. (2016). Understanding volunteer motivations to participate in citizen science projects: A deeper look at water quality monitoring. *Journal of Science Communication*, 15(3), 1–19.
- Ameka, F. (1992). Interjections: The universal yet neglected part of speech. *Journal of Pragmatics*, 18, 101–118.
- Anderson, D. (2015). A question of trust: Report of the investigatory powers review. *Independent Review of Terrorism Legislation*.
- Aristeidou, M., & Herodotu, C. (2020). Online citizen science: A systematic review of effects on learning and scientific literacy. *Citizen Science: Theory and Practice*, 5(11), 1–12.
- Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Planning Association*, 35(4), 216–224.
- Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion*. Rand McNally.
- Bakir, V. (2015). “Veillant panoptic assemblage”: Mutual watching and resistance to mass surveillance after Snowden. *Media and Communication*, 3(3), 12–25.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioural change. *Psychological Review*, 84(2), 191–215.
- Bauman, R., & Briggs, C. L. (1990). Poetics and performance as critical perspectives on language and social life. *Annual Review of Anthropology*, 19, 59–88.
- Baym, N. K. (1995). The performance of humour in computer-mediated communication. *Journal of Computer-Mediated Communication*, 1(2), 1–33.

- Bentham, J. (1791). *Panopticon; or, the inspection house*. T. Payne.
- Bimber, B. (2000). Measuring the gender gap on the Internet. *Social Science Quarterly*, 81(3).
- Birner, B. J. (2013). *Introduction to pragmatics*. Wiley-Blackwell.
- Boamah, O. (2024). What are Progressive Web Apps? pwa guide for beginners. *freeCodeCamp*.
- Bowser, A., Cooper, C., de Sherbinin, A., Wiggins, A., Brenton, P., Chuang, T.-R., Faustman, E., Haklay, M., & Meloche, M. (2020). Still in need of norms: The state of the data in citizen science. *Citizen Science: Theory and Practice*, 5(1), 1–16.
- Büchi, M., Festic, N., Just, N., & Latzer, M. (2021). Digital inequalities in online privacy protection: Effects of age, education and gender. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Büchi, M., & Latzer, M. (2016). Modelling the second-level digital divide: A five-country study of social differences in Internet use. *New Media & Society*, 18(11), 2703–2722.
- Bullen, M., & Morgan, T. (2011). Digital learners not digital natives. *La Cuestión Universitaria*, 7, 60–68.
- Bullen, M., Morgan, T., & Qayyum, A. (2011). Digital learners in higher education: Generation is not the issue. *Canadian Journal of Learning and Technology*, 37(1), 1–24.
- Carey, J. (1980). Paralanguage in computer mediated communication. *Association for Computational Linguistics*, 67–69.
- Cavalier, D., Hoffman, C., & Cooper, C. (2020). *The field guide to citizen science: How you can contribute to scientific research and make a difference*. Timber Press.
- Christozov, D., & Toleva-Stoimenova, S. (2015). Big data literacy: A new dimension of digital divide, barriers in learning via exploring "Big Data". In K. Berg, J. Girard, & D. Klein (Eds.), *Strategic data-based wisdom in the Big Data era*. Information Science Reference.

- Cooper, C., Shanley, L., Scassa, T., & Vayena, E. (2019). Project categories to guide institutional oversight of responsible conduct of scientists leading citizen science in the United States. *Citizen Science: Theory and Practice*, 4(1), 1–9.
- Correa, T., Pavez, I., & Contreras, J. (2021). Digital inequality and mobiles: Opportunities and challenges of relying on smartphones for digital inclusion in disadvantages contexts. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Coşkunserçe, O., & Aydoğdu, Ş. (2022). Investigating the digital skills of undergraduate students in terms of various variables. *Journal of Educational Technology and Online Learning*, 5(4), 1219–1238.
- Curtis, V. (2015). *Online citizen science projects: An exploration of motivation, contribution and participation* [Doctoral dissertation, The Open University].
- de Vries, M., Land-Zandstra, A., & Smeets, I. (2019). Citizen scientists’ preferences for communication of scientific output: A literature review. *Citizen Science: Theory and Practice*, 4(1), 1–13.
- Dean, J. (2018). Sorted for memes and GIFs: Visual media and everyday digital politics. *Political Studies Review*, 17(3).
- December, J. (1997). Notes on defining of Computer-Mediated Communication. *Computer-Mediated Communication Magazine*, 3, 1–14.
- Doudot, L. (2021). Étude ethnographique de 500,000 messages WhatsApp : Marqueurs paralinguistiques et expression d’émotions dans le corpus *What’s New, Switzerland ? Mémoire de maîtrise universitaire interfacultaire en humanités numérique*, 1–105.
- Durndell, A., & Haag, Z. (2002). Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an Eastern European sample. *Computers in Human Behaviour*, 18, 521–535.
- Eagly, A. H. (1987). *Sex differences in social behaviour: A social-role interpretation*. Psychology Press.

