Software
Sustainability
Institute

Understanding the software and data used in the social sciences

A study for the Economics and Social Research Council

Authors
Selina Aragon, Mario Antonioletti, Johanna Walker, and Neil Chue Hong

November 2023

Acknowledgements

The Software Sustainability Institute gratefully acknowledges the constructive feedback on this report from **Professor Andrew Stewart**, Head of the Department of Computer Science at the University of Manchester, and ESRC's **Ms Becci Hutchins**, Portfolio Manager, **Mr Richard Welpton**, Strategic Lead for Data Services Infrastructure, and **Mr Kieran Jarrett**, Senior Policy Manager.

We would also like to thank **Professor Andrew Stewart**, **Dr Nick Bearman**, **Dr Caitlin Bentley**,**Dr Chris Jochem**, and **Mr Nathan Khadaroo-McCheyne** for their responses and comments on a preliminary version of the survey.

Finally, we are indebted to all members of the research and innovation community who responded to the survey, participated in the interviews, and helped us circulate any relevant information pertaining to this study.

This study was supported by UKRI-ESRC through additional funding to grant EP/S021779/1 for The UK Software Sustainability Institute: Phase 3.

EXECUTIVE SUMMARY

BACKGROUND

Digital data, tools, and software are constantly evolving across the economic and social sciences community. This has led to changes in the methods used for data collection and analysis, and in the ways that data and software are managed, shared, and sustained for future generations.

The Economics and Social Research Council (ESRC) commissioned this study to map existing data and software mechanisms to identify what software should be considered "infrastructure" due to its widespread or foundational use and to establish how software is being supported and maintained, in order to determine whether there are any weaknesses in the current software infrastructure.

This study will help harness the information already available to the ESRC. Along with advances in integration and analysis, it aims to:

- expand the capability of researchers to link and integrate datasets using software and workflows, and to generate multilevel models incorporating social, economic, environmental, and physical data;
- improve the usability of models and simulations, allowing policy analysts, in conjunction with social sciences researchers, to plan for the future and devise what-if scenarios;
- encourage the use of increasingly available real-time data streams to improve the responsiveness, precision, and accuracy of simulations and models;
- encourage the use of machine learning to support the analysis of very large datasets.

RESEARCH METHODS

This work used a three-stage, mixed methods approach consisting of desk research to map the current social sciences research landscape, a survey of research practices in the social sciences research community, and a series of interviews with key stakeholders.

KEY FINDINGS

Below we present a summary of key findings & recommendations. Refer to Section 6 for all key findings and corresponding recommendations.

Research data practices

- Surveys (22%) and interviews (19%) are the most dominant forms of data used in social science research, followed by a long tail of data sources that include APIs, behavioural data, social media, human participants, new data, and questionnaires. 66% of interview respondents use the UK Data Service, commercially held data, or survey/interview data.
- A majority of survey respondents (53%) exhibit a tendency to both create and reuse datasets, rather than exclusively creating (28%) or reusing (19%) data.
- Most of the reused data comes from the UK Data Service (58%). This reflects the fact that the UK Data Service is the UK's largest collection of economic, social, and population data for research and teaching. The ESRC community is aware of its existence and confident in using the data it provides.
- Institutional repositories constitute the second most common source of data (36%). Other sources include the UK government (6%) and the Office of National Statistics Secure Research Service (5%). A long tail of smaller sources follows. Among the 55 data sources receiving <1% citations, only one ESRC-funded data service (CLOSER) appears.

Surveys (22%) and interviews (19%)

are the most dominant forms of data used in social science research

Data tends either

not to be shared (49%) or shared in an institutional repository (36%)

Sharing data

- Our research shows that data tends either not to be shared (49%) or shared in an institutional repository (36%). A significant minority of respondents (21%) reported promoting their data as an accessible resource in publications.
- 75% of those who exclusively create data refrain from sharing it. Just 34% of respondents funded by the ESRC in the last five years reported sharing their data, despite the ESRC's policy requiring that data be deposited.
- 73% of senior career researchers deposit their data. 82% of junior career researchers, meanwhile, decline to share any data at all (this statistic may reflect the fact that PhD students share their data only at the end of their programmes.) MCRs tend not to share their data (53%).

Research software practices

- Statistical analysis (89%) and spreadsheets (85%) are the most commonly used forms of research software reported in the survey. This corresponds with the widespread use of quantitative (survey) data in the social sciences research community. The third most commonly used software category is "qualitative data analysis tools" (52%), which corresponds with the second most used type of data: interviews.
- R's survey response rate (36%) is around double the response rates of SPSS (19%) and Stata (16%). It is notable that a significant majority of around two-thirds of respondents (67%) are using open source tools, though training isn't always readily available.
- Respondents value open source software for meeting their needs around cost, sustainability, and interoperability. Interoperability is a particular priority, since it allows researchers to engage in cross-institutional work and collaboration.
- A small number of software types is used very widely, while a much larger number (c. 130) is used by a relatively small set of researchers. This makes it difficult to create policies around developing and maintaining software and strategies for encouraging the use of open source software and data sharing.
- Around 64% of survey respondents report using multiple software types in combination in their research. We observe a tendency towards software being used in conjunction with other tools and methods, with each software category finding an application in a specific workflow or toolchain. We have also identified two main patterns of use: interview and survey. These correlate with the main types of data used in social sciences research.

Developing research software

- Most survey respondents (59%) do not develop software.
- Of the respondents who develop software, we found that most appear to be using R (41%) and Python (17%). Those who develop research software tend to share it widely (46%). The development and dissemination of software correspond with the widespread use of open source software (R and Python) in the social sciences community.
- Other software developed includes Excel functions, macros, and the production of charts (16%), Stata (11%), and SPSS (5%).
- Many respondents (42%) felt that developing and maintaining research software is not sufficiently rewarded or recognised. 37% of respondents agreed (against 23% who disagreed) that insufficient attention is paid to software funding, management, and licensing within the economic and social sciences research community.

Just 34%

of respondents funded by the ESRC in the last five years reported sharing their data

Statistical analysis (89%) and spreadsheets (85%)

are the most commonly used forms of research software

Around two-thirds of respondents (67%)

are using open source tools

Around 64% of survey respondents

report using multiple software types in combination in their research

Most survey respondents (59%)

do not develop software

Software policies

> Senior career researchers (63%) show more confidence in understanding software policies laid out by institutions and funders, while junior researchers (69%) have a greater tendency to be undecided, or to admit to not knowing relevant policies. Mid-career researchers (48%) feel that they are aware of relevant policies.

> Roughly 48% of respondents agree that there is not enough of an incentive to learn how their software is funded, managed, and licensed, while about 26% are undecided and 26% disagree with the statement.

> Fewer than 50% of survey respondents are aware of the publishing policies and software licensing arrangements relevant to their institutions, funders, or projects. A broad range of concepts and policies appear to vary according to institutions, departments, and publications.

Online courses (81%)

are the most popular way of acquiring skills, particularly for early career researchers (ECRs)

Infrastructure and hardware

> 68% of respondents use either their own hardware or hardware provided by their institutions to run software and tools. The complete absence of the use of Tier 1 and Tier 2 services in this population sample is notable.

> 51% of junior career researchers provide their own hardware. Interview data shows that this practice could be motivated in part by a desire to retain software settings. It may result in a researcher receiving less institutional support, and adhering less closely to institutional policies.

Skills and training

> Online courses (81%) are the most popular way of acquiring skills, particularly for early career researchers (ECRs). It is unclear, however, whether this enthusiasm for online learning is related to the pandemic.

> Self-led learning, through online examples (65%) and from peers and colleagues (58%), constituted the second and third most popular answers in the survey. Self-led learning was also seen as more appropriate for open source tools.

> Around 19% of respondents across all career stages had taken the NCRM courses listed in the survey.

The survey sample size for this study is broadly indicative; however, the total number of responses (164) is not statistically representative[1] of the social sciences community, based on our analysis of the HESA data.

Recommendations

The key findings from this study have informed the following list of recommendations made to institutions and funders.

1. ESRC should commission a specific study to understand the working relationship between researchers and data services, in particular, barriers to the reuse and sharing of data, and how good practice can be increased and incentivised.
2. ESRC Future Data Services strategy should take into account the "reluctance" to adhere to funder policies and guidance, e.g. around deposit and sharing of data and open data.
3. ESRC should mandate the inclusion of a metadata field when data is deposited that identifies the software used to generate/analyse that dataset.
4. To support open research and reproducibility, ESRC should support the adoption of widely used open-source software, such as the R ecosystem, by encouraging community engagement activities and recognising contributions to software communities. However, for important tools with smaller user

[1] 380 would have been a representative sample of the economic and social sciences community

bases, further investigation is required to understand what funding and recognition mechanisms would be most appropriate to remove barriers to encourage and support the adoption of such tools.

5. ESRC should provide targeted funding to support the maintenance and development of new features of open-source software used by the community.
6. Institutions should invest the same amount of resources in producing internal training and support for open source tools that they do for commercial tools used in social sciences.
7. ESRC should commission a study to understand the reasons for the lack of open source alternatives for qualitative data analysis.
8. Further research is needed to understand why there are differences in barriers to using research software amongst disabled people.
9. Institutions should embed RSEs within research departments to support the software needs of social sciences researchers.
10. ESRC should work with UKRI DRI to ensure researchers can access larger-scale computational infrastructure that has traditionally been seen as excluding ESRC researchers.
11. ESRC should consider how to fund computational research at the interface between ESRC and EPSRC domains.
12. More research is needed to investigate the impact of researchers using their own hardware vs institutionally provided hardware and their ability to use different software tools to conduct their research. ESRC should commission a specific study to understand the social, economic and research impact of people using their own computational hardware in preference to institutionally provided infrastructure.
13. ESRC should extend its research data policy to include sharing and publishing software.
14. ESRC/Institutions should support and encourage the uptake of existing external training (e.g. NCRM and The Carpentries), as well as invest in the development of more targeted internal training to help researchers write their own code following good practices.
15. Training providers should investigate whether learning attitudes are changing (in relation to the pandemic & use of online courses).
16. There is a potential role for a drop-in support model/mechanism, such as the UK Data Service Drop-in service, to fill the gap between the training offered and the practical applications in social sciences research. However, further investigation is needed to understand exactly what researchers are looking for.
17. ESRC should invest in supporting more courses to allow people to transition from commercial statistical analysis packages to in-demand open source alternatives that support open research practices and facilitate worldwide collaboration.
18. ESRC should encourage the development of good practice guidance at institutional and journal levels for sharing and publishing software and data following the FAIR principles.
19. Different forms of training should continue to be offered so that researchers with caring responsibilities can still access it. We also believe that additional asynchronous support will be required to make sure researchers keep practising the skills learned.
20. Further research is needed to understand how time to acquire new skills can be protected.
21. ESRC could improve collaboration & team research by incentivising interdisciplinary work.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AHRC	Arts and Humanities Research Council
BBSRC	Biotechnology and Biological Sciences Research Council
DOI	Digital Object Identifier
DTC	Doctoral Training Centre
DTP	Doctoral Training Partnership (with non-academic partners)
ESRC	Economic and Social Research Council
EPSRC	Engineering and Physical Sciences Research Council
GDPR	General Data Protection Regulations
GIS	Geographic Information Systems
HEI	Higher Education Institution
HESA	Higher Education Statistics Agency
JACS	Joint Academic Coding System
LERU	League of European Research Universities
MRC	Medical Research Council
NERC	Natural Environment Research Council
REF	Research Excellence Framework
RSE	Research Software Engineer
SCR	Senior Career Researchers
STFC	Science and Technology Facilities Council
UKRI	UK Research and Innovation

1. INTRODUCTION

Social sciences researchers have access to an extraordinarily rich, diverse, and growing repository of datasets, often produced by their colleagues in the research community at large. These data are found in publications, surveys, interviews, databases, data collections, longitudinal studies, administrative and judicial records, and online through social media platforms and tools like Google Trends. The sharing and redistribution of some of this data are facilitated by ESRC-funded services, as shown in Table 1.

Service	Link
UK Data Service	https://ukdataservice.ac.uk
The UK Household Longitudinal Study	https://www.understandingsociety.ac.uk
Administrative Data Research UK	https://www.adruk.org
Urban Big Data Centre	https://www.ubdc.ac.uk
Consumer Data Research Centre	https://www.cdrc.ac.uk

Table 1. ESRC-funded and supported services that provide access to data.

In 2022, the ESRC laid out its Data Infrastructure Strategy [9]:

> "Data can help us understand and respond to threats and challenges across the social, economic and political spectrum: from climate change and the impact of the pandemic to understanding and responding to local and regional inequality. The growth of new forms of data, and the new methods and tools by which it can be analysed, creates huge opportunities too."[2]

The increasing availability of different types of data means that tooling and algorithmic advances are necessary to tackle the increasingly difficult task of data processing. Researchers may need new software to support new types of data, or to implement new algorithms to process it.

Moreover, software encodes the processes that are applied to data and thus facilitates the transparency and reproducibility of findings, allowing others to uncover mistakes or faulty reasoning. It is therefore important to ensure that software is made available to those who wish to investigate the production of results or apply the work to their own projects. Currently, training is provided by the ESRC through the National Centre for Research Methods[3] (NCRM), and the implementation Q-Step[4] will ensure that future ESRC researchers are conversant in quantitative methods and the application or writing of software.

In 2020, work conducted by Daniela Duca and Katie Metzler of SAGE Publishing [2] showed that the number of research tools used in the social sciences has risen by 300% in the last 10 years, and more than one in five groups have employed someone specifically to develop software. We believe that this is the result of researchers seeking to expand their skill sets, and of DIY software development becoming more accessible through advances in training (e.g. The Carpentries[5]) and the open source software movement.

The UKRI Infrastructure Opportunity report [1] highlighted the key role that the social sciences research infrastructure plays in ensuring the health, wealth, and wellbeing of the nation; driving economic growth; and creating the conditions for greater social cohesion, equality, and inclusivity. Software is an integral part of that infrastructure, supporting users to discover, analyse, integrate, and visualise data and models.

The loss of key specialist software due to gaps in funding, or the proliferation of tools with poor support, risks decreasing the productivity of ESRC researchers and damaging the reproducibility of research. It also represents a significant loss of investment and opportunity for the UK research community.

In this report, we aim to gain an understanding of how software is used and maintained in the social sciences, with the hope that our findings will play a part in ensuring that high quality resources and training are available to researchers throughout the UK. To do this, we need to review the different ways that data and software are used across a variety of subgroups, map the diversity of the social sciences community, and explore the needs and priorities of researchers, practitioners, curators, administrators, and software developers.

[2] https://www.ukri.org/wp-content/uploads/2022/06/ESRC-090622-DataInfrastructureStrategy2022To2027.pdf (last accessed July 13 2022).

[3] https://www.ncrm.ac.uk/ (last accessed July 13 2022).

[4] https://www.nuffieldfoundation.org/students-teachers/q-step (last accessed July 13 2022).

[5] https://carpentries.org/ (last accessed July 13 2022).

2. BACKGROUND

In this section, we review some previous academic papers and reports to identify some key issues in the landscape of software use in the social sciences, in order to contextualise our research. We then summarise relevant existing ESRC investments.

2.1 LANDSCAPE

There are multiple drivers for the use of software in the social sciences. The first is open research, a key policy of the UKRI:

> "Open research improves research efficiency, quality and integrity through collaborative, transparent and reproducible research practices."[6]

Software is an enabler for all of these practices. Sharing code and data makes the research process transparent, while open source software allows researchers to carry out their work without the burden of expensive and restrictive proprietary software licences. The UNESCO Recommendation on Open Science states that users should

> "gain free access to open source software and source code in a timely and user-friendly manner, in human- and machine-readable and modifiable format, under an open licence."[7]

Another driver is the desire for connectivity in research. As the UKRI Strategy (2022-2027) states:

> "We need a more connected and agile system. We must capitalise fully on the breadth and depth of talent across the UK and create a nexus for global talent and investment. To achieve this goal, our strategy focuses on four 'shifts' to drive the necessary change: diversity, connectivity, resilience and engagement."[8]

Part of this connectivity is facilitated by software of multiple types:

> "The full benefits of diversity are only captured through connectivity and collaboration, bringing diverse ideas, skills and know-how together in novel combinations to catalyse discovery and innovation."[9]

A further key driver for the use of software in the social sciences is the sheer abundance of valuable existing and emerging data. This data is both 'digitally native' and digitisable, and includes demographic data, consumer data, sensor data, and a multitude of other types. While this is not always 'big data,' it is increasingly receptive to manipulation and investigation via computational techniques.

One last factor, closely related to the existence of this superabundance of data, is the rise of computational social science, described by Lazer et al. as a "quantitative modelling of these new kinds of digital traces" [4]. This has largely been spearheaded by private companies like Meta, Alphabet, Microsoft, Apple, and Amazon [4], which obtain customer data in vast quantities and develop new ways of analysing and exploiting it. In academia, research carried out by Duca and Metzler into the use of software in social sciences found 418 tools available to researchers. These were mostly concerned with textual analysis or surveys, however, and it is possible the authors would have discovered a greater number had they been looking at more quantitative approaches [2].

There are serious challenges for each of these drivers. In terms of collaborations, HEI structures and incentives often inhibit connectivity. Traditional disciplinary siloing, budgeting, and funding structures, along with employment pathways that fail to recognise interdisciplinary researchers, can work against the institutionalisation of cross-disciplinary projects [6].

Both the Australian Research Data Commons (2022) [3] and Duca and Metzler (2020) [2] identified poor citation practices as a problem for research software. Duca and Metzler also see challenges in what they term the "difficulty of navigating the ecosystem," referring to the problems researchers experience in attempting to establish which tools to use and where to find appropriate guidance. Lazer et al. (2020) also found that there are "inadequate data sharing paradigms" [6].

While there has been praise for advances in the sharing of administrative data, those studying the

[6] UKRI Strategy (2022–2027) p16, https://www.ukri.org/wp-content/uploads/2022/03/UKRI-210422-Strategy2022To2027TransformingTomorrowTogether.pdf (last accessed July 13 2022)

[7] Barker, M., Chue Hong, N. P., Katz, D.I S., Leggott, M., Treloar, A., van Eijnatten, J. & Aragon, S, "Research software is essential for research data, so how should governments respond?" Zenodo, 2021, https://doi.org/10.5281/zenodo.5762703

[8] Ibid, p8 (last accessed August 1 2022).

[9] Ibid, p9 (last accessed July 13 2022).

field have identified large gaps in access to privately held data. Writing in a 2016 SAGE Publishing White Paper [7], Metzler et al. found that of the 33% of survey respondents who had used big data in the past year, 55% had used administrative data whereas only 23% had used commercially controlled data. The academic disciplines principally using big data were Social Statistics & Research Methods, Economics, Demography, and Health Sciences. Duca and Metzler (2020) [2] also point out that not only do commercial companies control a great deal of data, but they are also able to attract high-calibre computational social scientists to work for them with the promise of a high-impact experimental environment.

Duca and Metzler (2020) [2] find that open source tools present both opportunities and challenges. Sustaining software that was developed for a bespoke purpose is possible if it can be open sourced, but this is not a "golden straitjacket". Building a sufficiently large community around a tool is a challenge in itself. In general, the sustainability of software remains an unsolved problem, with the Australian Research Data Commons (ARDC) finding that

> "software is an often invisible part of research, produced quickly within a funding window, often struggling to be maintained beyond that."[10]

The fact that research software is often invisible entails another serious challenge for its development and sustainability, which is that it is rarely recognised as a "first-class output of research" [3]. In the UK, this problem is exacerbated by the assessment criteria of the Research Excellence Framework (REF), which still largely focuses on publication-related outputs and lacks a systematic, peer review-style method of assessing software. Duca and Metzler (2020) [2] recommend introducing a standard model to correct this deficiency, noting that this might also help researchers navigate the ecosystem.

Finally, some authors have found that research ethics have failed to keep pace with the technical capabilities available to researchers, arguing that it is necessary to mandate

> "the ethical use of private data that preserves public values like privacy, autonomy, security, human dignity, justice, and balance of power to achieve important public goals."[11] [6]

In other words, when using data pertaining to the behaviour of human beings online, who aren't necessarily aware of the ways in which their data is being used, and where the benefit lies disproportionately with the user rather than the subject, simple compliance with current legal requirements such as the General Data Protection Regulation (GDPR) [8] is insufficient.

In June 2022, the ESRC published its Data Infrastructure Strategy (2022-2027) [9], setting out how it will invest in data infrastructure and associated capacity building. One of the five pillars of its strategy is ensuring that skilled researchers can effectively utilise data in their research. Software is necessary to achieve this aim (though the word is not specifically used in the delivery framework, nor mentioned once throughout the document except in reference to the Software Sustainability Institute.)

In the context of this capacity-building strategy, Metzler et al. (2016) [7] found that the biggest problem for educators trying to teach big data methods to students is that they tend not to have the appropriate level of programming or statistical knowledge.

2.2 ESRC INVESTMENTS

The ESRC has already made major investments in training and data availability and use. A list of 34 data availability and use sites is given in the ESRC centres: 2020 overview [11].

Together with the Nuffield Foundation, the ESRC funds Q-Step[12], a major programme dedicated to increasing the number of quantitatively skilled social sciences graduates in the UK. At the graduate level, it currently has two Centres for Doctoral Training (CDTs): the Soc-B Centre for Doctoral Training[13], and the Data Analytics and Society Centre for Doctoral Training[14]. It also funds 14 Doctoral Training Partnerships (DTPs) across 73 organisations[15].

[10] "A National Agenda for Research Software", Zenodo, March 28 2022, DOI: 10.5281/zenodo.6378082

[11] Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H. & Nelson, A. "Computational social science: Obstacles and opportunities", Science, 28 August 2020, Vol 369, Issue 6507, pp. 1060-1062, DOI: 10.1126/science.aaz81

[12] https://www.nuffieldfoundation.org/students-teachers/q-step (last accessed August 11 2022)

[13] https://www.ucl.ac.uk/soc-b-biosocial-doctoral-training/ (last accessed August 11 2022)

[14] https://datacdt.org/ (last accessed August 11 2022)

[15] https://www.ukri.org/what-we-offer/developing-people-and-skills/esrc/doctoral-training-partnerships (last accessed August 11 2022)

The NCRM[16] is a training body that the ESRC has built over the last few years to satisfy more advanced training needs. The NCRM delivers methodological training and resources on core and advanced quantitative, qualitative, digital, creative, visual, mixed, and multimodal methods. In its data strategy, the ESRC aims to continue strengthening relationships between data infrastructure investments and the NCRM.

Another key ESRC investment is the UK Data Service. This provides trusted access to the UK's largest collection of social, economic, and population data for research and teaching, and also colocates training, software, and tools resources such as the UK Data Service Learning Hub[17]. Some of this training is provided via ad-hoc drop-in sessions[18]. Other investments include the ESRC Data Centres[19] which often manage multiple datasets at different stages of the research process.

Further, in its Data Infrastructure Strategy, ESRC pledges to

> "review and update our core training requirements for doctoral students to embed data-driven research skills, particularly new digital methods."[20]

[16] https://www.ncrm.ac.uk/ (last accessed August 11 2022)

[17] https://ukdataservice.ac.uk/learning-hub/ (last accessed August 11 2022)

[18] https://ukdataservice.ac.uk/events/computational-social-science-drop-in-2022-07-12 (last accessed August 11 2022)

[19] https://www.ukri.org/publications/esrc-centres-2020-overview/ (last accessed August 11 2022)

[20] Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H. & Nelson, A. "Computational social science: Obstacles and opportunities", Science, 28 August 2020, Vol 369, Issue 6507, pp. 1060-1062, DOI: 10.1126/science.aaz81

3. METHODOLOGY

Our report summarising the findings from this study aims to

- map existing policies and mechanisms to support software in social sciences research;
- identify key software and supply information about licensing, funding methods, and issues affecting use;
- provide a current overview of how software is being used, identifying trends that might impact funder policy.

The aim of the report was to understand current patterns of software use and identify gaps. This study will also help the wider community of digital research infrastructure providers (including those outside the social sciences, whose services may be used by social sciences researchers or interdisciplinary teams) to understand how researchers use software to discover, access, integrate, and analyse their resources.

Our central goal is to assist the ESRC (and the broader social sciences community in the UK) in understanding how best to support the adoption of best practices in software use and maintenance. This work will help to ensure that the ESRC can design and deliver effective future interventions in the digital (software and data) arena as part of its data infrastructures and skills and careers portfolios. It will also assist the Software Sustainability Institute (SSI)[21] in providing services and resources that best support social science researchers. More widely, these contributions could benefit members of the public who are interested in social sciences research by improving the sustainability, accessibility, and usability of open source tools.

We took a three-stage, mixed-methods approach.

Stage one

Desk research to map existing policies, protocols, and mechanisms provided by the ESRC and other stakeholders in social sciences research, in order to support good practice in the development and maintenance of research software.

Stage two

Develop and implement a survey of current research practices among the UK social sciences research community, to identify current data practices, commonly-used software, and archetypical workflows or working patterns.

The research matrix in Table 2 shows how we approached the higher-level research question by breaking it down into parts, and which method (or methods) were used to address each question.

Stage three

Conduct interviews with a small number of key stakeholders (users and infrastructure providers identified in stages one and two, or suggested by the ESRC), designed to corroborate previously obtained data on critical software and workflows, and to identify priority barriers.

[21] https://www.software.ac.uk (last accessed August 11 2022)

Topic/Method	Desk Research	Survey	Interviews
Establishing which datasets are being collected/reused	✔	✔	✔
Creating a overview of existing NCRM software training provision	✔		
Identifying which NCRM software training is being taken up		✔	✔
Identifying what software is being used in funded ESRC projects			✔
Identifying what software is being used in ESRC projects		✔	✔
Establishing how software is currently being supported and maintained	✔	✔	✔
Examining existing relevant policies	✔	✔	
Establishing the diversity of the community	✔	✔	✔
Identifying priority barriers		✔	✔
Identifying work patterns/workflows		✔	✔
Providing background on the funding, updating, and licensing of key software	✔	✔	
Identifying possible key interventions			✔
Identifying previous reports to use as baseline/benchmarks	✔		

Table 2. Research matrix breakdown of the tasks undertaken.

A key risk lay in ensuring that each stage sampled a representative segment of the community. We managed this risk by ensuring that we understood the community we were trying to reach and the extent to which our sample group aligned with it.

An additional risk was presented by the short timescale within which we were required to complete the work. This was mitigated by existing work piloting a survey with SageOcean, and similar (but more extensive) work addressing the AHRC community that started in May 2021 and culminated in a November 2022 report [23].

3.1 STAGE ONE: UNDERSTANDING POLICIES AND THE COMMUNITY

Stage one of the research consisted of two separate desk research tasks. The first was to understand the software and data management policies of the ESRC and their implications. The second was to understand what the ESRC community looks like. We achieved these stage one goals by analysing Higher Education Statistics Agency (HESA) data.

Mapping policies

The ESRC does not have a specific software management policy. However, it is useful to examine the content and structure of the organisation's research data management policy. The seven UK research councils have data management policies for all applicants (though in the cases of Innovate UK and Research England, data management policies are applied on a grant-by-grant basis). The UKRI also has an Open Research Policy[22]. Six of the seven councils, including the ESRC, have their own bespoke policies (Table 3).

[22] www.ukri.org/about-us/policies-standards-and-data/good-research-resource-hub/open-research (last accessed July 31 2022)

Council	Policy Purpose
AHRC	Operates a UKRI-wide policy focusing on and supporting open research and open data [14]
BBSRC	Focuses on opening research data to stimulate new investigations in order to get the best value for the funds invested [15].
EPSRC	Deals with the management and provision of access to data, believing that funded research data should be made as widely and freely available as possible without damaging the research process [16].
ESRC	Operates a policy that explicitly supports grant holders who "collect, produce and reuse data", with an emphasis on roles and responsibilities [17].
MRC	Focuses on maximising research opportunities [18].
NERC	Supports open access and open research, with a particular concern for "enabling the tracking of [dataset] usage through citation and data licences". The NERC also has a legislative requirement to guide the management and distribution of environmental information [19]
STFC	Works to ensure that data is carefully managed and optimally exploited for economic impact. It covers both researchers and facilities [20].

Table 3. Purpose of research data management policies of UK research councils.

There is surprisingly little commonality among these policies in terms of structure. Their respective approaches are partly influenced by whether they have responsibility for, or relationships with, data institutions, as the ESRC does.

ESRC policy does not attempt to define exactly what is meant by 'data', but it defines roles and responsibilities in great detail. It consists of an Introduction, Definitions, ESRC Research Data Policy Principles, Ethical Data Considerations, Data Security, Responsibilities of the Grant Holders, Responsibility of the ESRC, Responsibilities of the Grant Holders' Institutions, Responsibilities of the ESRC Data Service Providers, and References.

Data use: Most guidance around data use focuses on downstream reuse by third-party non-applicants after collection or generation. However, the ESRC is explicit in stating that all grant applicants must make every effort to reuse existing data where feasible (upstream reuse).

Data access: Councils do not prescribe timelines for sharing, but require timely sharing of data and metadata. The ESRC provides for an exclusivity period of two years, which does not apply to metadata.

Data availability: As one of the councils with a repository function, the ESRC has strong guidance on data depositing. It requires data to be made available through the UK Data Service, UKDS Census Support, the Big Data Network (including the Administrative Data Research Network (ADRN), and the Business and Local Government Data Research Centres. Data must be deposited within three months of the end of the grant.

Data training and skills: Few councils include training and guidance in their data policies. The ESRC's policy provides the most detailed guidance in this area. It also specifies what should be included in a data management plan, asserting that data institutions are responsible for providing good data management and sharing practices training to ESRC grant holders, advising grant holders on improving DMPs, and providing institutional repositories with guidance on good practice.

Legal guidance and compliance: Although most councils focus on downstream (sharing) activities rather than upstream (collection) activities, the salient time to consider ethical and legal compliance is during data collection or generation. That this is largely unaddressed in research council policies is perhaps due to the fact that HEIs normally have their own ethics approval infrastructures that apply whether the research in question is UKRI-funded or not. The ESRC insists on compliance with the Copyright, Designs and Patents Act 1998, the Data Protection Act 1998, the Freedom of Information Act 2000, and the ESRC Framework for Research Ethics.

Skills and knowledge in data policies

The following skills and articles of knowledge are either explicitly required in data policies (such as when the ESRC encourages the use of secondary data) or implicitly required on the basis that they are generally expected by the research councils. They primarily focus on digital asset management (the curation and archiving of data to ensure their usability for future researchers) rather than digital data manipulation (the collection and analysis of data). This aligns with the focus of the policies on open research data and effective reuse.

- Digital data manipulation skills in the policies include:
  - Data search skills
  - Data citation skills
- Digital asset management skills include:
  - Familiarity with relevant legislation
  - Ability to create a data access statement
  - Familiarity with institutional or funder repositories/archives
  - Knowledge of metadata
  - Ability to identify risks of sharing, and, in some cases, to weigh them against the potential benefits
  - Ability to use standards such as the Data Documentation Initiative (DDI), SDMX, and INSPIRE
  - Copyright/IP knowledge
  - Data licencing understanding and creation
  - Development of documentation

Other relevant organisations are CLOSER, UK Data Service / UK Data Archive, NCRM, and CESSDA. All of these are dataset or training providers, which the ESRC endorses and has set goals to build on as part of their ESRC Data Infrastructure Strategy 2022.

Identifying the community

We used two data sources to help us understand the ESRC community from a socio-demographic point of view: Higher Education Statistics Agency[23] (HESA) data, which provides information on staff employed in higher education, and the Gateway to Research[24] (GtR) portal, which provides information on grants awarded by the UK Research Council.

HESA data

In order to gain an understanding of the ESRC community, we looked at data from HESA. This involved obtaining the numbers of full person equivalent (FPE) staff involved in research only and research and teaching (the data for those teaching only and those not involved in either teaching or research were not obtained). FPEs were used as this gave a better indication of the actual number of people involved, even if employed in a part-time capacity, than would have been the case if we had examined full-time equivalent (FTE) numbers[25]. As the survey did not distinguish between research only and research and teaching staff, the data for these two cohorts has been combined. HESA performs rounding[26] to preserve anonymity, introducing a degree of error, so numbers are indicative. We did not consider student numbers, which can also be obtained from HESA.

Research disciplines that correspond to the ESRC's criteria are encoded in different cost centres[27]. These cost centres are listed below (the numbers in the brackets refer to the cost centre codes):

1. (104) Psychology & behavioural sciences
2. (123) Architecture, built environment & planning
3. (124) Geography & environmental studies
4. (125) Area studies
5. (127) Anthropology & development studies
6. (128) Politics & international studies
7. (129) Economics & econometrics
8. (130) Law
9. (131) Social work & social policy
10. (132) Sociology
11. (133) Business & management studies

[23] https://www.hesa.ac.uk (last accessed August 3 2022)

[24] https://gtr.ukri.org (last accessed August 3 2022)

[25] For the difference between FPEs and FTEs, see https://www.hesa.ac.uk/collection/c20025/fte_vs_fpe (last accessed July 11 2022)

[26] Rounding is employed by HESA to the base numbers for a cost centre and given category after the column and row totals have been calculated (0-2 are rounded to zero, and all other numbers are rounded to the nearest multiple of 5). Halves are always rounded up. A full description of the rounding strategy is available at https://www.hesa.ac.uk/support/data-intelligence/general-performance-indicators-suppressions (last accessed July 27 2022)

[27] A list of the HESA cost centres is available at https://www.hesa.ac.uk/support/documentation/cost-centres/2012-13-onwards (last accessed July 27 2022)

12. (135) Education
13. (136) Continuing education
14. (145) Media studies

Using HESA's 2020/21 cost centre data (the latest year available is January 2022) we can ascertain that the economic and social sciences research community comprises approximately 43,335 FPEs across the 14 broad disciplinary groupings given above.

Community sizes across the available research disciplines are given in Fig. 1.

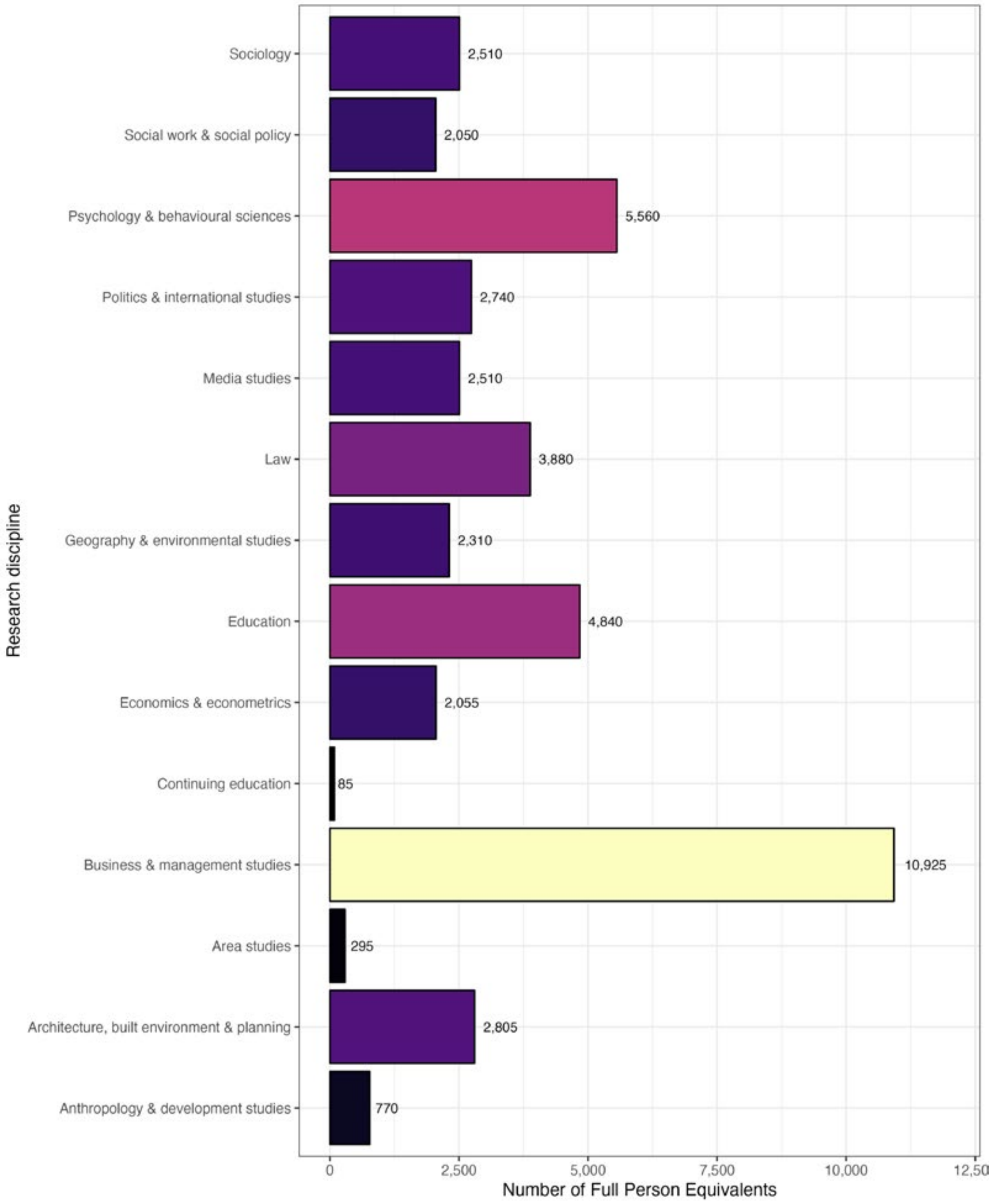


Figure 1. Full person equivalents (FPEs) per discipline in ESRC subjects for 2020/21 (HESA). Numbers at the end of each bar give the total FPE staff involved in that discipline.

It is not clear why the 'business & management studies' category shows a significantly higher number of FPE staff than other categories. Around 92% of these are engaged in teaching and research–but the same is true of law, which has far fewer FPE personnel. Another HESA dataset[28] allows us to partition the data by mode of employment (part-time/full-time). This shows nothing remarkable in terms of the number of part-time staff in business & management studies compared with the other research disciplines. Although this discipline has the greatest number of part-time staff, that alone does not explain the vast discrepancy shown in Fig. 1. Partitioning the data by institution and research discipline yields no further insight.

Table 4 provides a summary table of socio-demographic characteristics.

[28] Academic staff by cost centre and nationality https://www.hesa.ac.uk/data-and-analysis/staff/table-26 (last accessed July 27 2022)

<table>
<tr><th colspan="6">Gender</th></tr>
<tr><th colspan="2">Female</th><th colspan="2">Male</th><th colspan="2">Other</th></tr>
<tr><td colspan="2">49.5%</td><td colspan="2">50.4%</td><td colspan="2">0.1%</td></tr>
<tr><th colspan="6">Disability</th></tr>
<tr><th colspan="3">No disability</th><th colspan="3">Disability</th></tr>
<tr><td colspan="3">95.4%</td><td colspan="3">4.6%</td></tr>
<tr><th colspan="6">Ethnicity</th></tr>
<tr><th colspan="2">White</th><th colspan="2">Black, Asian, Mixed, or Other</th><th colspan="2">Unknown</th></tr>
<tr><td colspan="2">76%</td><td colspan="2">18%</td><td colspan="2">6%</td></tr>
</table>

Table 4. Summary of the ESRC observed population demographics as derived from HESA data.

Below, we provide a detailed summary of social-demographic characteristics across research disciplines. Table 5 gives a percentage breakdown for gender composition across each of the research disciplines.

Research discipline	Female	Male	Other
Business & management studies	43.4%	56.4%	0.2%
Architecture, built environment & planning	36.9%	63.1%	0%
Education	67.8%	32%	0.2%
Continuing education	61.1%	38.9%	0%
Area studies	44%	56%	0%
Psychology & behavioural sciences	59.3%	40.6%	0.1%
Geography & environmental studies	41.3%	58.4%	0.3%
Anthropology & development studies	49%	51%	0%
Politics & international studies	41.1%	58.2%	0.7%
Economics & econometrics	28.2%	71.8%	0%
Law	51%	48.7%	0.3%
Social work & social policy	65.4%	34.6%	0%
Sociology	56%	43.5%	0.5%
Media studies	45%	54.8%	0.2%

Table 5. Gender split across the different ESRC pertinent research disciplines (cost codes).

In summary, about 49.5% of FPE staff involved in ESRC subjects are women. Men constitute around 50.4%, while other genders account for around 0.1% across the ESRC subjects.

HESA also examines the prevalence of disabilities among ESRC subject staff in its data, covering physical impairment and mobility issues, specific learning difficulties, and so on. Fig. 2 shows multiple disabilities in aggregate and compares numbers of staff with disabilities against numbers of those without. In all disciplines, the proportion of staff with disabilities is 7% or lower.

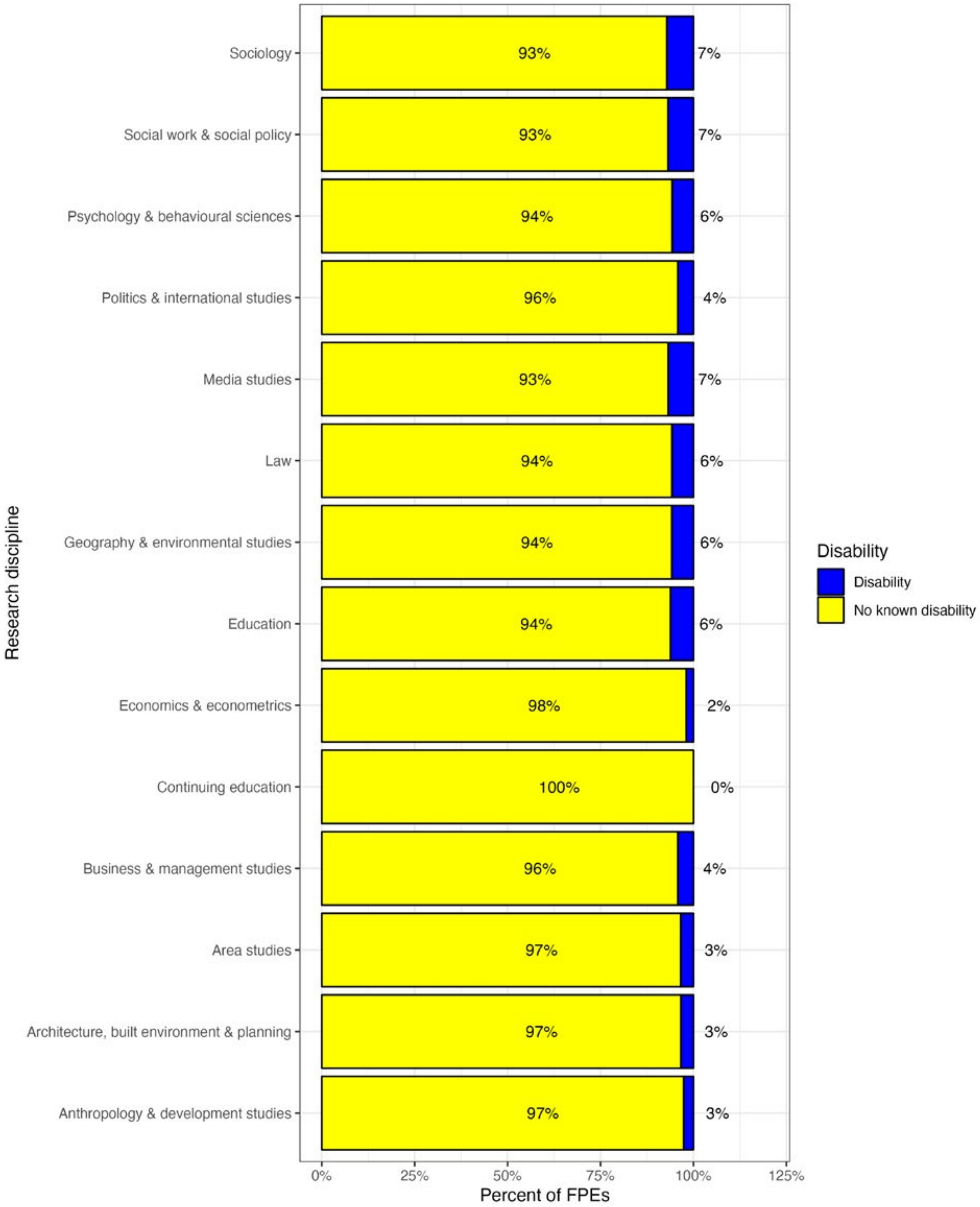


Figure 2. FPEs with disabilities across the different research disciplines (cost centres). The HESA data contains a fine-grained account of the different disabilities, but they have been aggregated together here.

In summary, about 4.6% FPE staff reported a disability, and 95.4% have no known disability across all the ESRC-related subjects.

Table 6 gives an ethnic breakdown for each of the ESRC-related research disciplines.

Research discipline	Asian	Black	Mixed	Other	Unknown NA	White
Business & management studies	19%	5%	2%	3%	6%	65%
Architecture, built environment & planning	11%	5%	2%	4%	6%	72%
Education	5%	2%	2%	1%	5%	85%
Continuing education	6%	0%	0%	0%	12%	82%
Area studies	8%	4%	4%	4%	12%	68%
Psychology & behavioural sciences	5%	1%	2%	2%	6%	84%
Geography & environmental studies	7%	2%	3%	1%	6%	81%
Anthropology & development studies	12%	3%	5%	3%	10%	67%
Politics & international studies	7%	1%	3%	3%	8%	78%
Economics & econometrics	17%	2%	2%	2%	10%	67%
Law	7%	4%	3%	2%	6%	78%
Social work & social policy	5%	3%	3%	1%	5%	83%
Sociology	6%	3%	3%	2%	7%	79%
Media studies	4%	1%	3%	2%	6%	84%

Table 6. Ethnic breakdown by research discipline.

As a baseline comparison[29], 87% of people in the UK are white, and 13% belong to a black, Asian, mixed, or other ethnic group (from 2011 census data). From the HESA data for ESRC-related subjects we can ascertain that 76% employees are white, 18%, belong to a black, Asian, mixed, or other ethnic group, and 6% have an unknown or unavailable ethnicity. Thus, there appears to be greater non-white representation among staff of ESRC-related subjects than among the general UK population.

Finally, we can see that ESRC FPE staff are spread across 135 different institutions in the UK. Fig. 3 shows institutions with more than 275 staff.

[29] From https://www.ethnicity-facts-figures.service.gov.uk/ (last accessed July 27 2022)

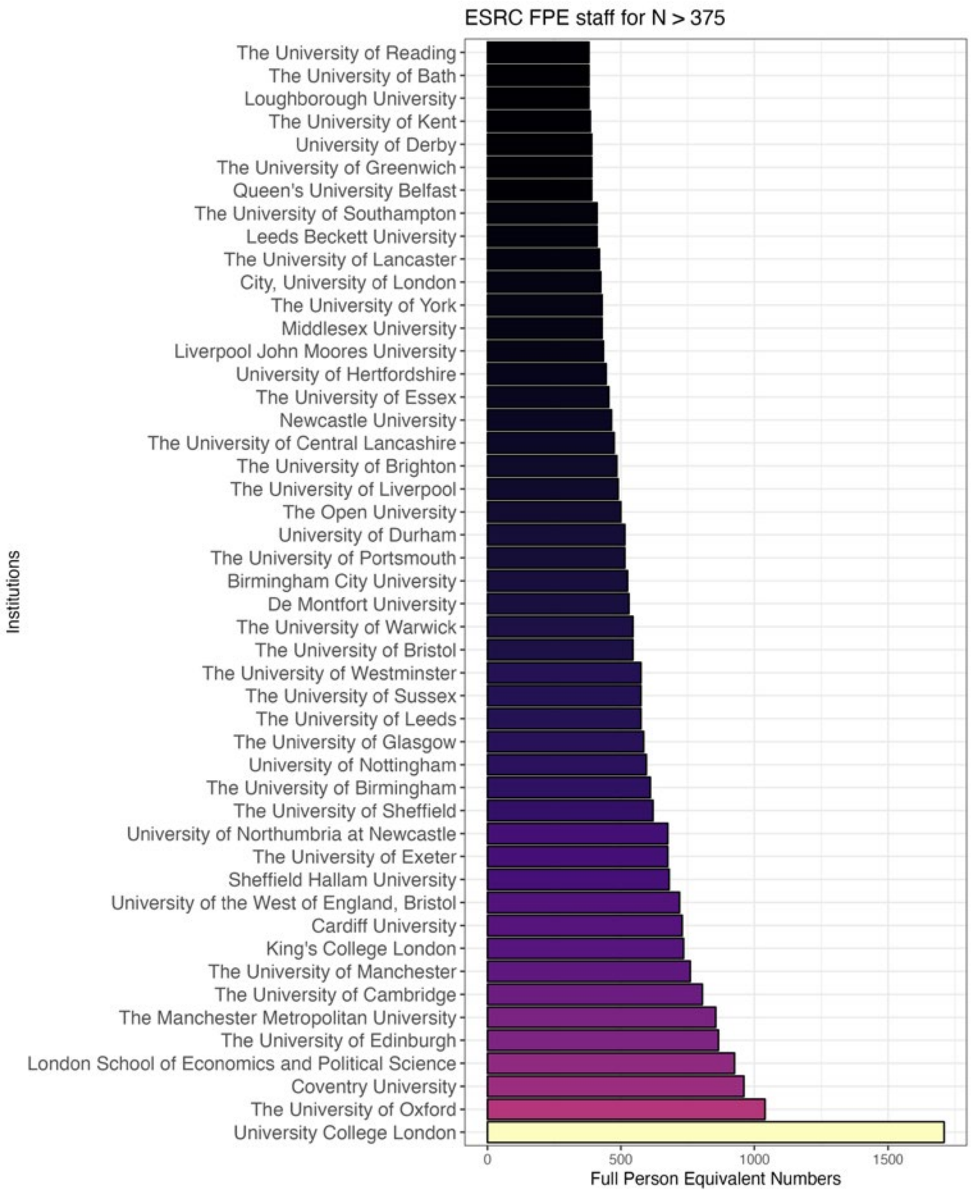


Figure 3. The distribution of the ESRC FPEs shows institutions which only have an FPE count greater than 375.

The mean number of FPE staff per institution is 377, with a median of 325. The institution with the greatest number of ESRC-related FPE staff is University College London, with 1,710 according to the HESA data.

Gateway to Research (GtR)

The GtR is a searchable record of research funded by the ESRC, allowing users to review current and past projects. The GtR amalgamates data from other sources, including research councils, the Joint Electronic-Submission system[30] (Je-S), and Researchfish[31]. The record is updated roughly every month, though this report draws on 24 February 2022 snapshot.

The ESRC data in the GtR classifies funded projects according to 70 research **subject** types and 399 different unique research **topics** (using data since 2008) with a preference to topics over subjects but as there are a lot more potential topics then subjects a project can be classified under it makes topics a little harder to work with. Moreover, each project can be allocated more than one research subject or topic. We can see a breakdown of research subjects and topics allocated to projects in the GtR ESRC data. Regardless, we need to map 70 subjects and 399 topics onto the 21 research disciplines (excluding the "Other") option. As each project can be classified under more than one topic or subject classification each of these will contribute a fractional part to the research discipline count, i.e. if there are N subject or topic classification then each will contribute 1/N to the research discipline these have been mapped to.

Fig 4 shows the number of subject/topic classifications that have been made to each project. However, there are not only more topic classifications but also more are used to classify a project.

[30] https://je-s.rcuk.ac.uk (last accessed June 23 2022)

[31] https://researchfish.com (last accessed June 23 2022)

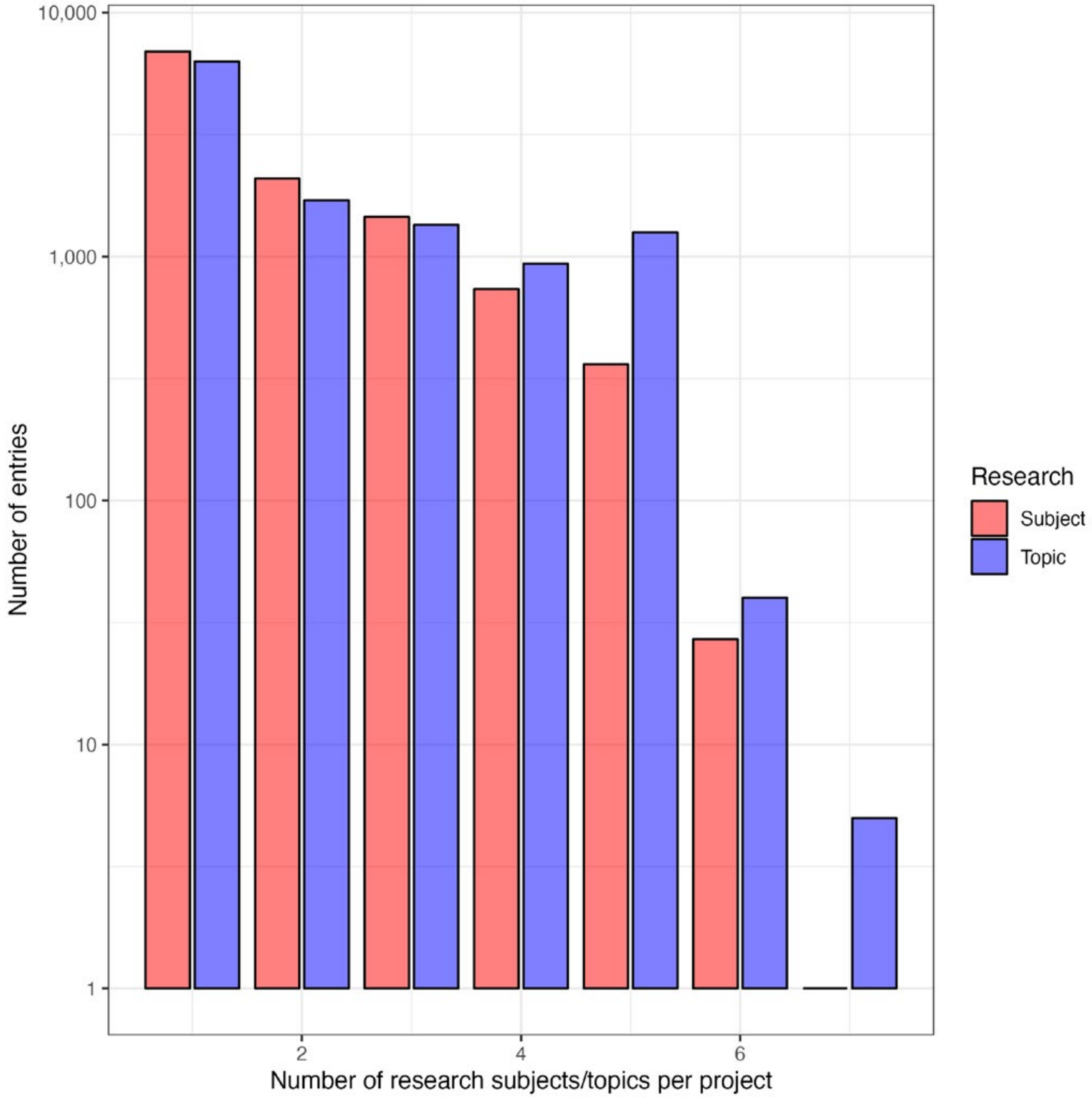


Figure 4. Research topics and subject comparison. There are more research subjects at the beginning but topics dominate towards the end.

We can use this data to see which subjects are mostly funded by the ESRC (using all data since 2008). Psychology is the most funded subject in terms of the number of awards. Social work is the least funded by the same measure.

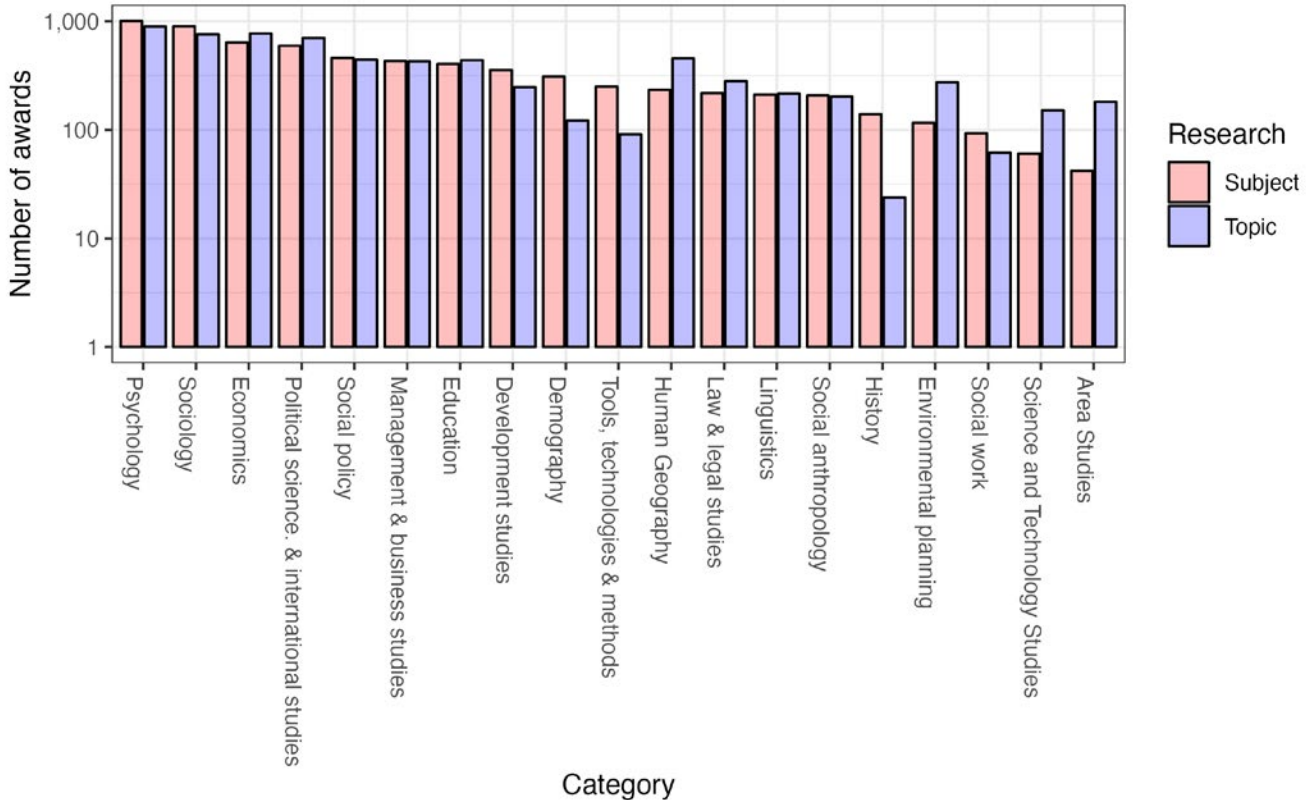


Figure 5. GtR projects are divided into research topics and subjects for the different research categories.