- Eleta, I., Clavell, G. G., Righi, V., & Balestrini, M. (2019). The promise of participation and decision-making power in citizen science. *Citizen Science: Theory and Practice*, 1–9.
- Elliott, K. C., & Rosenberg, J. (2019). Philosophical foundations for citizen science. *Citizen Science: Theory and Practice*, 4(1), 1–9.
- EPFL. (2020). Dhelta UNIL-EPFL. <https://www.epfl.ch/schools/cdh/education-2/dh-master/registered-students/dhelta/>.
- Foucault, M. (1975). *Surveiller et punir: Naissance de la prison*. Gallimard.
- Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Vintage Books.
- Gallardo-Echenique, E. E., Marqués-Molías, L., Bullen, M., & Strijbos, J.-W. (2015). Let’s talk about digital learners in the digital era. *International Review of Research in Open and Distributed Learning*, 16(3), 156–187.
- Ganzevoort, W., van den Born, R. J. G., Halffman, W., & Turnhout, S. (2017). Sharing biodiversity data: Citizen scientists’ concerns and motivations. *Biodiversity Conservation*, 28, 2821–2837.
- Gloria, A. M., Hird, J. S., & Navarro, R. L. (2001). Relationships of cultural congruity and perceptions of the university environment to helps-seeking attitudes by sociorace and gender. *Journal of College Student Development*, 42(6), 545–562.
- Goffman, E. (1956). *The presentation of self in everyday life*. University of Edinburgh.
- Goffman, E. (1974). *Frame analysis: An essay on the organisation of experience*. Northeastern University Press.
- Goffman, E. (1981). *Forms of talk*. University of Pennsylvania Press.
- Grey, F. (2011). Citizen cyberscience: The new age of the amateur. *CERN Courier*, 41–43.
- Gui, M., & Gerosa, T. (2021). Smartphone pervasiveness in youth daily life as a new form of digital inequality. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.

- Gumperz, J. J., & Cook-Gumperz, J. (1982). Introduction: Language and the communication of social identity. In J. J. Gumperz (Ed.), *Language and social identity*. Cambridge University Press.
- Haklay, M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., & Vohland, K. (2021). What is citizen science? The challenges of definition. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Hargittai, E. (2002). Second-level digital divide: Differences in people’s online skills. *First Monday*, 7(4), 1–20.
- Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social Science Computer Review*, 23(3), 371–379.
- Hargittai, E. (2009). An update on survey measures of web-oriented digital literacy. *Social Sciences Computer Review*, 27(1), 130–137.
- Hargittai, E. (Ed.). (2021). *Handbook of digital inequality*. Edward Elgar Publishing.
- Hargittai, E., & Shafer, S. (2006). Differences in actual and perceived online skills: The role of gender. *Social Science Quarterly*, 87(2).
- Hartzog, W., & Selinger, E. (2013a). Big Data in small hands. *Privacy and Big Data*, 66.
- Hartzog, W., & Selinger, E. (2013b). Obscurity: A better way to think about your data than ‘privacy’. *The Atlantic*.
- Hecker, S., Garbe, L., & Bonn, A. (2018). The European citizen science landscape – a snapshot. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science: Innovation in open science, society and policy*. UCL Press.
- Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., & Bonn, A. (Eds.). (2018). *Citizen science: Innovation in open science, society and policy*. UCL Press.
- Hilchenbach, B. (2023). Progressive web apps (pwa). *Tech Radar*.
- Hine, C. (2000). *Virtual ethnography*. Sage Publications.
- Hintz, A., Dencik, L., & Wahl-Jorgensen, K. (2019). *Digital citizenship in a datafied society*. Polity Press.

- Iacovides, I., Jennett, C., Cornish-Trestrail, C., & Cox, A. L. (2013). Do games attract or sustain engagement in citizen science? A study of volunteer motivations. *CHI*, 1–7.
- Irwin, A. (1995). *Citizen science: A study of people, expertise and sustainable development*. Routledge.
- Jenkins, H., & Clinton, K. (2006). Confronting the challenges of participatory culture: Media education for the 21st century. *Building the Field of Digital Media and Learning*, 1–62.
- Jenkins, H., Ito, M., & Boyd, D. (2016). *Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics*. Polity Press.
- Jennett, C., Kloetzer, L., Schneider, D., Iacovides, I., Cox, A. L., Gold, M., Fuchs, B., Eveleigh, A., Mathieu, K., Ajani, Z., & Tals, Y. (2016). Motivations, learning and creativity in online citizen science. *Journal of Science Communication*, 15(3), 1–23.
- Jones-Kavalier, B., & Flannigan, S. (2006). Connecting the digital dots: Literacy of the 21st century. *Educause Quarterly*, 2, 8–10.
- Kelan, E. (2007). Tools and toys: Communicating gendered positions towards technology. *Information Communication and Society*, 10(3), 357–382.
- Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10), 1123–34.
- Kiesler, S., & Sproull, L. (1992). Group decision making and communication technology. *Organisational Behaviour and Human Decision Processes*, 52, 96–123.
- Kieslinger, B., Schäfer, T., Heigl, F., Dörler, D., Richter, A., & Bonn, A. (2018). Evaluating citizen science: Towards an open framework. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science: Innovation in open science, society and policy*. UCL Press.
- Kim, S. J., & Hargittai, E. (2021). Looking back at millennials’ mobile transitions: Differentiated patterns of mobile phone use among a diverse group of young adults. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.

- Kimm, J., & Boase, J. (2021). Mobile media in teen life: Information, networks and access. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Kirkup, G. (1992). The social construction of computers: Hammers or harpsichords? In G. Kirkup & L. S. Keller (Eds.), *Inventing women: Science, technology and gender*. Polity Press.
- Kloetzer, L., Lorke, J., Roche, J., Golumbic, Y., Winter, S., & Jõgeva, A. (2021). Learning in citizen science. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Kloetzer, L., Schneider, D., Jennett, C., Iacovides, I., Eveleigh, A., Cox, A., & Gold, M. (2013). Learning by volunteer computing, thinking and gaming: What and how are volunteers learning by participating in virtual citizen science? *Changing Configurations of Adult Education in Transitional Times*, 73–92.
- Kucharska, W. (2021). Analysis of nonverbal cues based on *What's up, Switzerland?* corpus: The semiotics of emotion expression. *Mémoire de maîtrise universitaire interfacultaire en humanités numérique*, 1–105.
- Kullenberg, C., & Kasperowski, D. (2016). What is citizen science? – A scientometric meta-analysis. *PLoS ONE*, 11(1), 1–16.
- Lai, C.-H., & Katz, J. E. (2012). Are we evolved to live with mobiles? An evolutionary view of mobile communication. *Social and Management Sciences*, 20(1), 45–54.
- Land-Zandstra, A., Agnello, G., & Gültekin, Y. S. (2021). Participants in citizen science. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Lausanne, B. (2024). Renouvaud. <https://www.bcu-lausanne.ch/>.
- Lemmens, R., Antoniou, V., Hummer, P., & Potsiou, C. (2021). Citizen science in the digital world of apps. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.