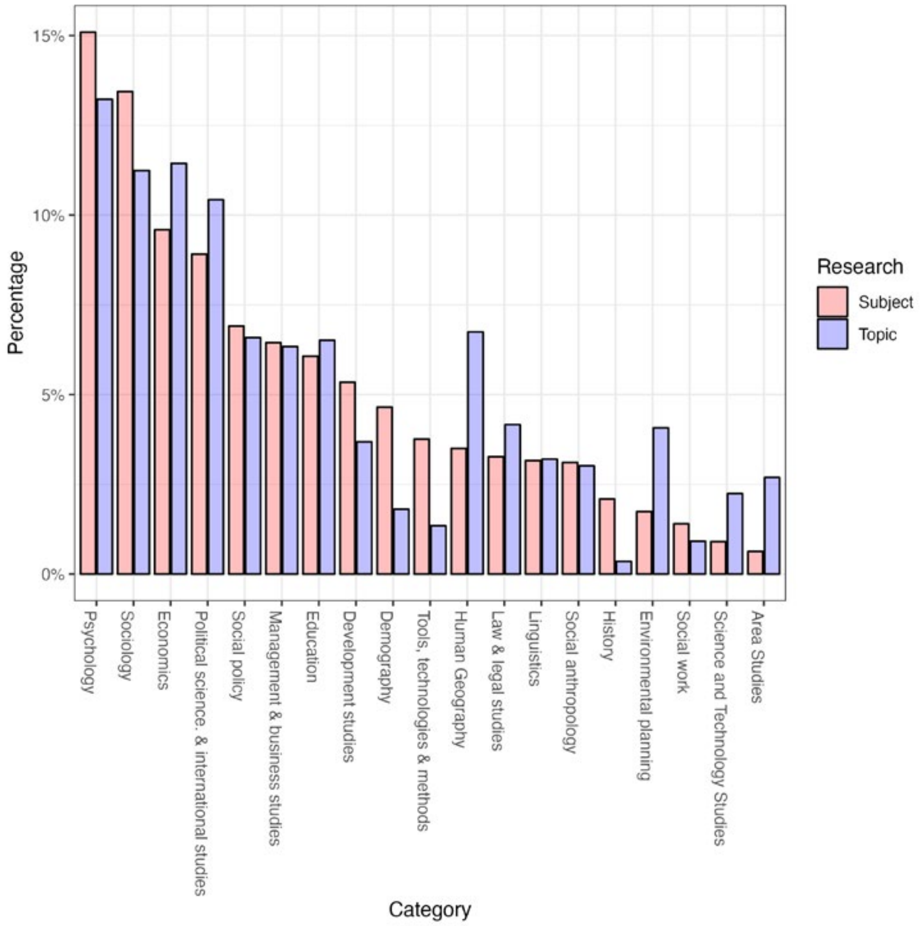


Figure 6. GtR projects are divided into research topics and subjects divided into different research categories, but only active projects from February 2022 are considered.

3.2 STAGE TWO: SURVEY

Stage Two of the research involved developing and distributing a survey to members of the economic and social sciences community. The final number of responses was 168.

Ethics

As our survey involved individuals, we followed the University of Edinburgh's Ethics Process and submitted our research tool for review by the EPCC's Director of Research [21].

Design

The survey was designed and distributed using the Jisc Online surveys platform [12]. Feedback on an initial draft was obtained from members of the community[32] and the survey was revised before launch on 3 February 2022. A complete version of the survey is available online[33] and in Appendix 1.

No questions were compulsory, though the survey required consent regarding the use of participant data. We incentivised participation with a prize draw to win shopping vouchers worth £30. We also offered respondents the opportunity to engage in follow-up research in the form of in-depth interviews. A hyphen (-) is used when no data for a particular question was obtained. The survey contained 28 questions, categorised as follows:

1. Data collection and reuse
2. Software practices and training
3. Support for and barriers to the use of software
4. Demographic information

Disciplinary categories were derived from a previous ESRC expenditure report entitled 'ESRC application and success rate data and analysis[34]. A couple of additional categories (data science and artificial intelligence and information science) were added at the suggestion of survey reviewers.

Recruitment

We identified the names of Principal Investigators (PI) from GtR records and created a customised database of 3,722 contact details. A further 989 emails were generated for students funded by the ESRC in 2019.

We were aware of the following issues:

- As we only had the names of successful PIs, we were less likely to reach ECRs. We mitigated this by approaching Doctoral Training Partnerships directly.
- Relatedly, our list tended towards successful projects, or topics funded by the ESRC, rather than those that are self-funded, institutionally funded, or funded by another body.
- We automated the production of email addresses, which entailed that we did not contact many people from universities whose email address systems do not follow a standard format, primarily the Universities of Oxford, Southampton, Cardiff, and Cambridge.

The survey was then publicised through various channels including Twitter, the SSI Fellows network, and a customised database of PIs and PhD students. Roughly 4,711 individuals from previously funded ESRC projects or studentships received emails about the survey.

It was open for 32 days, closing on 6 March 2022. We received a total of 168 responses, three of which were excluded from the analysis as they were submitted by individuals based outside the UK and with no connection to the ESRC. There was one duplicate entry, taking the total number of viable responses to 164.

[32] Thanks once again to Andrew Stewart, Nick Bearman, Caitlin Bentley, Chris Jochem, and Nathan Khadaroo-McCheyne.

[33] The survey can be accessed at: https://github.com/softwaresaved/esrc-software-study/blob/main/Docs/esrc-survey.pdf

[34] https://www.ukri.org/publications/esrc-application-and-success-rate-data-and-analysis, "Application and success rate data 2011-12 to 2017-18" spreadsheet in the "All_grants_by_discipline" tab (last accessed June 23 2022)

Analysis

The data were exported from the JISC online surveys[35] as Excel files and read by an R markdown script[36]. Open text boxes required some processing to make them tractable. Where necessary, Open Refine[37] was used to carry out the quantitative analysis. This not only facilitated the processing of text but also made the process reproducible.

Sample size, response rate and significance

The sample size for a survey can be calculated using the following formula:

$$Sample\ size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \frac{z^2 \times p(1-p)}{e^2 N}}$$

where: N = population size, e = margin of error, z = z score.

From the HESA data, we can estimate that the ESRC community consists of 43,335 staff researchers (N = 43,335). The industry standard for a survey is to aim for a 95% confidence level (generating a z = 1.96) and a 5% margin of error (e = 0.05). To achieve this level of significance, the survey would have required a sample of 381 responses[38]. The survey sample for this study is therefore broadly indicative: the total number of responses (164) is not statistically representative[39] of the total number of researchers in the social sciences community (43,335) based on our analysis of the HESA data.

3.3 STAGE THREE: INTERVIEWS

Stage Three of the research consisted of developing and conducting interviews with 21 members of the ESRC community.

Ethics

Having sought permission for our survey and questionnaire from the University of Edinburgh's Ethics Committee [21], we devised seven consent statements and produced individual participation information sheets for our interviewees.

Design

The interview design was based on

> questions we could not address in the survey,
> questions we wanted to investigate in more depth, and
> previously tested questions from the AHRC interview script.

The interviews were conducted over Zoom and audio recordings were made.

Recruitment

We recruited interviewees firstly by reaching out to key individuals to help us pilot the interviews and ensure they met our needs. We then invited select survey respondents to engage in follow-up research with us, and published a call for participation[40] on the SSI's website and social media channels, specifically targeting junior career researchers and early career researchers. We sought to engage with a diverse range of software users, non-users, and developers. We scrutinised HESA and GtR data to ensure we engaged researchers from across the full breadth of disciplines.

[35] https://www.onlinesurveys.ac.uk (last accessed May 5 2022)
[36] This is available at https://github.com/softwaresaved/esrc-software-study/blob/main/Src/SurveyResults.Rmd
[37] https://openrefine.org, (last accessed July 28 2022)
[38] We used https://select-statistics.co.uk/calculators/sample-size-calculator-population-proportion (last accessed August 31 2022)
[39] 380 would have been a representative sample of the economic and social sciences community
[40] www.software.ac.uk/news/call-participants-interviews-digital-methods-and-software-social-sciences (last accessed September 29 2022)

Analysis

Completed interviews were uploaded into Otter.ai for transcription and pseudonymisation using a randomly generated code. They were then downloaded onto the researcher's laptop and uploaded into NVivo for thematic analysis. We followed the Braun and Clarke (2006) approach [22], which involved the creation of an inductive coding framework based on key sections of the interview, resulting in top-level codes. We also ran queries over specific terms to understand how these were being explored in the interviews. These are all shown in Table 7.

Top Level Codes	Data; Software; Skills; Diversity; Multidisciplinarity; Research
Sub Codes	Data Management Plans; Open Source; Storage; Sharing
Queried Terms	Ethics; Licences/Licensing; GDPR; Policy; NCRM; Open source; Collaboration; Gender

Table 7. Interview Coding Framework.

The transcripts are the final data, and the audio will be destroyed after the project.

4. RESULTS

This section contains the results of our primary research. First, we present the characteristics of the community, including the samples reached via our survey and interviews (Section 4.1). We then look at practices around research data (4.2) and research software (4.3). Section 4.4 examines the policies related to data and software in more detail, and Section 4.5 reviews open source software practices. Section 4.6 examines research ethics considerations. Finally, Section 4.7 looks at skills and training issues in data and software.

4.1 SOCIO-DEMOGRAPHIC CHARACTERISTICS

In this section, we discuss the demographic characteristics of our survey respondents and interview participants. We also include the results of our interview questions regarding diversity and funding information.

Survey respondents

This section provides a comparison of the survey data demographics with the ESRC community as seen through the lens of the GtR portal[41]. The R markdown[42] used to generate the analysis and the markdown output[43] have both been made fully available.

The survey received 168 responses, of which 164 were valid. We asked demographic questions about career stage, gender, dis/ability, ethnicity, and discipline/institution. A detailed analysis has been made fully available in the project's repository[44]. The number of participants allowed us to segment our results by career stage and gender, although our respondents were primarily white and stated no disability so we did not have sufficient data to examine differences in these categories.

However, we can use this information to examine how representative our survey data is of the ESRC community as derived from the GtR data. We considered only active projects when comparing the data to the results of our survey (Fig. 7).

[41] https://gtr.ukri.org (last accessed May 5 2022)

[42] R markdown used to generate the analysis of the survey: https://github.com/softwaresaved/esrc-software-study/blob/main/Src/ESRC.Rmd (last accessed September 29 2022)

[43] Markdown output from survey analysis: https://github.com/softwaresaved/esrc-software-study/blob/main/Src/ESRC.md (last accessed September 29 2022)

[44] https://github.com/softwaresaved/esrc-software-study/blob/main/Src/ESRC.md (last accessed September 30 2022)

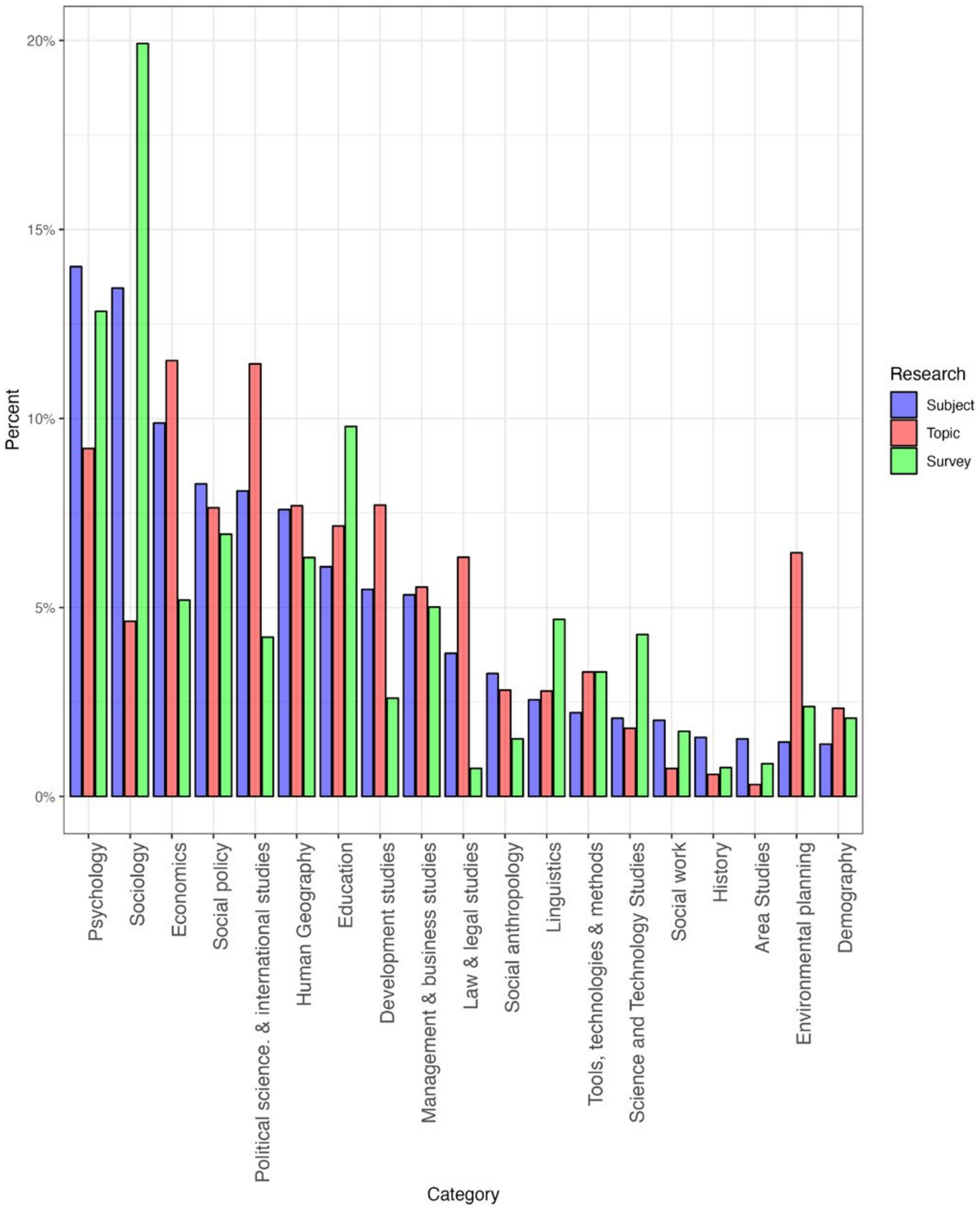


Figure 7. Representation of how the survey responses compare to the distribution of GtR data split into categories, using research subjects and topics for active projects only.

It is important to note that the division of the number of GtR awards partitioned into the survey categories gives us an indication of the size of the community, which is then compared to the survey data. This can only be used as indicative.

We can also use the GtR data to compare the total number of awards with the number of awards across institutions. Fig. 8 shows only GtR institutions that have more than seven awards to reduce the size of the distribution tail.

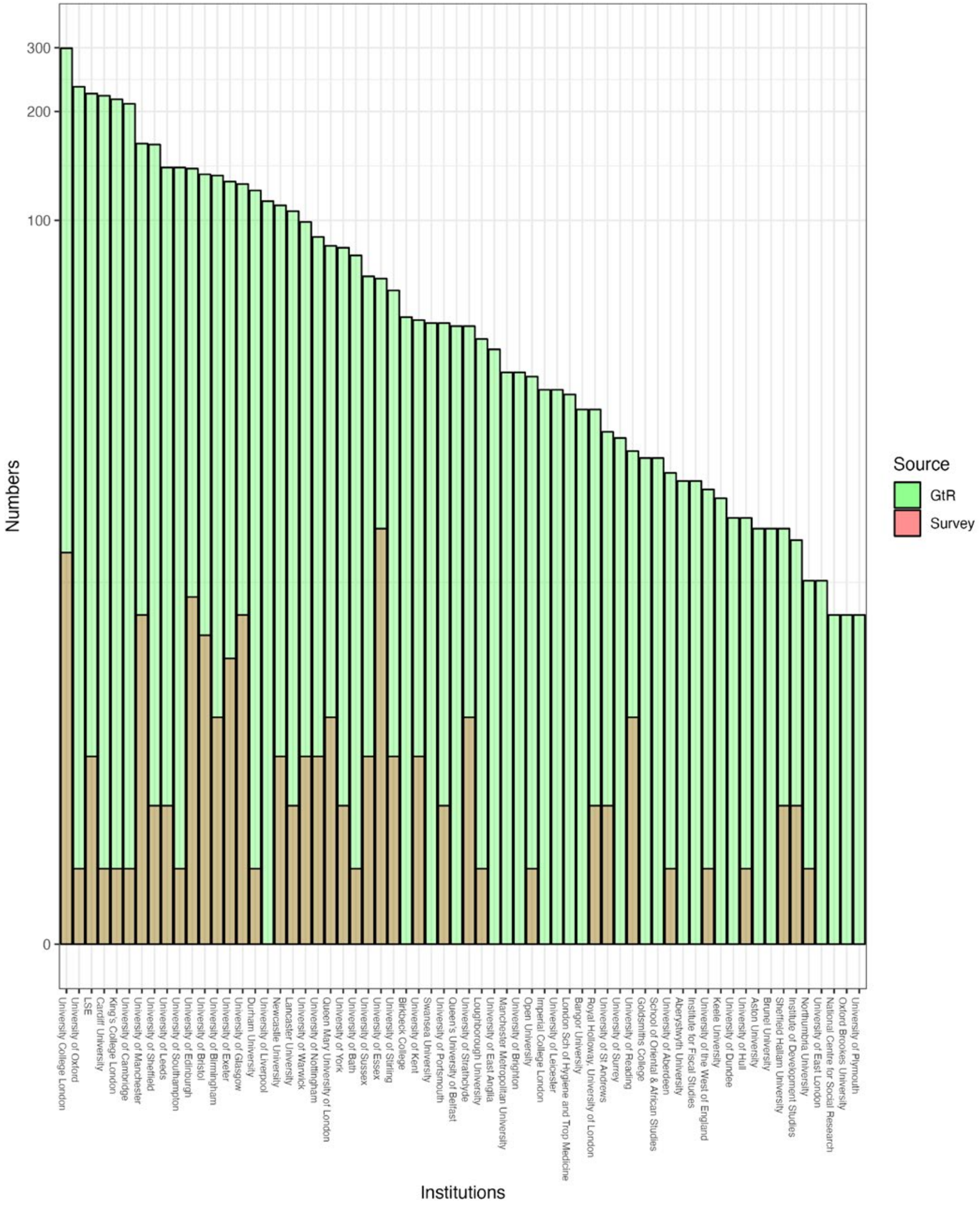


Figure 8. Distribution of GtR awards by institution restricted to values that have more than seven awards (upper green bar chart) and the corresponding survey results by institution (red bar chart or lower bars).

By comparing the HESA data with our survey participants we can see in Table 8 that the split between male and female is ~50:50, and the split between no disability and disability is roughly 95:5.

Category	Number of Valid Respondents (N = 164)
Disciplines[45]	Area studies: 4 Demography: 6 Development studies: 8 Data science and artificial intelligence: 16 Economics: 13 Education: 23 Environmental planning: 6 History: 3 Human geography: 16 Information science: 5 Law: 4 Linguistics: 11 Management and business studies: 10 Other: 24 Political science and international studies: 10 Psychology: 21 Science and technology studies: 13 Social anthropology: 4 Sociology: 43 Social policy: 22 Social work: 4 Tools, technologies, and methods: 11
Career Stage	Junior career researcher (JCR): 49 Early career researcher (ECR): 38 Mid-career researcher (MCR): 38 Senior (SNR): 33 Other/Not stated: 6
Gender	Male: 65 Female: 89 Other/not stated/not disclosed/non-binary: 10
Ethnicity	White: 130 Other ethnicity: 26 Undisclosed: 8
Disability	No disability: 133 Identified disabled: 25 Undisclosed: 6

Table 8. Summary of the survey demographics.

[45] Respondents were allowed to make more than one choice of research discipline. Here the raw count is shown, which has an N = 277 (about 69% higher than the actual number of respondents). The "Other" choice was included for those who felt that none of the labels applied to them.

Interview participants

By comparing the HESA data with our interview respondents we can see we have respondents from across our four career stages (Table 9). One respondent identifies as disabled. Women are slightly underrepresented. Towards the end of the recruitment process, we made a slightly greater effort to find respondents from as many disciplines as possible.

Category	Number of Respondents (total = 21)
Disciplines	Sociology, Geography, Psychology, Web Science, Economics, Maths, Architecture, Social Methods, Science Communication, Urban Studies, Political Science, Education, Peace Studies
Career Stage	JCR: 6 ECR: 5 MCR: 5 SNR: 5
Gender	Male: 12, Female/non-binary: 9
Ethnicity	White: 15, Other ethnicity: 6
Disability	No disability: 20, Identified disabled: 1

Table 9. Overview of the interview demographics.

Diversity issues

Our interviewees felt that the use of data and software was affected by issues such as age, ethnicity, being in the global north or south, gender, caring responsibilities, and a lack of diversity generally. The direct impacts were seen as follows:

- **Age:** Respondents cited age as a barrier to engaging in in-person activities: "I do most of my stuff online rather than going to places" [7227322280][46].
- **Ethnicity:** Interviewees expressed concerns that they might experience problems with data collection and access. One interviewee stated that they omitted their second name when applying for government data: "I'm a bit afraid over my second name, they might realise that I'm a foreigner. And that might, you know, slow the process." [7869268779]
- **Global south:** some interviewees noted that while researchers in the global north are most often associated with institutions and their associated infrastructure, researchers (and often collaborators) in the global south did not have access to proprietary tools and technologies. This "very strong financial and technological barrier is not conducive to sort of a global open world of research and data sharing" [5766299900]. Running international projects is problematic. However, open source software is perceived as a partial solution to this problem (although GDPR still creates problems around data sharing).
- **Gender:** Some interviewees felt that software in general, and the open source community in particular, is still "very, very, very male-dominated in every aspect" [5766299900]. This leads to "stigma" [5766299900]. This was a perception from outside the field. "Being a woman in the field of education in languages is perhaps not so much of a disadvantage compared to maybe the STEM subjects because we tend to have a larger, I think, representation of female researchers compared to engineering, for example, [which] tends to be male dominant." [7183719943] However, Communities of practice like R-ladies[47], which promotes the use of the R language among women, were seen as mitigating this and facilitating change.
- **Caring responsibilities:** Training was perceived to have become more accessible during the pandemic, as "things really opened up and it's much more acceptable to deliver things online now, which is brilliant." [5996360861]. Conversely, "sometimes I find it, it's like, really nice when you're learning a new skill to actually do that in person, and have someone experienced in the room with you." [5996360861]
- **Lack of diversity:** This was seen as limiting the perspectives, experiences, and even the theory, that are brought to bear on the collection or analysis of data.

[46] Numbers in square brackets refer to individual anonymised interviewees

[47] https://rladies.org (last accessed August 15 2022)

Funding

In the survey we asked (Q18, N = 153) who had funded the respondents' work over the last five years to get an idea of the funding streams that were used to finance their research. This was an open text box where respondents could list funding bodies, and as such some postprocessing was required to obtain a tractable result. The summary of this processing is shown in Fig. 9, which only shows results that have more than two entries. 11 people did not respond.

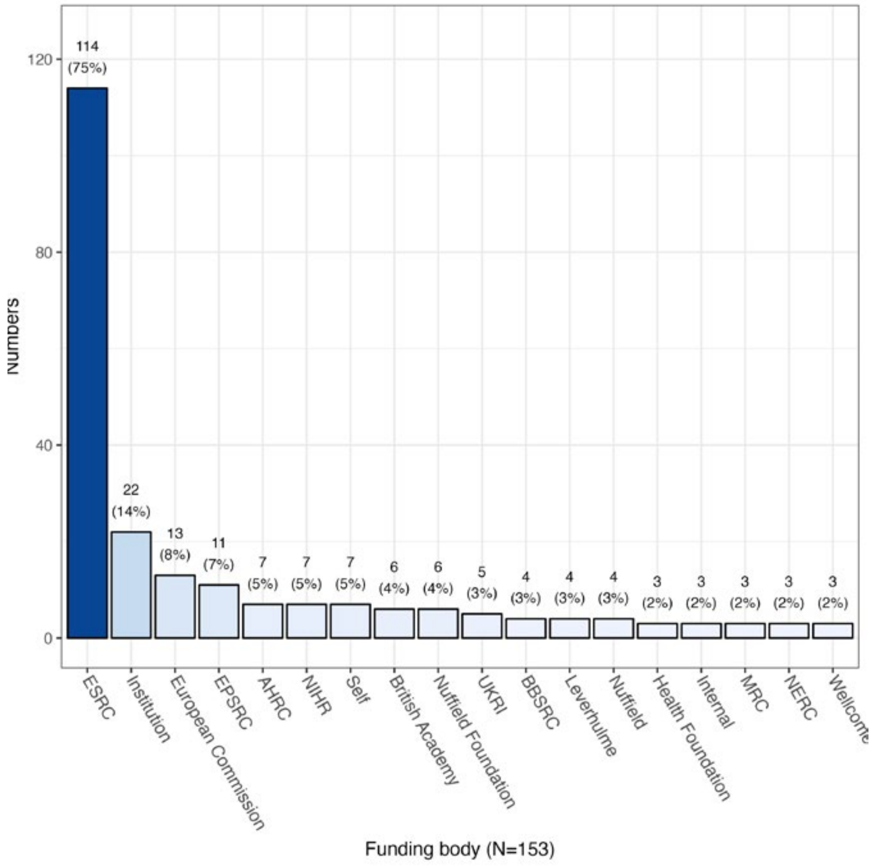


Figure 9. Funding sources obtained over the last 5 years. Only funding bodies with more than two occurrences are shown.

As might be expected in a survey aimed at the ESRC community, the ESRC dominates as the main funding body. Of those respondents who were not funded, in part or entirely, by the ESRC, five were exclusively funded by their institutions, four were exclusively self-funded, two were retired, and the remaining 28 were funded by a variety of combined funding lines that did not include the ESRC.

4.2 RESEARCH DATA PRACTICES

In this section we discuss key sources of data, and issues regarding its collection and access. We look at data storage and sharing, and the role of data management plans.

Using and creating data

Understanding what datasets researchers collect, create, or reuse informs our understanding of what kind of software might be required to collect, manipulate, analyse, and finally publish the data.

In our survey, we asked people to tell us what their most important sources of data were (Q2a, N = 144). All important data sources with more than two responses are shown in Fig. 10.

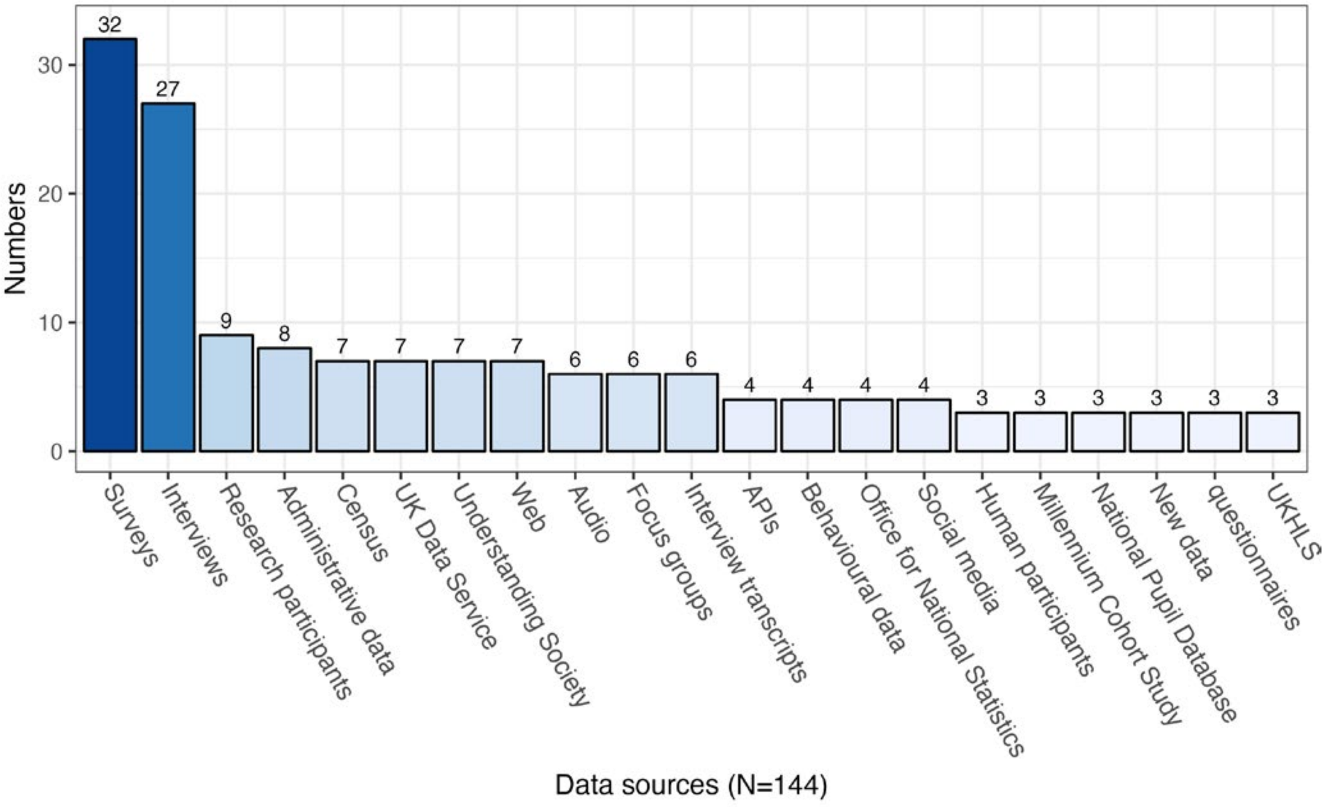


Figure 10. Indication of the most important data sources.

Survey (22%) and interview (19%) data were by far the most significant forms of data in use. One slightly confounding issue is the confusion of 'data format' with 'data type'; for instance, audio may well be recorded from interviews.

Interviewees reported using a wide range of datasets and data types, such as "archival data, images, videos, cartoons, interviews, social media" [6776362537]. Some were using data available from the UK Data Service. Some were using commercially held data, e.g. from mobile phones, social media, and app companies that provide APIs for accessing data held in restricted access secure environments. Not all data was natively digital. For instance, one respondent was working with UK Government data that had yet to be digitised, and was therefore manually inputting it from paper records.

In Fig. 11, the most important data sources are split by career stage. The biggest data source (surveys) is favoured by respondents in mid- (31%) and senior career stages (32%), followed by early career respondents and, to a lesser extent, junior researchers. When we look at the next most popular data source (interviews), the composition by career stage seems to be inverted, with junior researchers (25%) most likely to cite this as their most important data source, followed by ECRs and then their senior and mid-career counterparts.

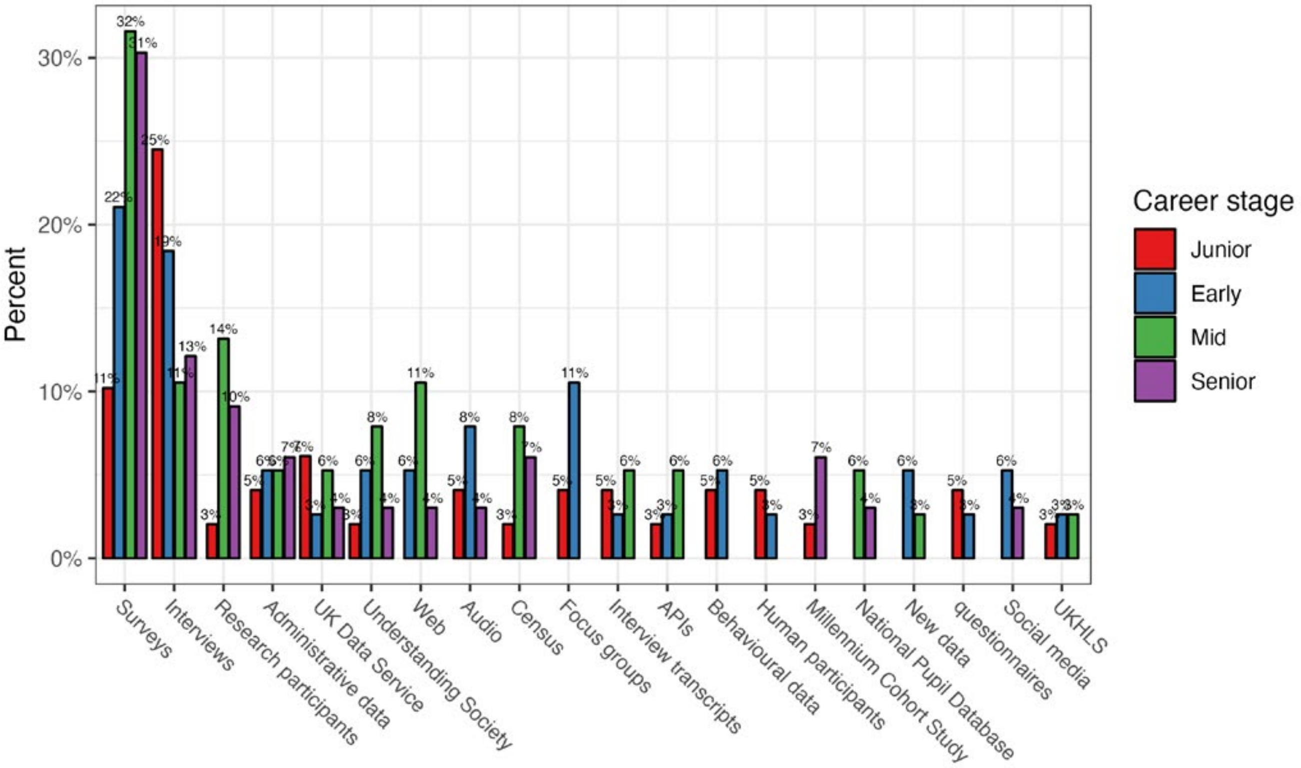


Figure 11. Most important data sources by career stage.

We can examine the data by gender (Fig. 12). It is important to note that junior researchers in our data set are mostly women (80%). Percentages here are given as a proportion of the corresponding gender population specifying that choice. Census and ONS data seem to be exclusively used by men while data from human participants and the National Pupil database are exclusively used by women.

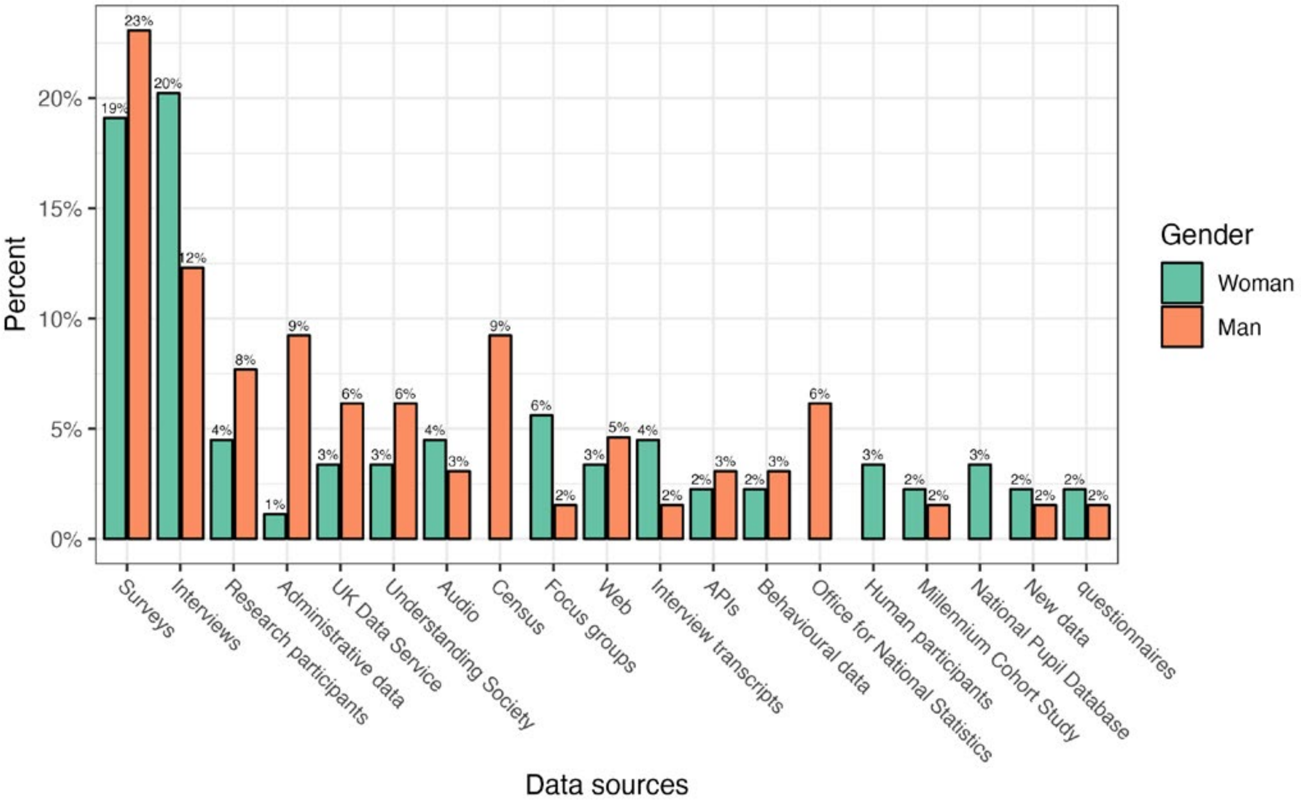


Figure 12. Most important data source by gender.

We also asked in the survey whether researchers created or reused existing datasets (Q2, N=163). This was a checkbox question, allowing respondents to select "create" and/or "reuse". This approach allows us to separate results into those who selected "create", "reuse", or both. Responses are summarised in Fig. 13.

Most survey respondents report both creating and reusing datasets (53%). Using this information, we were able to establish how much data is created, reused, or both for each of the most important data sources, as shown in Fig. 14.

Most data types that are created arise from interactions with people. They include surveys, interviews, research participants, and so on. Data creation and reuse take place across all data sources, other than the two sources that exclusively involve data reuse: the Millenium Cohort Study and the UK Household Longitudinal Study.

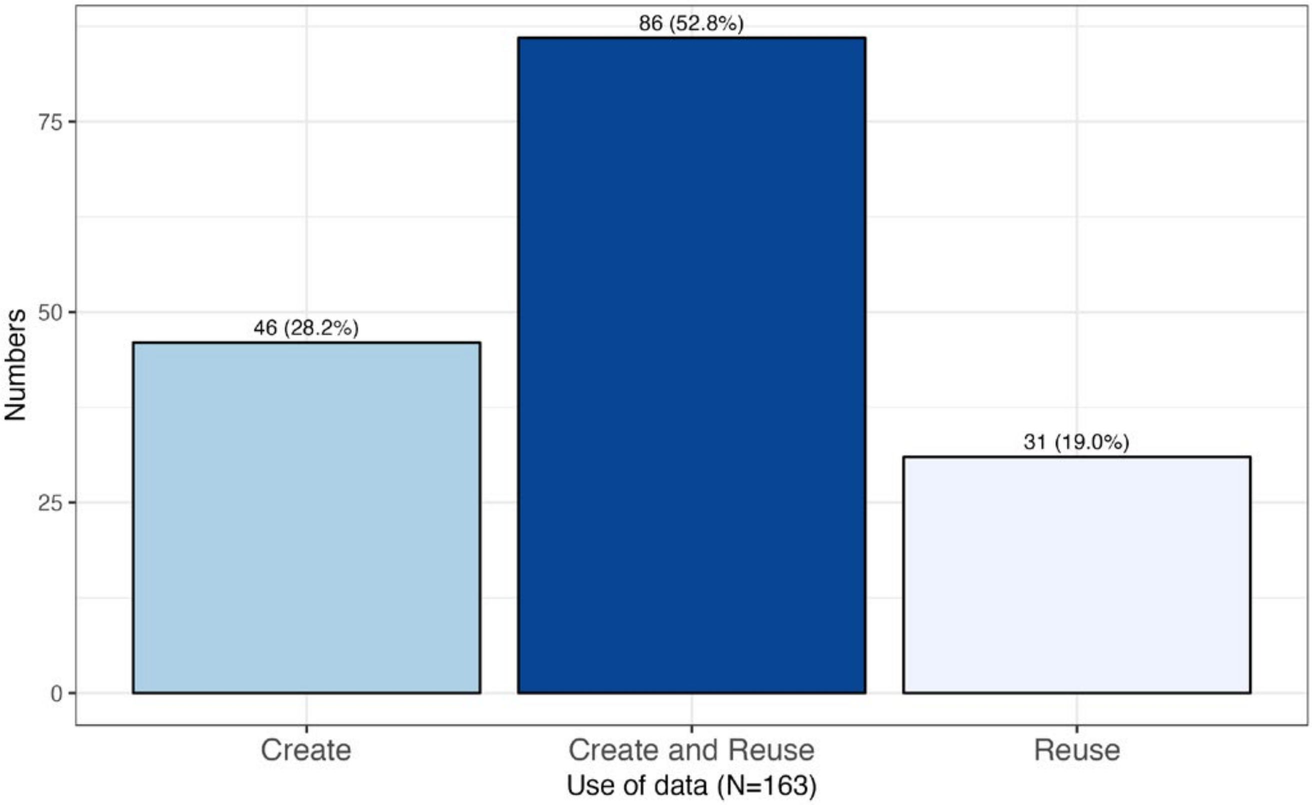


Figure 13. Responses indicating whether data is created-only, reused-only, or both created and reused.

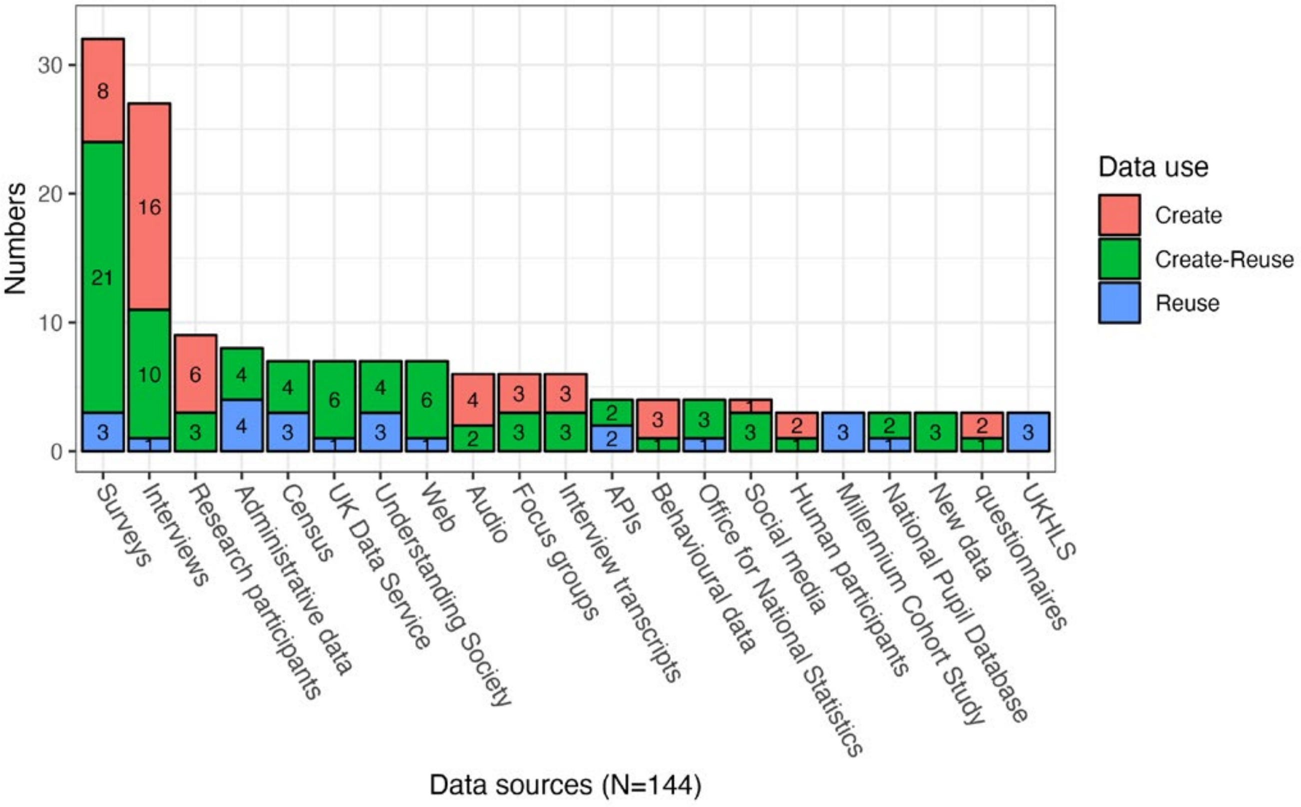


Figure 14. Most important data sources by whether data is created or reused.

It is also possible to order "create", "reuse", and "both" responses by career stage. Junior researchers (47%) appear far more likely to exclusively create data, while respondents at other career stages show a greater tendency to both create and reuse data. From Fig. 14 and Fig. 15, we can infer that the data junior researchers are creating is likely to be interview data.

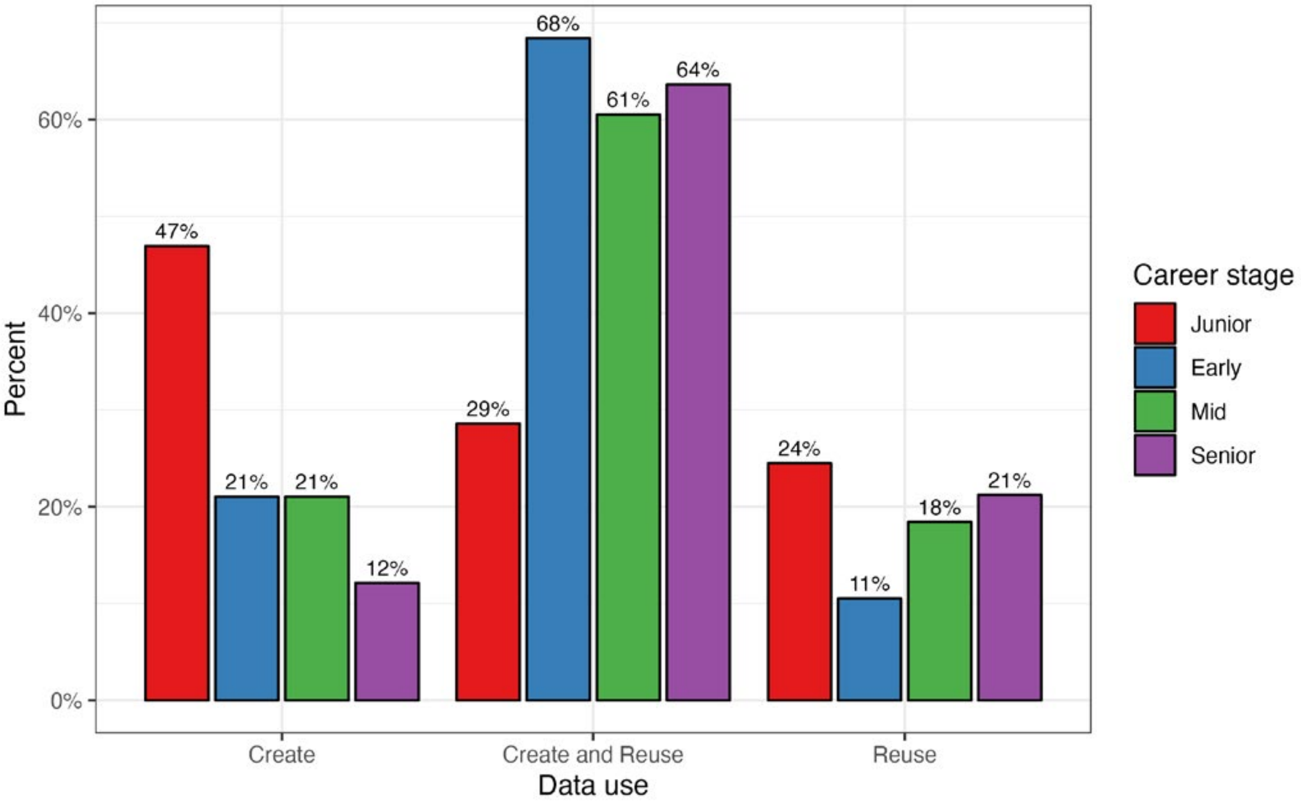


Figure 15. Created-only, reused-only, and created and reused by career stage. Percentages correspond to the number of responses at that particular career stage.

Although ESRC Research Data Policy encourages researchers to reuse data, the time it takes to request and obtain ESRC or government longitudinal data from certain sources remains a barrier. Many interviewees reported that access to more granular or secure datasets in these categories can involve a lengthy process.

> "In principle, as an academic, you could maybe apply, get clearance and after [three to] six months ... get ... access to [a] secure environment. [There are] maybe three or four in the country, and then you might get access for half a day a week if you're lucky." [1578450175]

Our interviewees revealed a diverse range of approaches to securing access. They include creating synthetic replicas of the datasets, seconding employees, and deputing project members to the specific task of navigating the protocols necessary to access secure data.

> "This particular colleague of mine is absolutely critical. Actually. I don't think ... any other member [of the team] would have had the diligence, the persistence, and the ability to do what she's done, she negotiated the access to [specialist government] data." [7336263536]

Despite this, some projects experienced fatal setbacks due to the onerous timing constraints and labyrinthine complexities of accessing data.

> "We spent six months, mainly ... waiting to hear back and being told ... we need this extra information. And it was just too complicated for us to even bother doing in the end." [4561769548]

In the survey, we asked those who were reusing data (Q2b, N = 126) about their sources. The options were as follows:

- UK Data Service (UK Data Service)
- University / Institutional Repository (University/Institutional Repo)
- General data repository (e.g. Dryad, Figshare, Open Science Framework, Zenodo) (General data repository)
- Shared by collaborators using a shared drive/folder (e.g. DropBox, Google Drive, ...) (Collaborator's shared drive/folder)
- Personal recommendation, eg from discussion with other researchers (Personal recommendation)
- Other (Other)

These were checkboxes, allowing respondents to select every option that applied to them. The responses are summarised in Fig. 16.

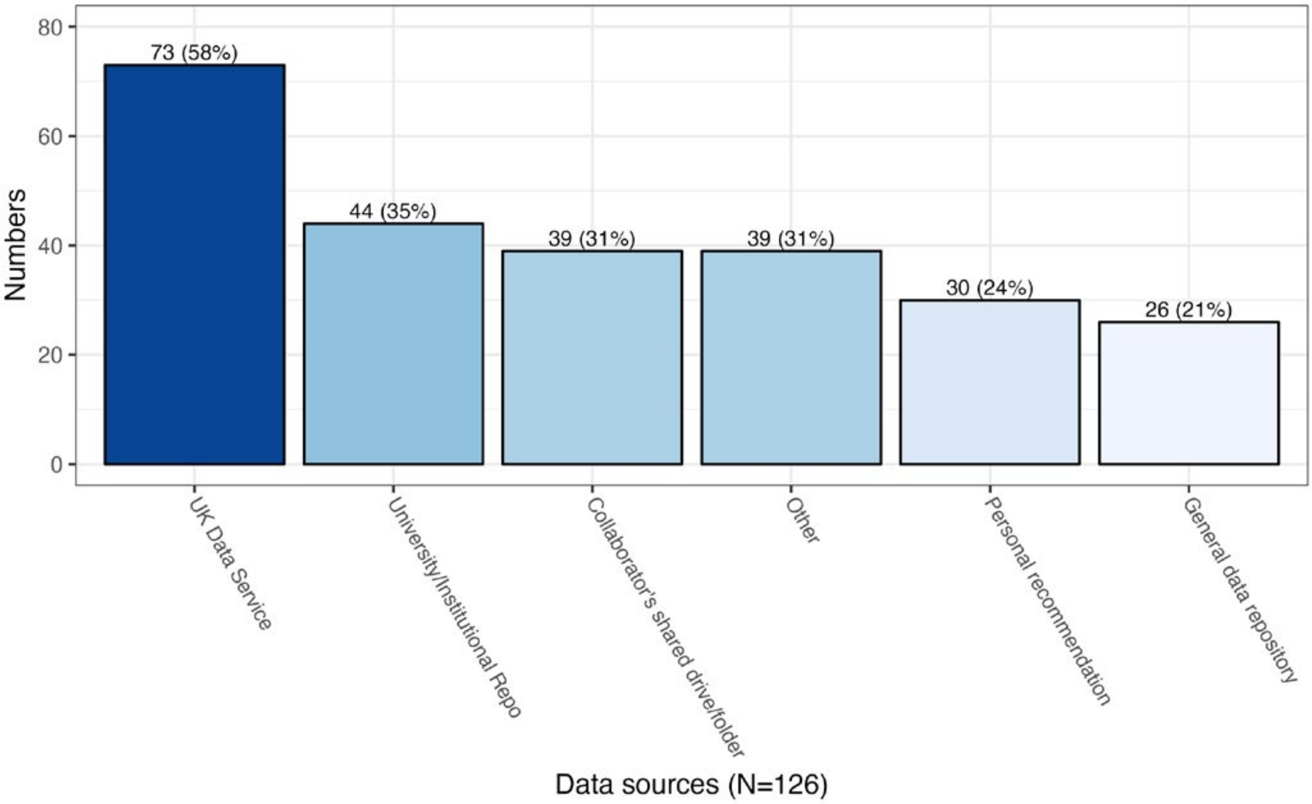


Figure 16. Where secondary data sources are obtained.

The main source for external data is the UK Data Service (58%) followed by institutional repositories (35%), shared drives (31%), and Other (31%).

Within the Other category, several additional sources were mentioned more than once: government data sources (5.5.%), the ONS and ONS-SRS (4%), researchers' own data (2%), and the NHS (1.3%). Of the remaining 55 suggestions, made by 39 of the respondents, the only one with ESRC funding is CLOSER[48].

Fig. 17 shows responses by career stage. What stands out in this diagram is that junior career stage researchers are using institutional and general repositories in insignificant numbers, though many still use the UK Data Service.

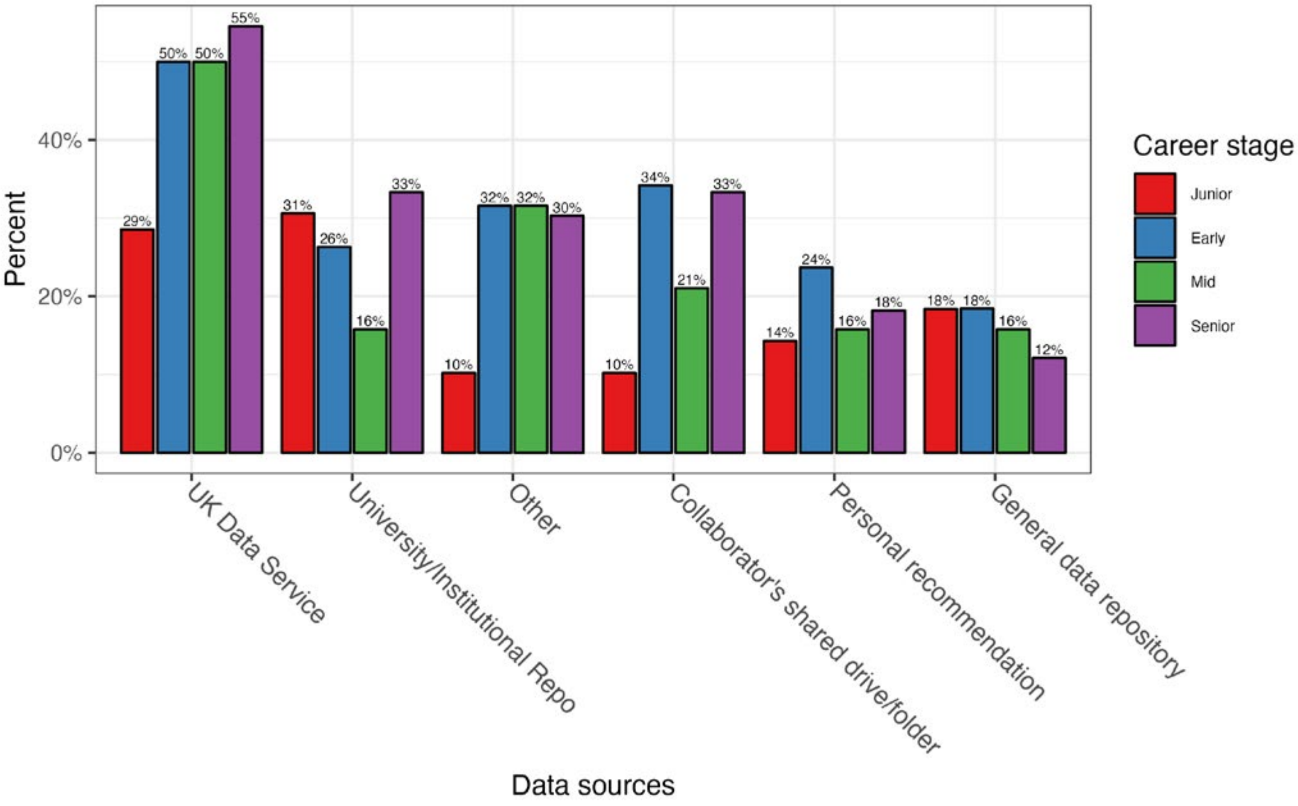


Figure 17. Data sources for data reuse segmented by career stage.

Fig. 18 shows service use by research discipline. Percentages correspond to normalised research discipline split across the different data sources. This approach allows us to see the extent to which each research discipline is making use of the different data sources.

[48] https://www.closer.ac.uk (last accessed August 29 2022)

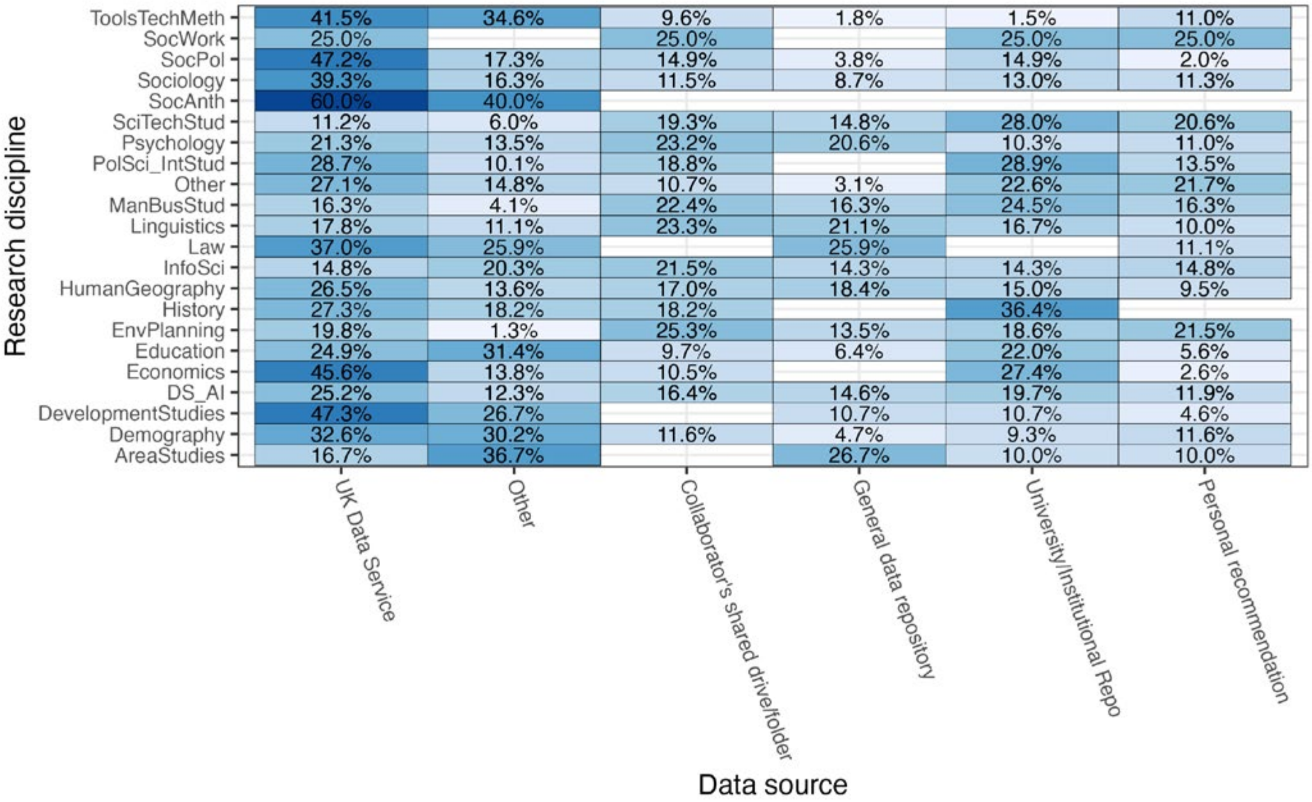


Figure 18. Use of data sources by research discipline.

Sharing data

We asked (Q3, N = 160) how respondents share their data. The following options were given:

- I have not yet shared my data (NotShared)
- I have licensed my data to allow it to be shared (Licensed)
- I have shared my data with individuals/groups that have requested access (RestrictedSharing)
- I created a DOI (or other unique identifiers) to make my data findable (CreatedDOI)
- I have promoted my data as an accessible resource in my publications (Publications)
- I have deposited my data in a repository (Repository)
- Other (Other)

In Fig. 19, we see that most survey respondents either do not share their data (49%), or have deposited data in an institutional repository (36%).

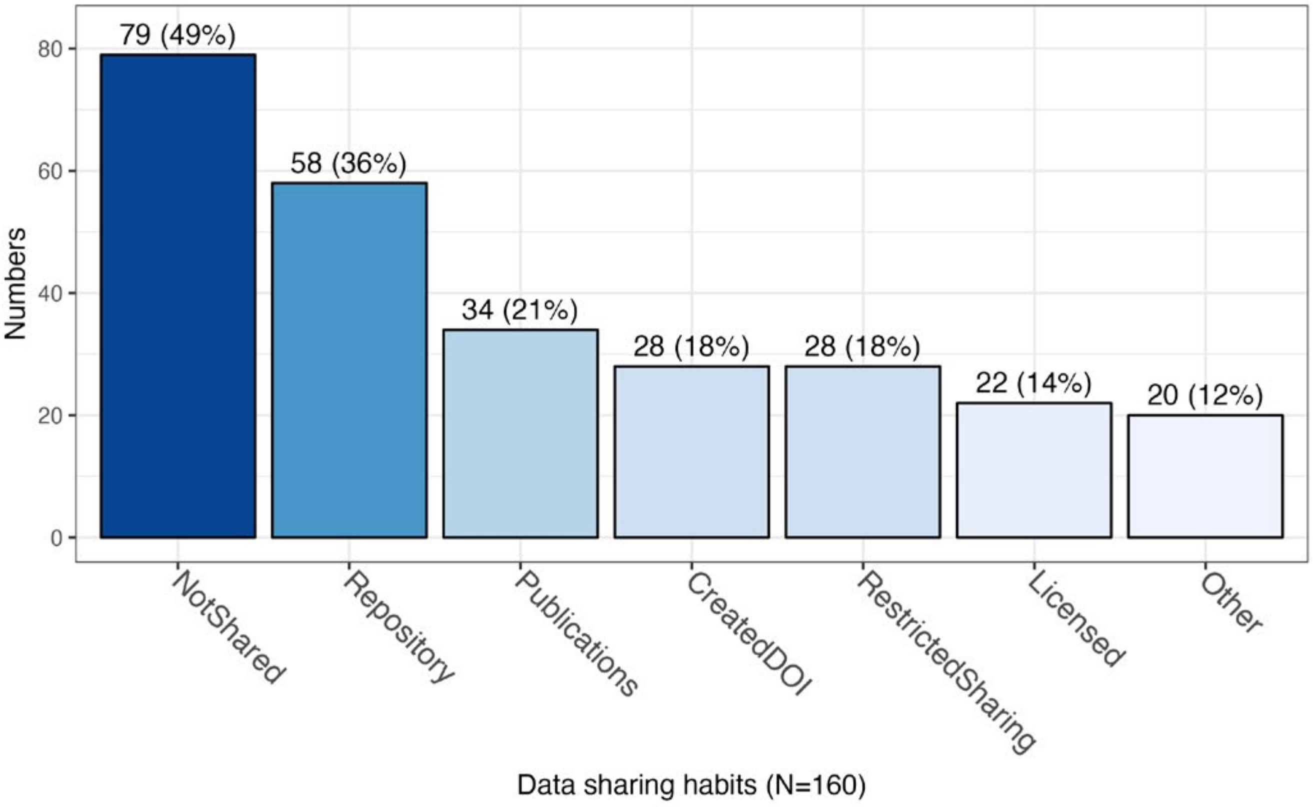


Figure 19. How data is shared.

We segment the choices above by career stage in Fig. 20. 73% of senior researchers deposit their data, whereas 82% of junior researchers do not. Junior career researchers often share their data at the end of their PhDs, which could explain the high number of respondents in this career stage not sharing data. However, early (37%) and mid-career researchers (53%) do not seem to share their data either.

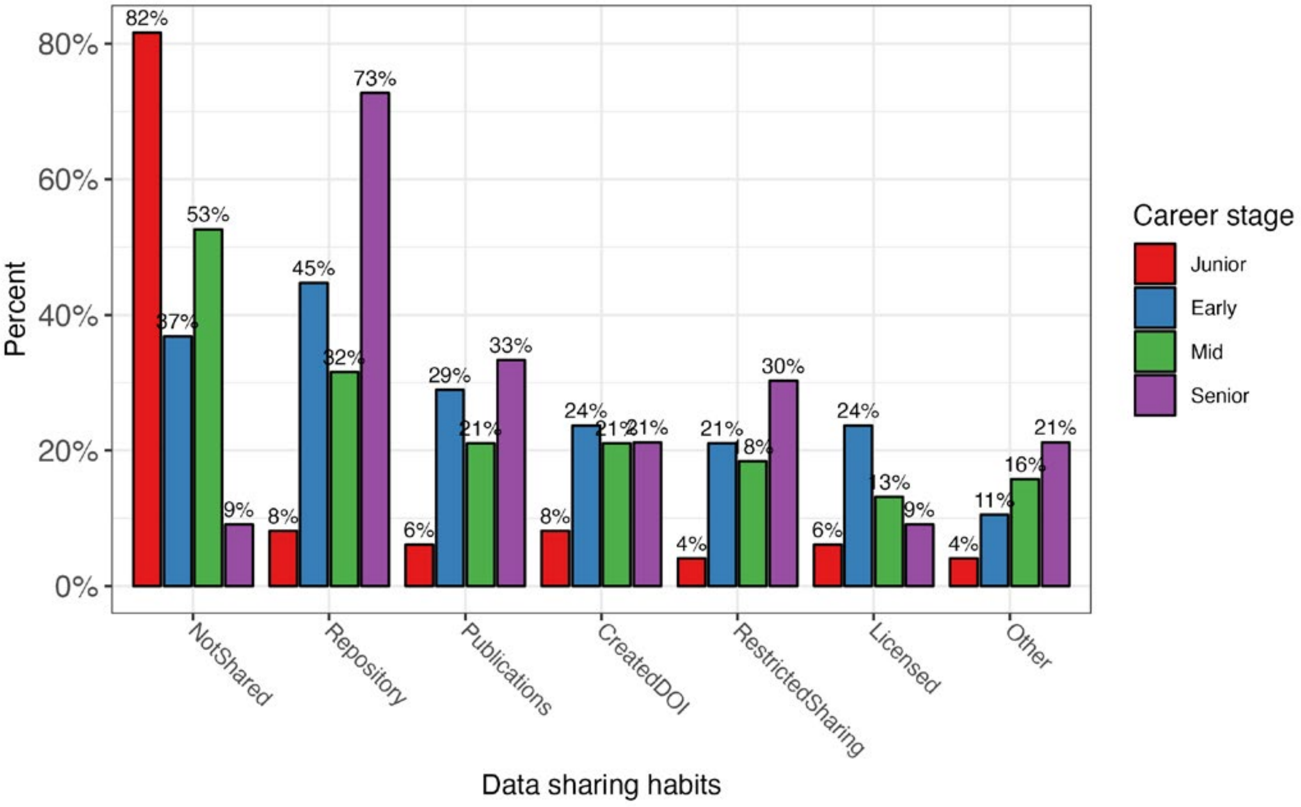


Figure 20. Data sharing by career stage.

From interviews, we learned that the limited nature of JCR and ECR contracts means that junior researchers can be cautious about putting all their data on institutional servers.

> "I'm conscious that when my funding runs out, or when your PhD finishes, you get immediately locked out of the university accounts that you have, and some of that's my work to collect, pulling all that data from different sources and stuff. So I'm conscious that I want to be able to have a backup somewhere else of the stuff that [I want] to hang on to." [5996360861]

Institutional repository choices also affect sharing and storage. Each institution has its own research storage space, and institutional storage choices are subject to change, creating complexity for researchers in managing their own data before additional sharing activities.

> "I have a paid subscription to Dropbox[49]. But now we're using Box[50]. And before COVID, we didn't. And I had to navigate between Box, Dropbox and my own computer when I'm working with data, which is a bit hectic." [7869268779]

Finally, inter-institution projects can be complicated by differing storage policies.

> "If you work with another institution, and they want to use Dropbox, but your university prohibits the use of Dropbox, ... then people end up defaulting to things like Google Drive." [4561769548]

As a result, there is a 'grey infrastructure' of research storage across the UK, much of which can be found on Google Drive. Although our research was focused on UK-funded projects, it is worth noting that many EU projects are officially managed from Google Drive, adding to the total sum of data and software stored there. When we asked one participant whether there essentially exists a mirror image of the entire research infrastructure in the UK on Google Drive, they replied, "Yeah, absolutely. Absolutely." [4561769548]. While this is anecdotal, it is illustrative of the way Google Drive is perceived. The widespread use of Google Drive exposes the UK research infrastructure to all potential risks associated with using unofficial, commercially owned products and services.

Fig. 21 shows data sharing choices by gender and career stage.

[49] A cloud storage system https://dropbox.com/ (last accessed August 18 2022)
[50] A cloud storage system https://www.box.com/ (last accessed August 18 2022)

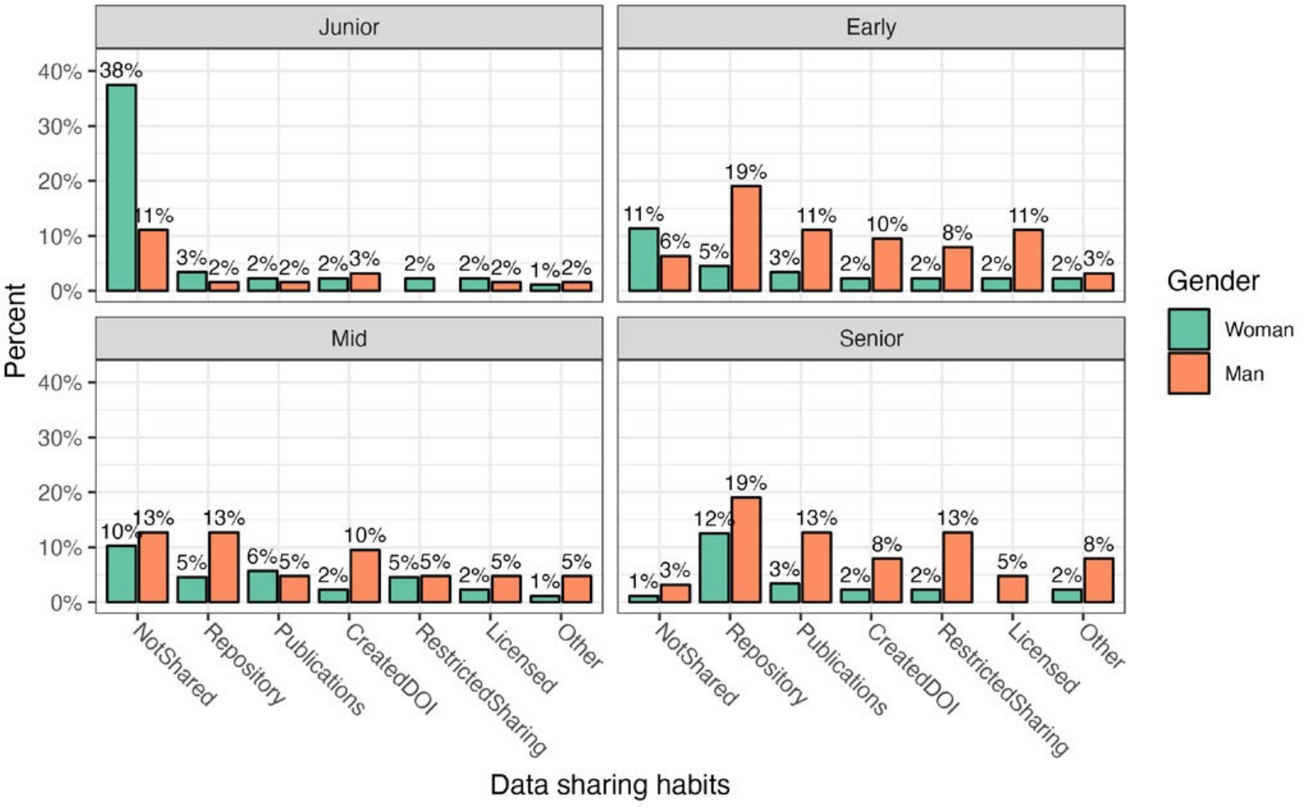


Figure 21. Data sharing by gender and career stage.

It is notable that women appear less likely to share data than men, even though in Q2 of the survey women demonstrated a tendency to create data more than their male counterparts (82% vs 77% respectively for each population component). One potential confounding issue is that women are more likely to be working with qualitative research data in the social sciences. This data is harder to anonymise (Fig. 32).

We analysed the data through the lens of Q2 responses, i.e. created, reused, or both, in Fig. 22.

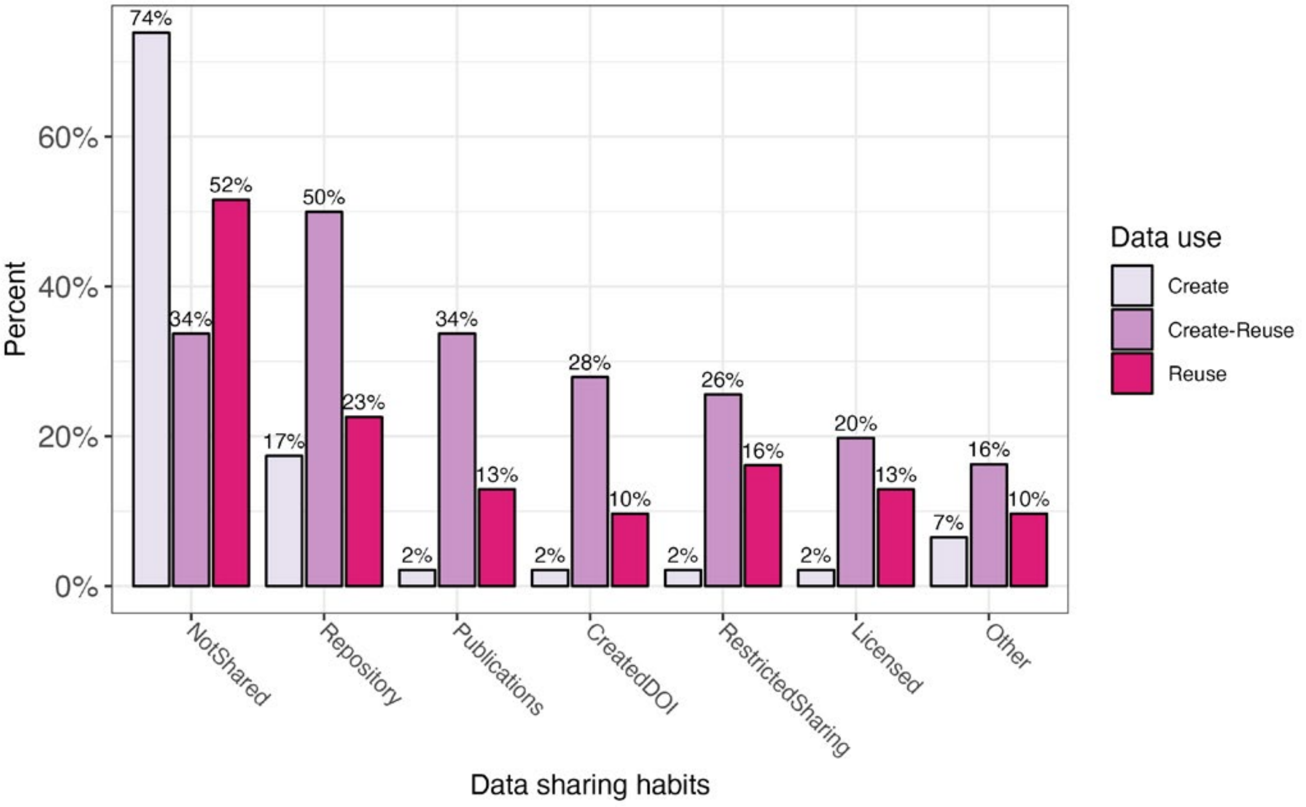


Figure 22. Data sharing through the creation-only/reuse-only/both lens established by Q2.

Interestingly, 74% of those who exclusively create data do not share it. This phenomenon may be partly attributable to the difficulty of anonymising the data in question. Our interviews show that there are major barriers to sharing data, and a general lack of clarity among respondents around institutional and funder policies, such as those relating to data management plans and open research.

Some of these barriers arise from GDPR restrictions and ethical concerns. This is particularly true of qualitative data to which computational methods such as differential privacy cannot be applied.

> "I think one of the questions that I have, especially with qualitative data is how do you make data publicly available but still keep the participants' identity anonymous." [7183719943]

There is also lack of clarity around the right to deposit third-party or reused data:

> "There's still some ethical like questions .... like, are we allowed to use this? In what format? Can we use it? In theory, I have an academic Twitter API access. But I got that on the basis of describing the particular project and what data I was expecting to get and use. Therefore, it seems to me like I'm probably not allowed to just share that." [7590937187]

> "I suspect the Strava data may have some conditions ... what are the conditions associated with it?" [4938919540]

Sharing of these kinds of data is also inhibited by a lack of awareness among researchers of an appropriate repository.

> "I haven't published them officially through the UK Data Service, because we don't have a good platform for things like Twitter data." [7590937187]

The absence of a perceived benefit is another disincentive for sharing data, particularly given the potentially time-consuming practical complexities of, say, preparing multiple interviews:

> "Whether I think anybody would ever use any of those datasets, is a very open question." [4561769548]

> "But when you're busy doing research, the last thing you think of is archiving it." [7227322280].

> "Transcripts certainly could be made public. I haven't made my set of data of transcripts public yet. [...] It ends up being quite a large set, because it's so many different files, and I've not made that public." [7183719943]

Even if the data is considered to be potentially valuable, it is not entirely clear that researchers know what to deposit:

> "If you don't know how somebody's going to use your data in the future, how do you know what to keep?" [4561769548]

A final constraining factor is the fact that the REF, along with institutional systems like workplace promotion pathways, reward publication rather than the amount of data a researcher has deposited.

We further analysed data sharing habits againsts respondents' main funding sources in the last five years (Q18, N = 153). 114 respondents had received some form of ESRC funding, 39 had not, and we were unable to obtain relevant information for the remaining 11. The results are shown in Fig. 23. The percentages relate to those who answered Q3.

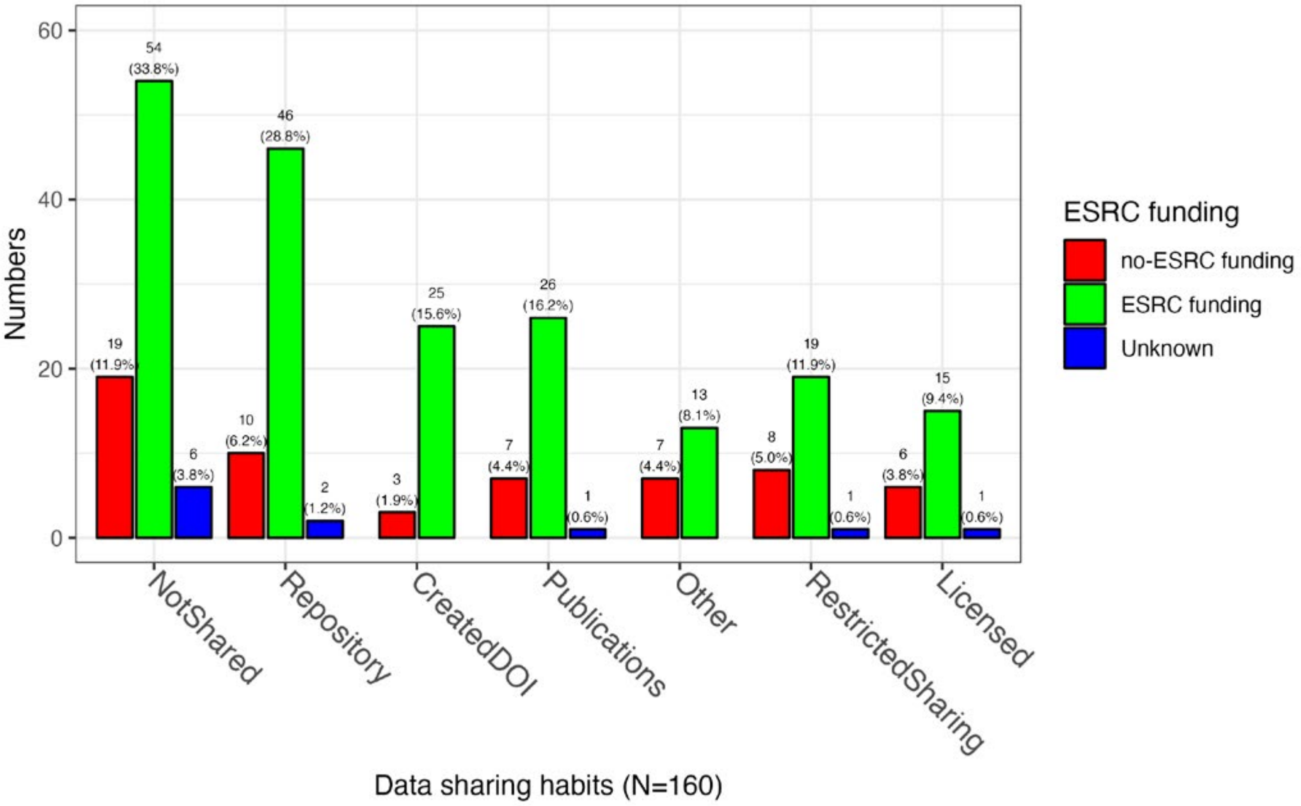


Figure 23. Sharing of data by funding source. Percentages are in relation to the number that answered Q3.

We see a similar split between not sharing and depositing in a repository. However, it is interesting to note that the majority of ESRC-funded respondents do not share their data (33.8%), while 28.8% of ESRC-funded respondents deposit their data in an institutional repository.

Among those selecting other methods for sharing their data, it was noted that the data in question was either not owned by the respondent or under embargo, and hence could not be shared. Others were sharing their data through a website, Zenodo[51], or a university server. A senior career researcher (SCR) respondent noted that

> "the local depositing process is intimidating and usually one gets on with new work rather than spending time making existing data more widely available. One hopes that younger members of staff are more flexible and responsive than I have been." [SCR]

But, as we have seen, the data sharing process is still eschewed by many researchers, though some interviewees felt that their institutions were supportive of sharing data:

> "I do think universities, in lots of ways, are trying to do good things ... around data. I think the library put a lot of time and resources into trying to help people deposit their datasets and I think there's some interesting stuff there." [4561769548]

This is also true of academic institutions in a wider sense: "I think the REF has tried to introduce some notions about that as well." [4561769548]. Some respondents belonged to a tradition of open science or computer science, and thus were in the habit of publishing on open repositories like the Open Science Framework[52] (OSF) or GitHub[53]. The OSF was seen as a resource where like-minded researchers can look for data. GitHub was seen as a more flexible repository for material that might be difficult to publish on the UK Data Service, such as Twitter data.

[51] https://zenodo.org/ (last accessed March 7 2023)

[52] https://osf.io (last accessed August 19 2022)

[53] https://github.com (last accessed August 19 2022)

Data management plans

Data management plans are the generally adopted method for establishing what will be done with data at each stage of research. Every research council with the exception of the EPSRC requires that submissions include a data management plan.

Many of our interview respondents reported that they were familiar with data management plans, but had never been called upon to write one.

> “So we have an IT and security manager who is dealing with these [...]. For bigger projects, they are doing these kinds of plans.” [7869268779]

A number of interviewees knew how to access help in their institutions for writing data management plans, or had received training.

> “I had, at [the] University where I am ... training us on how to write that management plan. And I just had ... training on data management and writing a data management plan in January from our university.” [6776362537]

> “The data management plan that you mentioned earlier, I went to a workshop on that. They talk about how to share data, they talk about how to make research more accessible. So yeah, in terms of data management, I think there is quite a lot of support about that.” [2964377624]

Some respondents were unsure whether they had written a data management plan or not, implying a lack of conviction as to their value. Other respondents, however, recognised the purpose and value of data management plans.

> “On one level, it does help you create a data management plan because you know what data you're applying for. And, you ... understand the characteristics of the data and where it's coming from and whom it's coming from and under what conditions it's held.” [5037840428]

Interviewees receiving funding from other councils had a much clearer idea of their purpose and requirements.

> “I mean, when you apply for an MRC programme part, you need to have a data management plan. So yes, lots of data management plans along with everything else.” [7336263536]

In many cases, data management plans are a component of the research ethics process, though not necessarily a welcome component or one with a clearly understood purpose.

> “I did a management plan as part of the ethics for my PhD. [...] It was part of the ethics application. And the ethics application was quite extensive. So I don't remember much about the specific data management part of it. I'm working on this project in [country], along with colleagues at the [university] and colleagues in [country]. And then, again, the data management plan is through the ethics application, and that's being submitted at [university] which is not my institution. So I'm contributing to it. But I've not had ... to do it if that's the question.” [7183719943]

> “I didn't have the time ... because I was scrambling to finish my thesis before the submission deadline. So what I didn't have was a formal data management plan. What I did have was in my ethics document, a description of where the data would be stored, how they would be accessed, and who would analyse them. So it was not a formal data management plan. But it sort of ticks a lot of the boxes of a data management plan”. [5766299900]

Often, they are driven by data protection considerations within ethics:

> “We had to create a data management plan for the data that we were collecting because we were doing focus groups, and it was being transcribed. There could have been some privacy concerns [so] we had to write up a data management plan as to how we would go about protecting the data, keeping it in an encrypted computer, and how we would transfer it between ourselves and when the data would expire, and all these things that had to do with data protection.” [4288358080]

Research data ethics and legalities

As shown above, fulfilling GDPR criteria and creating data management plans can end up essentially becoming one process.

> “I think at the submission point, sometimes you refer back to the data management plan to pull out the stuff when you're sticking your dataset in [an institutional repository]. It's often asking for metadata stuff that hopefully should be in your data management plan. And I think, thinking about data management plans, in terms of the whole curation lifecycle is useful from that perspective. So we often think ... in terms of the ethics at the start, so I need to tell ethics, what I'm going to capture, and where I'm going to store it. And that's your starting point for data management plan. But if you're serious about archiving this stuff, at the end, you need to look at that whole lifecycle.” [4561769548]

Data management plans can also be driven by a single major concern: to avoid accidentally identifying individuals in research data. This is a particular challenge when using data that is publicly available online.

> “If you put in an example of hate speech, that would then be personally identifiable, because someone could take that

> and search for it on Twitter." [9253010492]

This is both an ethical and a legal issue. Not only is it potentially unethical to publish information that could lead to the identification of a specific person, but the possibility of reidentification entails a potential breach of GDPR.

Inconsistency in ethics processes from institution to institution can also create problems.

> "It's the inconsistency which I do want to put on record, it's the outrageous inconsistency between ethics committees, so we might have an ethics committee around the corner that will say absolutely not. And then we get bounced up to somebody in [another institution] and they say, Yeah, sure, go ahead." [7336263536]

Crucially, institutional ethics processes are often incompatible with the UKRI's open data principles, which exist to promote data sustainability and open research, since these processes routinely require that data be destroyed after 10 years. One interviewee, faced with this stricture, told us,

> "I submitted ... an amendment to my ethics application with letters of support [from a collaborative partner] pointing out to the fact that I was not at all comfortable with my data being destroyed, they were a potentially unique source of data that might be worth examining at a future point or comparing with at a future point when you do some other research in this field." [5766299900]

4.3 RESEARCH SOFTWARE PRACTICES

This section reviews what researchers' software requirements are, what type of software they use, and whether they develop software for their research.

In our survey we gave the following definition:

> 'By "software", we mean any software or digital tool that you have used in the course of your research that has helped you undertake your research or produce a research output (e.g. a publication). This might be anything from a short script, such as one written in the Python or R computer languages, to help you clean your data, web/mobile apps, to a fully-fledged software suite or specialised toolset, whether you access this online or run it on your own computer. It includes code that you have written yourself and code written by someone else, especially or specifically for your project or a general tool for data, text or statistical analysis. It also includes the use and/or construction of spreadsheets that perform calculations or transformation automatically according to a set of pre-programmed rules, which are considered to be software.'

Use of research software

We asked (Q4, N = 161) what different software types they used in their research from the following list:

- Animation and storyboarding, e.g. Scratch, Storyteller (**AnimationStoryboarding**)
- Audio tools, e.g. Music Algorithms, Paperphone (**Audiotools**)
- Authoring and publishing tools, e.g. Twine, Oppia (**AuthoringPublishing**)
- Code versioning, e.g. GitHub (**CodeVersioning**)
- Content management systems (CMS), e.g. WordPress, Mura (**CMS**)
- Crowdsourcing, e.g. AllOurIdeas (**CrowdSourcing**)
- Exhibition/collection tools, e.g. Omeka, Neatline (**ExhibionCollection**)
- Internet research tools, e.g. Google tools or Wikipedia tools (**internetResearchTools**)
- Machine learning and artificial intelligence, e.g. leximancer (**ML_AI**)
- Mapping tools and platforms, geographic information systems, e.g. QGIS, CartoDB, ArcGIS (**GIS**)
- Mind-mapping tools, e.g. DebateGraph (**MindMapping**)
- Network analysis, e.g. GEPHI (**NetworkAnalysis**)
- Programming languages, e.g. Python, MATLAB (**ProgLanguagues**)
- Qualitative analyses, e.g. NVivo (**QualitativeAnalysis**)
- Simulation tools, e.g. NetLogo (**SimulationTools**)
- Spreadsheets, e.g. Excel, Google Sheets (**Spreadsheets**)
- Statistical analysis, e.g. R, SPSS, Stata, SAS (**StatisticalAnalysis**)
- Text analysis tools, e.g. Voyant, Linguistic Corpuses, Entity Recognizers (**TextAnalysis**)
- Text collation tools, e.g. Juxta Commons (**TextCollation**)
- Text Encoding, e.g. Oxygen XML (**TextEncoding**)

- Text and data wrangling, e.g. Overview, OpenRefine (**DataWrangling**)
- Topic modelling, e.g. Leximancer (**TopicModelling**)
- Transcription services, e.g. Otter.ai (**Transcription**)
- Video and film analysis, e.g. Cinemetrics (**VideoFilmAnal**)
- Visualisation tools, e.g. D3.js, Tableau (**Visualisation**)
- Other (**Other**)

This was a tick box survey, allowing respondents to select as many options as were applicable.

The summary of the responses is given in Fig. 24. We can see that statistical analysis software and spreadsheets make a strong showing, which is noteworthy. While survey data is substantially quantitative and amenable to statistical or spreadsheet manipulation, this is far less likely to be true of the second most used type of data: interviews. Nonetheless, qualitative analysis software was the third most used type.

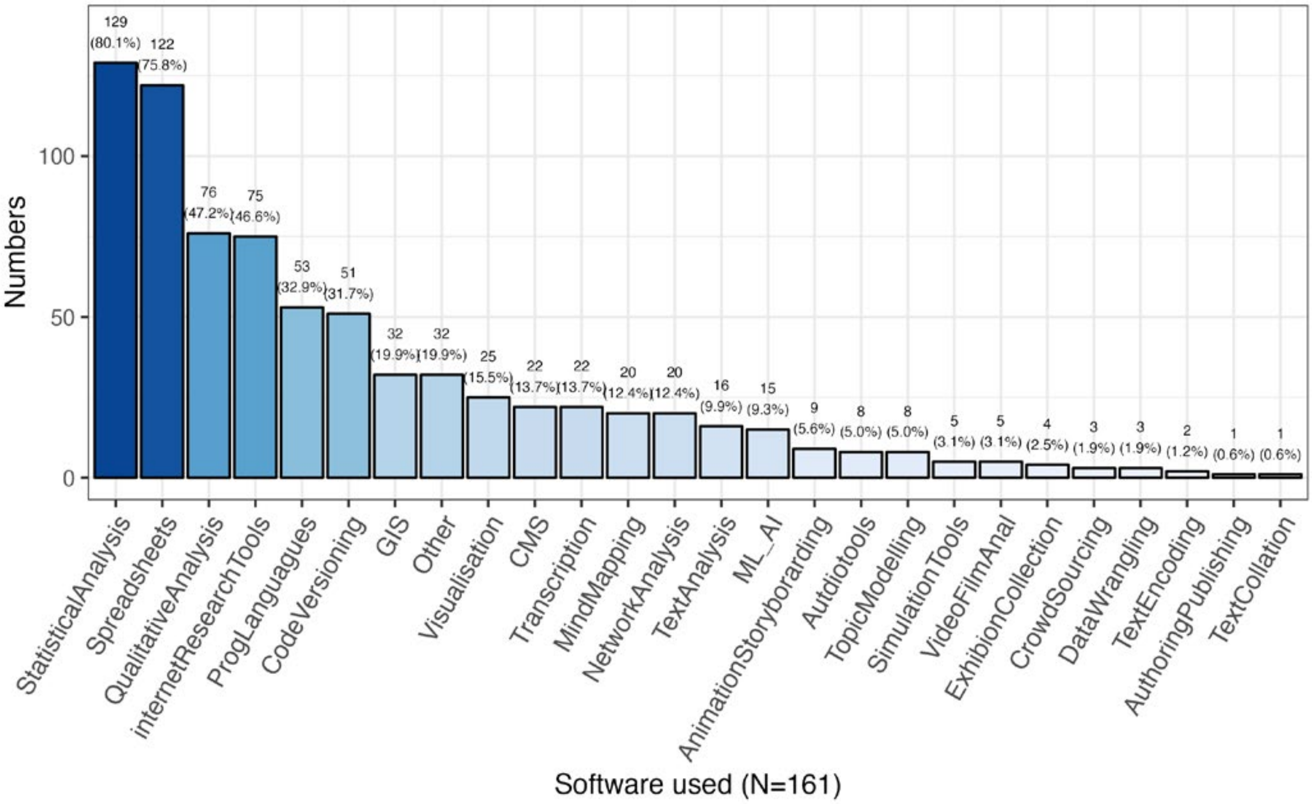


Figure 24. Software types used in research.

For the Other software category respondents could manually enter software not included on our list. 32 entries were recorded. Those that appeared more than once are as follows[54]: Word (2.5%), Zotero (2.5%), Zoom (2.5%), Photoshop (1.9%), PRAAT (1.9%), EndNote (1.9%), Miro (1.2%), online survey software (1.2%), PowerPoint (1.2%), Qualtrics (1.2%), R (1.2%), MS Teams (1.2%), MS Office (1.2%) and transcription tools (1.2%).

Fig. 25 only shows entries that have a count greater than 25, to make the diagram more tractable. For the top two categories, statistical analysis and spreadsheets, we see fairly uniform use across the career stages, though spreadsheets are used slightly less by JCRs. For qualitative analysis, JCRs seem to be using statistical analysis software the most.

[54] Software mentioned just once: 3D software, Adobe suite, Agisoft Metashape, social media for recruitment, audio/video recording software, Brainsight, Canva, database, DAWBA, Dropbox, Edraak, E-Prime, Express, Facebook, Figma, FutureLearn, GIMP, Google Drive, Google Translate, Gorilla, HOTGLUE, Hybrid, Illustrator, image processing, iMovie, InDesign for producing flyers, InqScribe, Jupyter, KNIME, Leonardo, Lucidchart, Lucidspark, Mentimeter, Mplus, Ngene, Notebooks - Jupyter, O-Tree, Obsidian MD, OneDrive, own software, Padlet, participatory systems mapper, ELAN, Premier Pro, PsychoPy/PsychoJS, PubMed, Qualtrics, reference management, Scrivener, search engines, similar systems mappers, SketchUp, SPSS, Stata, Steam, Tobii, Twitter, Unity, video editing software, VisiRule, Web of Science, Wix, word processing, and z-Tree.

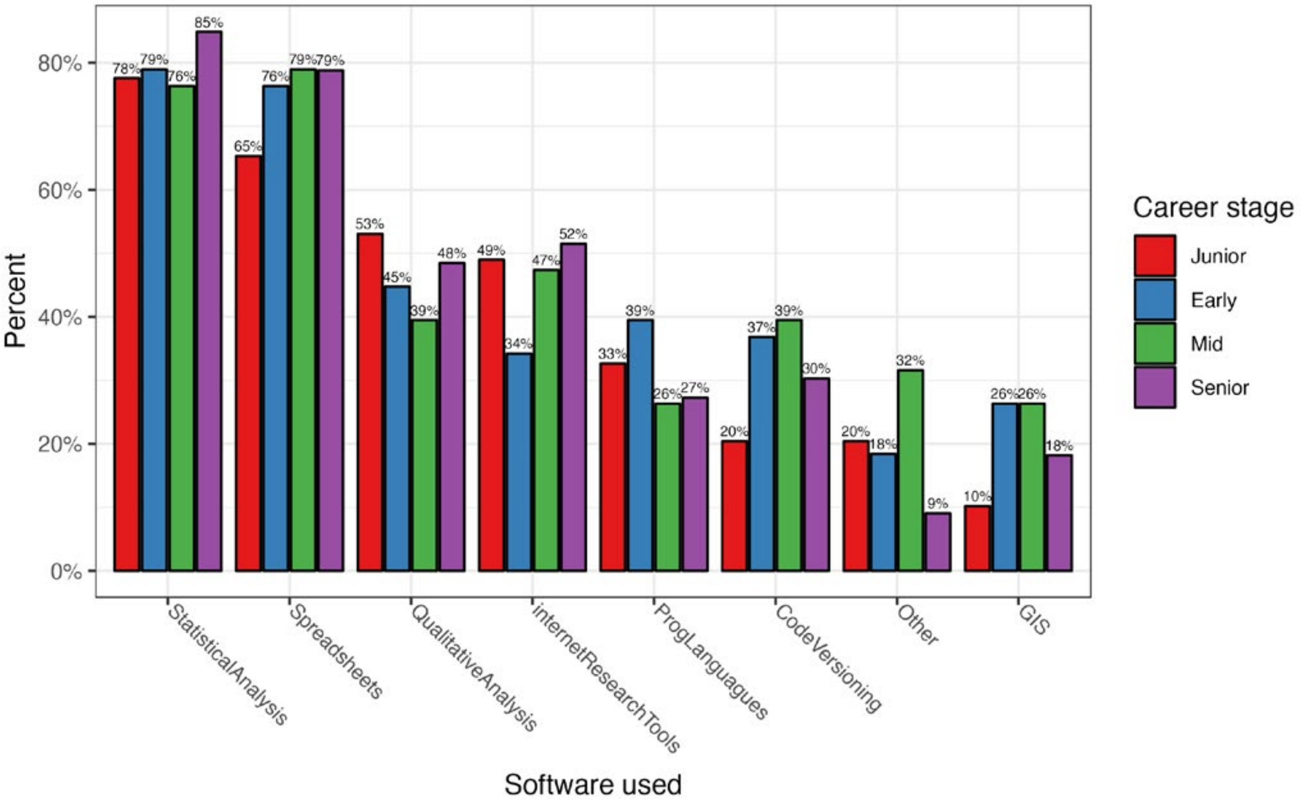


Figure 25. Software types used by career stage. Only show software categories with more than 25 entries.

Fig. 26 shows the data by gender and career stage (the junior career stage is composed of 80% women). There are some notable differences between genders. For instance, more women are using qualitative analysis tools throughout all career stages. The same is true for internet research tools, other than among senior career researchers. Men have a greater tendency to use programming languages, code versioning, and GIS software in every career stage except junior.

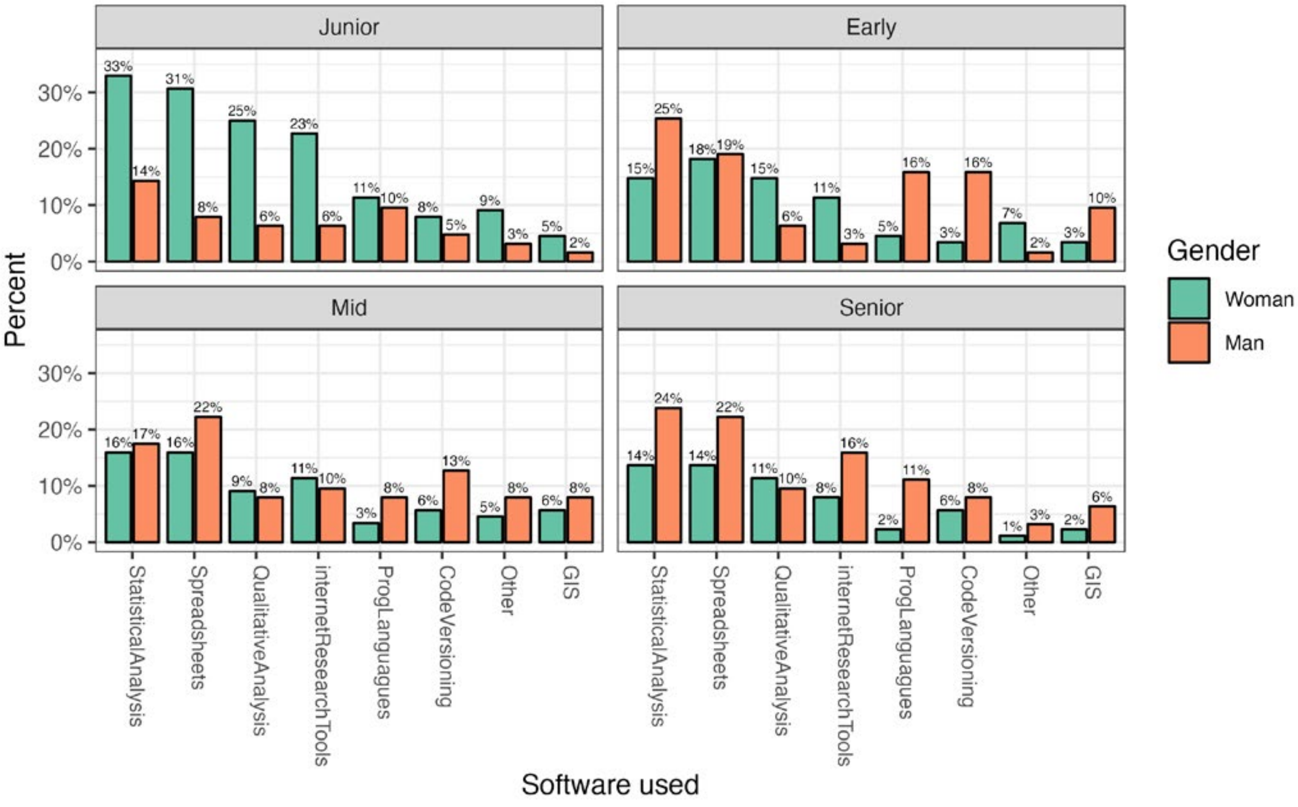


Figure 26. Software types are used by gender and career stage but only show the software types with more than 25 entries to make the diagram more tractable.

In Fig. 27 we can see which choices correlate.

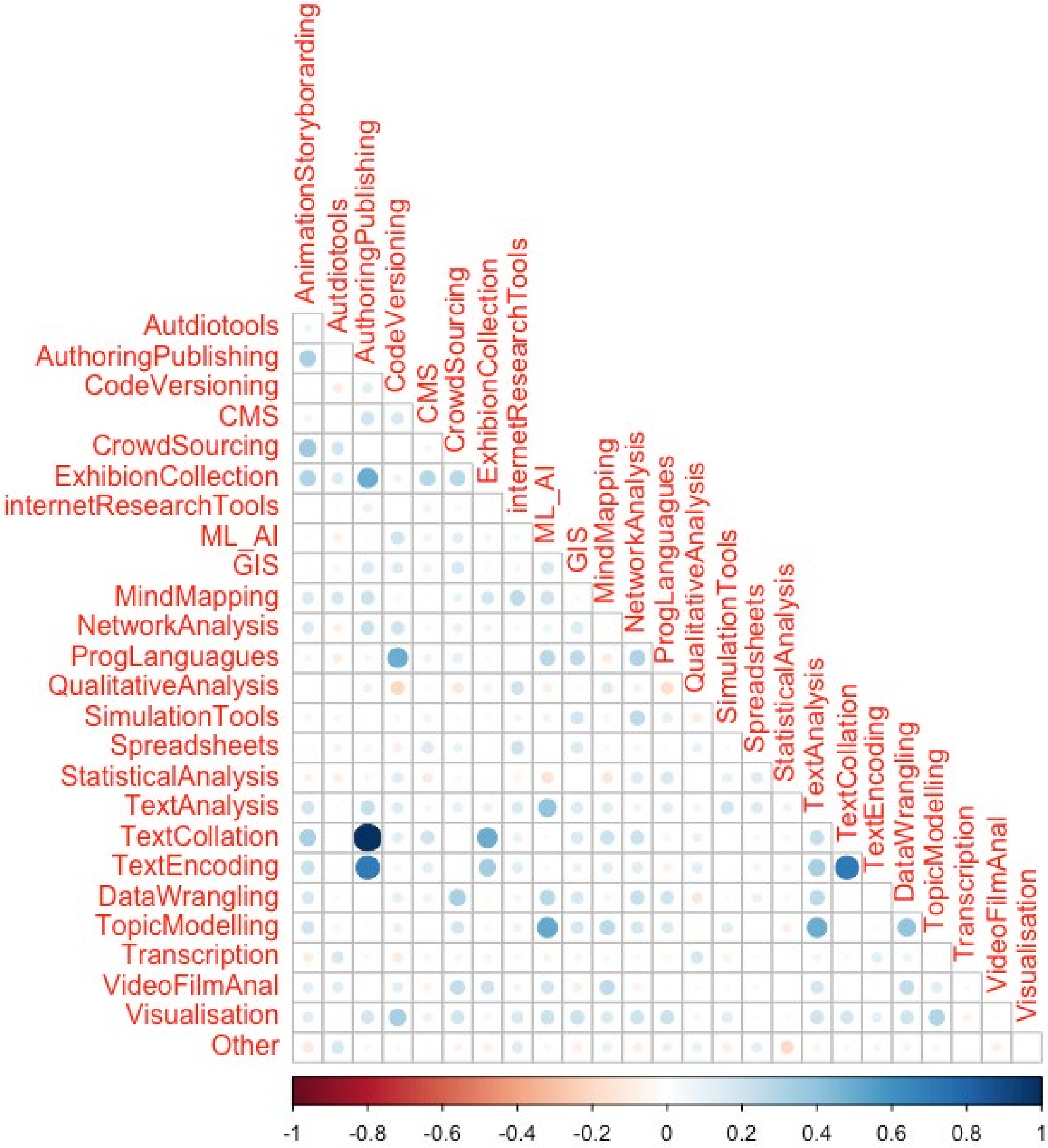


Figure 27. Correlations between the different software choices made.

There are some strong correlations between text encoding, audio tools, and text collation as these are all involved in the processing of interviews. This points to the fact that software is often used in specific workflows or toolchains, and seldom in isolation.

Finally, in Fig. 28 the software choices are split by normalised research disciplines and the percentages are taken across each research discipline. For instance, sociology makes the greatest use of statistical tools (24.3%), followed by spreadsheets (20.2%), followed by qualitative analysis tools (13%).

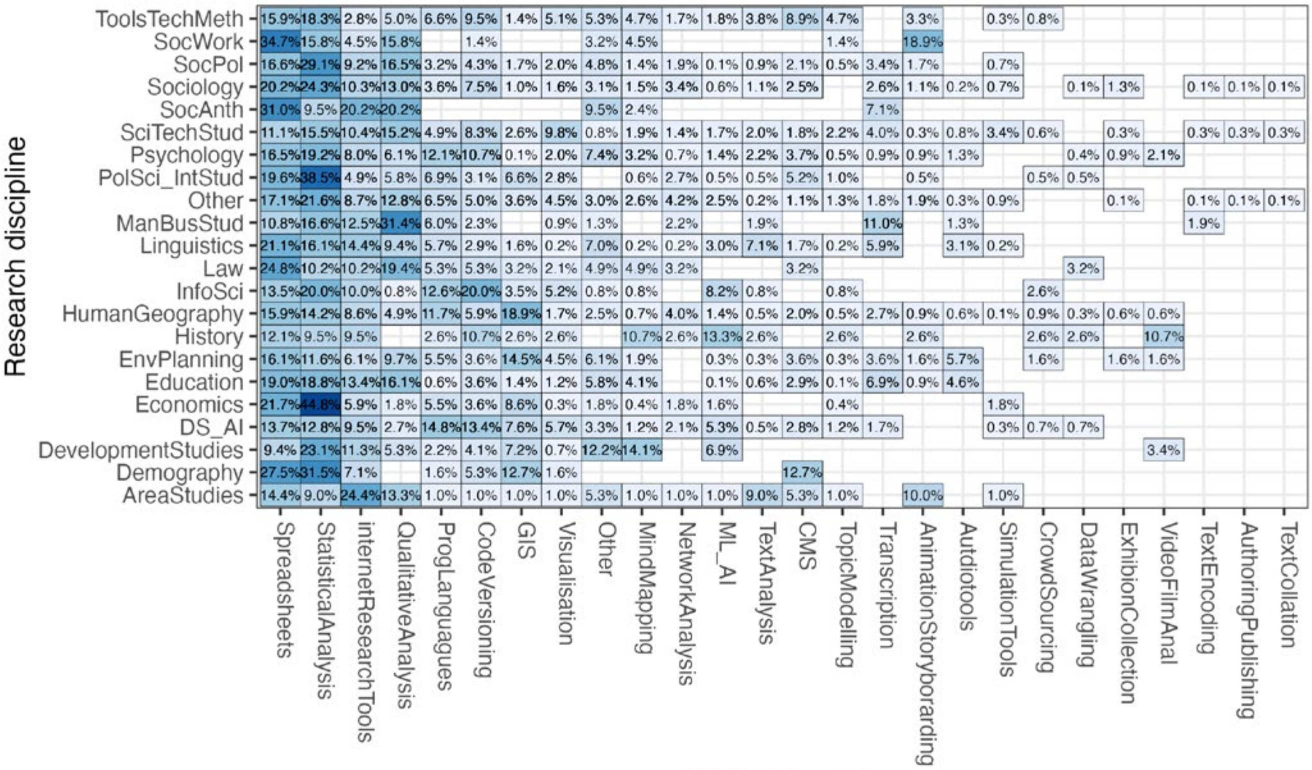


Figure 28. Percentage of normalised research disciplines by the software types used. The percentages are taken across each research disciplines.

Most important software used in research

We asked (Q5, N = 151) what software(s) were most important to their work. Responses were given in an open text box to allow respondents to name specific, rather than generic, software. The responses required some postprocessing. The results are summarised in Fig. 29.

From Fig. 29, we can see that R's response rate is almost double that of SPSS and Stata, which points to the community's appetite for using open source software. Facilitators and barriers to open source software are further explored in Section 4.5. It is interesting to note how many researchers favour open source software over proprietary equivalents. Given NVivo's lack of open source substitutes, we would expect it to make a strong showing in the qualitative analysis category.