- Lewandowski, E., Caldwell, W., Elmquist, D., & Oberhauser, K. (2017). Public perception of citizen science. *Citizen Science: Theory and Practice*, 2(1), 1–9.
- Luangrath, A. W., Peck, J., & Barger, V. A. (2017). Textual paralanguage and its implications for marketing communications. *Journal of Consumer Psychology*, 27(1), 98–107.
- Lynn, S. J., Kaplan, N., Newman, S., Scarpino, R., & Newman, G. (2019). Designing a platform for ethical citizen science: A case study of citsci.org. *Citizen Science: Theory and Practice*, 4(1), 1–15.
- Madden, M., Gilman, M., Levy, K., & Marwick, A. (2017). Privacy, poverty, and Big Data: A matrix of vulnerabilities for poor Americans. *Washington University Law Review*, 95(1), 53–125.
- Martek, A., Mucnjak, D., & Mumelaš, D. (2022). Citizen science in Europe — Challenges in conducting citizen science activities in cooperation of university and public libraries. *Publications*, 10(52), 1–15.
- Mazumdar, S., Ceccaroni, L., Piera, J., Hölker, F., Berre, A. J., Arlinghaus, R., & Bowser, A. (2018). Citizen science technologies and new opportunities for participation. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science: Innovation in open science, society and policy*. UCL Press.
- Mossberger, K., Tolbert, C. J., & McNeal, R. S. (2008). *Digital citizenship: The Internet, society, and participation*. Penguin Random House.
- Nuessle, T. M., McNamara, P. A., & Garneau, N. L. (2020). Planning and executing scientifically sound community science in a public-facing institution. *Citizen Science: Theory and Practice*, 5(1).
- Ono, H., & Zavodny, M. (2003). Gender and the Internet. *Social Science Quarterly*, 84(1), 111–121.
- Paleco, C., Peter, S. G., Seoane, N. S., Kaufmann, J., & Argyri, P. (2021). Inclusiveness and diversity in citizen science. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.

- Pandya, R. E. (2012). A framework for engaging diverse communities in citizen science in the us. *Frontiers in Ecology and the Environment*, 10(6), 314–317.
- Park, Y. J. (2015). Do men and women differ in privacy? Gendered privacy and (in)equality in the internet. *Computers in Human Behaviour*, 50, 252–258.
- Park, Y. J. (2021). Why privacy matters to digital inequality. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Payne, P. (2023). *Paralanguage Encoder* : Detection et recodage d’éléments paralinguistiques avec Python. *Projet en Informatique pour les sciences humaines et sociales*, 1–9.
- Phillips, T., Porticella, N., Conostas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in citizen science. *Citizen Science: Theory and Practice*, 3(2), 1–19.
- Pokhriyal, N., Nwogu, I., & Govindaraju, V. (2015). A large-scale study of language usage as a cognitive biometric trait. In V. Govindaraju, V. V. Raghavan, & C. R. Rao (Eds.), *Big Data analytics*. Elsevier.
- Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5), 1–6.
- Redmiles, E. M., & Buntain, C. L. J. (2021). How feelings of trust, concern, and control of personal online data influence web use. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Rees, H., & Noyes, J. M. (2007). Mobile telephones, computers, and the Internet: Sex differences in adolescents’ use and attitudes. *CyberPsychology & Behaviour*, 10(3), 482–484.
- Reisdorf, B. C., & Blank, G. (2021). Algorithmic literacy and platform trust. In E. Hargittai (Ed.), *Handbook of digital inequality*. Edward Elgar Publishing.
- Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., & Ansine, J. (2018). Ten principles of citizen science. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, & A. Bonn (Eds.), *Citizen science: Innovation in open science, society and policy*. UCL Press.

- Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., & Jacobs, D. (2012). Dynamic changes in motivation in collaborative citizen-science projects. *Conference Paper*, 1–10.
- Rudnicka, A., Gould, S., & Cox, A. (2022). Citizen scientists are not just quiz takers: Information about project type influences data disclosure in online psychological surveys. *Citizen Science: Theory and Practice*, 7(1), 1–13.
- Rüfenacht, S., Woods, T., Agnello, G., Gold, M., Hummer, P., Land-Zandstra, A., & Sieber, A. (2021). Communication and dissemination in citizen science. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Sharifian, F. (2009). *English as an international language: Perspectives and pedagogical issues*. MPG Books.
- Stapleton, A. (2024). PhDs? Number of people with doctoral degree. *Academia Insider*.
- Straub, M. C. P. (2016). Giving citizen scientists a chance: A study of volunteer-led scientific discovery. *Citizen Science: Theory and Practice*, 1(1), 1–10.
- Suman, A. B., & Pierce, R. (2018). Challenges for citizen science and the EU open science agenda under the GDPR. *European Data Protection Law Review*, 4(3), 284–295.
- Tauginienè, L., Hummer, P., Albert, A., Cigarini, A., & Vohland, K. (2021). Ethical challenges and dynamic informed consent. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, 28(3), 317–348.
- Tillotson-Chavez, K., & Weber, J. (2024). A new generation of citizen scientists: Self-efficacy and skill growth in a voluntary project applied in the college classroom setting. *Citizen Science: Theory and Practice*, 9(1), 1–18.
- Toft, J., Fore, L., Hass, T., Bennett, B., Brubaker, L., Brubaker, D., Rice, C., & Island County Beach Watchers. (2017). A framework to analyse citizen science data for

- volunteers, managers, and scientists. *Citizen Science: Theory and Practice*, 2(1), 1–11.
- Trumbull, D. J., & Bonney, R. (2000). Thinking scientifically during participation in a citizen-science project. *Science Education*, 265–275.
- Ueberwasser, S., & Stark, E. (2017). What’s up Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5), 105–126.
- UNIL. (2023). Nombre d’équivalents plein temps (ept) par nationalité principale. *Collaboratrices et collaborateurs de l’UNIL*.
- UNIL. (2024a). Nationalité par continent. *Répartition des étudiantes et étudiants par pays*.
- UNIL. (2024b). Pays de provenance. *Pays de provenance et destination des étudiantes et étudiants*.
- Upton, G., & Cook, I. (2014). *Oxford dictionary of statistics*. Oxford University Press.
- van Deursen, A. J. A. M., & Helsper, E. J. (2015). The third-level digital divide: Who benefits most from being online? *Communication and Information Technologies*, 10, 29–53.
- van Deursen, A. J. A. M., Helsper, E. J., & Eynon, R. (2014). Measuring digital skills: From digital skills to tangible outcomes. *Project Report*, 1–48.
- van Deursen, A. J. A. M., & van Dijk, J. A. G. M. (2014). The digital divide shifts to differences in usage. *New Media & Society*, 16(3), 507–526.
- van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, 12(2).
- Vohland, K., Göbel, C., Balázs, B., Butkevičienė, E., Daskolia, M., Duží, B., Hecker, S., Manzoni, M., & Schade, S. (2021). Citizen science in Europe. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The science of citizen science*. Springer.
- Warin, C., & Delaney, N. (2020). Citizen science and citizen engagement. *Achievements in Horizon 2020 and recommendations on the way forward*, 1–32.



- Wellman, B., & Gulia, M. (1999). Virtual communities as communities: Net surfers don't ride alone. In M. A. Smith & P. Kollock (Eds.), *Communities in cyberspace*. Routledge.
- West, S., & Pateman, R. (2016). Recruiting and retaining participants in citizen science: What can be learned from the volunteering literature? *Citizen Science: Theory and Practice*, 1(2), 1–10.
- Wharton, T. (2003). Interjections, language and the showing-saying continuum. *Pragmatics and Cognition*, 11(1), 39–91.


9 Annex

9.1 English survey

Questionnaire: Understanding Digital Citizens

MA / HN / ISH / Autumn 2024 / Public



 Not shared

* Indicates required question

Introduction

This questionnaire explores the links between demographic characteristics and digital literacy, as well as the motivations and concerns of potential participants in sociolinguistic research projects. The goal is to maximise future participation and, consequently, the effectiveness of this research.

Please be assured that all your responses will be treated with complete confidentiality as part of this academic study. Your data will only be used for research purposes and will remain anonymous.

This questionnaire will take approximately 5 to 12 minutes to complete.