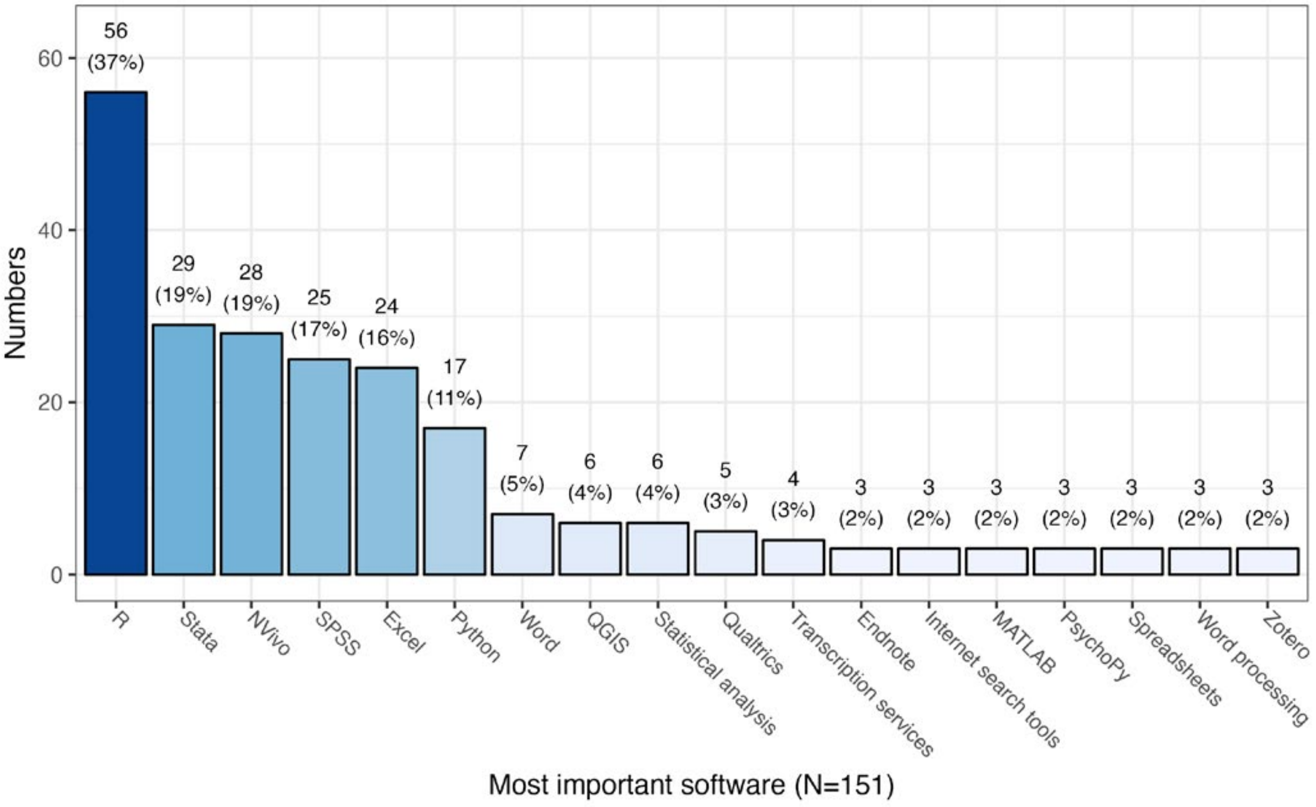


Figure 29. Most important software for respondents' own research (only categories with more than two entries are shown).

Fig. 30 shows the same data by career stage. It is notable that R is mainly used by mid-career researchers. It is possible that these researchers are no longer limited by the software skills and knowledge of their supervisors, or that they have had time during their career to acquire R skills. Once again, we see that qualitative analysis is largely carried out by junior career researchers and, to a certain extent, early career researchers.

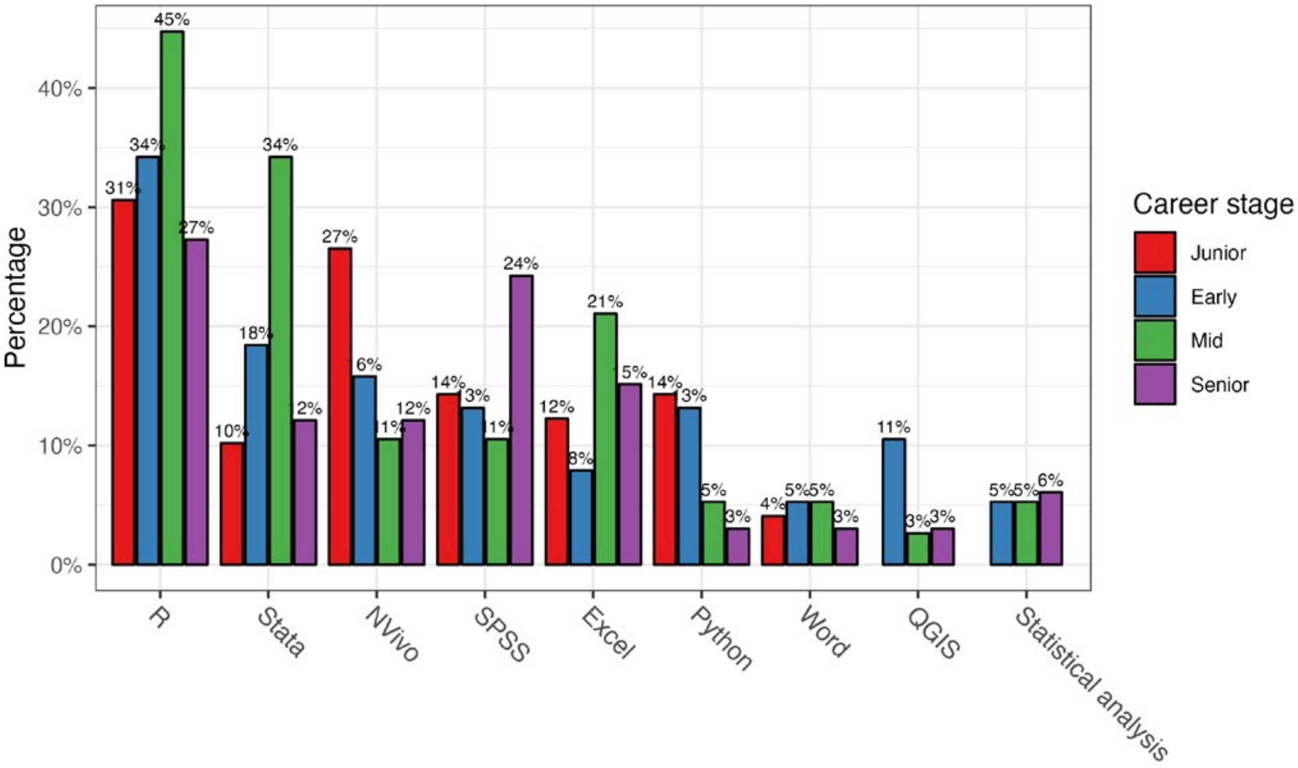


Figure 30. Most important software by career stage (only software with more than five entries is shown).

Fig. 31 shows the most important software divided by career stage and gender. Some interesting patterns arise: more men cite R as being important to them, but a significant number of women also cite R (about 40% vs 29%). NVivo is used more commonly by women than men (27% vs 5%). Stata is mentioned roughly the same number of times by both genders (16% for women vs 17% for men), but more women cite SPSS (19% vs 12%) and Excel (16% vs 12%). Python is more often cited by men (18% vs 3%), while QGIS is only cited by men and Word is more frequently cited by women (6% vs 2%).

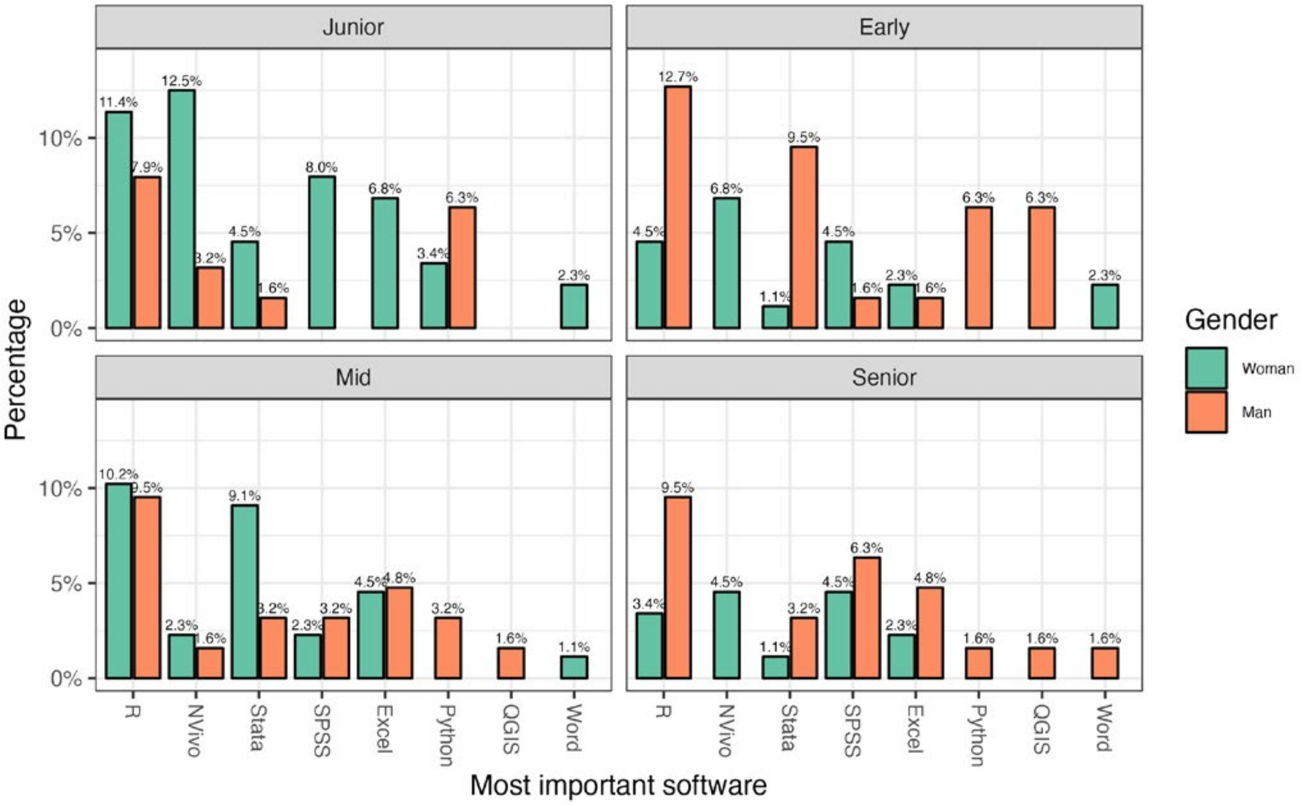


Figure 31. Most important software by gender and career stage.

Fig. 32 shows the data broken down by research discipline. The percentages are given in relation to each research discipline (split horizontally) to smooth out the variations in numbers. The values have been normalised before percentages are taken (if a respondent chose N disciplines then each contributes 1/N). From this we get an indication of what software is preferred by research disciplines. Sociology, Psychology, and Education have the highest representation in the survey.

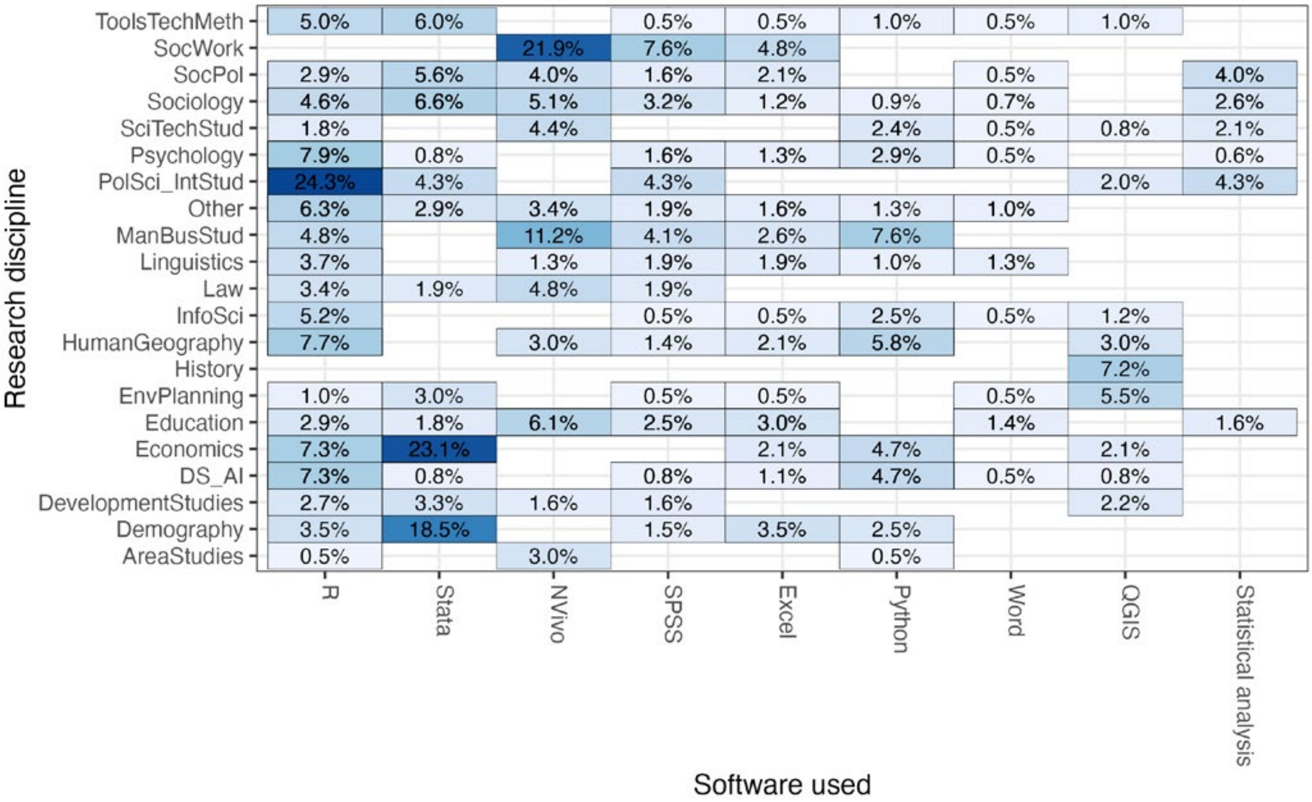


Figure 32. Most important software by research discipline. Percentages are by research discipline.

The top three software tools mentioned for each of these disciplines are: Sociology - Stata (6.6%), NVivio (5.1%) and R (4.8%); Psychology - R (7.9%), Python (2.9%) and SPSS (1.6%); Education - NVivo (6.1%), Excel (3.0%) and R (2.9%). It is important to bear in mind that small percentages are sensitive to relative changes in small numbers.

Regardless, software package usage varies from one research discipline to another. This makes it harder to a) develop policies for encouraging the maintenance of software, b) ensure resources are available for software development, and c) develop strategies around the use of open source software and publishing data.

We also asked interviewees for their thoughts on how the ESRC can facilitate innovation and best practice in software use. Training was high on the list, and we cover skills and training in Section 4.5.

> "I would say number one is providing the kind of training early on in postgraduate careers that would help people working in multidisciplinary groups, and developing some model protocols for doing some of the things that we've talked about. I wish I turned on some of this stuff a bit earlier." [3270614512]

It's worth noting that even 'basic' software can be used more effectively if training is available.

> "Google Drive, I feel like, again, it's not something that I've really had to use. It feels like it's now an embedded part. And the assumption is that everybody knows how to use these things. But, for me, it's something that ... I don't know, it's crept in while I've been working, you know, and it's not something that we've really used in workplaces before. So I'm probably not using it to its fullest effect." [1269794877]

Maintenance time can be saved and data protected through more coherent use of basic software.

Specific funding calls were also mentioned, at institution and funder levels.

> "[Digital Group] have different calls for different things across the uni. And all you need to do is be involved with that and collect digital data. So there's historians that need certain programmes or use certain things in museums ... like hardware, and they and the centre, they open calls. And then they distribute, they pay [for] stuff like that, across the university, as long as you're involved [in] what they call something digital [...] But it's also good for students because they can access funds that they normally don't get in their own schools. [4288358080]

Improving communication at the ESRC and academic output levels was also seen as key.

> "So to be honest, I used to get their (ESRC) emails every week around what they were doing with methods, but I never really found anything that they were doing with software, directly, something that directly impacted my area. So I never saw a call from them to like, purchase licences or any training. I remember, it didn't really directly talk about stuff that had

> to do with data, like directly data or data repository or software. So it's just maybe they have stuff just it didn't get directly communicated to me.” [4288358080]
>
> “And that's a big issue, you know, it's helping other researchers interpret your own process. So having blogging, for example, at the moment, ... I think it's quite important and especially [helps] qualitative researchers, [to] be a bit more concrete.” [3270614512]

Another important form of facilitator identified by our interviewees is the discipline-specific IT and software department.

> “I think it's very dependent on the IT people, we’re very lucky here, they have a lot of literacy, about academia about social science, the kind of strange things you might ask of them, which is very lucky. There have been stages when people I've worked with haven't been like that and it's really tricky, because they really don't understand what you're trying to achieve” [3270614512].

Software maintenance

We asked (Q7, N = 161) how the software our respondents use is supported and maintained. The choices were:

- Provided by my institution (**Institution**)
- Commercial or paid for external support (**Commercial**)
- Open source community initiative (**OpenSource**)
- By a software specialist/research software engineer in my institution (**RSE**)
- By me or my team (**MyTeam**)
- No longer supported/maintained (**NotMaintained**)
- Other

A summary of the responses is given in Fig. 33. Most of the software used for research in our sample is either supported by an institution or open source. This reflects the use of R (open source) and NVivo/Stata/SPSS (commercial but institutionally supported) that we see in the “Most important software used in research” section.

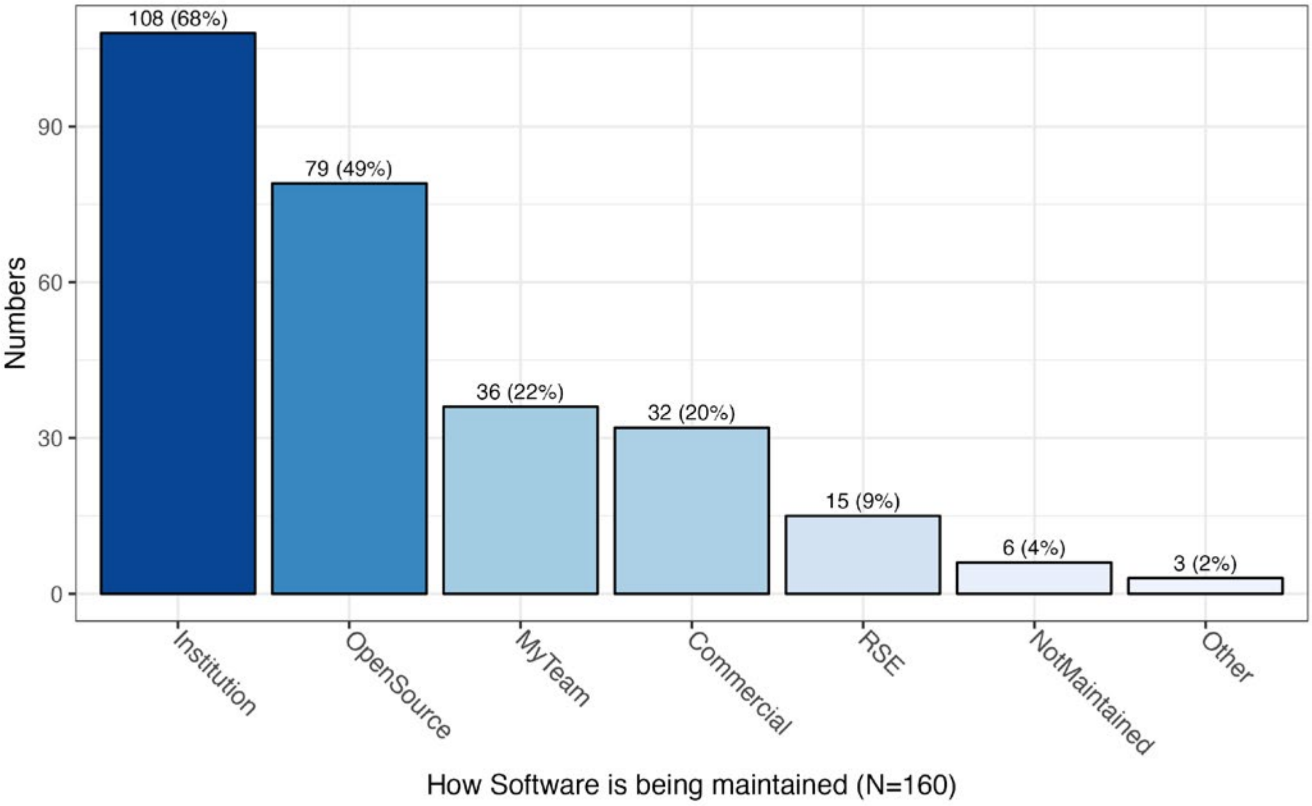


Figure 33. How the software used is supported/maintained.

Workflows: using software in combinations

In the survey we asked (Q8, N = 121) whether any typical workflows or pieces of software were used in combination to achieve a particular research end. The wording of the question is given below:

> Do you use different pieces of software in combination to achieve a specific goal or purpose? For instance, you could use audio software to record interviews, transcription software to transcribe the interviews, and qualitative coding software to bring documents, transcripts and photographs together for further analysis. If yes, please specify.

43 people declined to answer the question, while 16 denied using software in combination, stated that they did not understand the question, or stated that they were not at a stage of doing any analysis yet. As this was an open text box question, it required some postprocessing (using OpenRefine) to make the input more tractable.

The majority of respondents (64%) did use software in combination to do their research. Indeed, one of the survey respondents noted:

> "[I] use several tools within a research project as there is no one tool that enables everything and different projects require different tools, therefore exchange standards are very important" [SNR]

The example usage pattern given in the question (i.e. conducting and processing an interview) was recognised by many of the respondents, and they furthermore substantiated their responses by identifying the software they used at each stage of their research, as shown in Fig. 34. A count is included to take into account the number of mentions a piece of software received (e.g. if a particular piece of software was mentioned eight times it is appended with '* 8'. Specific pieces of software were not always mentioned in relation to a particular research stage, nor is the workflow depicted in Fig. 34 accurate for every response. While this was the usage pattern most frequently cited among respondents, it was also the one given as an example in the question, so its prevalence is to be expected.

Interviews tend to be conducted and recorded using unspecified audio recording software, though other tools such as Zoom and MS Teams became popular during the pandemic. These pieces of software have some capacity to perform the next stage of the process: transcription.

The number of tools for transcription appears to be wider ranging, from commercial software that can be installed on a laptop, such as Express scribe[55], to commercial cloud services such as Otter.ai[56], among many others. Alternatively, transcription may be outsourced to commercial transcribers. Our respondents indicated a preference for this approach when interviews involved languages other than English. For both audio and video, a tool such as Elan[57] might be used where annotations are required. There do not appear to be many open source choices for transcription.

In the next stage, where the transcriptions are processed, NVivo[58] was the most cited tool, though one piece of open source software, Qualcoder[59], earned a mention. In the analysis stage, open source programming language R received the most citations, followed by commercial tooling software like SPSS[60] and Stata[61]. An array of tools, such as Microsoft Word and Excel, were used throughout the processing stages.

For the usage pattern example shown in Fig. 35, qualitative tools were the third most important software type, and NVivo was by far the most frequently cited single piece of software for qualitative work. In the interview coding stage, NVivo was the most cited software overall. NVivo was also used to link time-separated interviews for a longitudinal study (not shown in the above diagram).

Statistical tools and spreadsheets play an important role in the analysis stage of this interview workflow pattern. We would expect them to be featured in the analysis of other workflows.

[55] https://www.nch.com.au/scribe/ (last accessed August 22 2022)
[56] https://otter.ai (last accessed August 22 2022)
[57] https://archive.mpi.nl/tla/elan (last accessed August 22 2022)
[58] https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software (last accessed August 22 2022)
[59] https://qualcoder.wordpress.com (last accessed August 22 2022)
[60] https://www.ibm.com/uk-en/products/spss-statistics (last accessed August 22 2022)
[61] https://www.stata.com (last accessed August 22 2022)

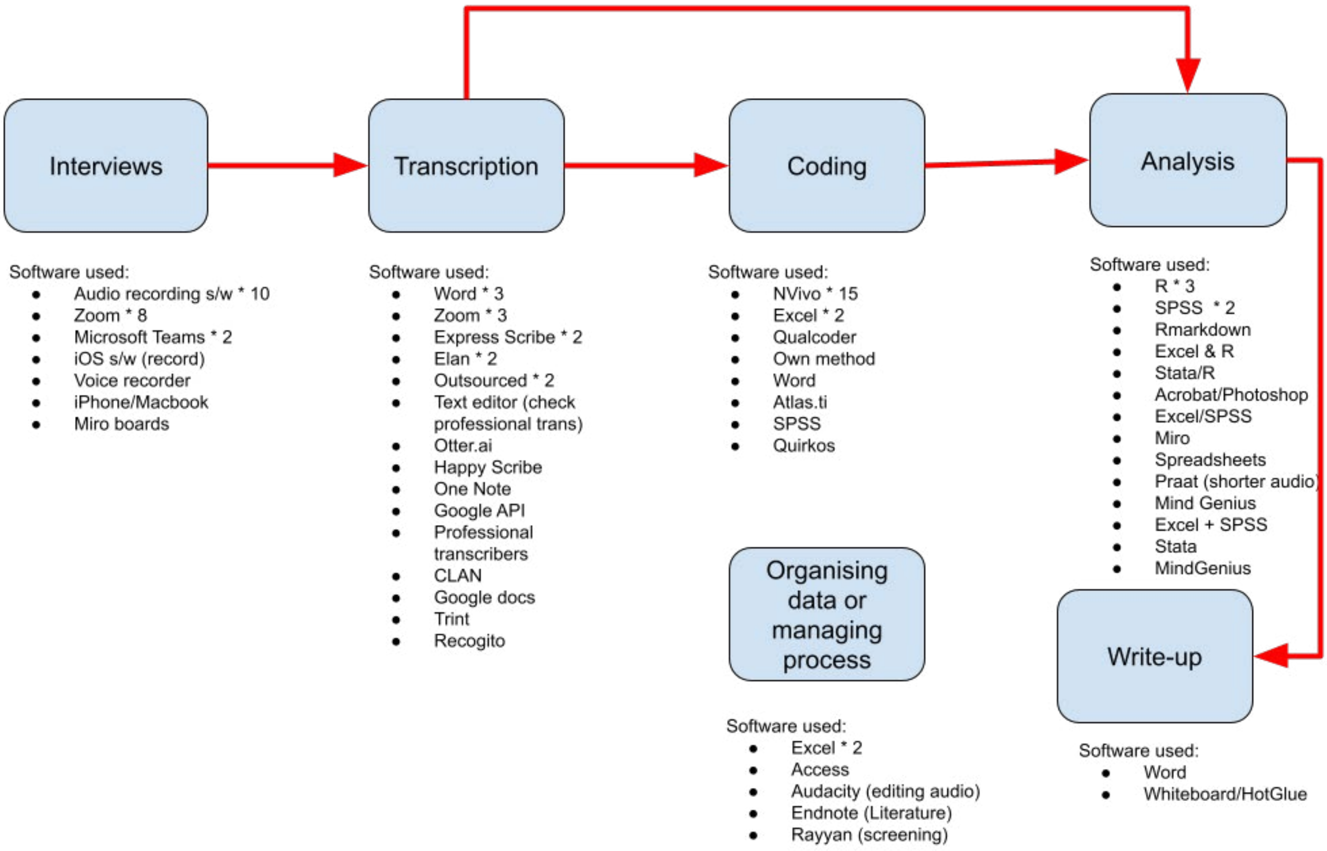


Figure 34. Interviews usage pattern.

The other main usage pattern that emerged from this question is the survey usage pattern illustrated in Fig. 35. Some of the software mentioned for the different stages is included, along with their counts.

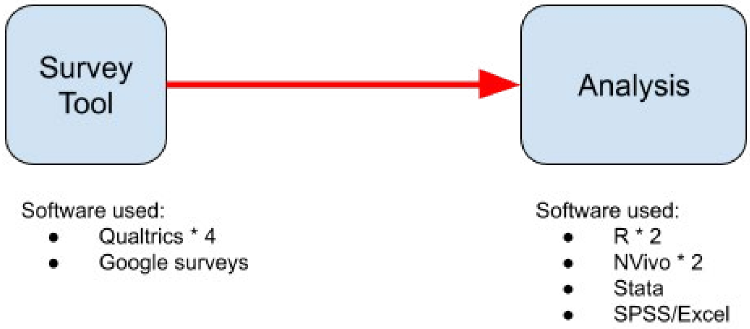


Figure 35. Survey usage pattern.

In some instances, no specific software was mentioned for a particular workflow stage, or it was mentioned only once as a separate stage in itself (as was the case with Facebook ads), meaning no item was not added to the workflow pattern. We found it unusual that more respondents did not cite software as a survey recruitment tool, but this may simply reflect the semantic oddity that tools like mailing lists are less commonly considered software.

There were other mentions of software used in combination:

- Video editing software or audio software (Audacity[62] [editing audio files] and/or Praat[63] [speech analysis in phonetics]) used in experiments performed on participants on Zoom. This software was used extensively in the preparation of online experiments, with PsychoPy[64] with Pavlovia[65], Gorilla[66], Sona[67] being used to run the experiments and Qualtrics[68] being used to manage surveys. This could have been a pattern, but it was not clear how the online experiments were run and there appeared to be significant variation.
- R, Stata, SPSS, and Python (also Jupyter notebooks/lab) are sometimes used with GIS software (QGIS[69] and ArcGIS[70]) where the data is geocoded.
- R and Python were both used for data collection, for instance when accessing the Twitter API and/or for the analysis of data, sometimes in combination with Stata or SPSS.
- Excel was used extensively for data cleaning and the production of charts (apparently not an uncommon practice when combined with Stata or SPSS), and sometimes for storing data and managing participants.
- GitHub (git) was used to manage software components together.
- Latex was used by some to produce a final output (paper/report).
- There was one instance of deployment of web servers and some data processing capabilities to a cloud platform (Azure).
- Gephi was used a couple of times to improve on network graphs produced by R.
- Databases (including Nexis Uni and PostgreSQL) were used for accessing and storing data.

These examples do not contain a single dominant toolchain.

From the analysis above, it is clear that software is being used in combination, and we have identified two main patterns: the interview pattern and the survey pattern. These correlate with the main types of data used in social sciences research. Different pieces of software are used to run the various stages of the workflow, and no single piece of software alone contains the functionality required to achieve a given end goal.

Open source software

Several high-level international policy changes have created a shift towards open science, open access, and open source. The UKRI has a set of seven principles on open research[71] and data management, derived from OECD[72] principles. We sought to establish how social sciences researchers use open source tools.

Our survey found that the majority of respondents use open source software in their research (Q6, N = 161). A summary of responses is given in Fig. 36.

[62] https://www.audacityteam.org (last accessed August 22 2022)
[63] https://www.fon.hum.uva.nl/praat (last accessed August 22 2022)
[64] https://www.psychopy.org (last accessed August 22 2022)
[65] https://pavlovia.org (last accessed August 22 2022)
[66] https://gorilla.sc (last accessed August 22 2022)
[67] https://www.sona-systems.com (last accessed August 22 2022)
[68] https://www.qualtrics.com (last accessed August 22 2022)
[69] https://www.qgis.org (last accessed August 22 2022)
[70] https://www.esri.com/en-us/arcgis/about-arcgis (last accessed August 22 2022)
[71] The RCUK Common Principles for Data: https://www.ukri.org/manage-your-award/publishing-your-research-findings/making-your-research-data-open/ (last accessed March 7 2023)
[72] https://www.oecd.org/sti/inno/38500813.pdf (last accessed March 7 2023)

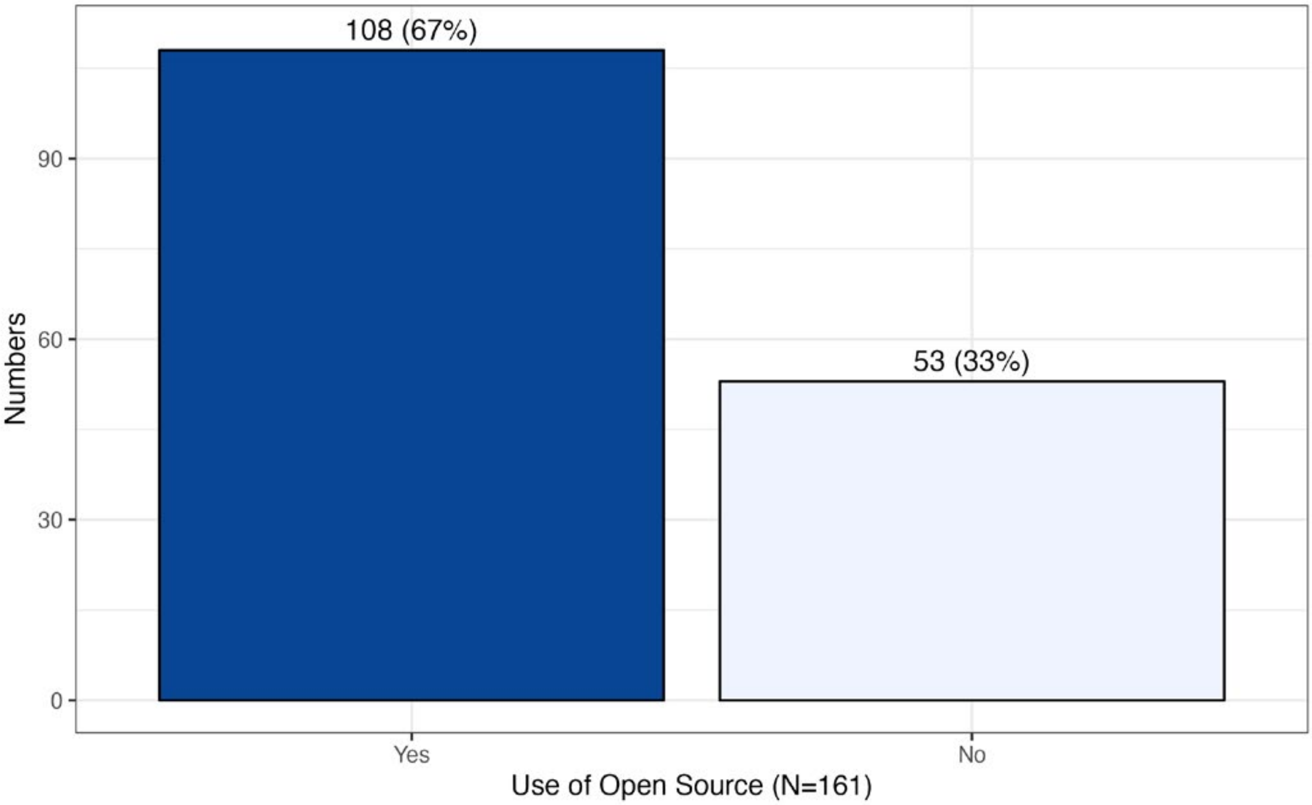


Figure 36. Whether open source software is used in research.

While most respondents report using open source tools, around a third, a significant minority, do not. We examine the use of open source by career stage in Fig. 37.

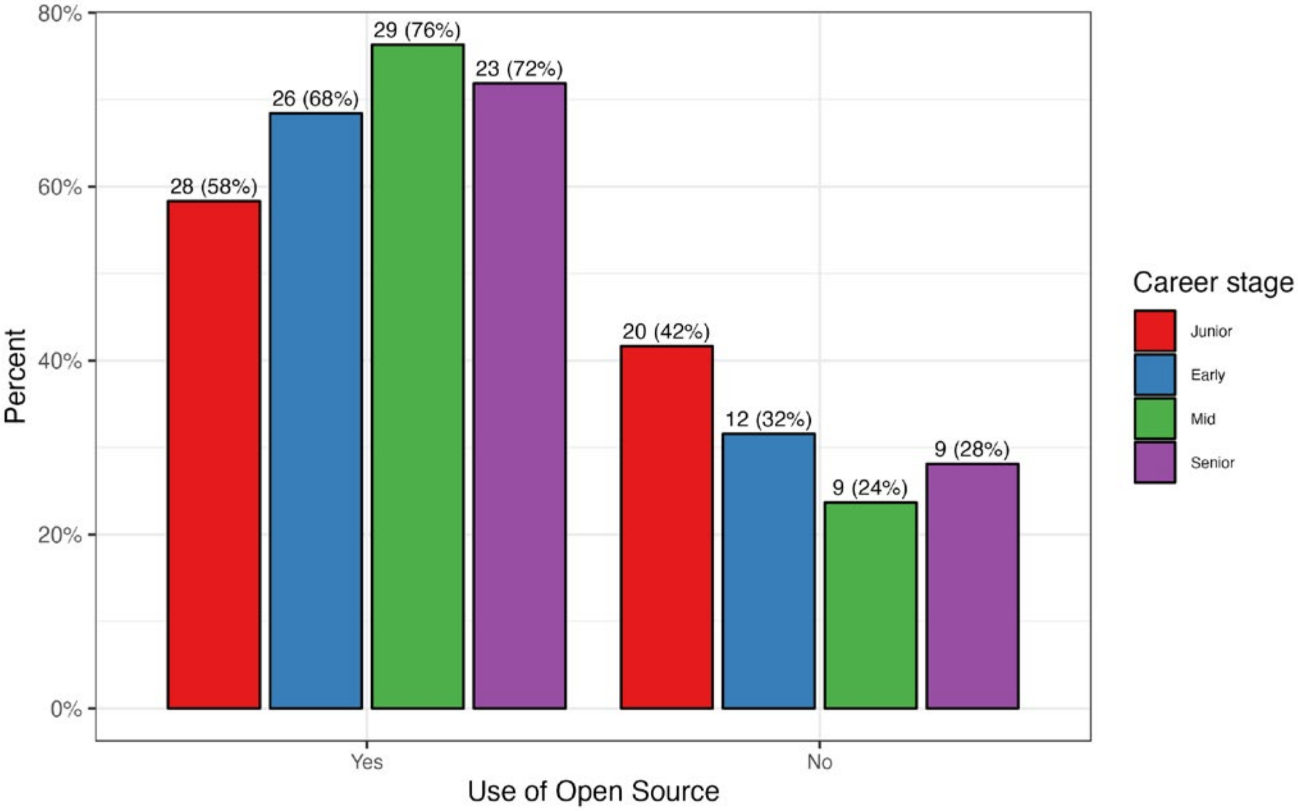


Figure 37. Use of open source by career stage.

As Fig. 37 shows, the use of open source is fairly uniform across the different career stages, though it is lowest among junior researchers. From our interviews, we can see that much of the teaching and training these researchers receive focuses on commercial packages:

> “So we need to break that cycle of just perpetuating the use of proprietary expensive software that only works within these limited contexts in academia. And as soon as you step out, it's not an option.” [5766299900]

Fig. 38 shows the data by gender and career stage. It indicates that open source is used more commonly by men, other than those at the junior career stage.

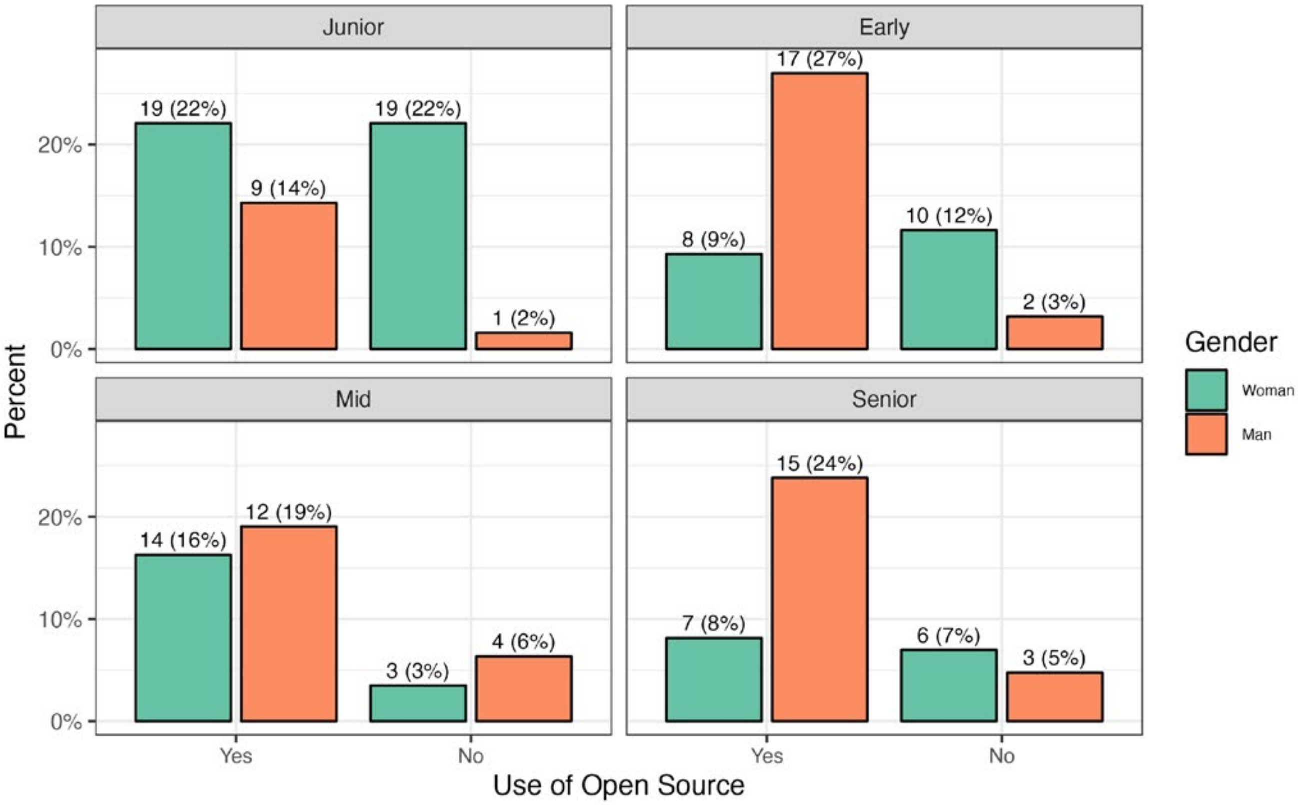


Figure 38. Use of open source by gender and career stage.

Q6a, N = 109, delved further into the reasons for using open source by eliciting an additional response from the 67% of respondents who use open source. The options were as follows:

- Institutional/funder policy **(Policy)**
- Quality of support **(Support)**
- Open standards/interoperability **(Interoperability)**
- **Cost**
- **Sustainability**
- **Licensing**
- Meets user needs/usability **(Usability)**
- Staff previous experience, no need for training **(Experience)**

Respondents could select as many options as were applicable to them. The results are summarised in Fig. 39.

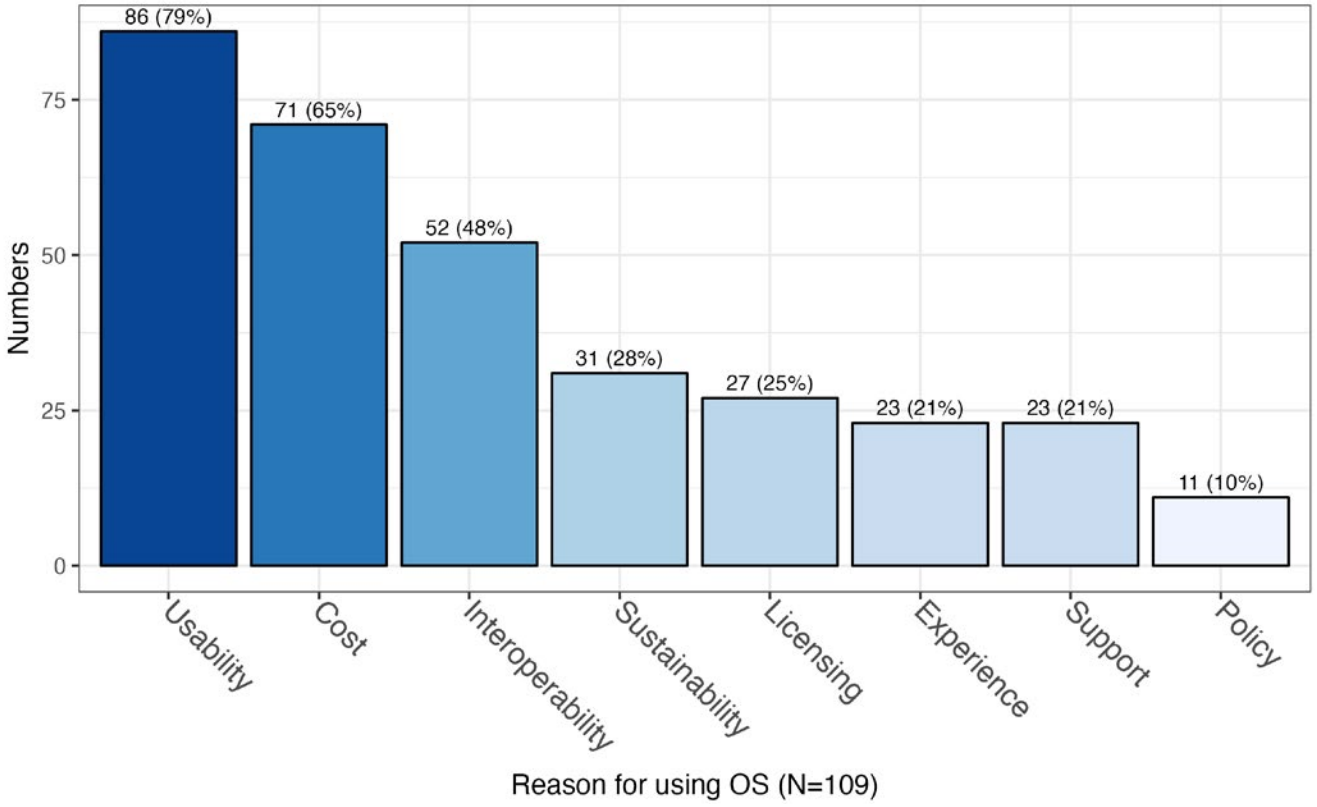
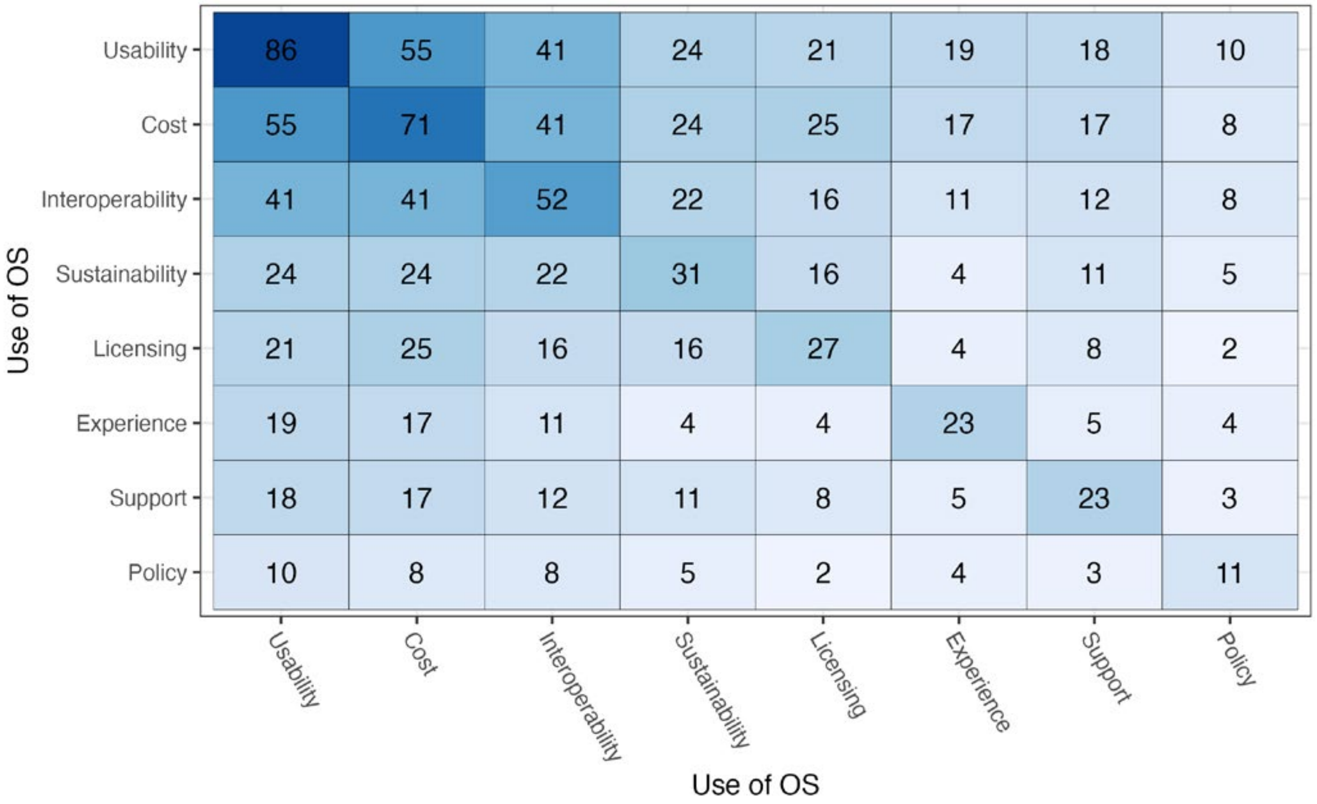


Figure 39. Reasons for using Open Source software in research.

Usability is the number one reason for using open source, followed by cost and interoperability. We can see how the various choices correlate in Fig. 40.



	Usability	Cost	Interoperability	Sustainability	Licensing	Experience	Support	Policy
Usability	86	55	41	24	21	19	18	10
Cost	55	71	41	24	25	17	17	8
Interoperability	41	41	52	22	16	11	12	8
Sustainability	24	24	22	31	16	4	11	5
Licensing	21	25	16	16	27	4	8	2
Experience	19	17	11	4	4	23	5	4
Support	18	17	12	11	8	5	23	3
Policy	10	8	8	5	2	4	3	11

Figure 40. How the reasons provided for using open source have been used together.

Here we see that interoperability, cost, usability, and to a slightly lesser extent sustainability seem to correlate. This pattern remains more or less stable across the career stages, as shown in Fig. 41.

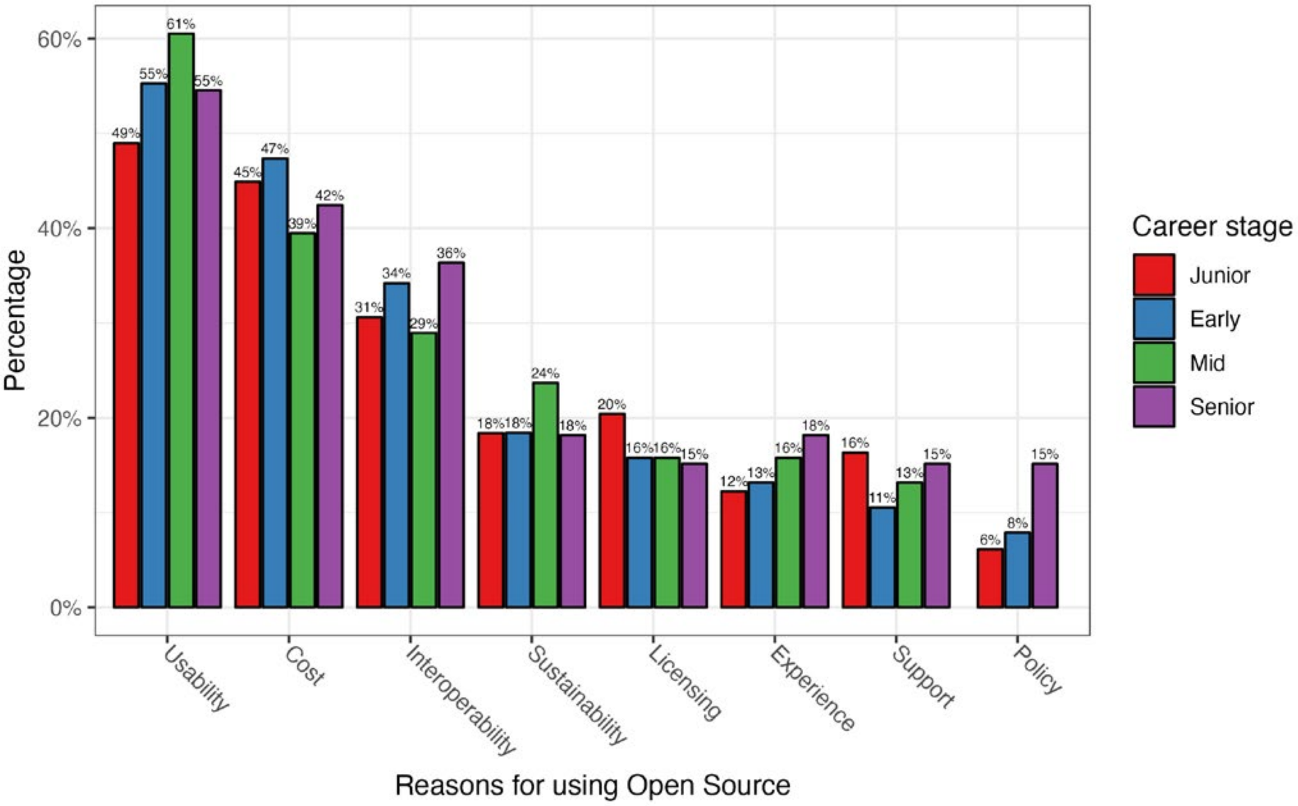


Figure 41. Reasons for using open source across the different career stages.

Fig. 42 decomposes the data by gender and career stage. The junior career stage is dominated by women, while the other three career stages are dominated by men.

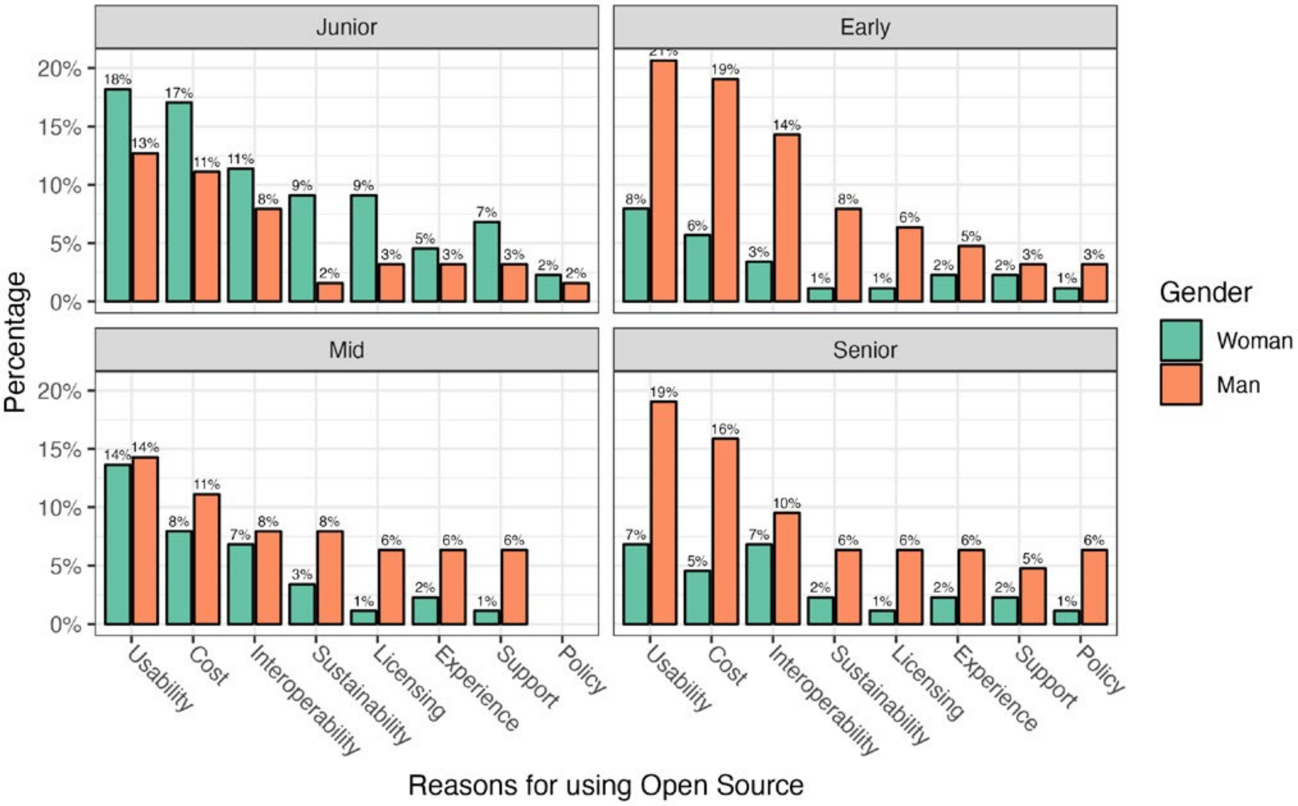


Figure 42. Reasons for using open source by gender and career stage.

We can see how many individuals answered in each category by gender and career stage in Table 10.

	Women	Men
Junior	22 (56.4%)	9 (50%)
Early	8 (44.4%)	17 (89.5%)
Mid	14 (82.4%)	12 (75%)
Senior	7 (50%)	15 (83.3%)

Table 10. Reasons for using open source by gender and career stage.

Among men, every career stage shows an answer rate of 50% or above. Among women, only early stage career researchers show an answer rate below 50%.

Institutional policy was seen as the least important facilitator. Our interview data suggests that researchers are not ignoring institutional policy in deciding to use open source software, but rather that few institutions have policies that support its use. We will discuss this further in Section 4.4.

Our interview data also indicates that researchers value open source software for meeting their needs, cost, sustainability, and interoperability, and because it provides free or cheap access to software that does not require institutional funding. This is true of both individual researchers and the research community in a global sense.

> “I've tried to learn how to use R especially because it's open source. So I wouldn't need to have to, you know, pay for it or rely on a university to provide me with the software.” [7183719943]

> “There's an element of the fact that we are in ... an increasingly global world with people from all around from other parts of the world, which do not necessarily have the financial privileges of the Global North.” [5766299900]

Respondents valued the interoperability and universal standards of open source software in particular when working across institutions and contexts. These qualities were perceived to provide flexibility for teams with different skills, and to prevent work in academic institutions from becoming too siloed.

> “If we all use different statistical programmes, then you've got to retrain somebody, you know, so it's helpful, I think, harmonisation across the different approaches.” [7336263536].

> “But if you're collaborating with people in the Global South, if you're collaborating with people, in my experience, in some parts of Asia or in Latin America, there are a lot of people who use free and open source operating systems.” [5766299900]

Open source also enables collaboration on a cross-institutional level.

> “With the [proprietary software] that I am using only people who have access to our university address can have access to the software right here.” [6776362537]

Sometimes that collaboration has been globally significant.

> “I mentioned COVID, that sort of thing, like, national, spatial, temporal epidemic modelling type thing, which actually turns out to be much more generally applicable ... that has been kind of developed and shared, you're open source to get to that ... kind of thing. And that's sort of a national collaborative project.” [6422621713]

In terms of sustainability, interviewees were concerned about retaining access to their data, and also about the sustainability of the tools they used. Data that resides within proprietary software, or can only be used with proprietary software, degrades once that software is no longer accessible. This might come about because the researcher has moved institutions, ceased to be a junior career researcher, or experienced a change of circumstance.

> “Even if I had access to these proprietary tools, once I left university, I would no longer have access to them and would have to pay 1,000 pounds or whatever to just use the software to run my own code” [5766299900]

Another compelling feature of open source software according to our respondents is the fact that it provides easy and free access to training, tutorials, and support.

> “And usually there are [resources] because it's open source. So the community is kind of fairly good in sharing resources”. [2964377624]

Open source tools also give researchers more control in terms of their access to software, since they do not require specific copies on university machines or “obtuse activation procedures” [5766299900].

In worst-case scenarios, researchers can be locked out of the proprietary software with which they have been working.

> “I feel like [licensed] softwares are more difficult for us to have access to, because we always have to get approval. And, and we ... do not control what we use, it has to go through admin.” [6776362537]

> “When I was at university, it worked fine. And I started doing some coding [...] I went back to [country], and a few months later, I could no longer use [proprietary software], because it's just like you need to be on campus regularly or something to ping the servers to keep the activation current.” [5766299900]

Furthermore, open source is seen as an ethical and principled approach, which can be important in itself to social scientists. This means that tools like QualCoder[73] and QGis[74]are sometimes favoured simply by virtue of being open source, despite having fewer functions than their commercial counterparts.

> “More importantly, having been a longtime supporter of the principles and practices of open science, I wanted to make sure that my data could be analysed by anyone else...And if they were to rerun my code, they should not themselves be presented with a financial barrier to do so.” [5766299900]

> “In my view, what should from a policy perspective be supported is the use of what I would like to just generally call ethical software. Free and open source software does a job but I think there's a very strong ethical component to it.” [5766299900]

While we heard many respondents discussing well-known open source tools such as R, there are some less well-known tools that research communities could benefit from exploiting more widely.

> “I think one of the things that [is] underexploited [is] ... open source intelligence tools. I really want to help people to have access to this and one way of doing that is through your report.” [6776362537]

One of our survey goals was to explore the reasons for not using open source (Q6b, N = 49). We gave respondents the following choices:

- Institutional/funder **(Policy)**
- Speed of access to support **(PoorSupport)**
- Lack of performance **(NotPerformant)**
- **(Legal)** reasons
- **(Continuity)**
- Does not meet user needs **(NotMeetNeeds)**
- Lack of expertise **(NoExpertise)**
- Requirement from collaborators **(CollabReqs)**

In Q6, 53 respondents said they did not use open source software or tools. This question only yielded 49 responses, summarised in Fig. 43. This diagram shows that lack of expertise is the most commonly cited reason for not using open source tools. This was corroborated by interviewees who cited the steep learning curve involved in using open source (not using Windows or Linux and a basic lack of coding knowledge were also given as reasons). We can also see the influence of discipline-based commercial tools in our skills analysis. As more than one reason could be given for not using open source, we can see how the choices correlate in the contingency table shown in Fig. 44.

[73] https://qualcoder.wordpress.com (last accessed September 14 2022)
[74] https://www.archer2.ac.uk (last accessed May 16 2022)

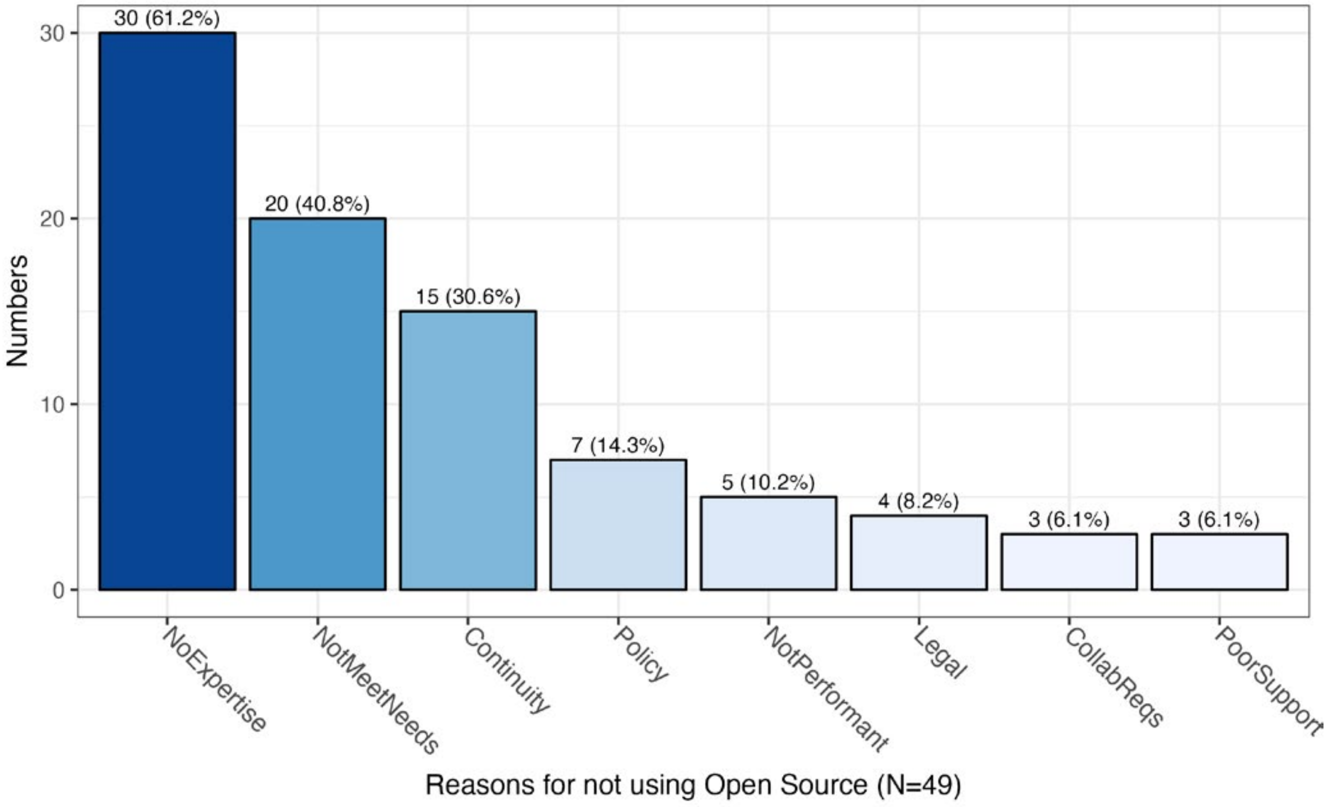
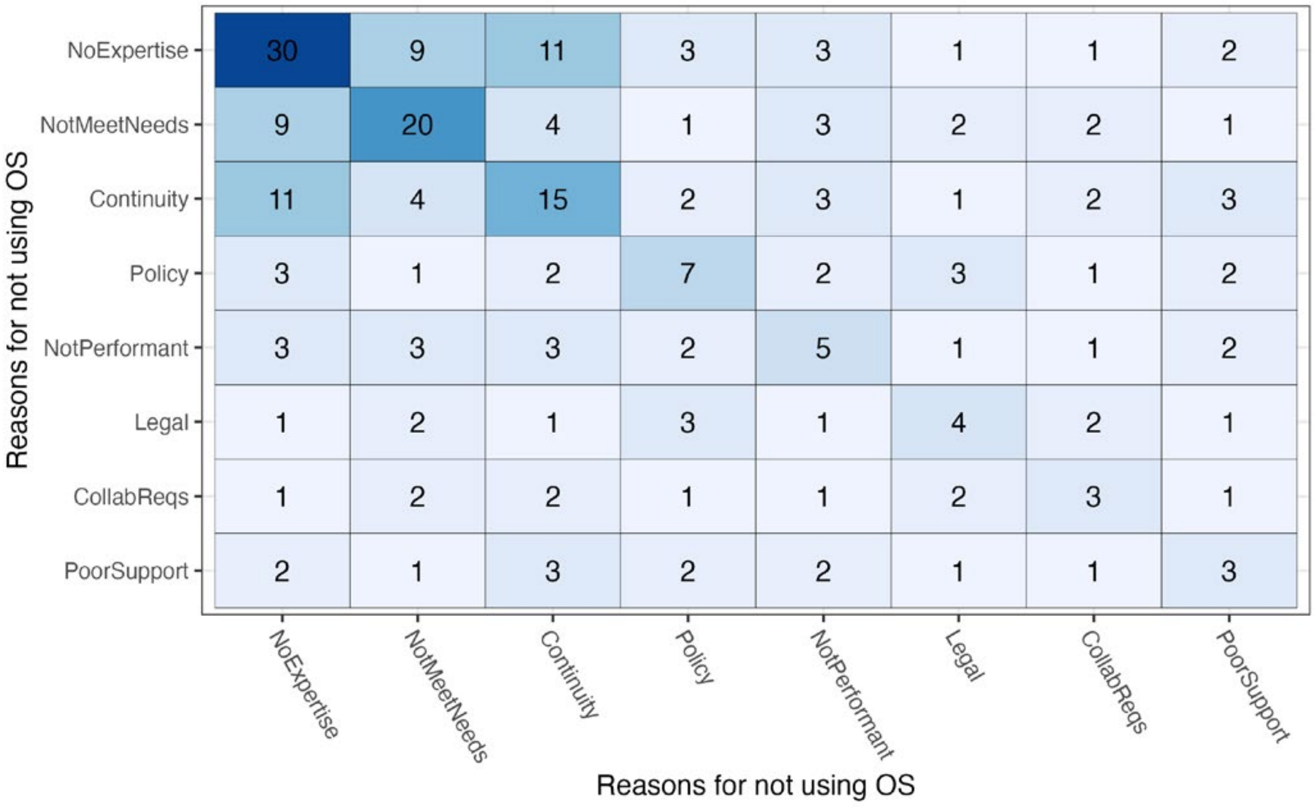


Figure 43. Reasons for not using open source.



	NoExpertise	NotMeetNeeds	Continuity	Policy	NotPerformant	Legal	CollabReqs	PoorSupport
NoExpertise	30	9	11	3	3	1	1	2
NotMeetNeeds	9	20	4	1	3	2	2	1
Continuity	11	4	15	2	3	1	2	3
Policy	3	1	2	7	2	3	1	2
NotPerformant	3	3	3	2	5	1	1	2
Legal	1	2	1	3	1	4	2	1
CollabReqs	1	2	2	1	1	2	3	1
PoorSupport	2	1	3	2	2	1	1	3

Figure 44. Contingency table for how the choices provided for not using open source.

It appears that the most common reasons for not using open source are lack of expertise, failure of the software to meet the user's needs, and lack of continuity. Other choices are less strongly indicated.

Fig. 45 shows reasons for not using open source by career stage. Senior career researchers are most likely to cite a lack of expertise, though this reason is well-represented by every career stage. Junior and mid-career researchers tend to feel that open source software does not meet their needs. Only ECRs cite legal and collaboration requirements.

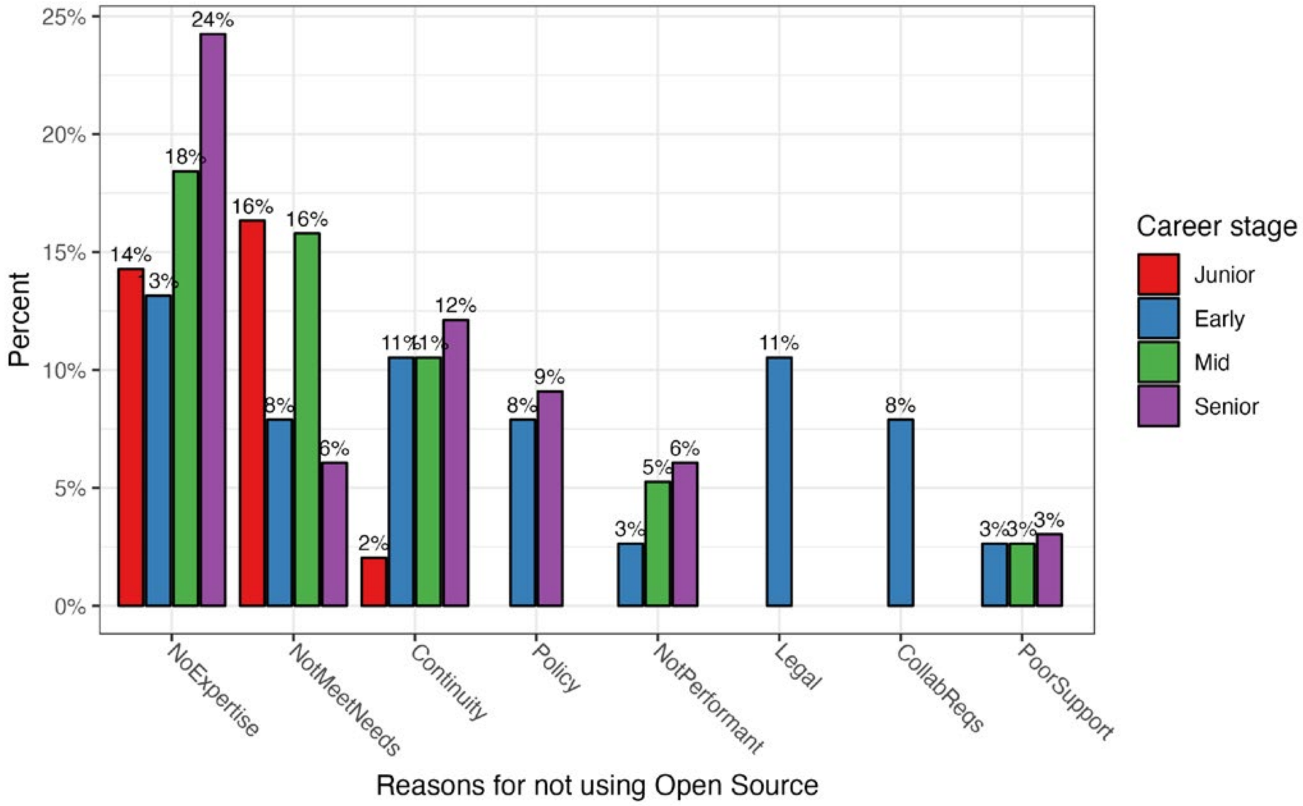


Figure 45. Reasons for not using open source by career stage.

Fig. 46 looks at the data by gender and career stage. Among junior career researchers, only women cite lack of expertise as a reason for not using open source, and the same proportion (8%) claim that it doesn't meet their needs. Conversely, these are the most commonly cited reasons among men at the mid-career stage.

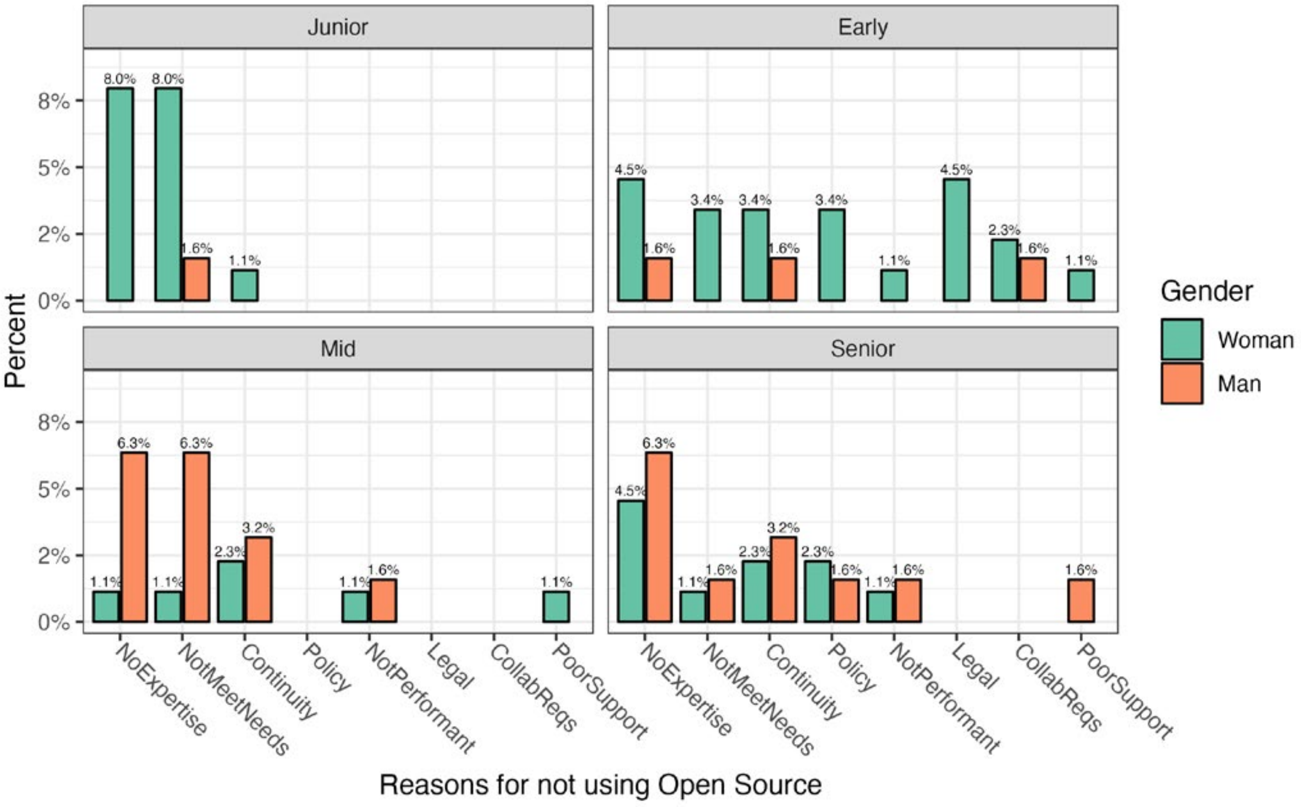


Figure 46. Reasons for not using open source by gender and career.

In our interviews, we frequently heard that even open source software supporters found it slightly onerous to use.

> "I have chosen personally from the beginning of my PhD to use a free and open source software as much as possible in my analysis ... which proved challenging." [5766299900]

There is of course a hardware challenge. Open source is not compatible with Macs, and it is difficult to use on shared machines.

> "Because, you know ... open source software, it's constantly updating, and it's like, every day things are changing, I have to download, I have to use web scraping or use complex tools, things are changing. And if I have to kind of use different computers or cloud computing, somebody else might have altered the settings that I have done." [4098334726]

Interestingly, our interviewees felt that,

> "while there's extensive open source tools for quantitative research, there are more limited options [for qualitative research]." [5766299900].

One interviewee told us,

> "the day after I tried [the qualitative open source software], and it worked, I even sent an email to the researcher thanking him for having written this tool. Because it was so rare to find something." [5766299900]

Open source software like R and Python appears to have a steep learning curve for those not familiar with coding or open source, "it's not obvious for people how to use it " [5766299900]. Compounding the problem further is the lack of formal support. In worst case scenarios, this can result in serious delays.

> "I lost months and months of my own time trying to learn how some of these tools work. Partly because there was no formal support structure, partly because the tools are not as trivial to use as a simple piece of software, that you just install and it runs on a Windows or Mac just fine." [5766299900]

Trying to carry out more advanced coding brings commensurate challenges. As we will explore further in Section 4.5, an ability to utilise existing functions in SPSS or Stata does not equip the user with sufficient skills, knowledge, or understanding to engage with the logic or coding requirements underlying R.

> "You need to be very comfortable with ... writing code to be able to do that, which is a challenge." [8031500357]

There was a perception among some of our respondents that R is used by "everyone", but we did not necessarily find this to be true. Furthermore, our respondents perceived some open source software as being widely used because it is popular in other countries, as RedCap is in the US.

We talked to researchers working on a variety of software development projects, some of which were coming to the end of their funding. We discussed whether going open source was a viable option, but the transfer between research software and open source presents a number of challenges.

> "The problem with open source is it needs an active community. It's amazing. But without the active community fixing bugs and updating the software and all that sort of stuff it will very quickly become unusable." [8031500357]

A particular issue is the question of who will form the community. If the community around a given piece of research software is focused on a specific discipline or purpose, it is hard to see where the incentive lies for people to become engaged in the community, or for software engineers to support the software gratis. Anyone engaged in supporting software requires a good understanding of the software in question, which is a barrier to growing specialist communities. We learned of one piece of software that's in a repository, fully licensed, but not available for use, "just because it would be very difficult for anybody to pick it up and maintain it." [4561769548].

Developing, extending, and sharing software

We asked (Q9, N = 161) whether participants develop or extend software themselves.

Responses are summarised in Fig. 47. Most users (59%) do not develop software for their research. We should take into account that there are many large-scale software development projects within the ESRC community, such as CLOSER, COSMOS, the Consumer Data Research Centre, and LifeGuide, and this may be reflected in our results.

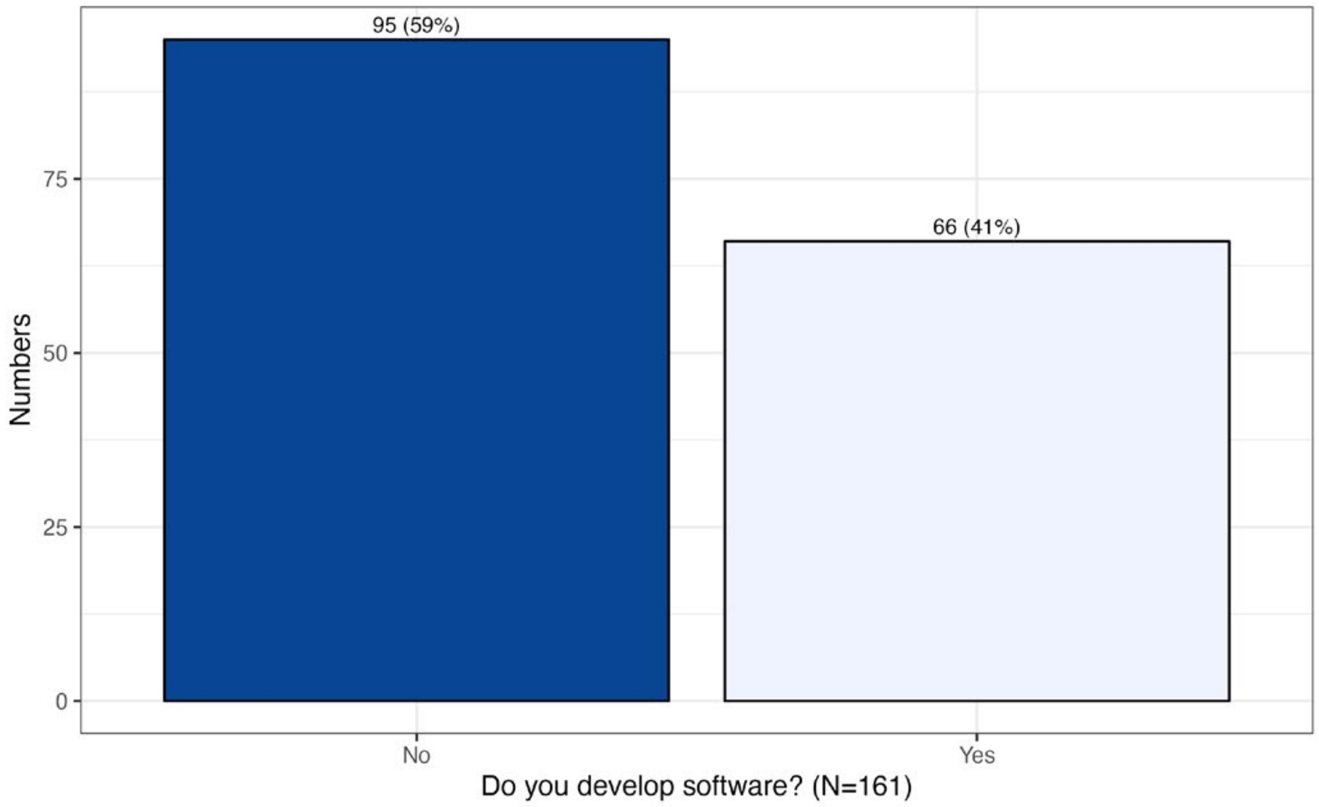


Figure 47. Whether users develop software.

We can see how responses vary by career stage in Fig. 48. The majority of junior (82%) and senior (58%) career researchers responding to the survey do not develop software, while just over half of early (53%) and mid-career (55%) researchers do.

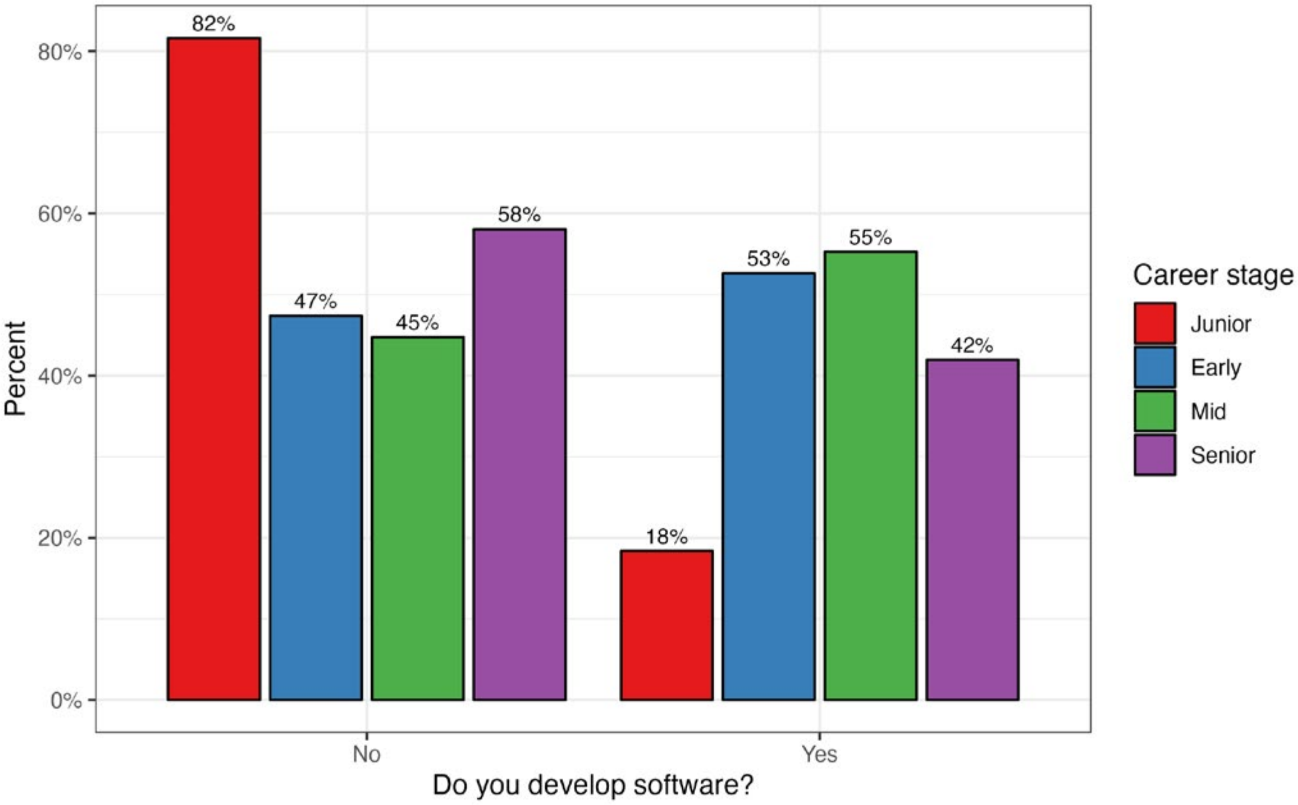


Figure 48. Whether software is developed by career stage.

Looking at gender and career in Fig. 49, we see that women are less likely than men to develop or maintain software across all career stages, though the discrepancy is marginal among junior career stage researchers.

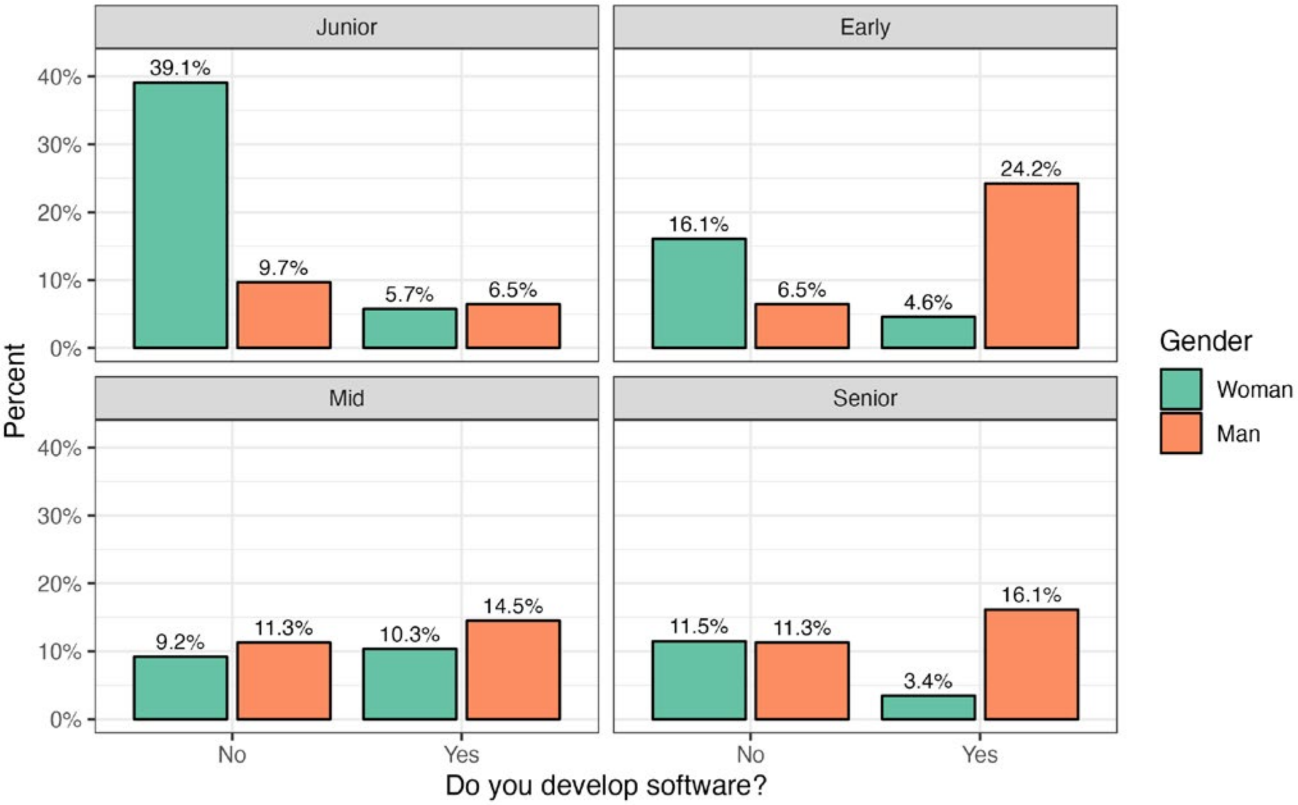


Figure 49. Whether software is developed is split by gender and career.

Q9a asked respondents to briefly describe the software they had developed or extended. 64 respondents answered this question, with the following results: R scripts (including R markdown and Shiny) (41%), Python scripts (17%), Excel functions, macros and charts (15%), Stata (10%), SPSS (5%), JavaScript (3%), and Perl (3%). These software platforms are followed in the results by a long-tail of programming languages and tools with just a single mention. R seems to be developed the most, followed by Python. Both of these are open source software, and basic training on their use can normally be accessed for free. As explored in the 'Workflows' section, these applications and languages are often used in combination.

We asked (Q10, N = 80) whether those who develop software (including scripts, applications, tools, codes, and formulae in spreadsheets) also share it, giving the following options:

- Only make available for your own use **(OwnUseOnly)**
- Only share within your research group **(RGUseOnly)**
- Only share with your collaborators (including at other institutions) **(CollaboratorsOnly)**
- Share with others on request **(ShareOnRequest)**
- Make widely available for use in your field/community **(ShareWidely)**
- Other **(Other)**

Respondents could choose as many options as applied to them.

The results are summarised in Fig. 50. We can see that the greatest number of respondents share their code widely (45%), though almost as many create software solely for their own use (44%).

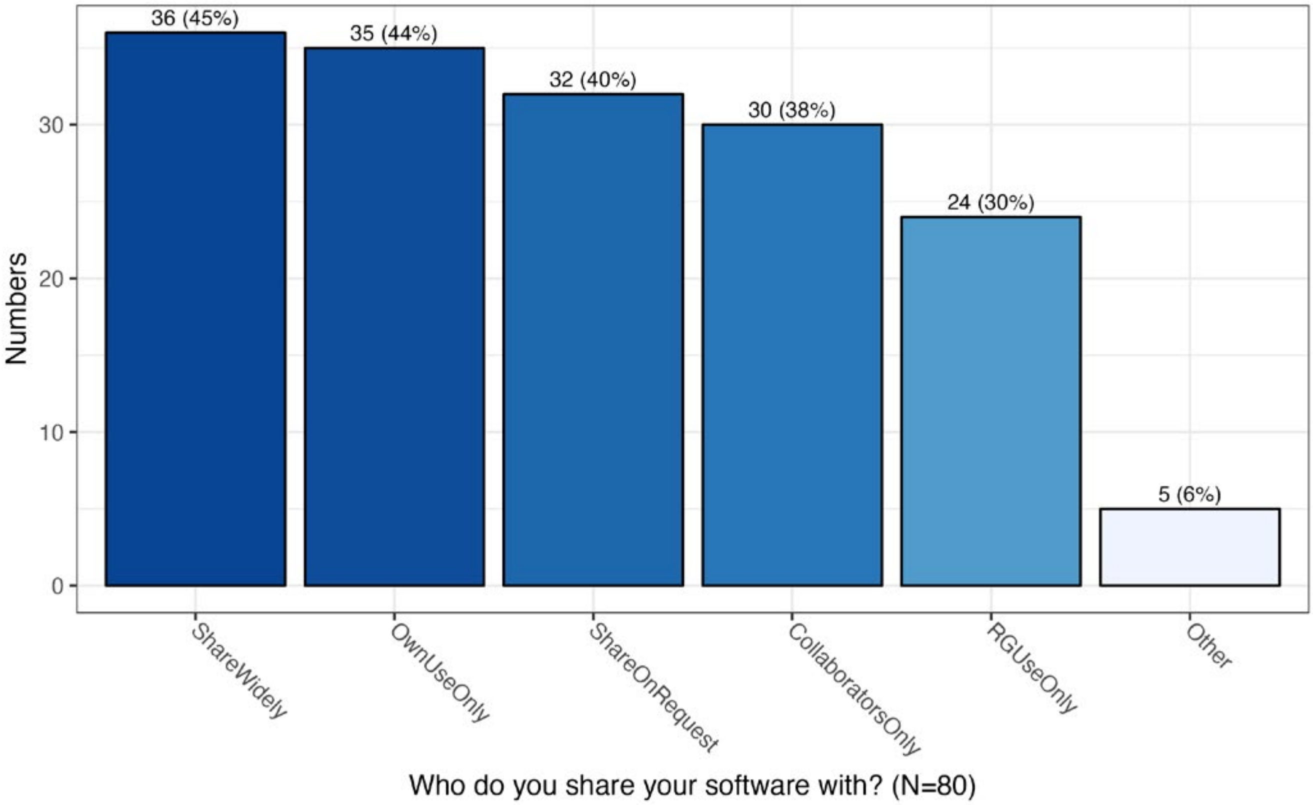
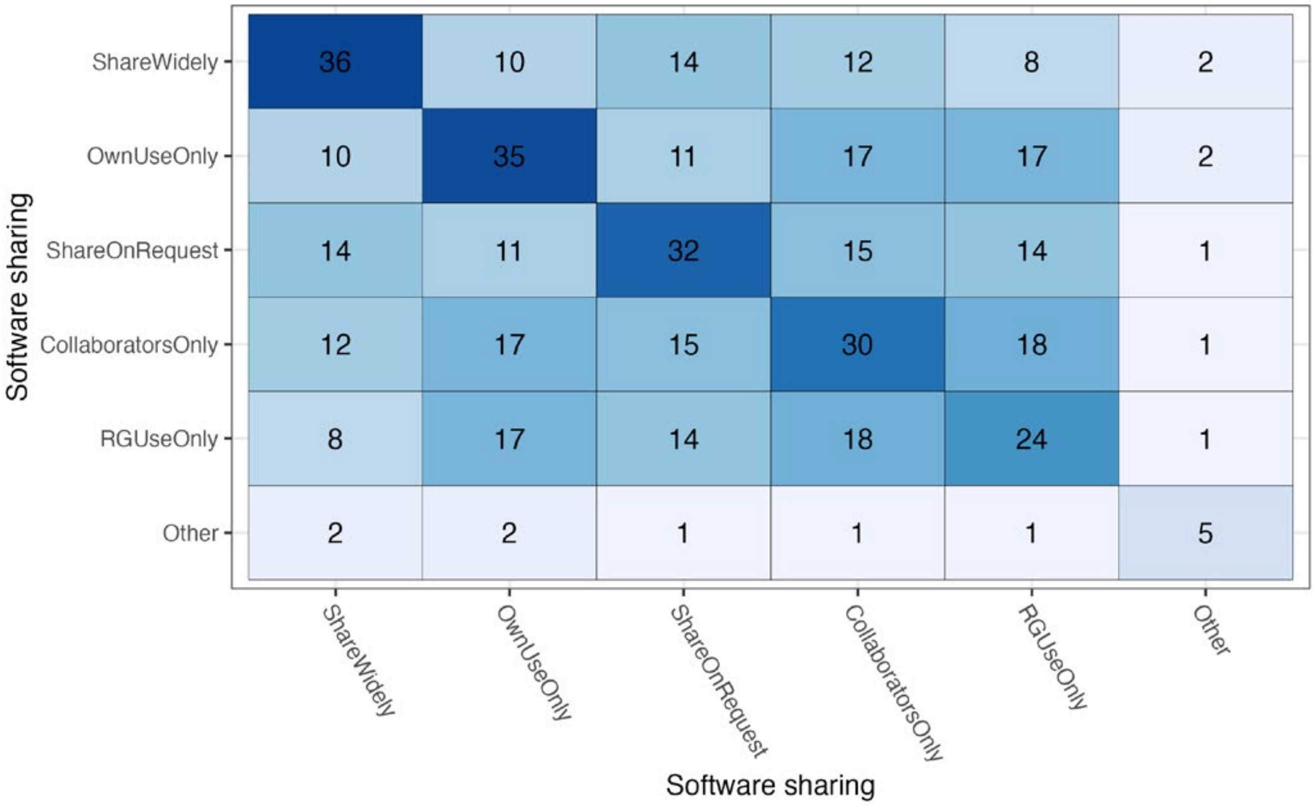


Figure 50. Sharing software. Percentages relate to the number of respondents.

However, Fig. 51, a contingency table showing which options were used together, makes it apparent that respondents did not necessarily answer in ways that we expected. For instance, we anticipated that OwnUseOnly would be seen as mutually exclusive of other options. Instead, respondents appeared to interpret it inclusively, claiming to simultaneously use code for their own use, make it available to other members or their research groups, and share it with anybody who asks.



	ShareWidely	OwnUseOnly	ShareOnRequest	CollaboratorsOnly	RGUseOnly	Other
ShareWidely	36	10	14	12	8	2
OwnUseOnly	10	35	11	17	17	2
ShareOnRequest	14	11	32	15	14	1
CollaboratorsOnly	12	17	15	30	18	1
RGUseOnly	8	17	14	18	24	1
Other	2	2	1	1	1	5

Figure 51. Contingency table showing which choices correlated.

From Fig. 51, when we examine the multiple-choice answers corresponding to Q10 i (excluding the Other option) to which a respondent could choose as many as applied to them, we notice that there is an increase in permissive sharing. For example, Own use only (OwnUseOnly) < Research Group Use Only (RGUseOnly) < For Collaborators only (CollaboratorsOnly) < Only share on request (ShareOnRequest) < Share widely (ShareWidely).

If we analyse these results by looking at the most permissive form of sharing respondents chose we get Fig. 52. The count of respondents here is reduced by one to N = 79 as one respondent chose the 'Other' option, which we haven't included in our analysis.

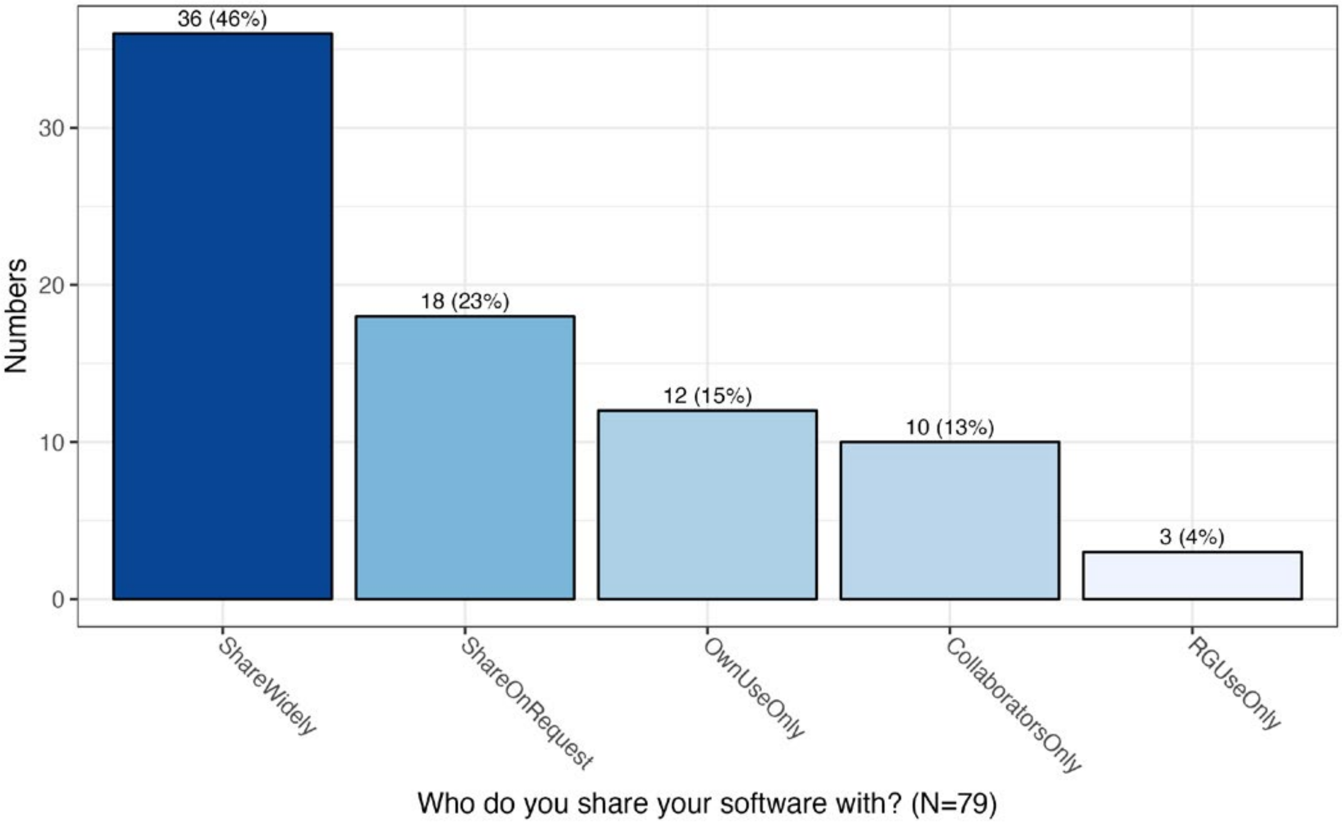


Figure 52. Sharing software (modified responses).

Fig. 52 shows that most of those who develop software appear to share it widely (46%), though responses to the original question suggest that for each respondent this applies to some pieces of software but not others. Fig. 53 splits the data by career stage. It is worth noting that choices are exclusive in this graph: the inclusion of a response in one of the categories precludes its presence in the others.

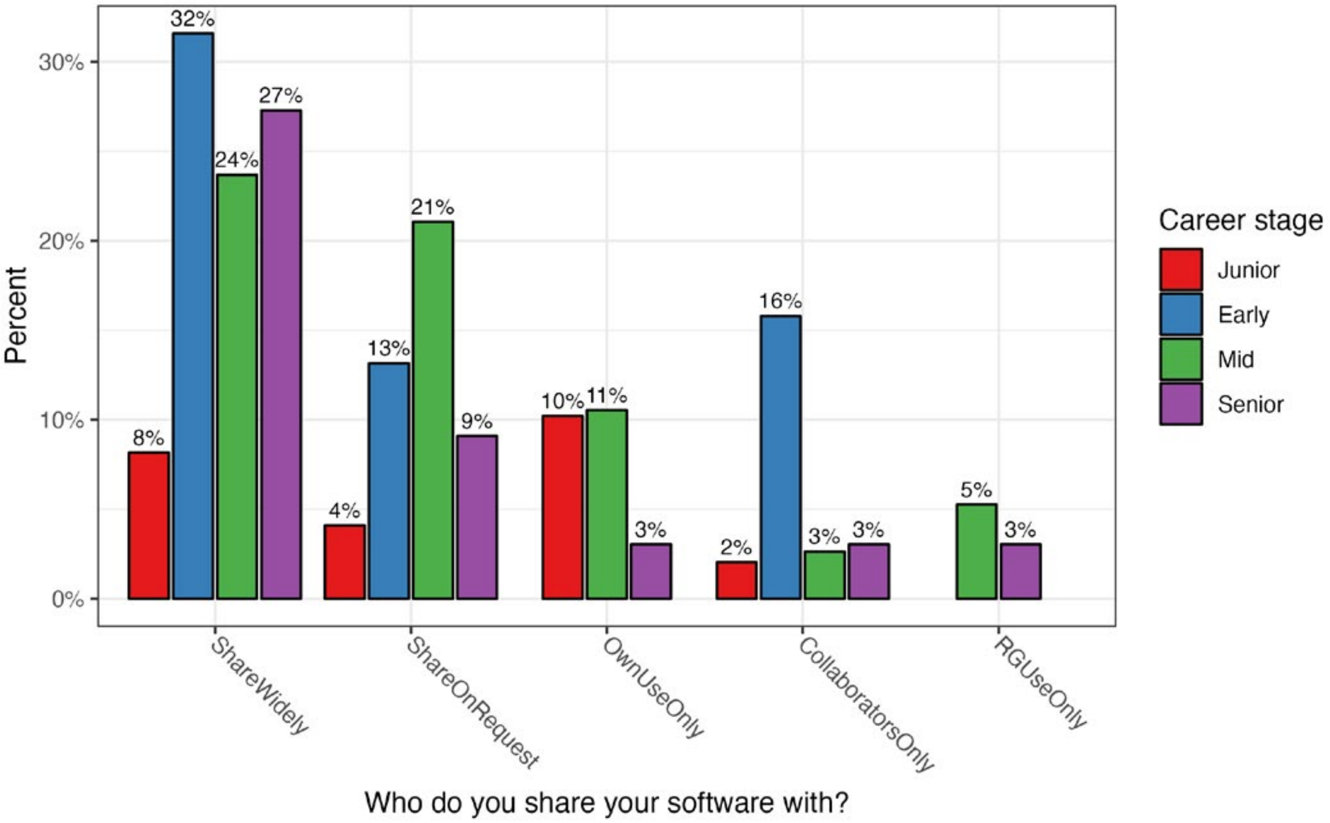


Figure 53. Sharing of software by career stage.

ECRs appear to share their software most widely (32%), closely followed by senior (27%) and mid-career (24%) researchers. The second most popular choice for mid-career researchers is to share on request (21%), while ECRs marginally prefer sharing with their collaborators (16%) to sharing on request (13%).

We can also examine the distribution by both career stage and gender. Fig. 54. shows a greater tendency among men than women to both share data widely and share data on request in early, mid-, and senior career stages. This pattern is not evident among ECRs.

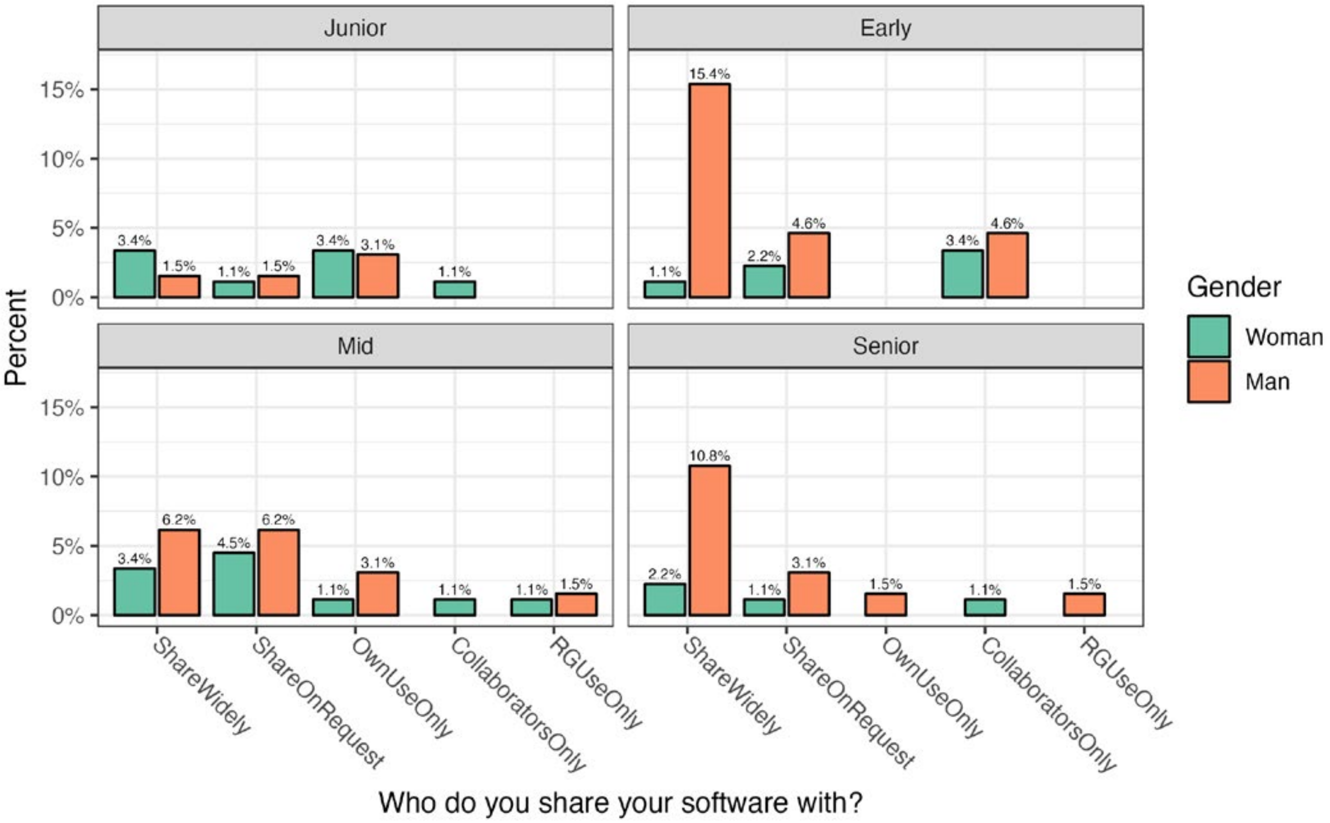


Figure 54. Sharing software by career stage and gender.

Recognition for software development

We asked (Q12, N = 161) participants whether they felt that the development and maintenance of research software is sufficiently recognised or rewarded in their fields. Respondents were only allowed to make one choice:

> Yes

> No

> Don't know

Responses are summarised in Fig. 55. The majority did not feel strongly enough to express an opinion, and replied that they did not know. 42% of respondents selected "No".

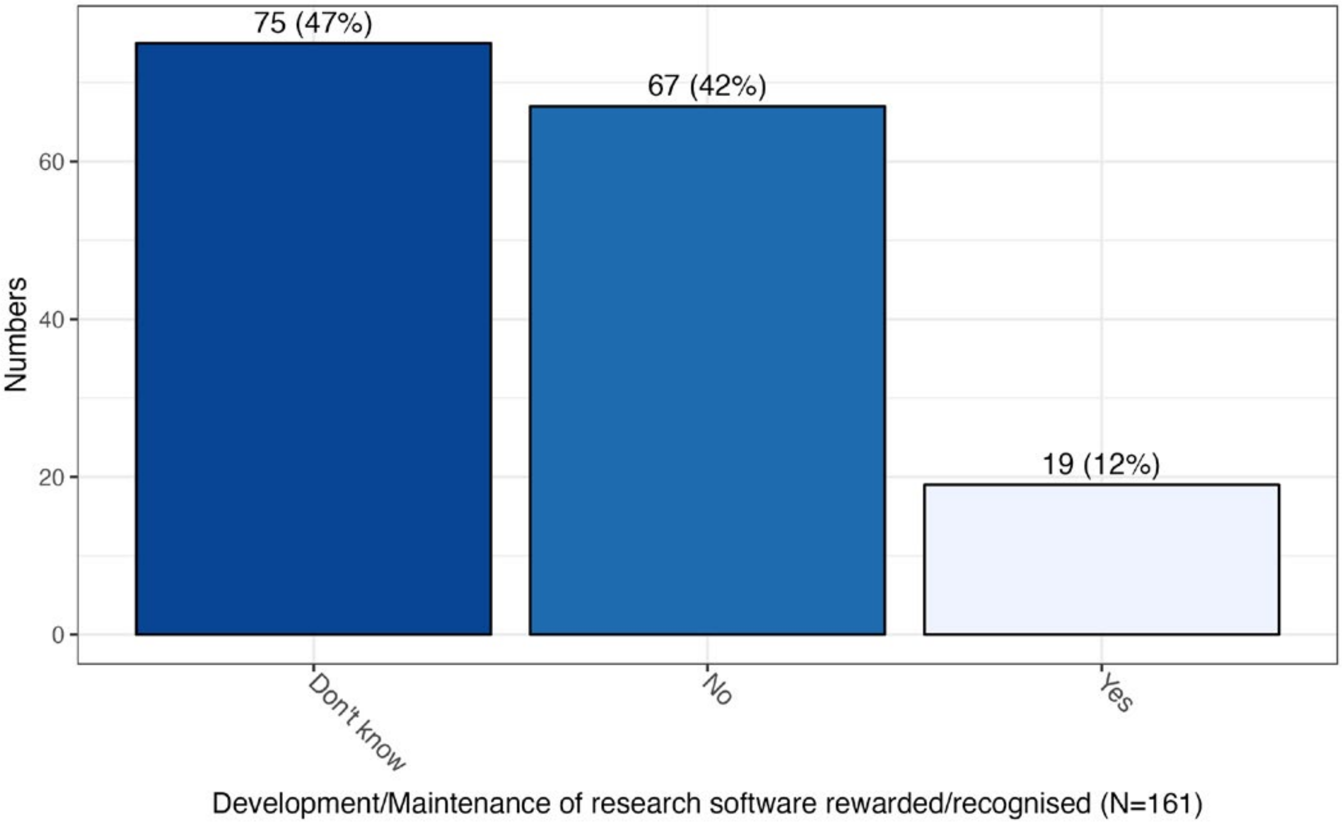


Figure 55. Whether respondents thought that software development was rewarded or recognised.

Fig. 56 shows responses by career stage. Early (61%) and mid-career (42%) researchers tended to believe that software maintenance and development is not sufficiently rewarded or recognised as a career path. Junior researchers (61%) in particular were unclear as to whether software maintenance and development is rewarded. Across all career stages, more respondents answered "No" than "Yes".