Section 1

1. Are you currently or have you ever been a student or staff member at the University of Lausanne? *

- ☐ Yes
- ☐ No

2. What is your gender identity? *

- ☐ Female
- ☐ Male
- ☐ Prefer not to say
- ☐ Other: _____

3. What is the highest level of education you have completed? *

- ☐ Compulsory education
- ☐ High school, or equivalent
- ☐ Bachelor's degree, or equivalent
- ☐ Master's degree, or equivalent
- ☐ Doctoral degree, or equivalent
- ☐ Prefer not to say
- ☐ Other: _____

4. What is your region of origin? *

Choose ▼

5. What is your age?

Your answer _____

Section 2

6. For approximately how many years have you used the internet? *

- ☐ Less than 1 year
- ☐ 1-2 years
- ☐ 3-5 years
- ☐ 6-10 years
- ☐ 11-15 years
- ☐ 15-20 years
- ☐ More than 20 years

7. How much do you perform the following activities on a computer, tablet (including e-readers) or mobile phone? *

	Never	Less than once a year	Yearly	Monthly	Weekly	Daily	Hourly
Searching information	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Making video or conference calls	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Learning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using GPS or navigation tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Emailing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Editing photos or videos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Watching films, TV or videos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Watching or making short reel videos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Live streaming	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sending or sharing photos or videos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Using calculators or currency converters	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using AI tools (chatbots, writing assistants, translators)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Setting alarms or timers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Browsing websites	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sending text or multimedia messages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using instant messaging applications	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Making lists	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Listening to music, podcasts or audiobooks	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Checking the time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reading books, blogs or articles	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Making GIFs or memes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sharing links or files	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Playing games	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participating in online forums or group chats	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sending voice messages or making voice calls	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Banking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using social media platforms	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shopping (objects, groceries, digital content, subscriptions)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Taking photos or videos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Using calendars or reminders	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

8. As a whole, how competent do you consider yourself in using digital tools and services (*i.e. the activities mentioned in question 7*)? *

1 2 3 4 5
 Not competent at all ☐ ☐ ☐ ☐ ☐ Extremely competent

9. How well do you understand how algorithms are used on your data online? ** *

1 2 3 4 5
 Do not understand at all ☐ ☐ ☐ ☐ ☐ Fully understand

Section 3

10. Have you ever contributed data, time, or skills to a research project? *

☐ Yes

☐ No

11. If you have contributed to a research project, to which of the following organisations have you contributed? *

☐ A university or educational institution (ex. University of Lausanne)

☐ A hospital or medical research centre (ex. CHUV)

☐ A government organisation (ex. Federal Statistical Office)

☐ A charity or non-government organisation (ex. Red Cross)

☐ A commercial or for-profit company (ex. Google)

☐ I have not contributed to a research project.

☐ Other: _____

Section 4

12. How concerned are you about the following when sharing data of online communications with research? *** *

	Not concerned at all	Slightly concerned	Moderately concerned	Very concerned	Extremely concerned
My personal information will be exposed.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My data will be stored incorrectly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My data will be used outside of research.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My data will be misused.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My identity will not be protected.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I will not have control over my data.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My data will be used for a long time.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Researchers will not be transparent.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

13. What would motivate you to share data of online communications with research? *

I would be motivated by...

- ☐ .. none of the above.
- ☐ ... learning about myself.
- ☐ ... having a more active role in the project.
- ☐ ... networking opportunities.
- ☐ ... the possibility to share with friends and family.
- ☐ ... co-authorship in publications.
- ☐ ... learning about science or research.
- ☐ ... acknowledgement in citations.
- ☐ ... competitive aspects (gamification).
- ☐ ... learning a skill.
- ☐ ... financial compensation (money or vouchers).

14. Would you be interested and comfortable in using a mobile application that facilitates the collection and analysis of your online communication data, if your concerns and motivations were met? *

- ☐ Yes
- ☐ No

Section 5

15. How familiar are you with the following terms? *

	Don't recognise and don't understand	Recognise but don't understand	Recognise and understand
Computer-Mediated Communication	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parts of Speech	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Paralanguage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interjections	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Citizen Science	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

References

* Kim, S. J., & Hargittai, E. (2021). Looking back at millennials' mobile transitions: Differentiated patterns of mobile phone use among a diverse group of young adults. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 114-130). Edward Elgar Publishing.