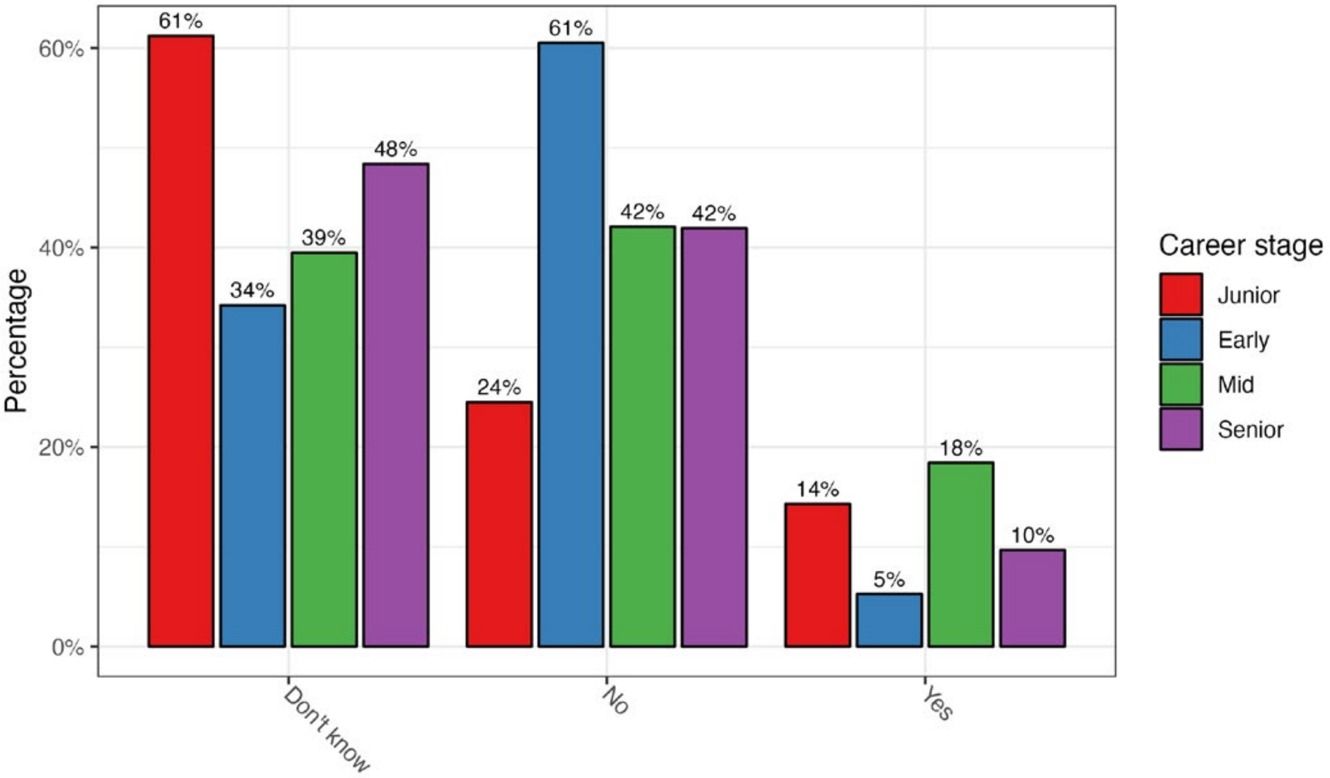


Figure 56. Whether software maintenance and development is rewarded or recognised, by career stage. Percentages are relative to each career stage.

We invited respondents to elaborate on their answers. A lack of space in the academic recognition system for software development was frequently cited. In particular, respondents lamented the difficulty of getting software development recognised in the REF alongside other outputs like publications, and bemoaned the impact of this lack of recognition on their careers.

> "I spend a lot of time developing technical skills and learning to develop software, but my end of year evaluation asks about whether I've published papers and how many colloquia I've attended. Job posts require publications in high impact journals and grants. No-one cares about open source software because it doesn't bring in any money." [JCR]
>
> "A lot of effort and research can go into developing a package yet it is not always deemed as important as an academic publication". [JCR]
>
> "It is not widely rewarded at all in terms of career progression. Outputs including papers that describe access to data made openly available are not recognised as 'high-impact' research output. But it does give ... skills desired in a team." [ECR]
>
> "Approximately all that matters for retention/promotion decisions is published journal articles, and often only in certain journals. Most importantly, institutions demand outputs that are REF-returnable. Whilst there are a few journal outlets for papers about software packages (e.g., the Journal of Statistical Software), it remains incredibly difficult to get to a state where software that you have developed is recognised as a REF-returnable output." [ECR]
>
> "It is only captured as a Researchfish output, not visible within the institution. Limited career paths for developers" [MCR]

The problem is not insuperable in principle, and our respondents made suggestions as to how software development might be more closely aligned with existing reward frameworks.

> "If a paper is written about that software then the answer is yes, in terms of being equivalent to other fields of work. Software should operate within peer review (at least for a journal publication) if it is to be recognised in academia." [MCR]
>
> "Somewhat dependent on software. Stata has the Stata journal and R packages can be published in the Journal of Statistical Software. However, these resources are somewhat overlooked by colleagues not actively using said software." [MCR]

Poor integration of software developers within teams or departments also was put forward as a reason that software development is not recognised within core academia.

> "It is still [feels] like software development can be outsourced and a small group of researchers should be dedicated to such a task. Collaborative software development is not yet seen as a transversal activity. Software maintenance is not yet included as part of the research tasks. It is usually perceived that external software can always supply project needs." [ECR]

Respondents also felt that there was little support for software maintenance and development in terms of funding and time from institutions and the ESRC.

> "In my area, we have little to no support in relation to research software. It is hardly mentioned." [JCR]

> "'Field' is vague. Some institutes are better than others, and I'd say mine is one of the less-good ones as there is no funding/incentives for continued support of software, leading to lots of quickly-written code that is published, but then not maintained and essentially gets discarded. Research grants are poor for maintaining software as institutional memory is almost completely missing." [ECR]

> "Often it is funded excluding overheads, which is not attractive to the institution." [MCR]

Our respondents felt that user communities go some small way towards mitigating this issue, operating as a 'shadow' recognition system.

> "Recognised in that you remain engaged by the user community, which is vital, but not recognised by institution/REF/etc. Not recognised by funders like ESRC either." [SNR]

> "People are definitely not often ready to create, maintain or share research software on their own. Not all are supported by colleagues (whose contracts do not dedicate time or money for this work) or dedicated institutional support teams. I am not aware of any efforts within the institution to recognise or reward such efforts. There are some community-led efforts to recognise but not reward (as far as I know)." [ECR]

Other respondents felt that the skills academia requires but does not recognise are valued by industry, leading to the loss of talent.

> "Researchers are expected to develop open source software in their free time, with no funding and no recognition. If this is going to continue, academics will be moving more and more into the private sector". [JCR]

> "In my area, the sorts of things people make that should be recognised and rewarded are R packages. There is basically no recognition for the creation of packages, even if these are widely used in the research community, the example that comes to mind is [software engineer], who used to be at the LSE and created many widely used R packages but left academia and now works for Facebook, in part because the immense service he provided for the research community was not valued by his institution!" [ECR]

Some respondents felt that their discipline in particular was poor in this respect.

> "Expert colleagues within and beyond my discipline appreciate the work necessary in creating new software and the challenges of sharing (and especially maintaining and supporting it), but levels of awareness and recognition within my own discipline and department more broadly are low." [SNR]

> "There is no recognition of how hard it is to create software as a research tool in education, and no understanding of its value within a design-based research methodology. The ESRC has often not even mentioned education much as a key area for research. The value of creating and testing software does not seem to be recognised in the way either funding, support or evaluation is done." [SNR]

Infrastructure and hardware

To investigate infrastructure and hardware use in the social sciences, and to establish whether high-performance computing is used in particular, we asked our respondents (Q14, N = 161) where their digital tools or software are normally run.

- My own laptop/desktop **(Own)**
- Institutionally provided laptop/desktop **(InstitutionallyProvided)**
- Server operated by research group or department **(RG/DepServer)**
- Institutional central service **(InstitutionalCentralService)**
- Individual data centre or at a data safe haven **(DataCentre/SafeHaven)**
- UK Tier 2 high-performance computing service **(Tier2)**
- UK Tier 1 high-performance computing service **(Tier1)**
- The cloud, e.g. Microsoft Azure or Amazon Web Services **(Cloud)**
- Other **(Other)**

The UK Tier 1 system provides supercomputing facilities such as ARCHER2[75], and the UK Tier 2 system[76] provides computational services at a level between institutional and Tier 1. Respondents were able to choose as many answers as were applicable to them. A summary of the responses is given in Fig. 57.

Most respondents reported using either their own hardware (68%) or institutionally provided hardware (68%) to run their software and tools. As respondents were able to select multiple answers, we can separate the responses as follows: 37.9% used both institutionally provided hardware and their own hardware, 30.4% exclusively used their own hardware, and 29.8% exclusively used hardware provided by their institutions.

The use of departmental or research group hardware was reported, but to a lesser extent. It is notable that no respondents in our sample reported using Tier 1 or Tier 2 services.

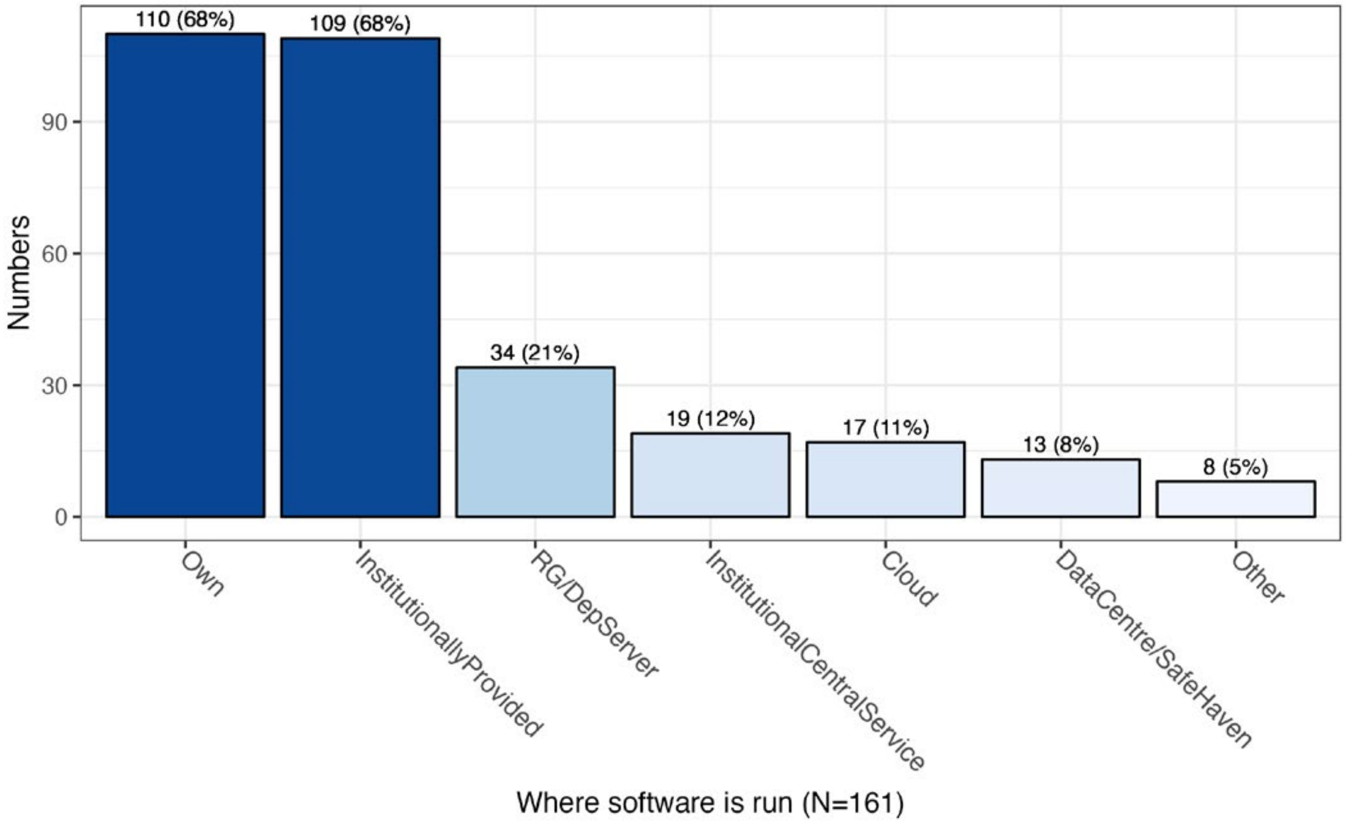


Figure 57. Where software is run. Respondents were able to choose multiple answers.

Fig. 58 shows an overview of hardware use by career stage.

[75] https://www.archer2.ac.uk (last accessed on May 16 2022)

[76] http://www.hpc-uk.ac.uk/facilities (last accessed on 16 February 2022)

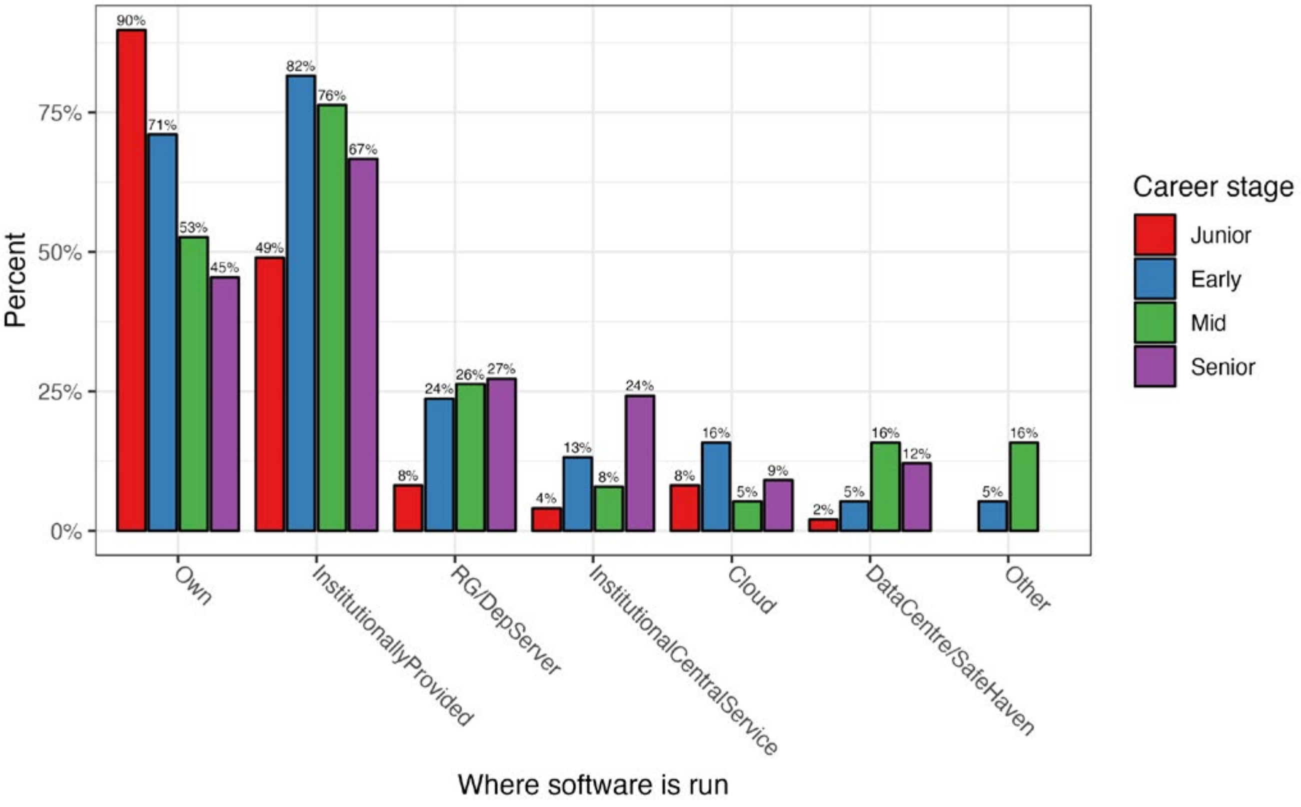


Figure 58. The hardware used to run tools and software by career stage.

Fig. 58 shows that respondents are most likely to use their own hardware or institutionally provided hardware across all career stages. However, the fact that our questionnaire responses could be chosen either alone or in combination with other responses complicates these results. The data is disambiguated in Table 11, which breaks down respondents' answers into the categories of 'Institutional', 'Own', and 'Both'. It shows that 51% of junior career researchers are mostly using their own machines. ECRs rely mostly on a combination of institutionally provided hardware and their own hardware (53%), while mid-career (42%) and senior career (45%) researchers mostly rely on institutionally provided hardware.

	Junior	Early	Mid	Senior
Institutional	5 (10.2%)	11 (28.9%)	16 (42.1%)	15 (45.5%)
Own	25 (51%)	7 (18.4%)	7 (18.4%)	8 (24.2%)
Both	19 (38.8%)	20 (52.6%)	13 (34.2%)	7 (21.2%)

Table 11. Use of hardware by career stage.

The data shows that there is a trend towards greater reliance on institutionally provided hardware as seniority increases. Senior researchers are the most frequent users of institutional services (45%). ECRs (16%) use cloud services the most, while mid-career (16%) and senior career (12%) researchers tend to use data centres or safe havens.

The tendency of researchers to favour institutional facilities as they become more senior may arise from their having been exposed to differences in institutional policies during their careers. One respondent noted,

> "In the past I have used HPC facilities, but my current university does not have these facilities. National services only available for EPSRC-funded projects, not ESRC. This has massively impacted upon the analysis I can undertake. Note that I do not require huge amounts of resources, but more than can be done on a laptop and without being able to run code in parallel." [ECR]

Barriers to research software practices

We asked our respondents (Q16, N = 149) to help us identify the main barriers to the use of software in economic and social sciences research by selecting as many of the reasons given below as applied to them:

- Lack of expertise within your team **(LackOfExpertise)**
- Unsure how to best engage with research technical specialists such as software engineers, digital archivists, etc. **(UnsureHowToEngageSpecialists)**
- Training is not available **(NoTraining)**
- Training is available but you have no capacity to engage **(NoCapacity)**
- Infrastructure is not available (Infrastructure could include: digital archives; computers with access to specialist software; datasets not digitised; etc) **(NoInfrastructure)**
- No disciplinary tradition for digital methodology and tools **(NoTradition)**
- Concern that less value is attached to digital publications and other outputs of research, etc. **(ConcernAboutValue)**
- Format of data and assets you work with make them less amenable to technology **(AssetsNotAmenable)**
- Lack of funding to support research projects/components of projects focusing on digital data and digital assets **(LackOfFunding)**
- Previous bad experience **(BadExperiences)**
- I have not found software that is fit for my purpose **(NoPertinentSoftware)**
- Lack of time to learn how to best use research software **(LackOfTime)**
- Other **(Other)**

The results are summarised in Fig. 59.

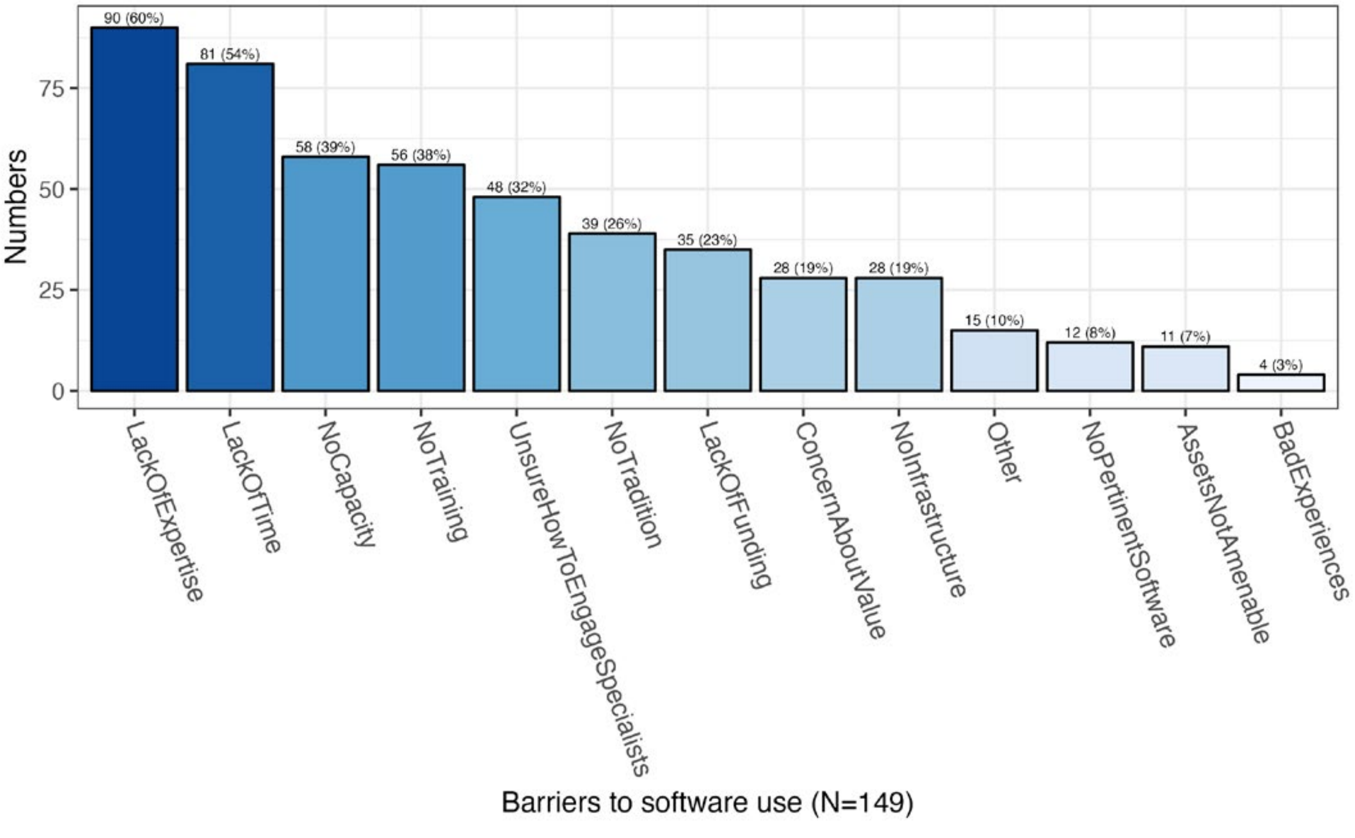
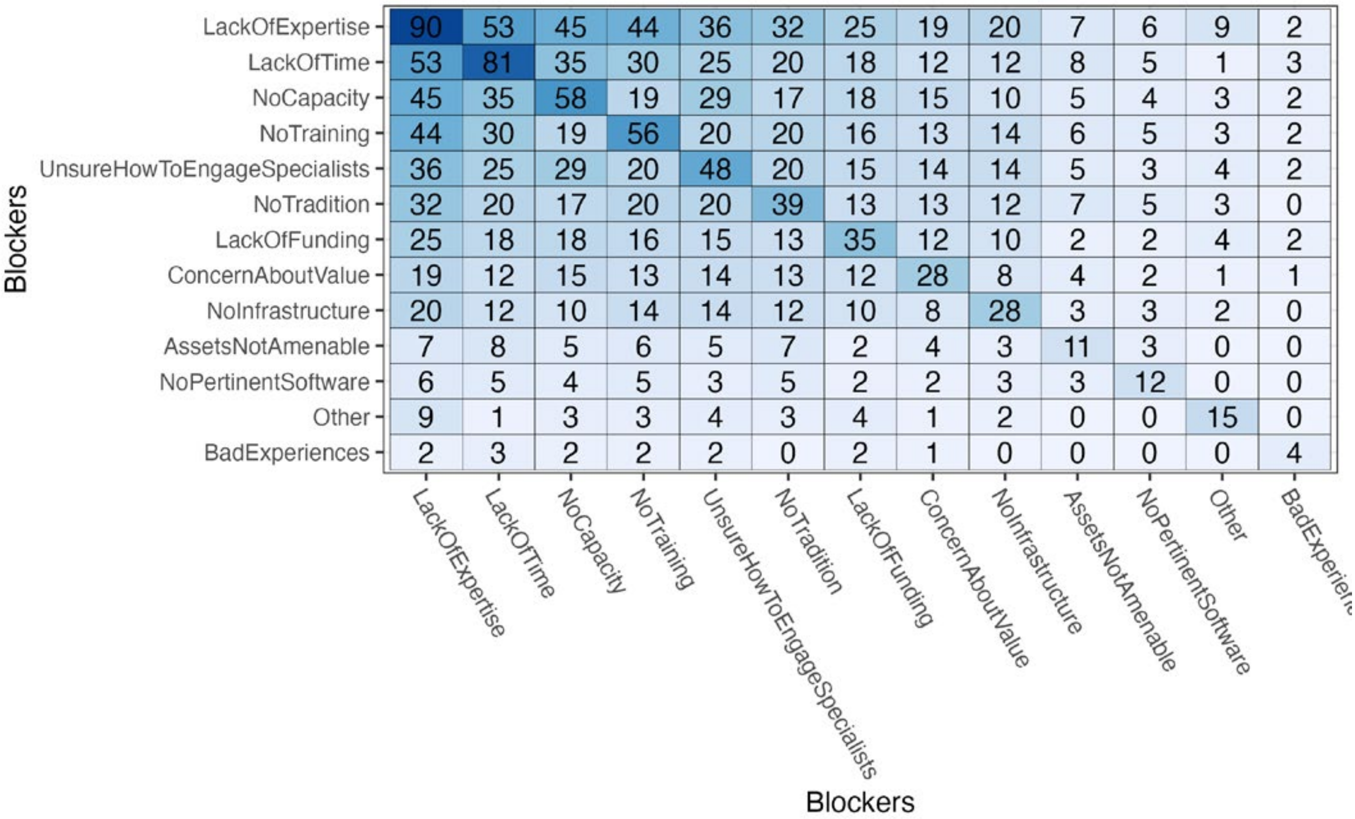


Figure 59. Barriers to software use.

The most commonly cited barriers were lack of expertise (60%), lack of time (54%), lack of capacity (39%), and lack of training (38%). Fig. 60 shows choices in combination, allowing us to see that these four responses were frequently chosen together, along with, to a lesser extent, lack of certainty in engaging a specialist and lack of existing tradition.



Blockers	LackOfExpertise	LackOfTime	NoCapacity	NoTraining	UnsureHowToEngageSpecialists	NoTradition	LackOfFunding	ConcernAboutValue	NoInfrastructure	AssetsNotAmenable	NoPertinentSoftware	Other	BadExperienc
LackOfExpertise	90	53	45	44	36	32	25	19	20	7	6	9	2
LackOfTime	53	81	35	30	25	20	18	12	12	8	5	1	3
NoCapacity	45	35	58	19	29	17	18	15	10	5	4	3	2
NoTraining	44	30	19	56	20	20	16	13	14	6	5	3	2
UnsureHowToEngageSpecialists	36	25	29	20	48	20	15	14	14	5	3	4	2
NoTradition	32	20	17	20	20	39	13	13	12	7	5	3	0
LackOfFunding	25	18	18	16	15	13	35	12	10	2	2	4	2
ConcernAboutValue	19	12	15	13	14	13	12	28	8	4	2	1	1
NoInfrastructure	20	12	10	14	14	12	10	8	28	3	3	2	0
AssetsNotAmenable	7	8	5	6	5	7	2	4	3	11	3	0	0
NoPertinentSoftware	6	5	4	5	3	5	2	2	3	3	12	0	0
Other	9	1	3	3	4	3	4	1	2	0	0	15	0
BadExperiences	2	3	2	2	2	0	2	1	0	0	0	0	4

Figure 60. Contingency table showing how responses correlate.

Fig. 61 further breaks down the responses by career stage. It shows that senior researchers (58%) consider lack of time to learn how to use software and lack of capacity to engage with it (58%), together with lack of expertise (55%), to be their greatest barriers to using software. They are also the group with the greatest level of concern (30%) about the perceived value of software in research.

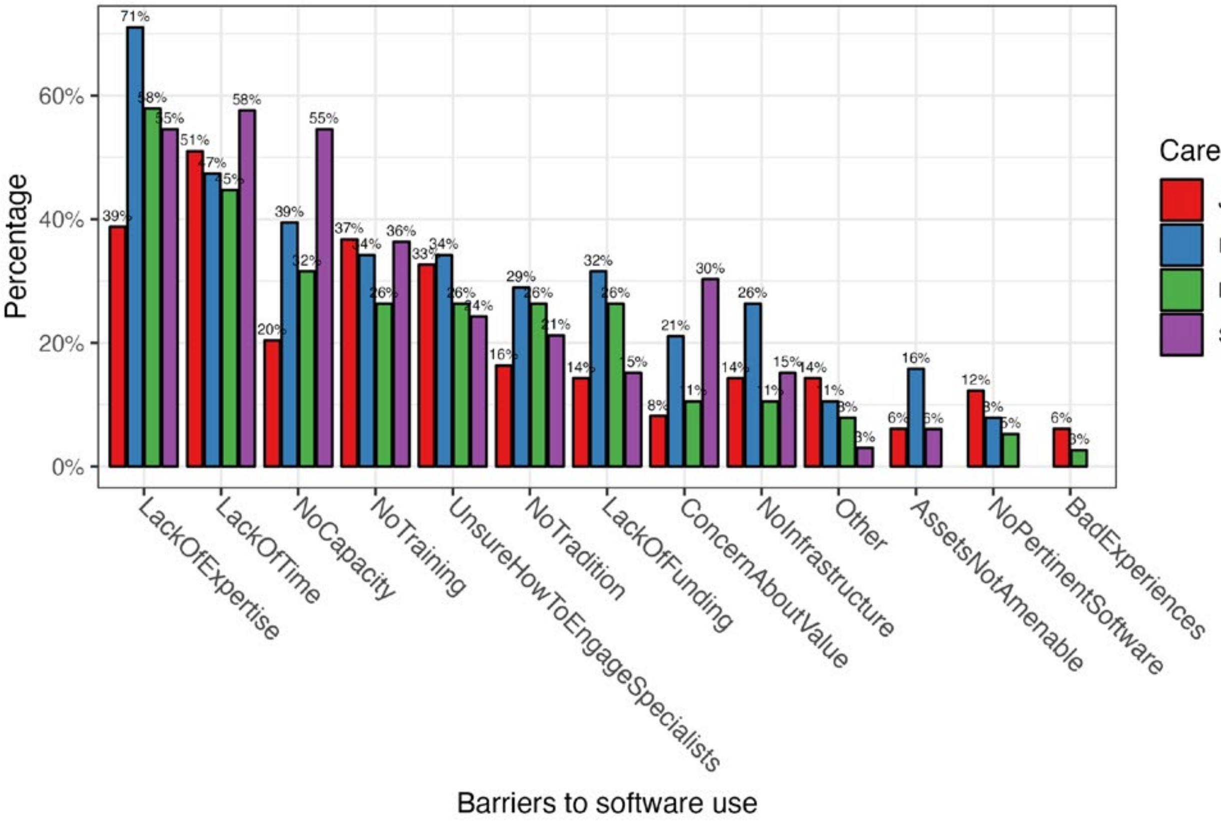


Figure 61. Barriers to the use of software by career stage.

Early career stage researchers are the respondents most constrained by lack of expertise (71%), and most concerned about the lack of funding available to support their research projects (32%). Mid-career researchers also cite lack of expertise (58%), which correlates with answers to survey questions shown in Section 4.5.

Interestingly, junior researchers, i.e. postgraduate researchers, cited lack of time as the biggest impediment to their engagement with research software. We examine this sentiment more closely in the interview responses given below.

'Lack of time' is a problematic answer because it aggregates a host of underlying reasons, including time to learn the software and time that might potentially be lost if there is a problem with the software. The learning curve for non-programmers also creates time pressure.

> "Not having enough time to actually learn how to use it. That seems to be one of the biggest barriers, I think, because you do have to dedicate quite a lot of time to actually learning how to use it. That's one of the things that we do with [software package]. We have a wiki where we actually explain how to use it, and we try and make things as easy to use as possible. Often some people will struggle with things like the programming side of things which are a little bit different from what they are used to." [179250558]

> "Lack of time is definitely a big one. And that was also one of the reasons why, when I was talking about the module that I audited is I think if I wanted to, there is the opportunity to actually take the module formally, but the kind of time pressure that that would put me under would be ridiculous, or I would not have time to engage with assessments and actually pause them, it'll be just so much stress. So often it feels whenever I need to learn something, it's more like, okay, let's just learn the basics and try and get the job done, rather than learning ... really, really well, because that will take a lot of time. And also, I have to actually make progress with the project rather than just spending time learning." [2964377624]

> "And it ... comes down to time because it's like, I feel like there are so many ways to get things wrong ... one time, I erased about two weeks of progress by accident". [2964377624]

Lack of expertise, meanwhile, can be compounded by a perception of software as challenging.

> "You know, I'm more of a qualitative researcher ... over there is computer scientists, people who will be assembling things from different elements." [3270614512]

Crucially, lack of expertise can affect both individual researchers and entire departments.

> "My supervisor doesn't know R and she said that if I used R, my data analysis would be really hard for her to support, she'd have to learn it as well. And then I didn't know I didn't feel confident enough to be able to...Because it's one thing

> if you get things wrong in an assignment, but it's another thing if you get them wrong in your data analysis. And if you haven't got the other people there to help check it for you, then that makes me a little bit more concerned. So I feel a bit ashamed for saying that I've used SPSS because it isn't what I want to use. It's because of not having the expertise around me to support me." [1269794877]

> "I would say the lack of expertise in the team. I have a supervisor, and he's really technically skilled and he knows R really well. But at the same time I sometimes need to do stuff that he has never done before. And then whenever I hit that point, I'll have to try and find some resources to learn it." [2964377624]

> "There's a confidence thing around using some of these new softwares and packages. And this kind of feeling like we just need to spend a bit more time getting our heads around it and awareness that maybe students who have done their undergraduate more recently have maybe had more time to get used to using those things through the course of their undergraduate or have been exposed to it." [1269794877]

Fig. 62 shows the data by career stage and gender. It is worth noting that 80% of the junior researchers in our sample were women, and 43% of the women who responded to the survey were junior researchers.

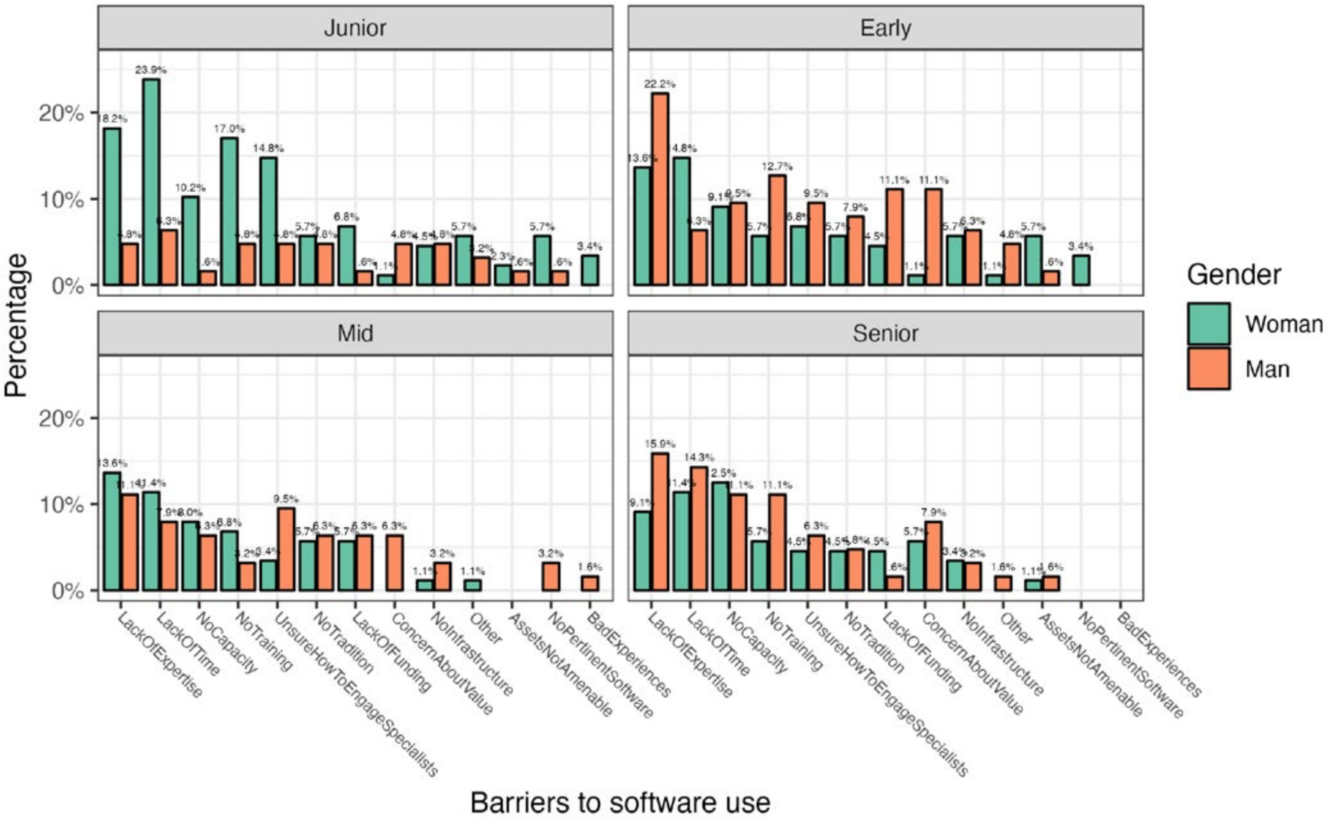


Figure 62. Barriers to the use of software by gender and career stage.

Women in the junior career stage specified lack of time (23.9%) as the main barrier to engaging with software. Among ECRs, men cited lack of expertise (22.2%) while women cited lack of time (14.8%). Both male and female mid-career researchers cited lack of expertise (13.6% and 11% respectively), while male senior researchers cited lack of expertise (15.9%). Female researchers at the same career stage cited lack of capacity (12.5%).

Although representation from respondents reporting disabilities was small (about 15%, or 25 out of the 164 survey respondents), we were keen to know whether their experiences differed from those of the rest of the cohort. Results are shown in Fig. 63.

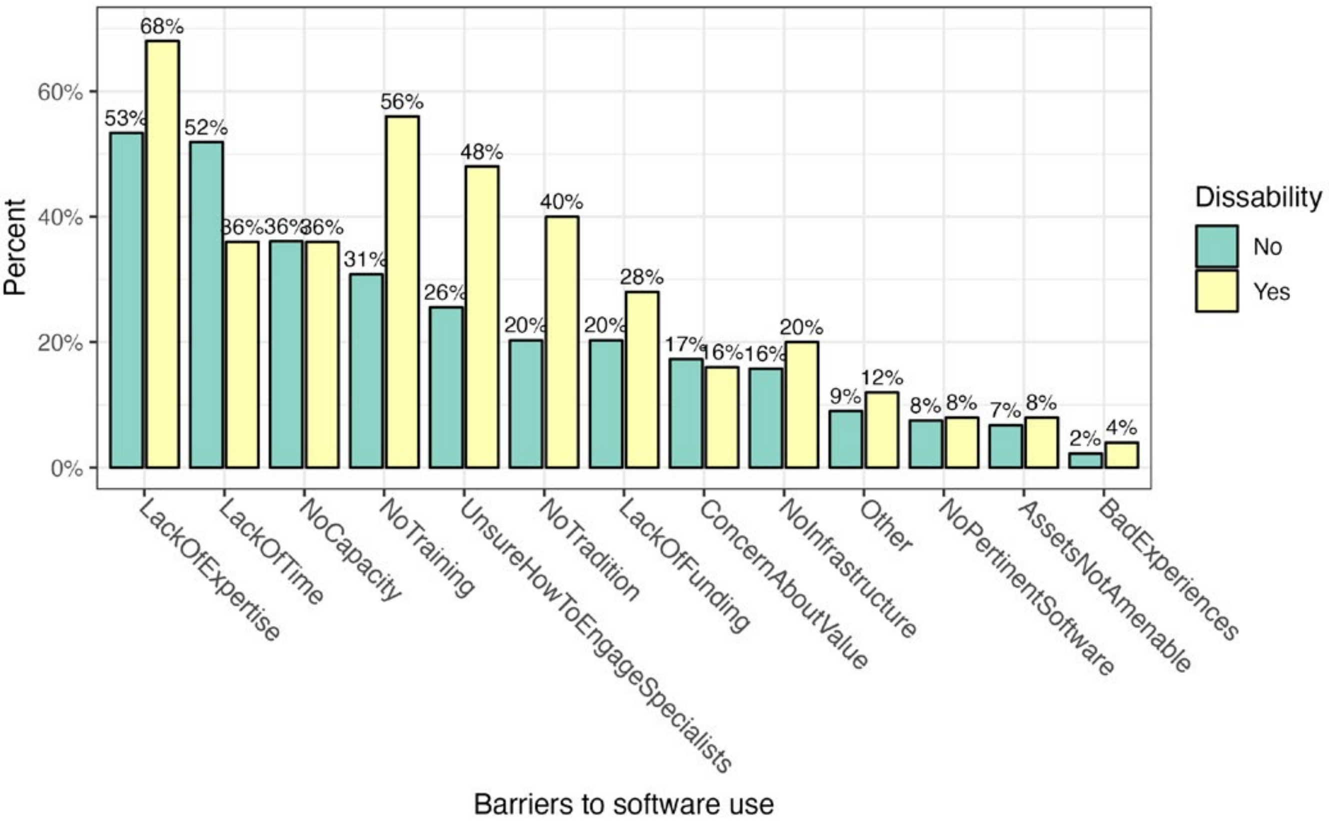


Figure 63. Barriers to software by respondents reporting and not reporting a disability.

Among those reporting a disability, lack of expertise (68%) was given as the main barrier to using research software. Lack of time (36%) and lack of capacity (36%) were perceived to be less significant barriers than, for example, lack of training (56%). This might indicate that appropriate, accessible training is not available.

Where respondents chose ‘Other’, our survey invited them to provide additional information. Junior career researchers cited privacy and the use of data, the inhibitive cost of courses, lack of personal expertise, and lack of confidence, characterised by the general belief that they were not technologically capable.

There are of course many software training courses, so it’s important to know what respondents mean when they claim no training is available. They may be referring to a lack of training within their own organisation or department, or a lack of training at an appropriate level.

> “[Training] was not part of my PhD programme, and my faculty has not explicitly provided training for lecturers that I am aware of.” [ECR]

> “I am aware that there is ESRC training that focuses on Python, but ... because it's social sciences, it's usually kind of focused on very basic stuff. So it will be an introduction to inference or kind of things that I typically know.” [2964377624]

Respondents who wrote about the difficulty of engaging with specialists noted that the challenge was in knowing not only whom to work with, but also how to work with them.

> “I’m often not sure where to go to find someone who could provide the technical expertise for some of the things I want to do.” [MCR]

Different understandings of concepts and processes can also complicate this.

> “I think I've noticed a lot with the game, because programmers, or at least what I've experienced, ... tend to programme stuff in a big data mindset. So it was a big mess in order to actually get the programmer to have the logic of ‘I just need a simple Excel file’. It makes it a little bit complicated to do stuff between different disciplines. That's for sure.” [4288358080]

Lack of funding was cited in several contexts, including training, infrastructure, and licensing.

> “No financial support for ongoing data infrastructure maintenance and operation; having to scrabble cap in hand every five years to overcome technology debt.” [ECR]

> “I would like to complete an intense face-to-face course in R over several days. Although available, these courses are too expensive. Although I use R, I know my skills need further development.” [JCR]

> "So Stata is provided by the [school]. But there's an interesting catch there, not every student in the school has access to Stata. I initially had it because I was tutoring in a course that utilised it. I think not just any random student can get it, you need to be signed up to a course that utilises it from what I understand. I'm guessing because of the price difference between Stata and SPSS, they don't make it widely available to all students." [4288358080]

Similarly, relevant answers for the response 'Infrastructure is not available' spanned topics like the lack of VPNs for security, the lack of support for long-term maintenance and sharing, the limited choice of software, and the labour required to troubleshoot software.

> "Institution requiring use of Microsoft tools for cloud storage which are not compatible with how Stata handles file paths (requires Static). Secure storage within university system is required for secondary datasets. During campus closures for COVID this meant using VPN to securely access data stored in university servers, but they are no longer supporting VPN, even though people are still needing to work from home." [JCR]

> "No longer-term support for maintaining software is available, leading to lots of quickly-written code snippets that are not maintained/tested/supported and then disappear." [ECR]

> "And it depends on availability, also, which tools are available to me when I start with my project, which tools are available at my university? So it totally depends on that." [4098334726]

> "The main challenge is just ensuring uptime, particularly the scraping, software gets broken a lot. So it's actually quite a bit of labour to manage it." [3270614512]

There is often a combination of issues at play, including a lack of understanding of potential value, expertise, expense, the speed of software change, pressure on capacity, and a perceived propensity for qualitative rather than quantitative research in the UK.

> "It's mainly people. Given the current stress on teaching, staff have little time, or inclination, to learn anything new, including software. When the Q-Step programme[77] has had some effect, things may improve, but at the moment people seem to be more interested in flowery words than in detailed relevant data. There are opportunities, but few senior university managers have the background needed to see that such software is desirable." [SNR]

> "Colleagues with historic experience of different software systems. For example, joining a research project where the other members have long experience of using Stata, when I have more experience of R. (I personally find Stata to be particularly problematic, both because of the highly idiosyncratic language design of .do files and because of being an expensive piece of software, limiting potential future use; anyone you want to share your code with needs to have access to a Stata licence.)" [ECR]

> "First of all, it's always learning the software, it's always a steep learning curve with the software. The challenge is sometimes troubleshooting, if it happens out of the blue, and then you spend hours and hours searching Google for how to resolve the issue. Sometimes your computer shuts down, because it's overheated [...] And I'm not trained in that. So [you need to] find a solution and where to go, and you lose your data, you lose your information, you lose the work that you've done. Other than that, it's advancing so fast, it's advancing so fast, it's difficult to catch up." [4098334726]

Of the barriers to software use not explicitly mentioned in the survey, lack of perceived value within academia came up most frequently in interviews.

> "In an ideal world, I would love for [my software] to be widely available ... neat scripts that anyone can just use, but at the same time, if I have to prioritise putting this into a paper and writing it up and then making the script nice and clean, then [I'm] definitely prioritising the paper because I know that's what is more valued in academia." [2964377624]

> "In terms of longevity of stuff, if they don't want everything to be prototype and throwaway, then they need to have visible routes to allow ... successful projects that demonstrate impact, and [are] worth pursuing to help support future research ... there needs to be some visible mechanism about how that can happen, because I just don't see it at the moment. It doesn't seem to be in that plan." [4561769548]

As we saw above, the majority of junior and ECRs are using their own hardware in order to ensure that they can retain software settings, which can be a barrier to using more intensive software.

> "I had to buy a new computer and it's kind of a pricey computer. That's what was advised by two of our professors here as well, who were teaching me software-related courses, they said that it's okay, if you use lab computers, they are good. But eventually, you should be able to do it on your system. Because it's constantly, constantly updating, and every day things are changing, I have to download, I have to use web scraping or use complex tools, things are changing. And if I have to kind of use different computers or cloud computing, somebody else might have altered the settings that I have done." [4098334726]

In the next section, we will review researchers' familiarity or otherwise with relevant institutional and funder software policies.

[77] https://www.nuffieldfoundation.org/students-teachers/q-step (last accessed June 22 2022)

4.4 PERCEPTIONS OF SOFTWARE POLICIES

We sought to understand the degree to which researchers are aware of how the software they use is funded, licensed, and managed, as a means of ascertaining their level of engagement with it. We set out to measure the awareness of licensing using a Likert scale. Respondents were invited to choose from Strongly Agree, Agree, Undecided, Disagree, or Strongly Disagree in response to a series of statements, with the following guidance:

> When answering this question, think about the most important piece of software you use for research that you couldn't live without. To what level do you agree/disagree with the following statements? (Non-mandatory)

Awareness of software funding, management, and licensing

Q13a, N = 160, asked whether respondents agreed with the following statement:

> I am aware of how the software I use is funded, managed and licensed.

Fig. 64 shows that most respondents (32% for Strongly Agree, 37% for Agree) felt they understood how the software they use is funded, managed, and licensed. However, a more nuanced picture emerges when we examine the responses by career stage.

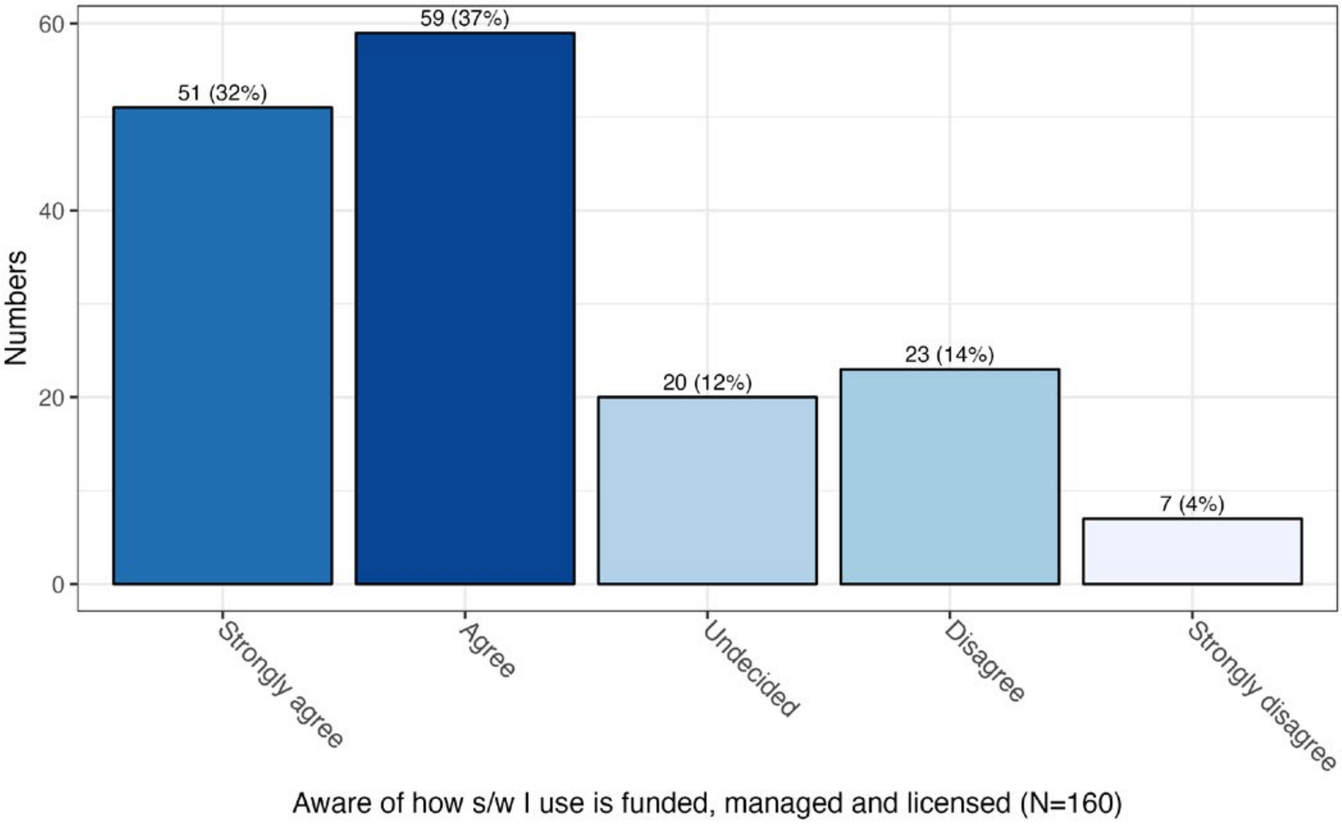


Figure 64. Awareness of how the software I use is funded, managed and licensed.

Fig. 65 refines the data by career stage. It shows that early career, mid-career, and senior career researchers feel sufficiently well-informed, with 74% or more agreeing with the statement and fewer than 11% disagreeing. But junior career researchers appear more ambivalent, with roughly 47% agreeing with the statement and roughly 35% disagreeing.

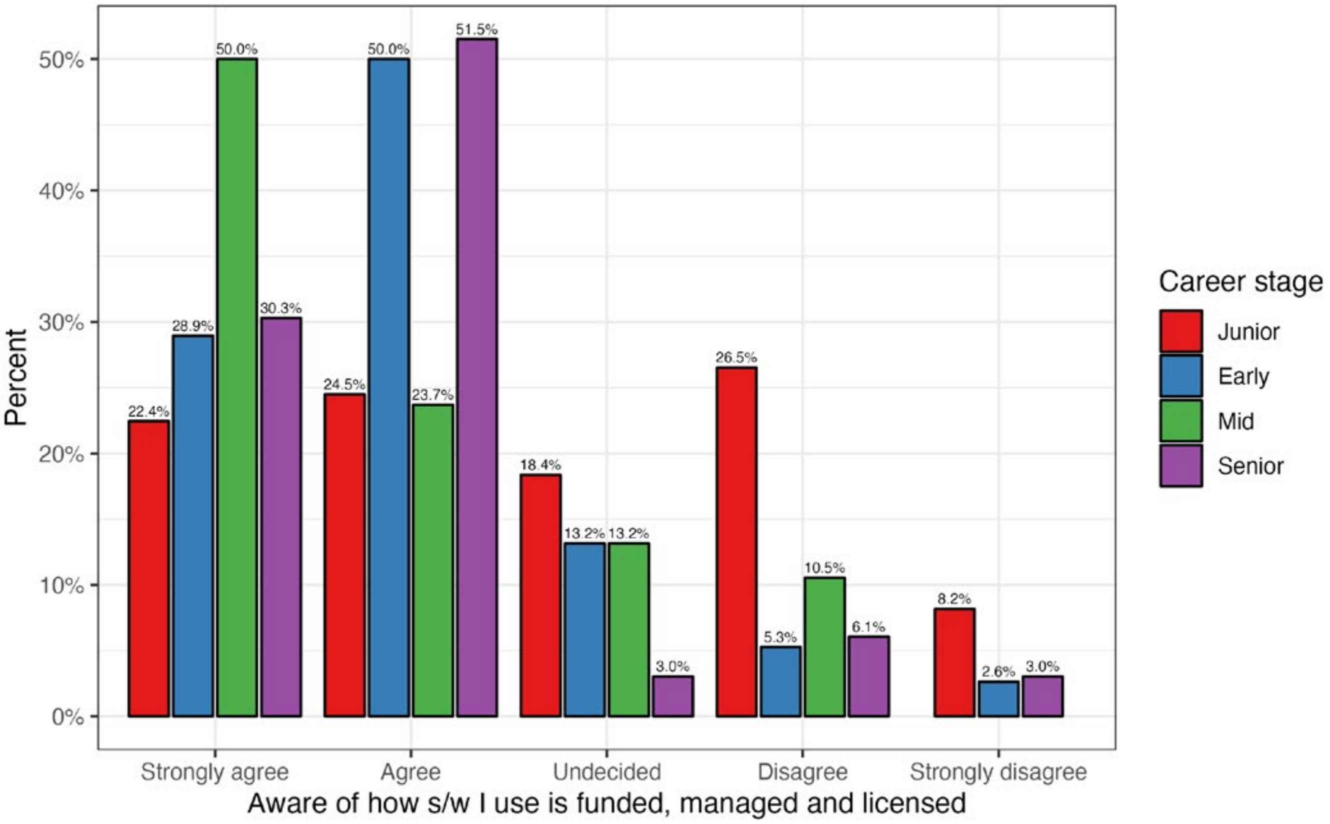


Figure 65. Awareness of how the software I use is funded, managed, and licensed segmented career stage.

Awareness of institution and funder policies

We further asked (Q13b, N = 160) how familiar respondents were with their institution's or funder's policies on software. The underlying hypothesis for this question was that researchers unfamiliar with these policies are less likely to be compliant with them.

> My institution/funder's policies on how software is funded, managed and licensed are clear to me.

In widening the scope of the statement from a personal understanding of the management, funding, and licensing of software to a broader understanding of the institutional perspective, we elicit less confidence from our respondents. The results are summarised in Fig. 66, where only 28% of respondents agree with the statement, 24% are undecided, and 25% disagree.

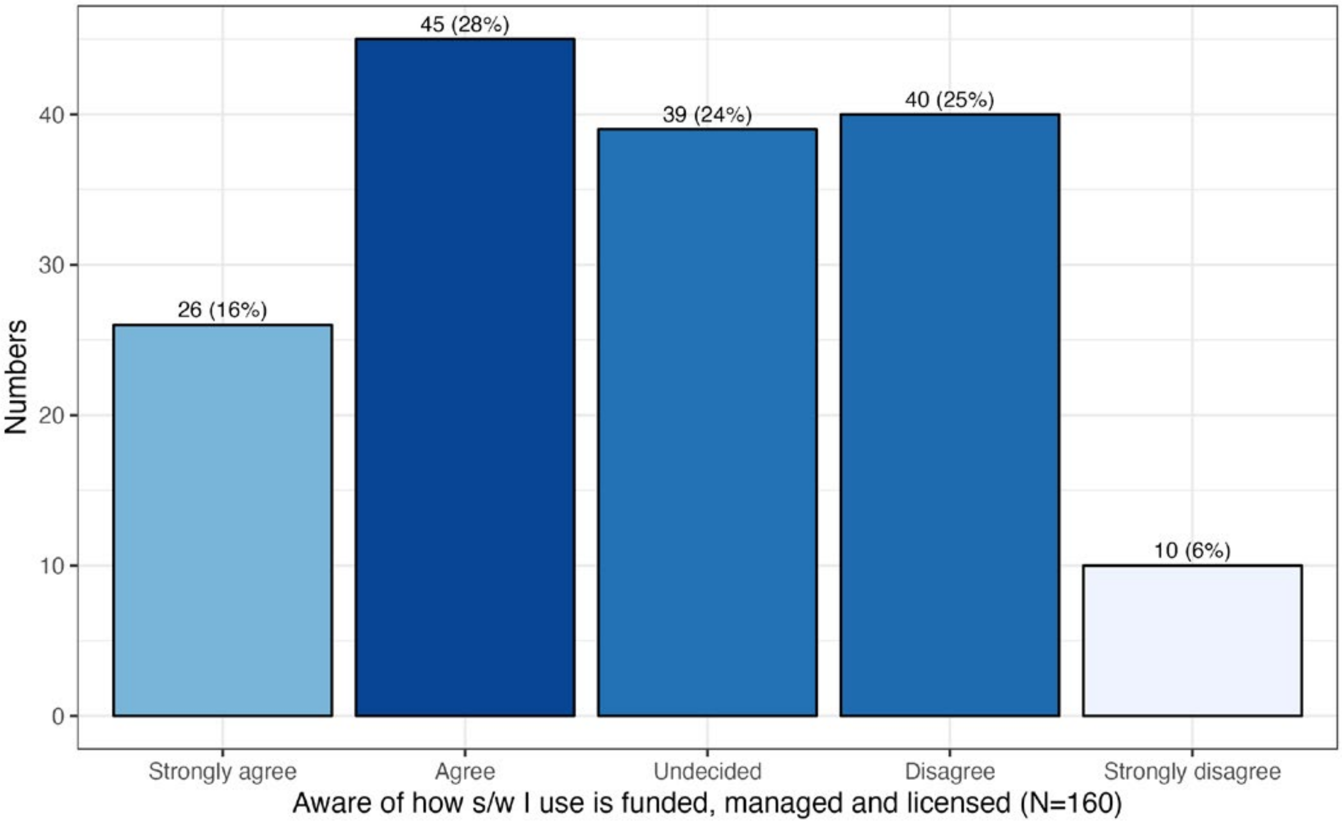


Figure 66. Understanding of institution/funder's policies on how software is funded, managed, and licensed.