** Reisdorf, B. C., & Blank, G. (2021). Algorithmic literacy and platform trust. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 341-357). Edward Elgar Publishing.

*** Redmiles, E. M., & Buntain, C. L. J. (2021). How feelings of trust, concern, and control of personal online data influences web use. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 311-325). Edward Elgar Publishing.

Submit

Clear form

9.2 French survey

Questionnaire : Comprendre les *digital citizens*

MA / HN / ISH / Automne 2024

Not shared

* Indicates required question

Introduction

Ce questionnaire explore les liens entre les caractéristiques démographiques et la littératie numérique, ainsi que les motivations et préoccupations des participants potentiels aux projets de recherche en sociolinguistique. L'objectif est de maximiser la participation future et, par conséquent, l'efficacité de ces recherches.

Soyez assuré.e que toutes vos réponses seront traitées de façon entièrement confidentielles dans le cadre de cette étude universitaire. Vos données ne seront utilisées qu'à des fins de recherche et resteront anonyme.

Ce questionnaire prendra environ 5 à 12 minutes.

Section 1

1. Êtes-vous actuellement ou avez-vous déjà été étudiant.e ou membre du personnel à l'Université de Lausanne ? *

- ☐ Oui
- ☐ Non

2. Quelle est votre identité de genre ? *

- ☐ Féminin
- ☐ Masculin
- ☐ Je ne souhaite pas le préciser
- ☐ Other: _____

3. Quel est le plus haut niveau d'éducation que vous avez atteint ? *

- ☐ Éducation obligatoire terminée
- ☐ Diplôme de maturité, ou équivalent
- ☐ Licence, ou équivalent
- ☐ Master, ou équivalent
- ☐ Doctorat, ou équivalent
- ☐ Je ne souhaite pas le préciser
- ☐ Other: _____

4. Quelle est votre région d'origine ? *

Choose ▼

5. Quel est votre âge ?

Your answer _____

Section 2

6. Depuis environ combien d'années utilisez-vous Internet ? *

- ☐ Moins d'un an
- ☐ 1 à 2 ans
- ☐ 3 à 5 ans
- ☐ 6 à 10 ans
- ☐ 11 à 15 ans
- ☐ 15 à 20 ans
- ☐ Plus de 20 ans

7. À quelle fréquence réalisez-vous les activités suivantes sur un ordinateur, une tablette (y compris les liseuses électroniques) ou un téléphone mobile ? *

	Jamais	Moins d'une fois par an	Annuellement	Mensuellement	Hebdomadairement	Quotidiennement	Toutes les heures
Envoi de messages texte ou multimédia	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utilisation d'applications de messagerie instantanée	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Envoi de messages vocaux ou appels vocaux	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Appels vidéo ou visioconférences	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Envoi de courriels	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Partage de liens ou de fichiers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participation à des forums ou des chats de groupe en ligne	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utilisation de plateformes de médias sociaux	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jouer à des jeux	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Regarder des films, des émissions de télévision ou des vidéos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Regarder ou créer des vidéos courtes (reels)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Écouter de la musique, des podcasts ou des livres audios	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Écouter de la musique, des podcasts ou des livres audios	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lire des livres, des blogs ou des articles	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Rechercher des informations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Naviguer sur des sites web	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Faire du shopping (objets, produits d'épicerie, contenu numérique, abonnements, etc.)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Effectuer des opérations bancaires	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utiliser des outils GPS ou de navigation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Apprendre	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Prendre des photos ou des vidéos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Éditer des photos ou des vidéos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Envoyer ou partager des photos ou des vidéos	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Faire du streaming en direct (live streaming)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Créer des GIFs ou des mèmes	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Régler des alarmes ou des minuteurs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Utiliser des calendriers ou des rappels	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Consulter l'heure ☐ ☐ ☐ ☐ ☐ ☐ ☐

Faire des listes ☐ ☐ ☐ ☐ ☐ ☐ ☐

Utiliser des
calculatrices ou
des
convertisseurs
de devises ☐ ☐ ☐ ☐ ☐ ☐ ☐

Utilisation
d'outils d'IA
(chatbots,
assistants
d'écriture,
traducteurs) ☐ ☐ ☐ ☐ ☐ ☐ ☐



8. Dans l'ensemble, comment évaluez-vous vos compétences dans l'utilisation des outils et services numériques (*étant donné les activités mentionnées dans la question 7*) ? *

1 2 3 4 5

Pas compétent du tout ☐ ☐ ☐ ☐ ☐ Extrêmement compétent

9. Dans quelle mesure comprenez-vous comment les algorithmes sont utilisés sur vos données en ligne ? ** *

1 2 3 4 5

Ne comprends pas du tout ☐ ☐ ☐ ☐ ☐ Comprends parfaitement

Section 3

10. Avez-vous déjà contribué des données, des interprétations ou des compétences à un projet de recherche ? *

- ☐ Oui
- ☐ Non

11. Si vous avez contribué à un projet de recherche, à laquelle des options suivantes avez-vous contribué ? *

- ☐ Une université ou une institution éducative (ex. Université de Lausanne)
- ☐ Un hôpital ou un centre de recherche médicale (ex. CHUV)
- ☐ Une organisation gouvernementale (ex. Office fédéral de la statistique)
- ☐ Une organisation caritative ou non gouvernementale (ex. Croix-Rouge)
- ☐ Une entreprise commerciale ou à but lucratif (ex. Google)
- ☐ Je n'ai pas contribué à un projet de recherche.
- ☐ Other: _____

Section 4

12. À quel point êtes-vous préoccupé.e par les éléments suivants lorsque vous partagez des données, concernant vos communications en ligne, avec la recherche ? *** *

	Pas préoccupé.e du tout	Légèrement préoccupé.e	Modérément préoccupé.e	Très préoccupé.e	Extrêmement préoccupé.e
Mes informations personnelles seront exposées.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mes données seront stockées de manière incorrecte.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mes données seront utilisées en dehors de la recherche.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mes données seront mal utilisées.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mon identité ne sera pas protégée.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Je n'aurai pas de contrôle sur mes données.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mes données seront utilisées pendant longtemps.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Les chercheurs ne seront pas transparents.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

13. Qu'est-ce qui vous motiverait à partager des données de communications en ligne avec la recherche ? *

Je serais motivé.e par...

- ☐ ... avoir un rôle plus actif dans le projet.
- ☐ ... apprendre sur la science ou la recherche.
- ☐ ... aspects compétitifs (gamification).
- ☐ ... aucune des réponses ci-dessus.
- ☐ ... la possibilité de partager avec des ami.e.s et la famille.
- ☐ ... apprendre une compétence.
- ☐ ... compensation financière (argent ou bons).
- ☐ ... co-auteur dans les publications.
- ☐ ... opportunités de mise en réseau.
- ☐ ... apprendre sur moi-même.
- ☐ ... reconnaissance dans les citations.

14. Seriez-vous intéressé.e et à l'aise d'utiliser une application mobile qui facilite la collecte et l'analyse de vos données de communication en ligne, si vos préoccupations et motivations étaient satisfaites ? *

- ☐ Oui
- ☐ Non

Section 5

15. Quelle est votre familiarité avec les termes suivants ? *

	Ne reconnaît pas et ne comprend pas	Reconnaît mais ne comprend pas	Reconnaît et comprend
Communication médiée par ordinateur	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parties du discours	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Paralangage	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interjections	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Science citoyenne (Citizen Science)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Sources

* Kim, S. J., & Hargittai, E. (2021). Looking back at millennials' mobile transitions: Differentiated patterns of mobile phone use among a diverse group of young adults. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 114-130). Edward Elgar Publishing.

** Reisdorf, B. C., & Blank, G. (2021). Algorithmic literacy and platform trust. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 341-357). Edward Elgar Publishing.

*** Redmiles, E. M., & Buntain, C. L. J. (2021). How feelings of trust, concern, and control of personal online data influences web use. In E. Hargittai (Ed.), *Handbook of digital inequality* (pp. 311-325). Edward Elgar Publishing.

Submit

Clear form