Fig. 67 shows the data decomposed by career stage. It appears to demonstrate once again that researchers' confidence grows as their seniority increases. Among ECRs, 33% agree with the statement and 26% disagree, while 48% of mid-career researchers agree and 24% disagree. At the senior career stage, where confidence is highest, 63% agree and 15% disagree.

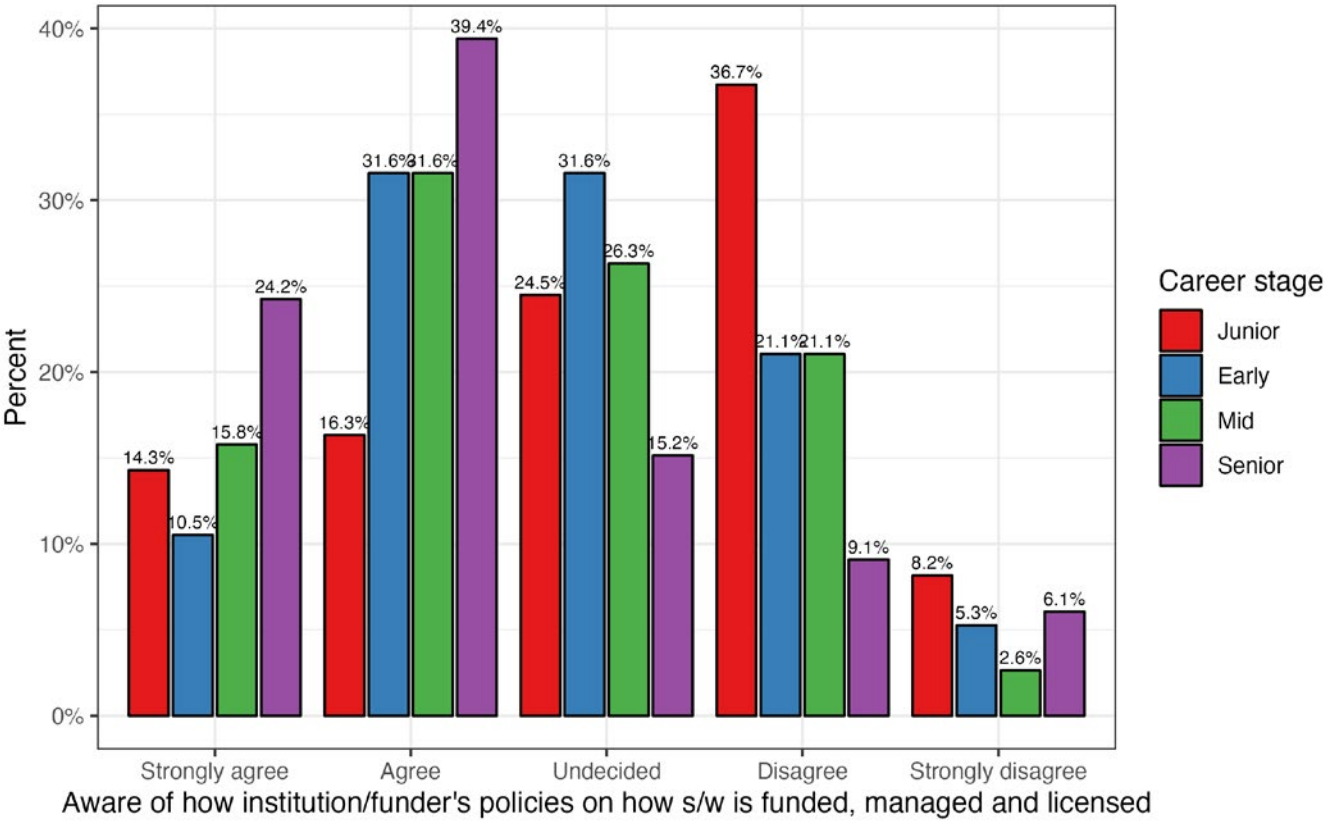


Figure 67. Understanding of institution/funder's policies on how software is funded, managed, and licensed, segmented by career stage.

Views on institutional attention paid to software

We asked (Q13c, N = 160) whether respondents agreed with the following statement.

> There is insufficient attention paid to software funding, management and licensing by the economic and social sciences research community.

The responses are summarised in Fig. 68. Relative to the preceding two questions there is a decrease in the number of respondents agreeing, and a significant increase in those who are undecided. The picture is mixed among those disagreeing. Agreement still outnumbers disagreement, but the margin has narrowed.

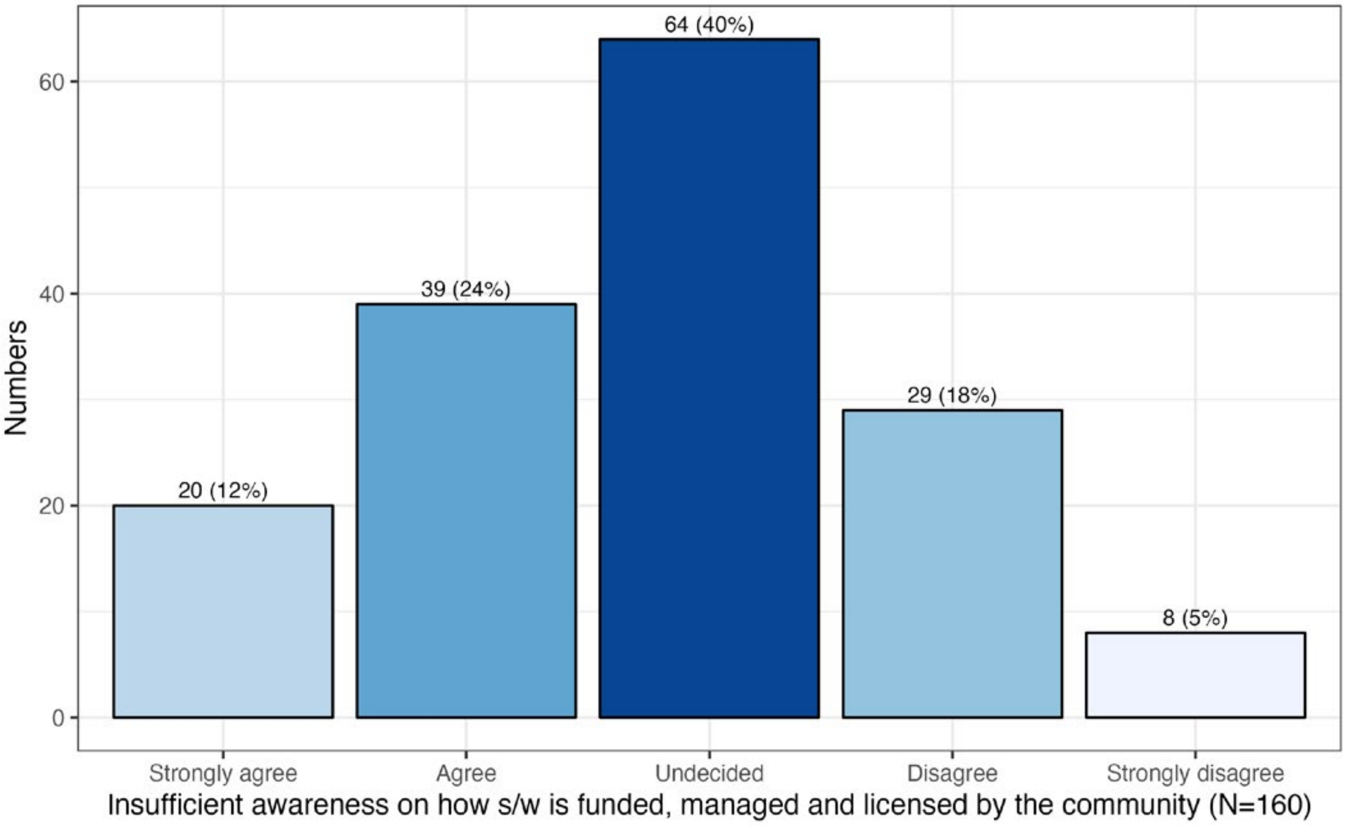


Figure 68. Whether insufficient attention is paid to software funding, management, and licensing by the economic and social sciences research community.

The data is segmented by career stage in Fig. 69. Respondents were relatively ambivalent, with Undecided proving to be the most popular response for every career stage. Junior career researchers reported the greatest uncertainty, with 45% choosing Undecided.

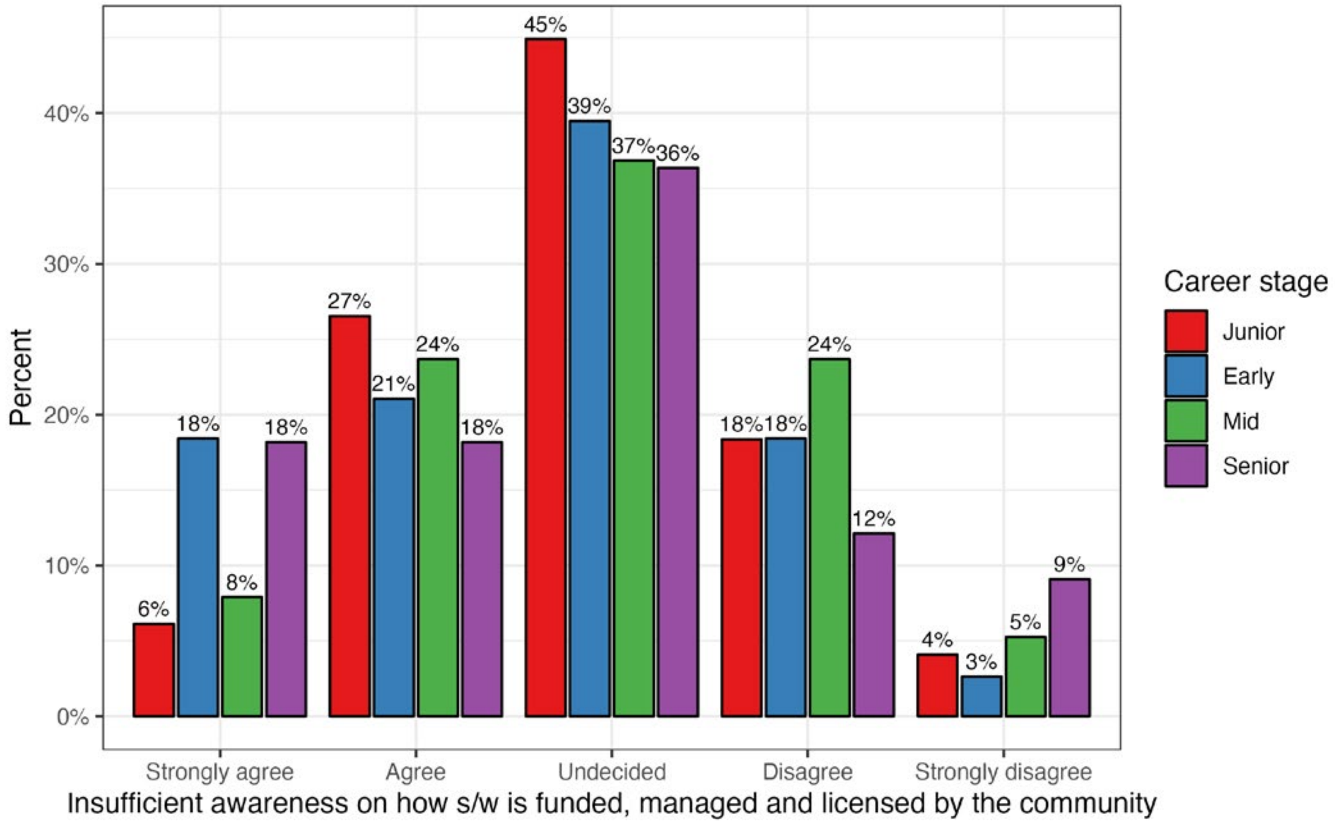


Figure 69. Whether insufficient attention is paid to software funding, management, and licensing by the economic and social sciences research community, segmented by career stage.

Incentives to engage with software policy

Finally, we asked (Q13d, N = 159) whether respondents agreed with the following statement:

> There is insufficient incentive for me to learn how my software is funded, managed and licensed.

Roughly 48% of those who answered felt that there was not enough of an incentive for them to learn how their software is funded, managed, and licensed, while around 26% were undecided, with the same proportion disagreeing with the statement (Fig. 70).

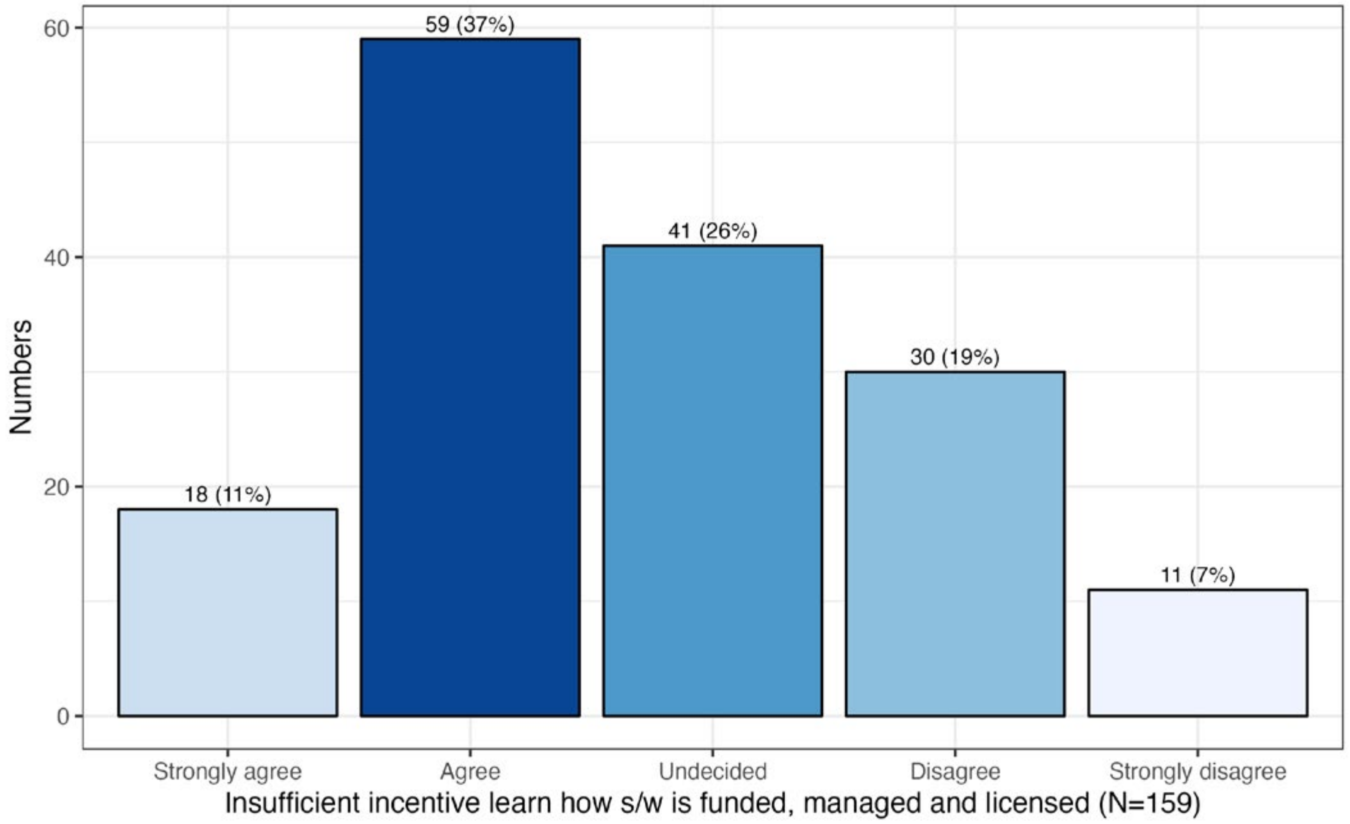


Figure 70. Whether there is insufficient incentive to learn how respondents' software is funded, managed, and licensed.

The responses split by career stage in Fig. 71 show that junior career (43%) and mid-career (42%) researchers agree with the statement most markedly, while early career (24%) and senior career (21%) researchers show the greatest level of disagreement. Less variation is evident among the other career stages.

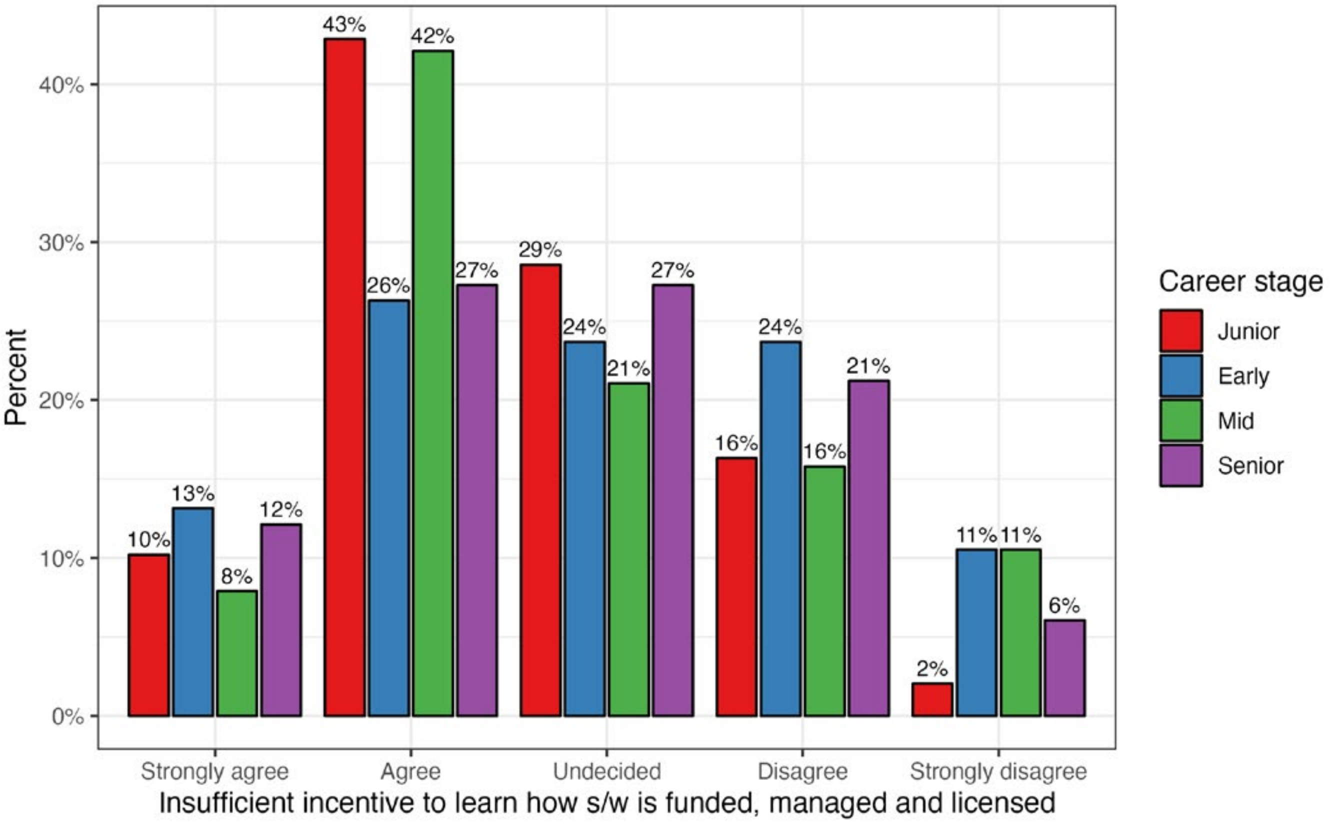


Figure 71. Whether there is insufficient incentive to learn how respondents' software is funded, managed, and licensed, segmented by career stage.

Comparisons

We ran an analysis of the previous statements:

- 13a: I am aware of how the software I use is funded, managed and licensed.
- 13b: my institution/funder's policies on how software is funded, managed and licensed are clear to me.
- 13c: there is insufficient attention paid to software funding, management and licensing by the economic and social sciences research community.
- 13d: there is insufficient incentive for me to learn how my software is funded, managed and licensed.

It is notable that the results appear to show an inherent contradiction. 36.9% of respondents agree that they are aware of how software is funded, managed, and licensed, and yet almost exactly the same proportion of respondents also agree that there is no incentive to learn how their software is funded, managed, and licensed (Fig. 72).

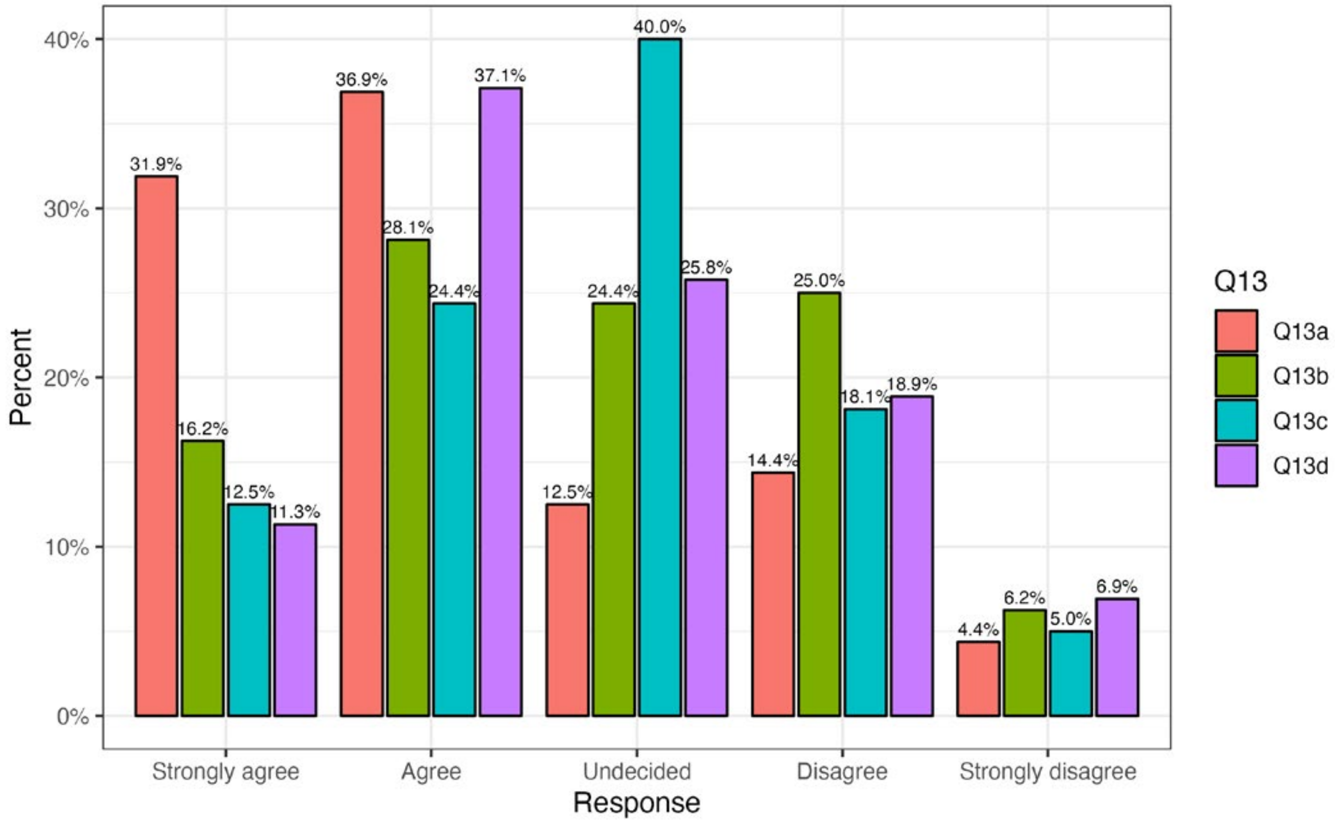


Figure 72. Combined results to question 13.

Licensing and publishing policies

Q15, N = 90 concerned the awareness of licensing and publishing policies for respondents publishing their own software. We provided the following definitions:

- A licensing policy provides instruction or guidance on what software licences are preferred for the release of software.
- A publishing policy provides instruction or guidance on how software should be made available to others.

In order to collect responses, we provided 'licensing' and 'publishing' checkboxes for the following categories:

- My institution's
- My funder's
- My project's

A summary of responses is provided in Fig. 73. In all cases other than institutional licensing and project licensing, fewer than 50% of respondents reported being aware of relevant policies. Awareness of licensing policy is greater than awareness of software publishing policy in relation to institutions and projects, but marginally lower in relation to funders. Awareness of both policies is the poorest in relation to funders.

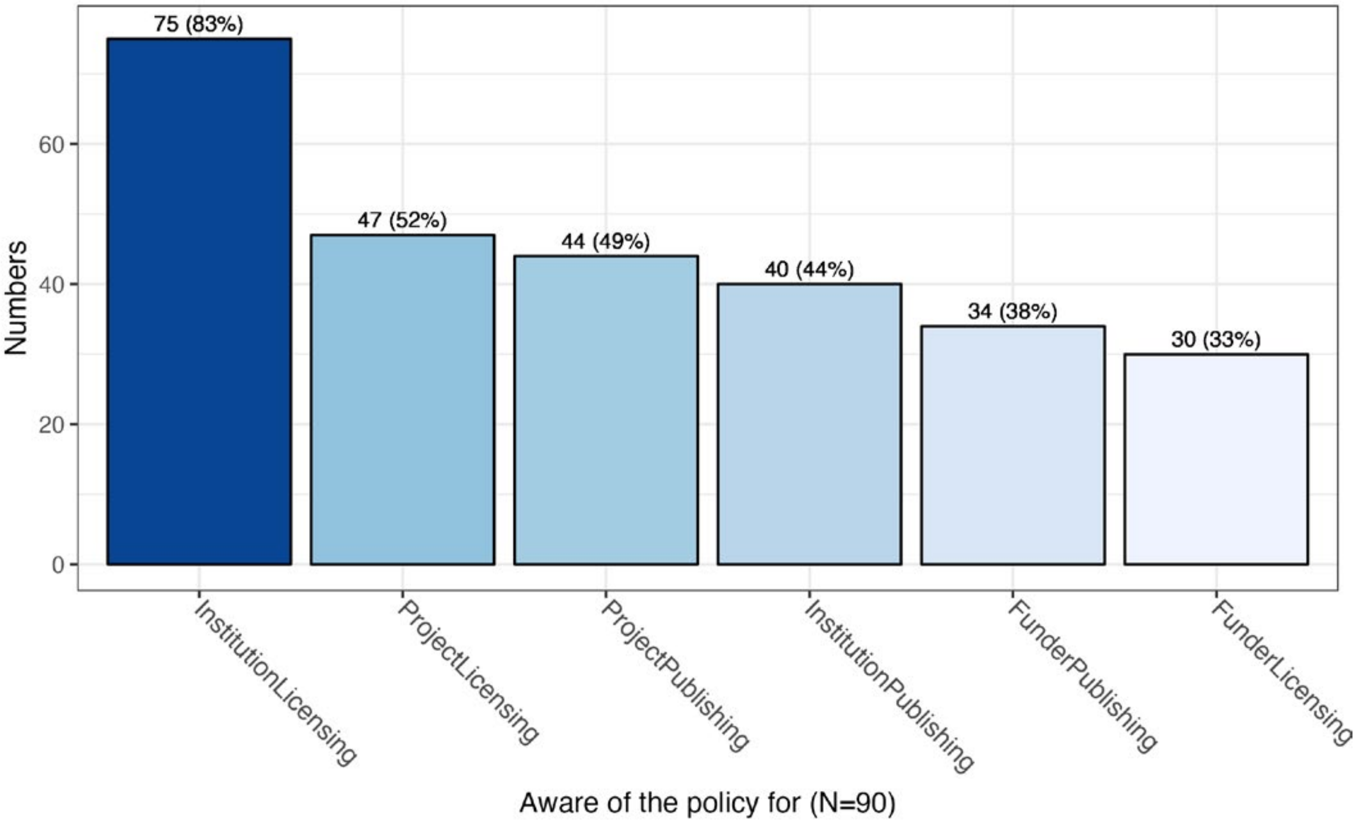


Figure 73. Awareness of software licensing and publishing policies at different organisation levels. Percentages are given in relation to the number of answers used for the survey analysis.

Fig. 74 presents the data by career stage. Senior career researchers show the highest level of awareness across all categories other than project publishing. However, institutional licensing is the only category in which more than 50% of this cohort report awareness. In terms of project licensing, ECRs show the highest level of awareness. Junior career researchers come second for institutional licensing awareness, but fare poorly across every other category excluding institutional publishing, where they come in third after senior career and ECRs. In most instances, early career and mid-career researchers are bracketed by their junior career and senior career counterparts.

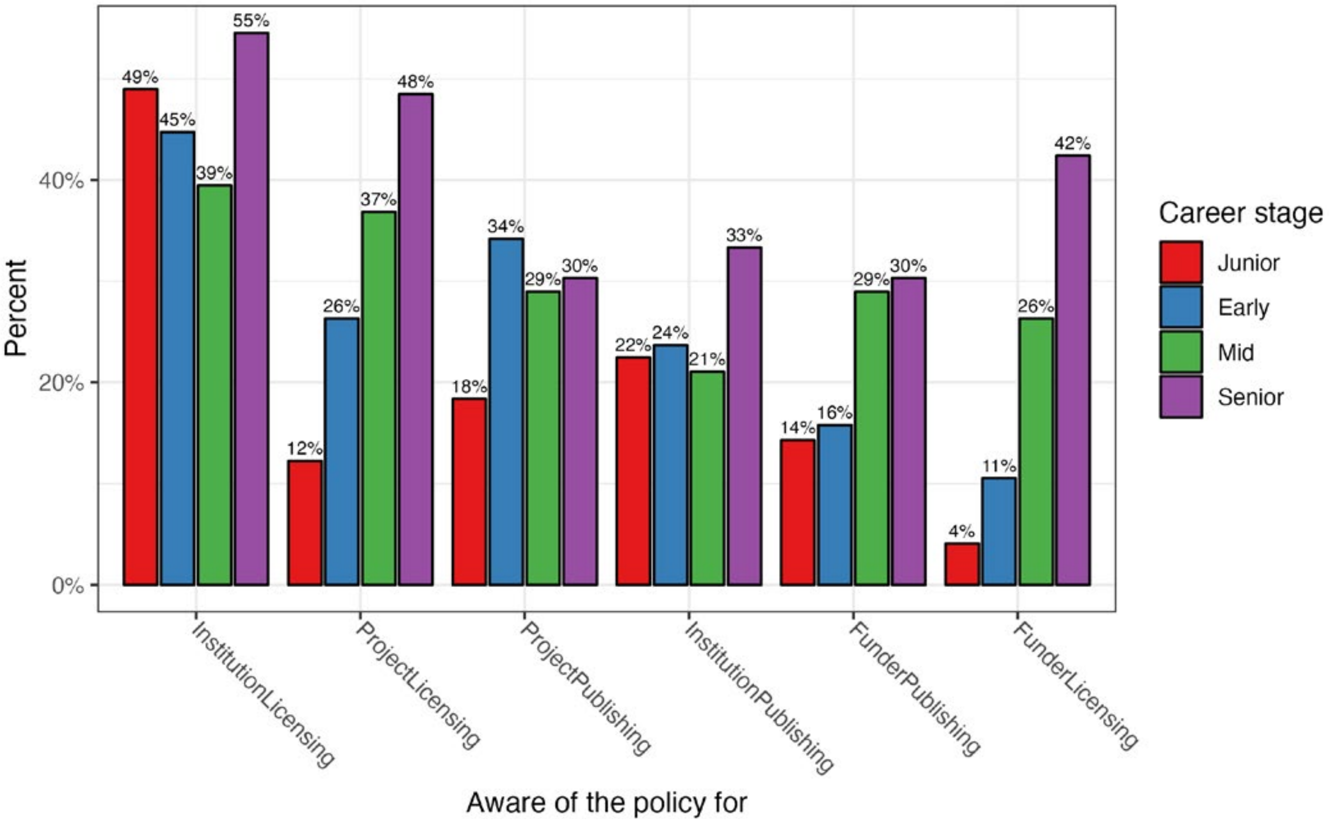


Figure 74. Awareness of software licensing and publishing policies at different organisational levels. Percentages are given in relation to each of the career stages.

Fig. 75 shows the data by career stage and gender. Men appear to report greater awareness than women at all career stages other than junior, where women dominate.

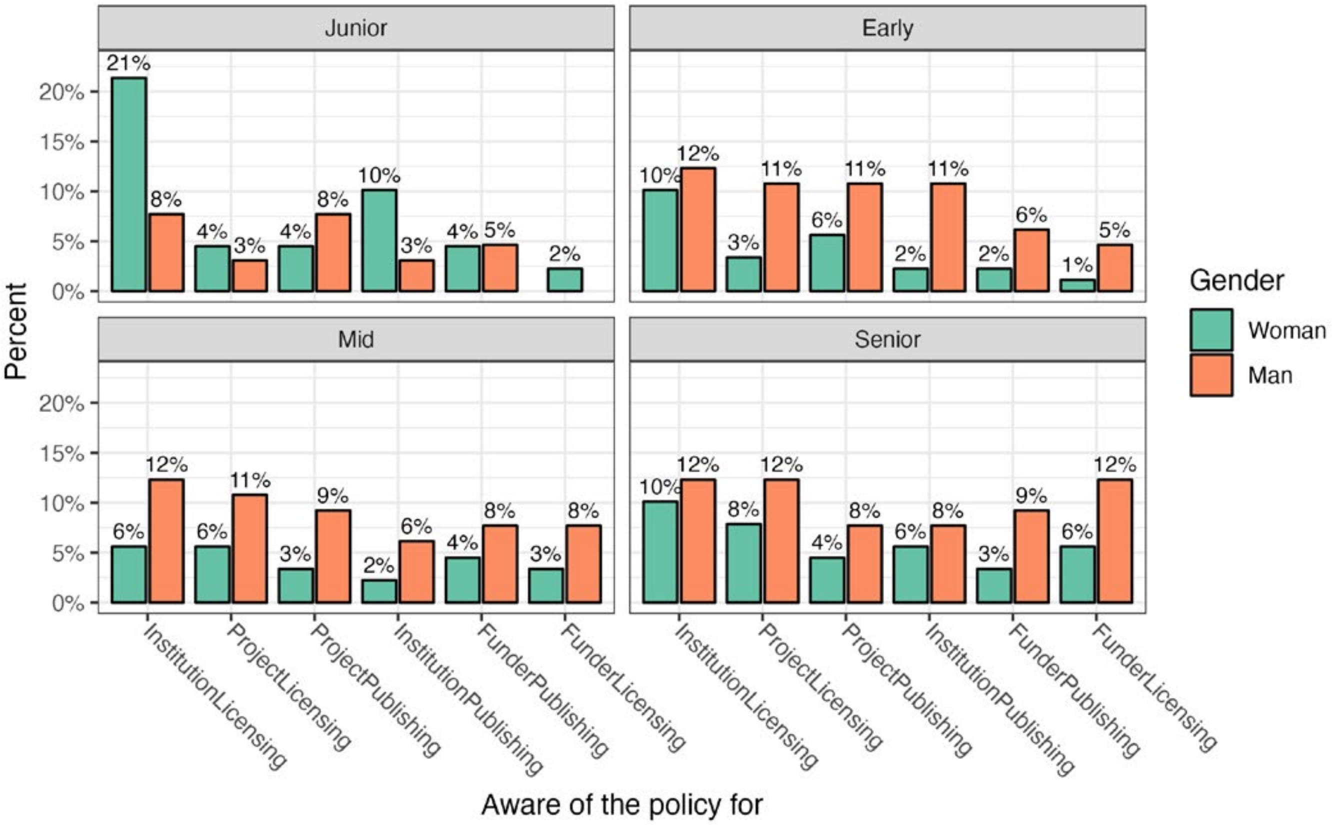


Figure 75. Awareness of software licensing and publishing policies at different organisational levels by gender and career stage. Percentages in this instance are concerning each gender.

We invited respondents to elaborate on their responses in a free text field. Their remarks illustrate the complexity of the issue owing to the range of stakeholders, interpretations, and concepts involved.

> “I'm aware of some details for licensing from my institution, but had to request specifically for more detailed information to get access to its full functionality.” [ECR]

> “All publishing guidelines that I am aware of talk about journals, books, conference proceedings, etc. These sometimes have data or code publishing options. My research group has our own policy (which is essentially put it on GitHub with good documentation). I am not aware of any publishing or licensing guidelines relating to software for my institution or funder.” [ECR]

Unsurprisingly, respondents were also guided by GDPR data management policies, along with concepts like copyright law and “open source product policies” [ECR].

Creative Commons (CC) licences were the most popular and well-understood. Those using CC licences were aware of the different types and their applications. For instance, one interviewee told us that for their actual code, they would use the CC-BY licence.

> “But a lot of the stuff, I would put CC zero, just because they expect people to mess about with it, and then not publish anything with it. So it just seems like I don't want to create a barrier to people messing about.” [7590937187]

Another acknowledged that while CC licences are useful for content, database rights differ across countries and it may make sense to use another kind of licence, which demonstrated a nuanced awareness of the situation.

In the next section, we discuss the skills and training necessary to use software in the economic and social sciences.

4.5 SKILLS AND TRAINING

We sought to ascertain how researchers are acquiring the necessary skills to utilise software in their research, and to establish what barriers they face. We presented our respondents (Q11, N = 158) with the options given below. Respondents could choose as many answers as were applicable to them.

- I am still trying to acquire the skills **(StillTrying)**
- Self-led online material **(OnlineCourses)**
- From examples and posts found online **(OnlineExamples)**
- Free community-led online workshops (e.g. Riot Science Club, ReproducibiliTea) **(OnlineWorkshops)**
- Conference workshops and tutorials **(ConferenceWorkshops)**
- Peers and colleagues **(PeersColleagues)**
- Undergraduate/masters course **(UG/MasterCourse)**
- Postgraduate training as part of my PhD/CDT/DTC **(PGTraining)**
- Institutional training course **(InstitutionalCourse)**
- National Centre for Research Methods (NCRM) training **(NCRM)**
- Third-party training course (not NCRM) **(ThirdPartyCourse)**
- Other **(Other)**

The responses are summarised in Fig. 76. We also present the answers by career stage in Fig. 77.

Our responses show that online courses (81%) are the most popular means of acquiring skills. This is most markedly the case for ECRs, but it is also true for other career stages. NCRM courses (18%) are the ninth most popular means of acquiring software skills overall, though their placement varies by career stage.

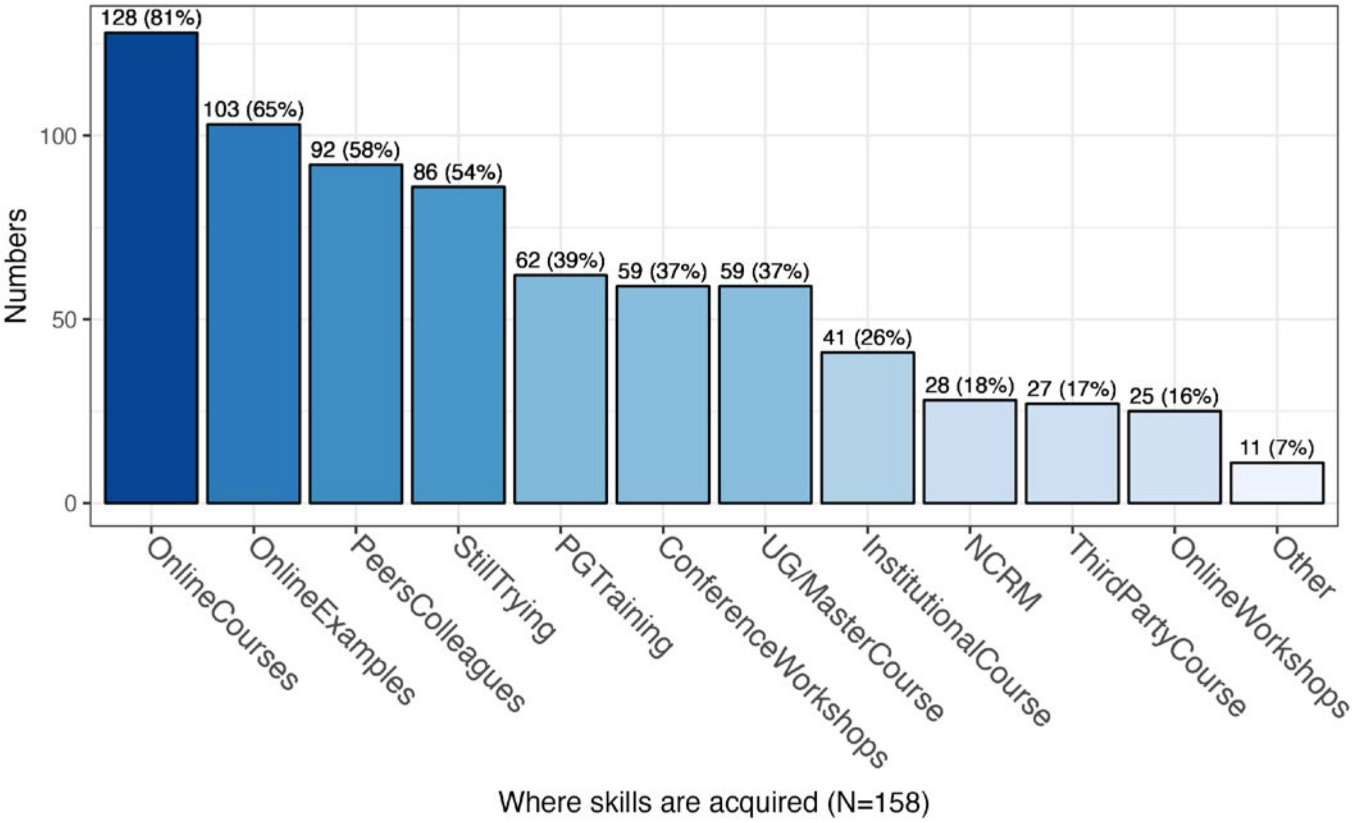


Figure 76. How the skills to utilise software are being acquired.

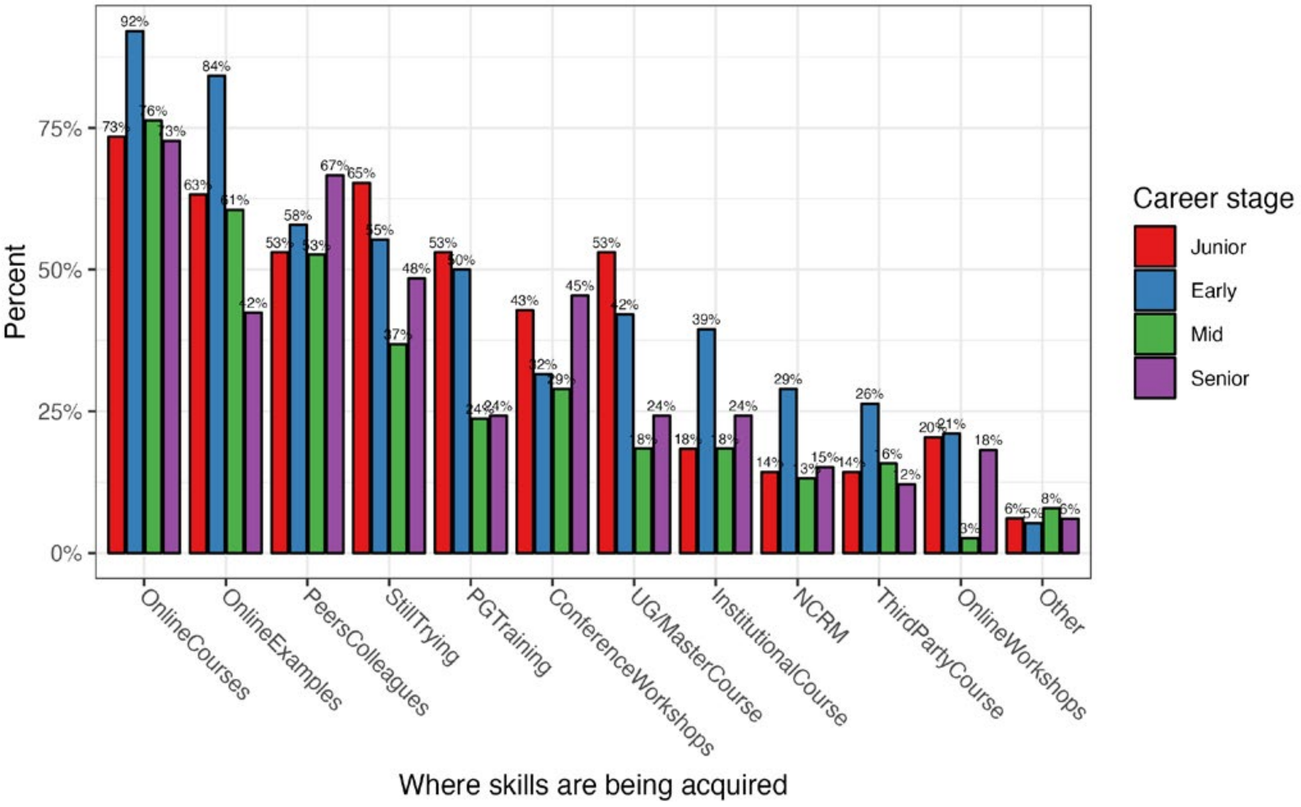
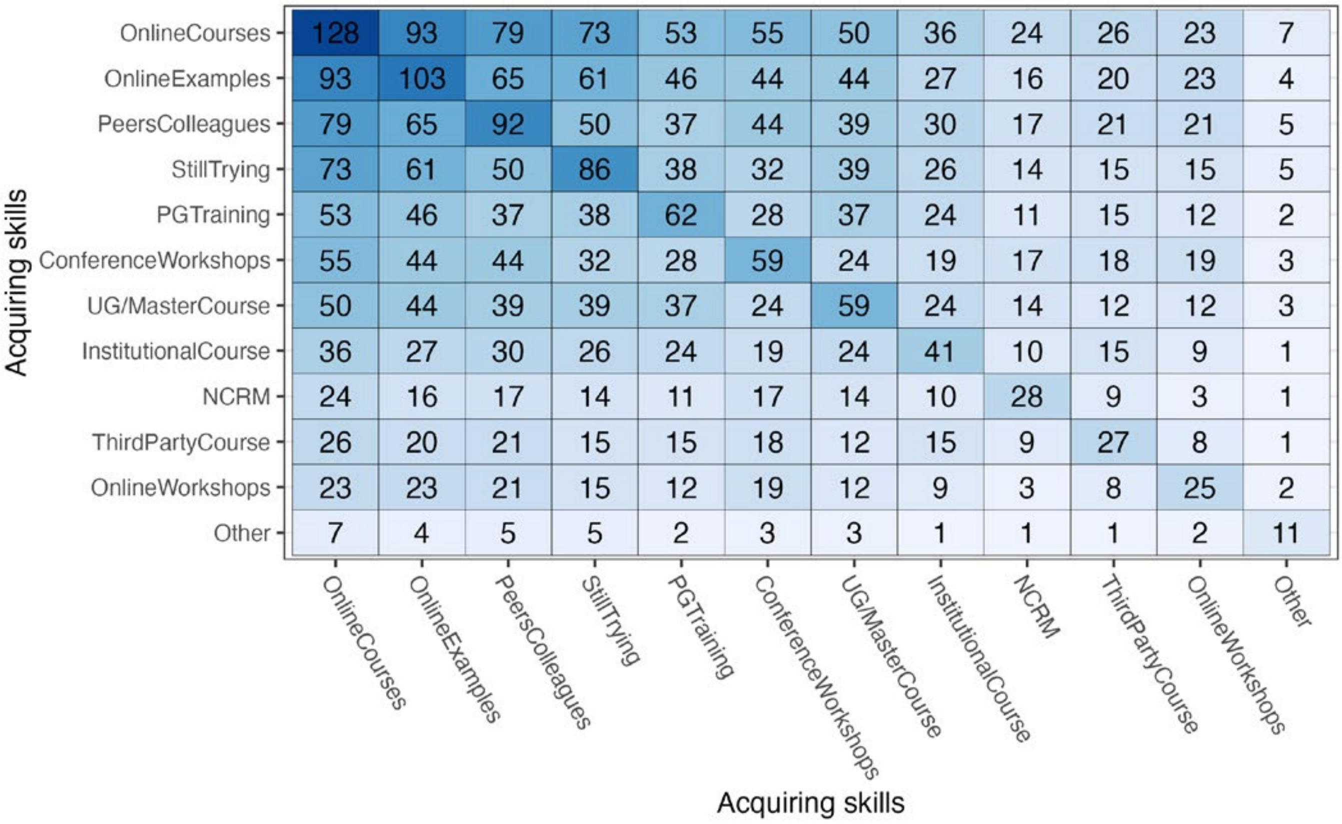


Figure 77. How the skills to utilise software are being acquired, by career stage.

It is not clear whether this enthusiasm for online learning is a consequence of the pandemic, with respondents having become accustomed to new ways of working, or whether it reflects a desire to pursue a more self-directed approach to learning.

The survey was carried out in March 2022, two years after training and research started to migrate into virtual environments. It is possible that the previous model of work will eventually return. It is also possible that the research community has experienced a permanent paradigm shift, embracing a hybrid virtual and physical approach towards research.

We can see how plural choices made by individual respondents correlate in Fig. 78. Online courses were often chosen alongside online examples and access to peers and colleagues. And there appears to be a second tier of methods for acquiring skills that includes post-graduate training, conference workshops, and undergraduate and master's courses.



Acquiring skills	OnlineCourses	OnlineExamples	PeersColleagues	StillTrying	PGTraining	ConferenceWorkshops	UG/MasterCourse	InstitutionalCourse	NCRM	ThirdPartyCourse	OnlineWorkshops	Other
OnlineCourses	128	93	79	73	53	55	50	36	24	26	23	7
OnlineExamples	93	103	65	61	46	44	44	27	16	20	23	4
PeersColleagues	79	65	92	50	37	44	39	30	17	21	21	5
StillTrying	73	61	50	86	38	32	39	26	14	15	15	5
PGTraining	53	46	37	38	62	28	37	24	11	15	12	2
ConferenceWorkshops	55	44	44	32	28	59	24	19	17	18	19	3
UG/MasterCourse	50	44	39	39	37	24	59	24	14	12	12	3
InstitutionalCourse	36	27	30	26	24	19	24	41	10	15	9	1
NCRM	24	16	17	14	11	17	14	10	28	9	3	1
ThirdPartyCourse	26	20	21	15	15	18	12	15	9	27	8	1
OnlineWorkshops	23	23	21	15	12	19	12	9	3	8	25	2
Other	7	4	5	5	2	3	3	1	1	1	2	11

Figure 78. Contingency table on how skills to utilise software are being acquired.

For the Other category, which allowed respondents to describe possibilities not covered by the examples, we provided a blank text box. Responses included: trial and error, working collaboratively with PhD students, self-taught (including not using online methods), banging head against a wall, and books.

Interview data appeared to show that the biggest barriers to the acquisition of skills are time and cost. These were not linear: the time barrier was perceived to worsen in mid-career and the cost barrier was greatest amongst ECRs.

> "[Training] can cost several 100 pounds, sometimes perhaps at some place different, you also have to pay for travel and accommodation to go there and it's a three-night [or] longer course." [7183719943]

However, responses varied greatly. Some researchers had to consider whether to spend their budget on training or other activities:

> "[Do you spend it] on conferences or do you spend it for example on some other practical things that you might need for your data analysis or collection" [7183719943]

Other researchers (especially JCRs) had ringfenced training budgets.

> "I do know that the funding we have does give us access to any kind of training, we do the research development plan or training needs analysis every year. So that kind of helps us identify what we need. And if there is something that is going to be necessary, then we do have funding available for it." [2964377624]

As shown in Fig. 76, online resources were popular. Provision was seen as strong.

> "There's tonnes more online resources now. So I think it's UCLA in the States have a massive online repository now of online tutorials, basically in multiple different software. So they've got it in Python, [they've] got it in SPSS, they've got it in STATA. [They] basically walk you through how to do all them." [8031500357]

DataCamp[78] was frequently cited as a useful resource, as were the Data Carpentries provided by the SSI.

Online resources were considered particularly useful in cases where researchers had a degree of context or some existing knowledge.

> "I've taken on an online course for six weeks on Introduction to Python for data analysis. And it was pretty similar to R to me, so it was also good for my R skills." [7869268779]

[78] https://www.datacamp.com (last accessed August 31 2022)

However, the disadvantages of online courses were shown to be their lack of interactivity and the lack of support available whenever students encountered problems.

Self-led learning (in the form of online examples and learning from peers and colleagues) constituted the second and third most popular options in the survey, and was widely deemed to be effective by our respondents. This method is free, time-efficient (it can be accessed when needed), and suitable for general training as well as for addressing specific issues. Googling, user community content, and YouTube were all cited.

> "It's mostly kind of self-taught ... YouTube so if it's something that is completely new to me, like a completely new area that I've never done before, if it's kind of smaller things that are you have a little bit of a background but need a bit of help, then it tends to be just random websites, whatever pops up on Google." [2964377624]

However, barriers to self-led learning include a lack of initial confidence in understanding the content and an inability to define learning goals. In addition, self-led learning was deemed more appropriate for open source tools than proprietary ones, for which training materials are commercially controlled and the community is less active.

> "Whenever I need to do something in R that I've never done before, I just try to google it and see how complicated it is, if there are any tutorials out there. And usually there are because it's open source. So the community is fairly good in sharing resources." [2964377624]

The survey also shows that women may be less likely to adopt self-led learning approaches than their male counterparts, even accounting for career stage.

In some cases, researchers are aware that software might be useful, but have no idea how to access it.

> "I think there are times when I don't know where else to go. And those might be for things that are a bit foreign to my discipline." [7183719943]

As a respondent who delivered training told us,

> "we do get people, sometimes asking very basic questions, but they're not necessarily asking me to do it. They're kind of just asking me like, I'm out to sea here. How do I get started?" [7590937187]

In these cases, researchers have insufficient knowledge to access self-led learning.

Training in the 'traditional' proprietary softwares of social science, such as NVivo and SPSS, or STATA for economics, was perceived to be well-established. Where universities invested in licences for software, they also invested in training.

> "These are tools that are bought in by the university. NVivo is an example of this [...] through the staff training and post-grad training, there are courses on how to use [NVivo]. So I think universities can be quite good at training for those sorts of bespoke software packages, provided they have the licence for them." [4561769548]

Furthermore, using software provided by an institution typically means enjoying access to institutional support.

> "It's the university's formal support for certain tools and how they offer what they offer ... in terms of licences, but also in terms of training. [They have] already paid for ... licences for, for NVivo and for SPSS, on the one hand, which incentivises installing it on your machine, because it's supported in theory, so you can ... ask them questions. Whereas if you're installing some weird R package, no one's gonna help you from university." [5766299900]

This remark is interesting in light of Fig. 29, which shows that R was twice as likely to be crucial for research as SPSS. Despite its reported efficacy, R requires a degree of basic knowledge. As a result, respondents suggested it was sometimes hard to move away from traditional tools. As well as institutional barriers, summarised above, and a lack of alternatives for thematic analysis, which we discuss in our analysis of open source, there were a number of training-specific barriers. A significant proportion of teaching and training uses commercial packages like SPSS and Stata. Junior career researchers may want to use R, but do not have adequate support from their supervisors.

> "I think the majority of researchers in my discipline still use SPSS. And that in itself is a barrier, right? Because if everyone is not trying to build up skills in R, it's really hard to justify building up skills in R for yourself." [7183719943]

This is a particular issue in cases where team members or supervisors are not experienced in a specific, non-disciplinary tool.

> "My supervisors are extremely supportive of my decision to [use R]. But they did warn me that they themselves did not possess the expertise to help me so it was very clear, they said, 'We understand where you're coming from, we understand your reasons for it, and we support you, in principle, but we cannot give you any practical support, because we don't use these tools ourselves.'" [5766299900]

In other words, because training in academia involves learning the tools and methods employed by supervisors, the acquisition and application of new technologies is curtailed. One respondent described this as,

> "path dependence...Because the lecturers and professors have been trained with using certain tools, they're in a position where they can train their students with those tools. And then unless you actively think about, is this what I want to do

> it? Does this fit in with my worldview and my ethics and whatever, you might not even question the tools that ... you are using, as a result of which, it's just that path dependence. This is what my supervisors use, this is what I end up using." [5766299900]

Related to this was the idea that training is generally focused on a specific project or discipline, and therefore developing a broader, more future-oriented set of skills is not prioritised.

> "I would like to see training available that doesn't directly relate to someone's immediate project. So for example, as a PhD student, I found that the training that I took related to, you know, using SPSS or NVivo, because those were the tools I was going to be using. And then as I said, that made it really hard to take these courses on R, especially to a level where I could actually have enough competence to use R." [7183719943]

Acquiring these 'future-oriented skills' was a major priority for our respondents.

> "The opportunity to say, you know, in the long term, I will want to also have these skills for future projects. How can I get those skills now? And so it would be great if there were opportunities to build on skills that you might want in the future, even if you don't need them right now." [7183719943]

The main barriers to achieving this were lack of time and lack of courses that deliver enough knowledge to be useful. The time barrier became harder to overcome as our respondents advanced into mid-career. Acquiring future-relevant skills during a PhD was seen as a way of mitigating this issue in advance of assuming academic roles with administrative, teaching, and research responsibilities.

> "I would love to be able to enrol in a class that would be able to show me how to use R, and I've enrolled in two short courses. But they're very short. So you know, this kind of gives you the introduction and not enough skills to be able to do it yourself. With R there's a lot of material online, so you could teach yourself how to do it. But it's just a matter of setting aside the time to actually do it. And because it's not central to my own research, at least at the moment, I've not really been able to justify setting aside enough time to teach myself how to use it, and then actually using it." [7183719943]

Our respondents did provide examples of either institutions or individuals driving the acquisition of future-relevant skills, but these were not consistent or appealing enough to reinforce best practices across the board.

One serious objection to the notion of future-oriented training is the 'riding a bike' problem. While the ability to ride a bike can famously never be forgotten, the ability to use a given piece of software requires constant use or it will atrophy.

> "You know, you don't just take a one-snap, photo, you need to keep going back and saying, What do you want? And then actually doing something about it." [2964377624]

There was a perception among our respondents that UK-based social sciences training is more qualitatively oriented than social sciences training in, for instance, the US.

> "The US is much more quantitative in its approach...in training social scientists in the use of data manipulation, large scale data analysis, quantitative methods, even using things like SPSS and tools like that at a more advanced level for advanced quantitative analysis involves writing algorithms and code." [8031500357]

This approach confers the confidence to use other code and algorithm-based tools. Without quantitative training, researchers face serious barriers. One ECR who wanted to begin working with R told us,

> "I started looking at tutorials online, I started installing packages that came with tutorials to see if I could use them. It was a major stumbling block for me because I've never really programmed before in my life. I've never written any code for any, I mean, very basic code in undergrad as part of our computer science modules, but that's like from a lifetime ago. So it was a little challenging for me to not even know where to start. Some of the terminologies didn't make a lot of sense I did not know. For so long I used functional programming language without really using functions in them because I just didn't understand where to start." [5766299900]

A more balanced approach to quantitative and qualitative research might equip researchers with a knowledge base from which to build vital software and data manipulation skills.

One researcher also cited inconsistent hardware amongst JCRs and ECRs as a barrier to training (and, by extension, the use of software).

> "We find that's a real hurdle when you're trying to train people on things ... getting them to instal the right version, and then we don't know how powerful the machines are. So ... people will technically have it installed, but it just won't run very well." [7590937187]

This aligns with the survey finding that most junior career researchers are using their own laptops, which necessarily vary in terms of specifications. Throughout our research, we repeatedly heard about the value of general software and data management skills. As one respondent told us,

> "I think it would be good if we worked out what the basic concepts of data management were, and taught those before we just had loads and loads of bespoke training on very specific tools, because I think sometimes that gets missed.

> And that includes all the stuff around ethics and GDPR and all of those things, because it doesn't matter what tool you're using, you should understand that stuff." [4561769548]

Learning generalised skills was also seen to mitigate the issues created by the movement of ECRs between institutions.

> "[S]ome of the training that we used to do on the MSc... tries to give people the sort of underlying philosophy of research methods and using data and managing data before you jump straight into the tools. Because the danger is if you just teach people the bespoke way of using specific tools, they go to a different institution where they use a different set of tools. And they get really confused." [4561769548]

Acquiring this level of competence may also open doors for effective self-led learning, as explored above.

There was a range of views on whether JCRs and ECRs have an appropriate generalised skills base. These were the survey cohorts most likely to report that they were attempting to acquire software skills, but more senior researchers felt that the ECRs they worked with had an appropriate level of software literacy.

> "[T]hese new researchers are coming in fully immersed in this environment. So they're well used to, for example, signing up for an API account, because they need to be able to connect one app to another or an app to their PC, they're quite savvy with it. So they understand the terminology." [8031500357]

On the other hand, JCRs in particular appeared less confident in their skills, which may be a reflection of the fact that the junior career researchers who gain skills and exhibit confidence in them tend to work with more senior researchers on larger projects as ECRs.

On a couple of occasions when interviewing JCRs we found ourselves suggesting ways in which relatively accessible, non-specific (facilitating) technology could be used to make data collection and analysis more efficient. These included using transcription software, basic web scraping, and scanning text using a simple mobile phone function rather than transcribing it manually. (Interestingly, Fig. 30 shows that many different types of transcription software are used in interview workflows, whereas Fig. 23 shows that respondents do not consider it to be important.) As we heard from one respondent who was typing in text,

> 'It sounds brilliant. But ... it's two extra things to learn, isn't it?... I'll just type in text and that's fine. Because that's the level of wisdom you have." [5996360861]

This illustrates the dual problem of not knowing that software is available to address research issues, and not having the time to acquire new skills during the research process rather than in advance.

NCRM courses

The ESRC is one of the councils that invests most in methods training, primarily via the NCRM. We asked respondents whether they had taken NCRM courses (Q11b, N = 31). A checklist with a sample of 22 current software-related NCRM courses (as of early 2022) was provided in the survey

Of the 164 individuals in the survey, 31 had taken a total of 44 NCRM courses. If we include the Other category, where NCRM courses not listed in our survey were given, then the number of NCRM courses taken rises to at least 55, which means that 18.9% of respondents had taken an NCRM course in the past.

In Fig. 79 we see the total number of NCRM courses taken by individuals at each career stage. For each career stage the number of courses is on the left and the number of individuals is on the right, in brackets.

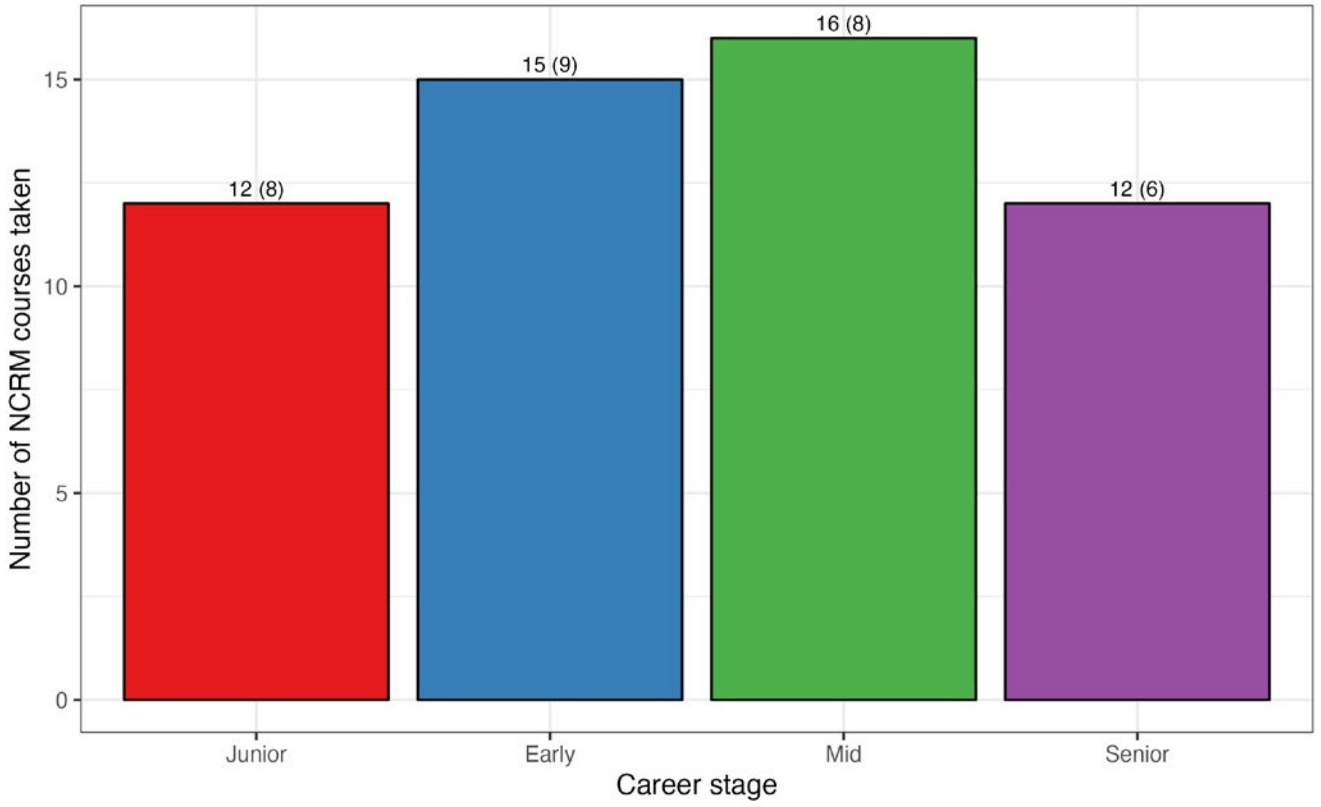


Figure 79. The total number of NCRM courses taken by respondents at each career stage. The number in brackets refers to the number of individuals involved.

Fig. 80 shows the number of NCRM courses respondents at each career stage have taken:

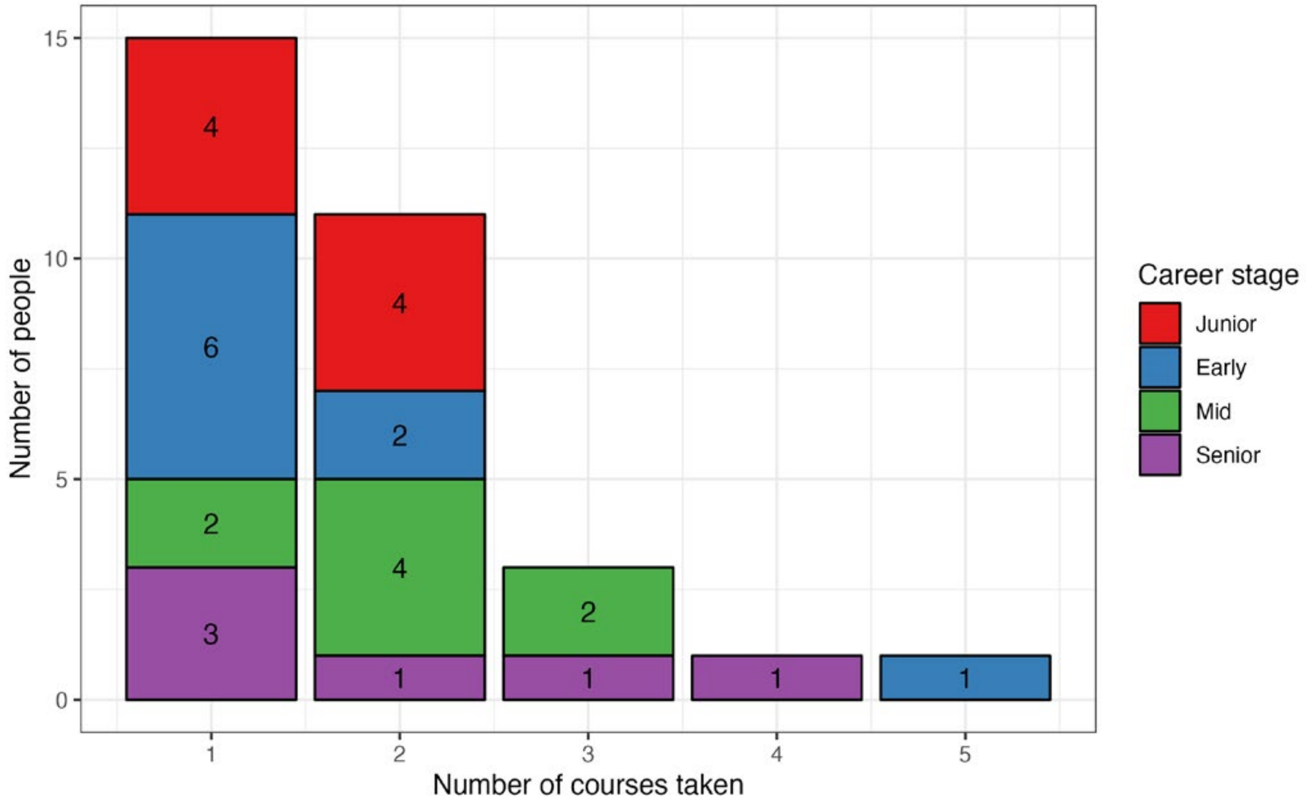


Figure 80. The number of NCRM courses respondents at each career stage have taken.

4.6 COLLABORATION AND INTERDISCIPLINARITY

Collaboration reduces the necessity for domain experts to also be software experts, but it requires its own form of expertise. Those who set out to collaborate on research projects need to learn how to communicate effectively so as to ensure that they understand what stakeholders with varying skills can bring to the project.

> “It varies project by project. But for this specific example ... the team is actually half and half, so the economists have expert domain, they know what they're looking for, they have better intuition [as to] what's relevant from the economics perspective. And we have a group of statisticians that really know that the whole networks, the graphs, statistics, they are able to build much more sophisticated models and techniques, and they better understand, you know, what's possible, and they just talk to each other. I think it's very important to collaborate. And for this collaboration to be possible, we need incentives, right.” [1578450175]

As the previous interviewee notes, incentives are vital for collaboration. There is currently a perception that research councils and the REF are not doing enough to encourage interdisciplinarity. Many respondents, particularly those developing software for particular disciplines, felt they had fallen into a funding gap.

> “There's EPSRC and ESRC. And the councils do not necessarily work together. So, you know, where do people apply for some funding?” [1578450175]

> “So in the UK, if I feel like we're quite siloed, maybe it's because of the REF process. Whereas in the US, if you're a computer scientist, and you write a paper that's got really interesting social insights, that's brilliant, because now we're, you know, really understanding society.” [8031500357]

One of the side effects of remaining within a ‘silo’ is a diminished awareness of the need to develop new software skills, and a reduced ability to access them, as noted in Section 4.6.

> “I think the majority of researchers in my discipline still use SPSS. And that in itself is a barrier, right? Because if everyone is not trying to build up skills in R, it's really hard to justify building up skills in R for yourself.” [7183719943]

> “I, you know, when I want to talk to other people in the department that have done it, there aren't as many yet who are experienced in it talk to.” [1269794877]

Interviewees (at senior and junior levels) felt it would be useful to receive training for interdisciplinary work, and to create models of how it can work early on in careers.

> “I would say number one is providing the kind of training early on in postgraduate careers that would help people working in multidisciplinary groups and developing some ideal sort of model protocols for doing some of the things that we've talked about. And I think that'd be very helpful in terms of planning and de-risking. I think they'd certainly have been helpful to me, I kind of wish I turned on some of this stuff a bit earlier.” [3270614]

Overall, our respondents welcomed more opportunities for sharing practice, especially in light of the fact that the ESRC already facilitates sharing and promotes epistemological methods through the NCRM and other mechanisms.

> “I think that there could be a lot more sharing ... practice, I mean, people develop their own ways of doing things. Methods is a thing in the ESRC. And we do a lot more about sharing the methods for this. So I think that that's an important point.” [100530]

In terms of infrastructure, not every software package lends itself to collaborative work. This is related not only to licensing and skills, but also to the fact that the software in question may be designed for the lone user.

> “I also think that NVivo is really difficult if ... you have multiple users who are going in and then coding it. I think you have to pay for sort of the advanced package, which I do not have access to. So it's really hard to share an NVivo set between multiple users. And so on big projects, I think I think there's a lot of limitations in terms of qualitative data analysis with big packages, or with big teams and big projects.” [7183719943]

One thing to consider in multidisciplinary teams is who is responsible for the long-term sustainability of the software that has been developed.

> “I think it is an issue. Who’s responsible for the software tool? Is it the engineers? Or is it the people that want to use it? There’s nowhere near enough resources put into this.” [4561769548]

When our respondents were asked how they found other researchers with whom to collaborate, it was very clear that face-to-face interactions had a big role to play, and that the move to more online engagement had created a challenge for collaborative working.

> “I've just had chance meetings, like at conferences or events [...] And then sometimes it's people that I've worked with indirectly, as in, you know, they are a couple of teams over but we just got to chatting over the course of months of working together infrequently.” [7590937187]

> “Obviously, we've all had, you know, we've all had online opportunities to meet up. But you don't, in an online opportunity,

> have the chance to just have little side chats and little conversations with people about what they're up to and what they might have seen or experienced." [5996360861]

It should be noted that working with data that is personally identifiable (or has the potential to be personally identifiable) can be a challenge for international collaboration between institutions.

> "Certainly, collaborating with colleagues in Europe is not at all unusual. In my field, that raises all sorts of issues about data security, and different interpretations of GDPR. I'm sure it's a good thing, you know that we have that sort of protection. But unfortunately, it is open to interpretation. And different countries interpret it in different ways. And that has created a massive headache, actually." [7336263536]

In the next section, we discuss some of the main points arising from our findings.

5. CONCLUSION

Throughout this research, a picture has emerged of the ways in which the drivers identified in Section 2.1 have affected, and continue to affect, social sciences research.

Until recently, researchers in the social sciences have been working with comparatively limited varieties and quantities of data. The software and methods used for gathering and analysing these data have tended to vary by discipline, and have been handed down through the traditional apprenticeship model of doctoral training, which sees supervisors transmitting knowledge and skills to doctoral researchers. The outputs of the research have almost exclusively taken the form of publications.

In contemporary social sciences research, there is both a greater volume and a greater variety of data, even in small scale projects. Likewise, there is greater variation in the outputs of research and the ways in which they are published.

New software tools are required to deal with the growing abundance of data within each discipline. This typically means employing not one tool but many, pulled together into a single workflow. To meet this challenge, there are new software tools that embody a wide range of research methods, especially in the realm of machine learning. Researchers must not only identify the tools they need and acquire them (which can involve overcoming technical and financial barriers), but they also need to understand how to use them effectively in social sciences research. Importantly, not all of these tools were built with scholarly research in mind.

These factors have combined to create a steep learning curve for researchers, exacerbated by a lack of what Milliken et al. (2021) describe as a "mental model" for understanding software in terms of its infrastructure and utility [13]. This frequently leads to sub-optimal use of tools.

The difficulty of navigating this new environment is a major issue facing social sciences research. The key questions are: what technical grounding do researchers need in order to deal with this explosion of data and tools? And are they receiving this grounding? Strong investments in training, particularly around machine learning (through the Alan Turing Institute[79]), R (through the NCRM), and data (through planned data training[80]) are evident, but there remains a widespread lack of fundamental software literacy among academics in the social sciences. Researchers in this field frequently lack an understanding of basic computer science and programming concepts, and have no pathway for obtaining this knowledge, or face too many barriers to acquiring the software skills they need.

In the next section, we present the key findings from our results and provide recommendations to institutions and funders of social sciences research.

[79] https://www.turing.ac.uk/ (last accessed December 5 2022)

[80] https://www.nuffieldfoundation.org/students-teachers/q-step (last accessed December 5 2022)

6. KEY FINDINGS AND RECOMMENDATIONS

6.1 RESEARCH DATA PRACTICES

Key findings

Use of data

Survey (22%) and interview (19%) data seem to be the dominant types [Q2b]. Behind them is a long tail of data sources that includes APIs, behavioural data, social media, human participants, new data, and questionnaires.

Most interviewees are using the UK Data Service, commercially held data, or survey/interview data. None of our respondents reported using other datasets providers funded by the ESRC.

The split by career stage shows that mid-career (31%) and senior career (32%) researchers predominantly employ surveys, while junior career researchers prefer interview data (25%) over survey data (11%).

Data reuse/creation of data

53% of our survey respondents indicated that they both create and reuse data [Q2]. This is significantly greater than the proportion who only create (28%) and only reuse (19%) datasets. Junior researchers seem disproportionately likely to only create data (47%). This might be due to their preference for using interview data, as previously shown [Q2b].

The main source of reused data [Q2b] is the UK Data Service (58% of respondents). This points toward the fact that the UK Data Service is the UK's largest collection of economic, social, and population data for research and teaching. The second most common source is institutional repositories (36%). Other data sources used include government data sources (6%) and the ONS and ONS-SRS (Secure Research Services) (5%).

Our survey recorded 55 suggestions accounting for <1% of responses. Of these, only CLOSER is an ESRC-funded data service. Interview participants cited the Urban Big Data Centre and commercial data as significant sources.

ESRC Research Data Policy encourages researchers to reuse data where possible. However, the significant time investment involved remains a barrier. Other barriers to reusing data include having to create synthetic replicas of commercially sensitive or protected datasets (to allow collaborators to develop and test their software), secondment of employees, and the necessity of assigning project members to the task of navigating secure data access protocols.

Sharing data

Our survey found that data tend to be either not shared (49%) or shared in a repository (36%) [Q3]. Furthermore, 75% of those who only create data do not share it, and only 34% of respondents funded by the ESRC in the last five years do so, even though ESRC data policy stipulates that data be deposited. In short, the majority of researchers who responded to the survey and were funded by the ESRC did not adhere to ESRC policy.

73% of senior career researchers deposit their data, whereas 82% of junior career researchers do not, perhaps because they intend to do so once they have completed their PhDs. Many mid-career researchers appear similarly disinclined to share their data (53%), however, so there may simply be a general lack of incentive to comply with ESRC policy.

Our interviewees reported that some institutional repositories mandate the use of particular infrastructures, such as SharePoint and OneDrive, affecting sharing and storage. Short-term junior and early career contracts and differing policies for depositing and sharing data across institutions are also potential barriers.

As a result, there is a 'grey infrastructure' of research storage across the UK. Data is routinely stored on commercial platforms like Google Drive or Dropbox, with all the risks this practice entails.

Our interviewees also cited confusion over policies on depositing third-party reused data and data management plans, a lack of awareness of appropriate repositories, a lack of incentive (REF and other recognition systems reward publications, rather than the depositing of data), and the fact that policies around data ethics seem to be incompatible with the principles of data sustainability, open research, and reproducibility, since they often require that data be destroyed after a set period of time.

Recommendations

1. The ESRC should commission a study to understand the working relationship between researchers and data services, explore barriers to the reuse and sharing of data, and identify strategies for incentivising good practice.
2. The ESRC Future Data Services strategy should address the lack of incentives to adhere to funder policies and guidance, e.g. around the depositing and sharing of data and open data.
3. The ESRC should mandate the inclusion of a metadata field when data is deposited that identifies the software used to generate/analyse that dataset.

6.2 RESEARCH SOFTWARE PRACTICES

Key findings

Use of software/tools

Statistical analysis (89%) and spreadsheets (85%) are the most used software types reported by survey respondents [Q4]. This corresponds with the wide use of quantitative (survey) data in the social sciences research community shown in [Q2b].

The third most commonly used type of software is "qualitative data analysis tools" (52%), which corresponds with the second most used type of data: interviews. Interviewees cited NVivo most frequently. [Q5].

The use of certain tools highlights the importance of open source software in the social sciences research community: The reported use of R (36%) is almost double that of SPSS (19%) and Stata (16%) [Q5]. Around two-thirds of respondents (67%), a significant majority, use source tools [Q6], but, as explored in Section 4.5, instruction in the use of these tools isn't always available.

Our survey respondents rely on a long tail of around 130 other software tools, each one employed by a small subset of researchers. This makes it harder to develop policies around the maintenance, development, and sharing of software.

Barriers to software use [Q16] include:

- Lack of expertise (56%), both in terms of individual researchers and departments
- Lack of time to learn how to use research software (50%) and lack of capacity to engage in training (36%). Junior researchers in particular felt they did not have enough time to engage with research software, which relates to the steep learning curve and lack of institutional or departmental support explored in Section 4.5.
- Lack of training options (35%), in general and for specific skill levels.
- Not knowing how to engage with specialists (30%), not just in terms of who to work with, but also how to work with them.

From interviews, we found the following barriers: lack of understanding of potential value, expertise, expense, the speed of software change, pressure on capacity, and an overall perceived propensity for qualitative rather than quantitative research in the UK.

Among those reporting a disability, lack of expertise (68%) was selected as the main barrier to using research software. Lack of time (36%) and capacity (36%) were perceived to be less important than lack of training (56%), which is the second biggest barrier. This may indicate that suitable, accessible training is not sufficiently available.

Software workflows

Around 64% of survey respondents reported using software in combination in their research processes [Q8]. Different pieces of software are used in specific workflows or toolchains. We have also identified two main patterns: the interview pattern and the survey pattern. These correlate with the main types of data used in social sciences research.

However, no single piece of software was universally employed for any specific stage of an identified workflow, except for NVivo, which dominated qualitative processes. More research on this is needed to understand additional common workflows used by social sciences researchers, and the software tooling these require.

Views on software policies

Mid-career, early career, and senior career researchers reported feeling sufficiently well-informed on software policy, with 74% or more agreeing that they were aware of how software is funded, managed, and licensed. However, junior career researchers appeared less sure, with roughly 47% agreeing that they were aware how software is funded, managed, and licensed, and about 35% disagreeing [Q13a].

Junior researchers were also more likely to admit to not knowing their institution's or funder's software policies, making them less likely to be compliant [Q13b].

37% of respondents agreed that insufficient attention is paid to software funding, management, and licensing by the economic and social sciences research community, while 23% disagreed [Q13c].

Roughly 48% of respondents felt there is not enough of an incentive to learn how software is funded, managed, and licensed while about 26% were undecided and 26% disagreed with the statement [Q13d].

Open source software

Two-thirds of survey respondents reported using open source software. Further investigation from both the survey [Q6a] and interview questions showed that researchers value open source software for meeting their needs, cost, sustainability, and interoperability, which enables cross-institutional work and collaboration.

Barriers to using open source software include:

- Lack of skills and expertise (61%) [Q6b], with further analysis of interview data showing a steep learning curve for acquiring open source software skills.
- Open source software does not meet researchers' needs (40%).

Just 24% of senior researchers and 18% of early career researchers are reluctant to use open source software due to lack of expertise, while only 38% of junior and mid-career researchers feel that open source software does not meet their needs.

Developing research software

Most survey participants (59%) did not report developing software [Q9]. Among those who did [Q9a], we found that most appear to be using R (including R markdown and Shiny) (41%) and Python (17%). It is important to note that respondents who develop research software are sharing it widely (46%) [Q10]. Both the development and sharing of software correspond to the widespread use of open source software (R and Python) among the social sciences community.

Other software being developed includes Excel functions, macros, and the production of charts (16%), Stata (11%), and SPSS (5%).

Licensing & publishing

Fewer than 50% of survey respondents reported being aware of publishing or licensing policies held by their institution, funder, or project [Q15]. A broad range of concepts and policies seems to exist across institutions, departments, and publications.

Respondents appear to be guided by GDPR and data management policies, as well as general policies such as copyright law. Creative Commons licences were the most popular and well-understood types of licence among our interviewees.

Recognition of research software development and maintenance

Many respondents (42%) felt that the development and maintenance of research software is not sufficiently rewarded or recognised [Q12].

In particular, early career researchers (62%) felt that software development and/or maintenance did not contribute to career progression, while junior researchers (61%) were particularly unclear as to whether software development and/or maintenance is rewarded. [Q12]

This is notable in light of the reported awareness of how software is funded, managed, and licensed. 37% of respondents agree (vs 23% who disagree) that there is insufficient attention paid to software funding, management, and licensing by the economic and social sciences research community [Q13c].

Early career, mid-career, and senior career researchers (74%) felt sufficiently well-informed about how software is funded, managed, and licensed, while junior career researchers (47% agree vs 35% who disagree) expressed ambivalence about this statement. [Q13a].

Perceived barriers to the recognition of research software development and maintenance include:

- Lack of recognition of software development at institutional and funder levels, leading to limited career options.
- Lack of integration of software developers within teams or departments.
- Lack of support for software maintenance and development in terms of funding and time from institutions and the ESRC.

Infrastructure and hardware

68% of respondents reported using either their own hardware or institutionally provided hardware to run their software and tools [Q14]. The complete absence of the use of Tier 1 and Tier 2 services in this population sample is notable. This could reflect the fact that these are mainly regarded as EPSRC resources, while links in the survey pointed towards EPSRC-funded resources only.

Respondents were more likely to use institutional facilities as they progressed in seniority. This may indicate that they have been exposed to, and incentivised to learn, differences in institutional policies during their careers.

Our survey showed that 51% of junior career researchers provide their own hardware. Interview data shows that this could in part reflect a desire to retain their own software settings. The potential consequences of this practice are less institutional support, and less adherence to institutional policies.

Recommendations

1. To encourage open research and reproducibility, the ESRC should support the adoption of widely used open source software, such as the R ecosystem, by encouraging community engagement activities and recognising contributions

to software communities. However, for important tools with smaller user bases, further investigation is required to understand what funding and recognition mechanisms would be most appropriate to encourage and support the adoption of such tools.

2. The ESRC should provide targeted funding to support the maintenance and development of new features of open source software used by the community.
3. Institutions should invest as much in producing internal training and support for open source tools as they invest in commercial tools.
4. The ESRC should commission a study to understand the reasons for the lack of open source alternatives for qualitative data analysis.
5. Further research is needed to understand why there are differences in barriers to using research software among those reporting disabilities.
6. Institutions should embed RSEs within research departments to support the software needs of social sciences researchers.
7. The ESRC should work with the UKRI Digital Research Infrastructure to ensure researchers can access larger-scale computational infrastructure that has traditionally excluded ESRC researchers.
8. The ESRC should consider how to fund computational research at the interface between ESRC and EPSRC domains.
9. More research is required on the impact of researchers using their own hardware vs institutionally provided hardware, and the use of different software tools to conduct research. The ESRC should commission a specific study to understand the social, economic, and research impacts of people using their own computational hardware in preference over institutionally provided infrastructure.
10. The ESRC should extend its research data policy to include the sharing and publishing of software.

6.3 SKILLS AND TRAINING

Key findings

Among our respondents, online courses (81%) are the most popular means of acquiring skills, particularly for early career researchers. It remains unclear whether the enthusiasm for online learning is a consequence of the pandemic. [Q11].

Online examples (65%) and learning from peers and colleagues (58%) were the second and third most popular forms of self-led learning identified in the survey. This form of learning was seen as appropriate in relation to open source tools. Stated barriers to self-led learning included a lack of initial competence and confidence in understanding the content.

Key barriers to acquiring general software and data management skills include:

- Lack of support at an institutional and funder level. While there is training available for traditional proprietary software (NVivo, SPSS, STATA), a lack of curricular and senior researcher support appears to be limiting the horizons of junior and early career researchers who want to build a broader, more future-oriented set of skills.
- Lack of initial confidence and competence. The ability to perform functions in SPSS or Stata does not equip researchers with the skills and knowledge to understand the coding requirements underlying R or Python. They face a steep learning curve.
- Lack of time. Unless learning a specific tool is necessary for their research, researchers seem unable to take the time to learn new skills. This time barrier increases with seniority into mid-career.
- Lack of appropriate courses delivering sufficient knowledge to be useful.

Software and data management skills could mitigate, for instance, the issues created by the movement of early career researchers between institutions, and could potentially instil the competence required to take on self-led training.

NCRM courses

Roughly 19% of respondents across all career stages had taken the NCRM courses listed in the survey [Q11b].

Recommendations

1. The ESRC and institutions should support and encourage the uptake of existing external training (e.g. NCRM and The Carpentries), and invest in the development of more targeted internal training to help researchers write their own code following good practices.
2. Training providers should investigate whether learning attitudes are changing (in relation to the pandemic and the use of online courses).
3. There is a potential role for a drop-in support model/mechanism, such as the UK Data Service Data Drop-in[81], to fill the gap between the training offered and the practical applications in social sciences research. However, further investigation is needed to understand exactly what researchers are looking for.
4. The ESRC should invest in supporting more courses to allow people to transition from commercial statistical analysis packages to in-demand open source alternatives that support open research practices and facilitate worldwide collaboration.
5. The ESRC should encourage the development of good practice guidance at institutional and journal levels for sharing and publishing software[82] and data[83] following FAIR principles.
6. Different forms of training should continue to be offered so that researchers with caring responsibilities can still access it. We also believe that additional asynchronous support will be required to make sure researchers keep practising the skills they have learned.
7. Further research is needed to understand how time to acquire new skills can be protected.

[81] https://ukdataservice.ac.uk/events/uk-data-service-computational-social-science-drop-in/ (last accessed December 5 2022)

[82] FAIR principles for research software: https://doi.org/10.15497/RDA00068 DOI: 10.15497/RDA00068

[83] FAIR principles for data https://www.go-fair.org/fair-principles/ (last accessed December 5 2022)

6.4 COLLABORATION AND INTERDISCIPLINARITY

Key findings

Interview data shows a number of perceived barriers to collaboration and interdisciplinarity, such as

- Lack of recognition. Incentives are vital for collaboration, and there is currently a perception that the research councils and the REF do not encourage interdisciplinarity.
- One of the side effects of not collaborating with other disciplines is the reduced need for developing new software skills.
- There is currently no training for carrying out interdisciplinary research, nor are there examples of how this can work early on in careers.

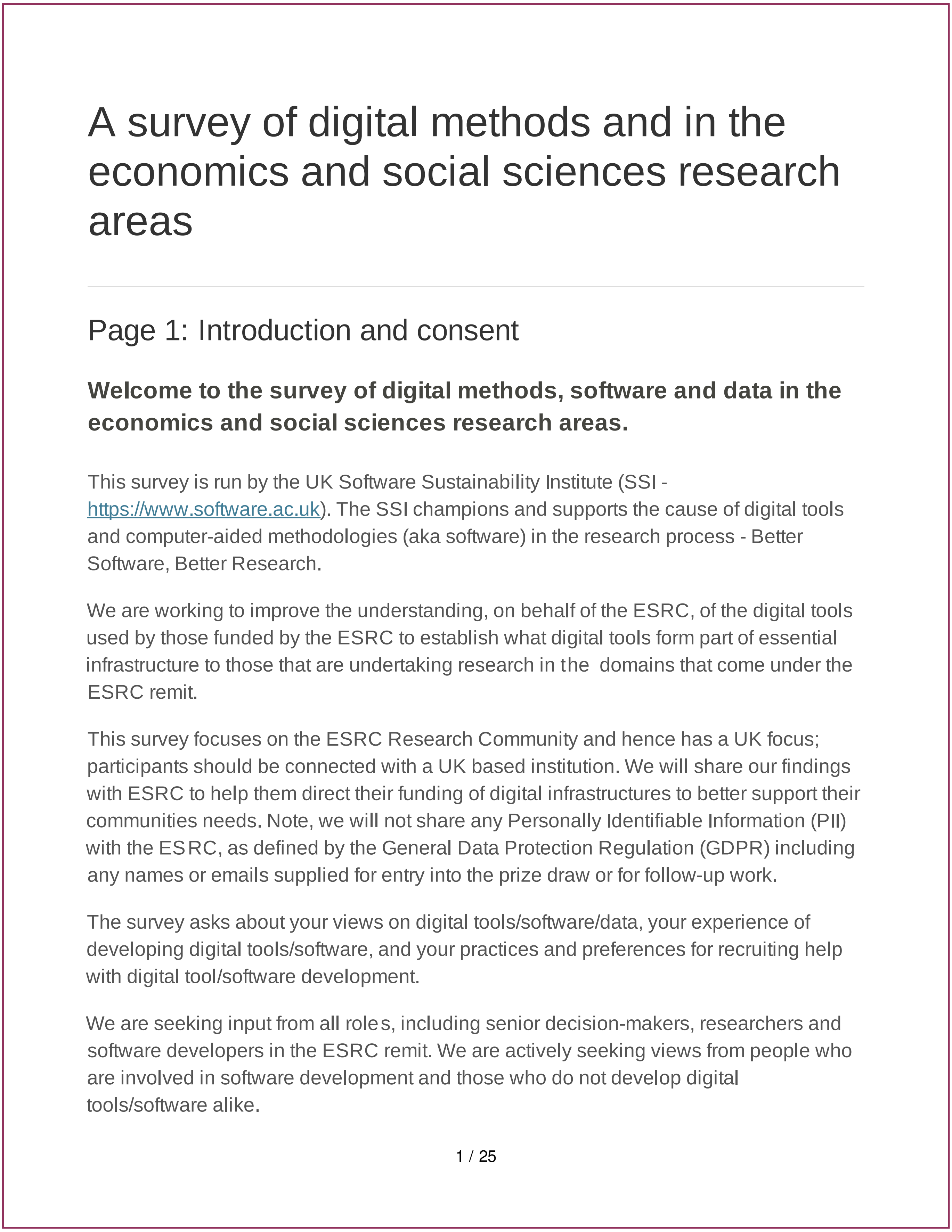
Recommendations

1. ESRC could improve collaboration and team research by incentivising interdisciplinary work.

7. REFERENCES

1. "The UK's research and innovation infrastructure: opportunities to grow our capability", UKRI, October 2020, https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf.
2. Duca, D. & Metzler, K., "The Ecosystem of Technologies for Social Science Research", SAGE, May 29, 2020, DOI: 10.4135/wp191101.
3. "A National Agenda for Research Software", Zenodo, March 28 2022, DOI: 10.5281/zenodo.6378082.
4. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann M., Jebara, T., King, G., Macy, M., Roy, D. & Alstyne, M. V. "Computational Social Science", Science, Feb 6 2009, Vol 323, Issue 5915, pp. 721-723, DOI: 10.1126/science.1167.
5. Barker, M., Chue Hong, N. P., Katz, D. S., Leggott, M., Treloar, A., van Eijnatten, J. & Aragon, S., "Research software is essential for research data, so how should governments respond?", Zenodo, December 9 2021, DOI: 10.5281/zenodo.5762703.
6. Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H. & Nelson, A. "Computational social science: Obstacles and opportunities", Science, 28 August 2020, Vol 369, Issue 6507, pp. 1060-1062, DOI: 10.1126/science.aaz81.
7. Metzler, K., Kim, D. A., Allum, N & Denman, A., "Who is doing computational social science? Trends in big data research", SAGE, 2016, DOI: 10.4135/wp160926.
8. Heeks, R., & Shekhar, S. "Datafication, development and marginalised urban communities: An applied data justice framework", Information, Communication & Society, May 13 2019, Volume 22, Issue 7, pp. 992–1011, DOI: 10.1080/1369118X.2019.1599039
9. "ESRC Data Infrastructure Strategy", UKRI, June 2022, https://www.ukri.org/wp-content/uploads/2022/06/ESRC-090622-DataInfrastructureStrategy2022To2027.pdf (last accessed July 13 2022).
10. "UKRI Strategy 2022–2027 - Transforming tomorrow together", UKRI, March 2022, https://www.ukri.org/wp-content/uploads/2022/06/ESRC-090622-DataInfrastructureStrategy2022To2027.pdf (last accessed July 13 2022).
11. "ESRC centres: 2020 overview - A listing of accredited Economic and Social Research Council (ESRC) centres in 2020", UKRI, https://www.ukri.org/publications/esrc-centres-2020-overview (last accessed August 10 2022).
12. Jisc Online surveys, https://www.onlinesurveys.ac.uk/ (last Accessed March 9 2022).
13. Milliken, G, Nguyen, S. & Steeves, V. , "A Behavioural Approach to Understanding the Git Experience", Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/c25852a0-505f-4027-87d8-da104f4147c6/content (last accessed July 28 2022)
14. Publishing your research findings, UKRI, https://www.ukri.org/manage-your-award/publishing-your-research-findings/#contents-list, (last accessed: August 8 2022).
15. BBSRC data sharing policy, UKRI, https://www.ukri.org/publications/bbsrc-data-sharing-policy (last accessed August 18 2022).
16. EPSRC policy framework on research data, UKRI, https://www.ukri.org/about-us/epsrc/our-policies-and-standards/policy-framework-on-research-data/ (last accessed August 18 2022).
17. ESRC research data policy, UKRI, https://www.ukri.org/publications/esrc-research-data-policy/ (last accessed August 18 2022).
18. MRC data sharing policy, UKRI, https://www.ukri.org/publications/mrc-data-sharing-policy/ (last accessed August 18 2022).
19. NERC data policy, UKRI, https://www.ukri.org/about-us/nerc/our-policies-and-standards/nerc-data-policy/ (last accessed August 19 2022).
20. STFC scientific data policy, UKRI, https://www.ukri.org/publications/stfc-scientific-data-policy/ (last accessed August 18 2022).
21. Research ethics and integrity policy, College of Science and Engineering, University of Edinburgh, https://www.ed.ac.uk/science-engineering/research/research-ethics (last accessed September 29 2022).
22. Braun, V. & Clarke, V., "Using thematic analysis in psychology", Qualitative Research in Psychology, July 21 2008, Volume 3, Issue 2, pp. 77-101, DOI: 10.1191/1478088706qp063oa.
23. Taylor, R., Walker, J., Hettrick, S., Broadbent, P., De Roure, D., "Shaping Data and Software Policy in the Arts and Humanities Research Community", UKRI, https://www.ukri.org/wp-content/uploads/2022/10/AHRC-011122-SSIReport-ShapingDataAndSoftwarePolicyInTheArtsAndHumanities.pdf (last accessed December 5 2022).
24. Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A., Martinez, Carlos, Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., Struck, A., Lee, A., Loewe, A., van Werkhoven, B., Jones, C., Garijo, D., Plomp, E. & Genova, F., "FAIR Principles for Research Software (FAIR4RS Principles)", Zenodo, May 24 2022, DOI: 10.15497/RDA00068

APPENDIX 1.
SURVEY



A survey of digital methods and in the economics and social sciences research areas

Page 1: Introduction and consent

Welcome to the survey of digital methods, software and data in the economics and social sciences research areas.

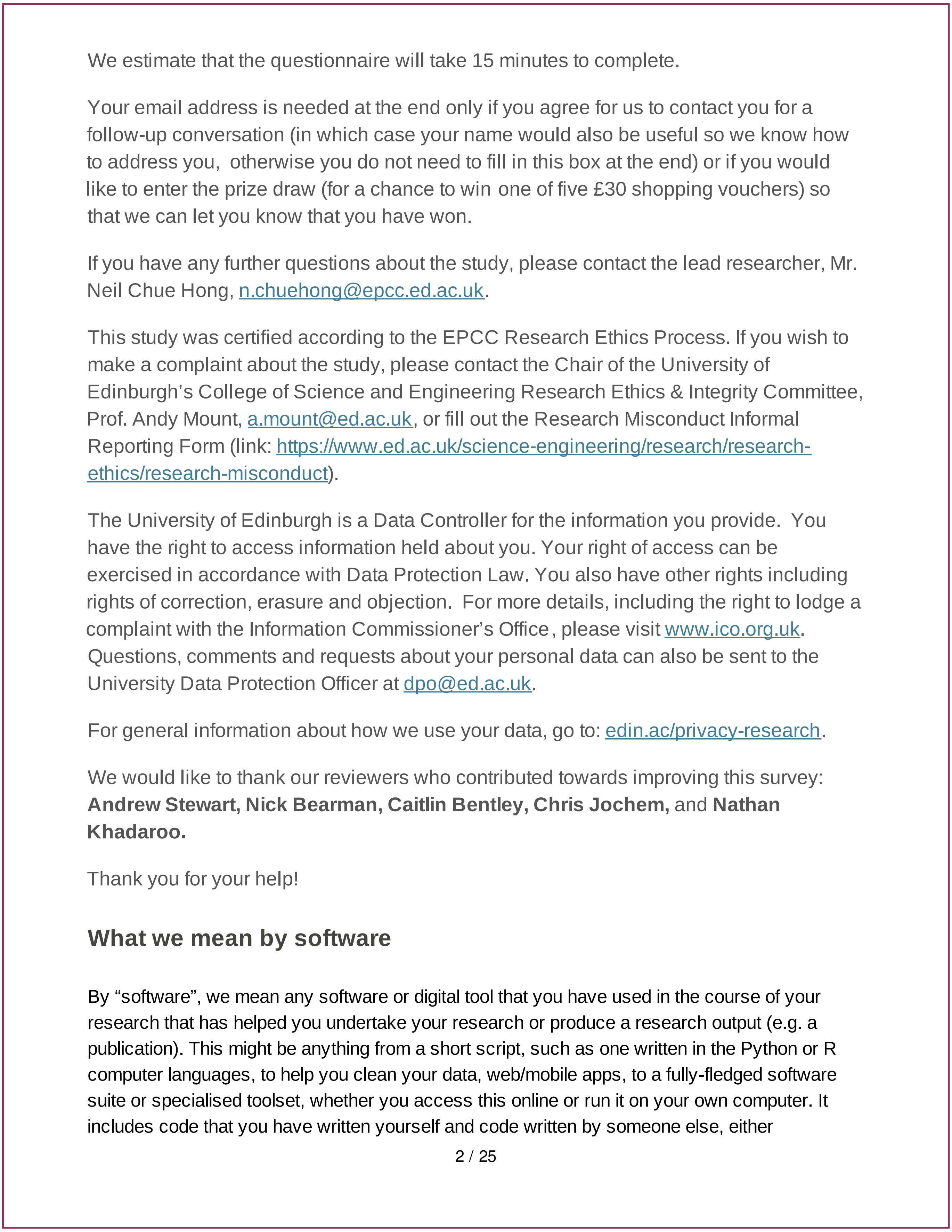
This survey is run by the UK Software Sustainability Institute (SSI - https://www.software.ac.uk). The SSI champions and supports the cause of digital tools and computer-aided methodologies (aka software) in the research process - Better Software, Better Research.

We are working to improve the understanding, on behalf of the ESRC, of the digital tools used by those funded by the ESRC to establish what digital tools form part of essential infrastructure to those that are undertaking research in the domains that come under the ESRC remit.

This survey focuses on the ESRC Research Community and hence has a UK focus; participants should be connected with a UK based institution. We will share our findings with ESRC to help them direct their funding of digital infrastructures to better support their communities needs. Note, we will not share any Personally Identifiable Information (PII) with the ESRC, as defined by the General Data Protection Regulation (GDPR) including any names or emails supplied for entry into the prize draw or for follow-up work.

The survey asks about your views on digital tools/software/data, your experience of developing digital tools/software, and your practices and preferences for recruiting help with digital tool/software development.

We are seeking input from all roles, including senior decision-makers, researchers and software developers in the ESRC remit. We are actively seeking views from people who are involved in software development and those who do not develop digital tools/software alike.



We estimate that the questionnaire will take 15 minutes to complete.

Your email address is needed at the end only if you agree for us to contact you for a follow-up conversation (in which case your name would also be useful so we know how to address you, otherwise you do not need to fill in this box at the end) or if you would like to enter the prize draw (for a chance to win one of five £30 shopping vouchers) so that we can let you know that you have won.

If you have any further questions about the study, please contact the lead researcher, Mr. Neil Chue Hong, n.chuehong@epcc.ed.ac.uk.

This study was certified according to the EPCC Research Ethics Process. If you wish to make a complaint about the study, please contact the Chair of the University of Edinburgh's College of Science and Engineering Research Ethics & Integrity Committee, Prof. Andy Mount, a.mount@ed.ac.uk, or fill out the Research Misconduct Informal Reporting Form (link: https://www.ed.ac.uk/science-engineering/research/research-ethics/research-misconduct).

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance with Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

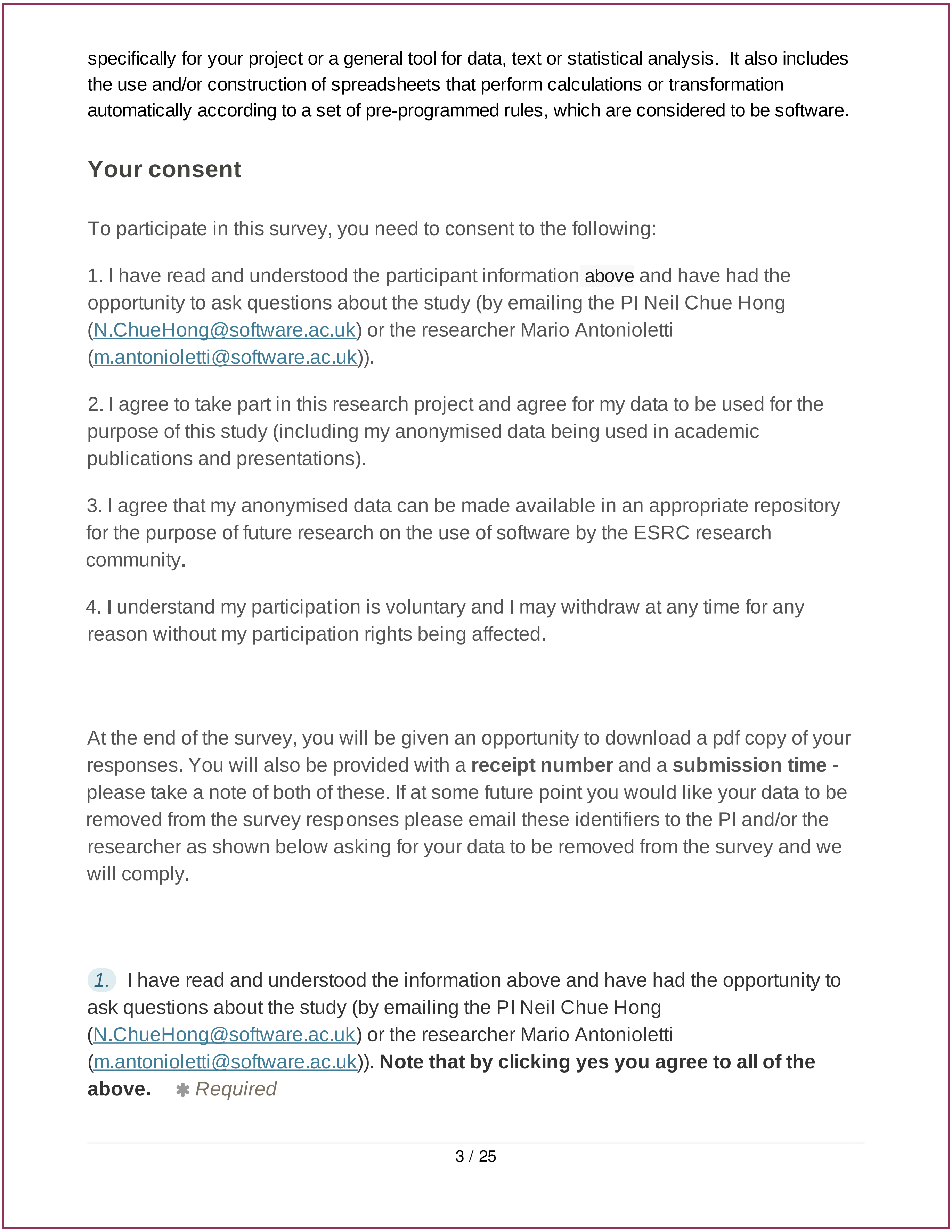
For general information about how we use your data, go to: edin.ac/privacy-research.

We would like to thank our reviewers who contributed towards improving this survey: **Andrew Stewart, Nick Bearman, Caitlin Bentley, Chris Jochem,** and **Nathan Khadaroo.**

Thank you for your help!

What we mean by software

By "software", we mean any software or digital tool that you have used in the course of your research that has helped you undertake your research or produce a research output (e.g. a publication). This might be anything from a short script, such as one written in the Python or R computer languages, to help you clean your data, web/mobile apps, to a fully-fledged software suite or specialised toolset, whether you access this online or run it on your own computer. It includes code that you have written yourself and code written by someone else, either



specifically for your project or a general tool for data, text or statistical analysis. It also includes the use and/or construction of spreadsheets that perform calculations or transformation automatically according to a set of pre-programmed rules, which are considered to be software.

Your consent

To participate in this survey, you need to consent to the following:

1. I have read and understood the participant information above and have had the opportunity to ask questions about the study (by emailing the PI Neil Chue Hong (N.ChueHong@software.ac.uk) or the researcher Mario Antonioletti (m.antonioletti@software.ac.uk)).

2. I agree to take part in this research project and agree for my data to be used for the purpose of this study (including my anonymised data being used in academic publications and presentations).

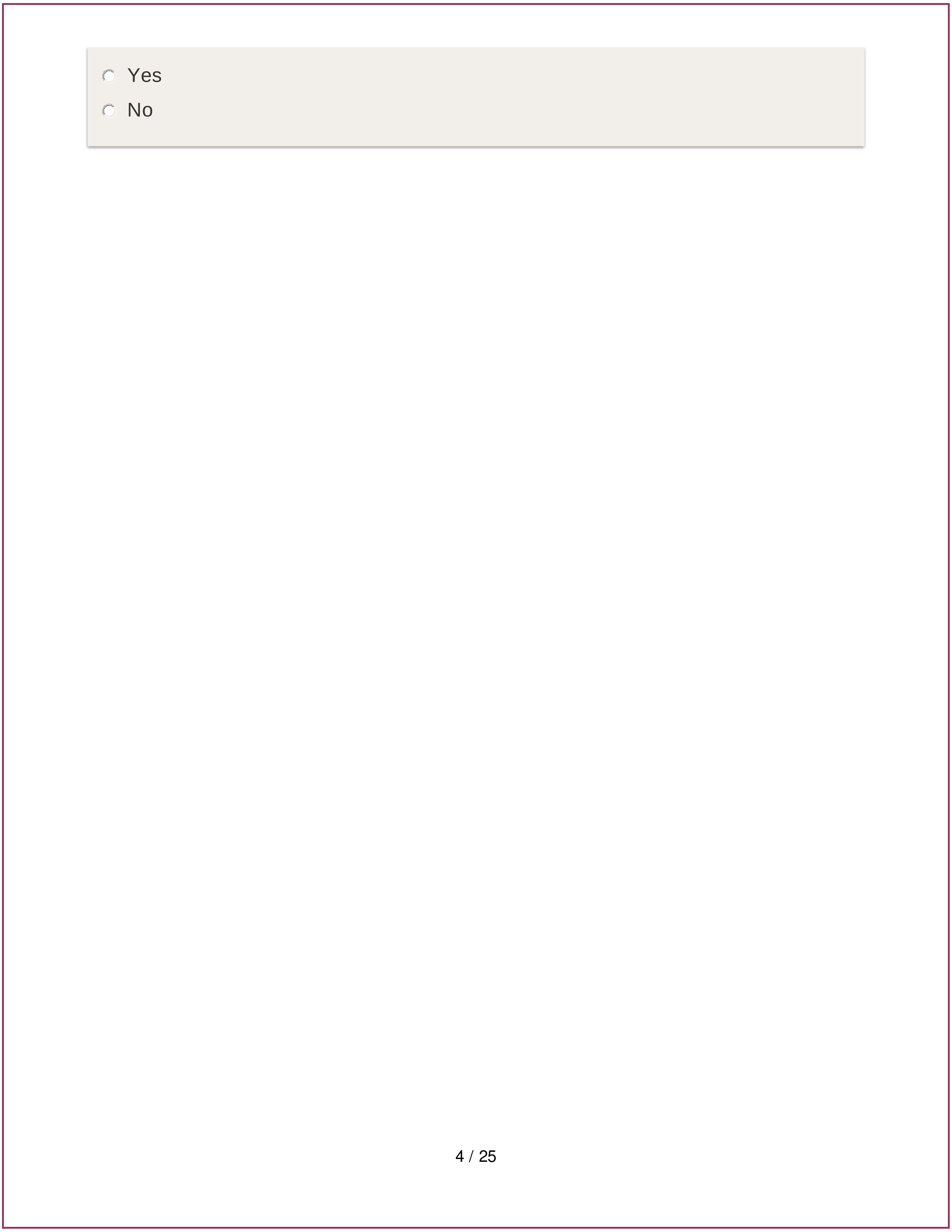
3. I agree that my anonymised data can be made available in an appropriate repository for the purpose of future research on the use of software by the ESRC research community.

4. I understand my participation is voluntary and I may withdraw at any time for any reason without my participation rights being affected.

At the end of the survey, you will be given an opportunity to download a pdf copy of your responses. You will also be provided with a **receipt number** and a **submission time** - please take a note of both of these. If at some future point you would like your data to be removed from the survey responses please email these identifiers to the PI and/or the researcher as shown below asking for your data to be removed from the survey and we will comply.

1. I have read and understood the information above and have had the opportunity to ask questions about the study (by emailing the PI Neil Chue Hong (N.ChueHong@software.ac.uk) or the researcher Mario Antonioletti (m.antonioletti@software.ac.uk)). **Note that by clicking yes you agree to all of the above.** ✱ *Required*

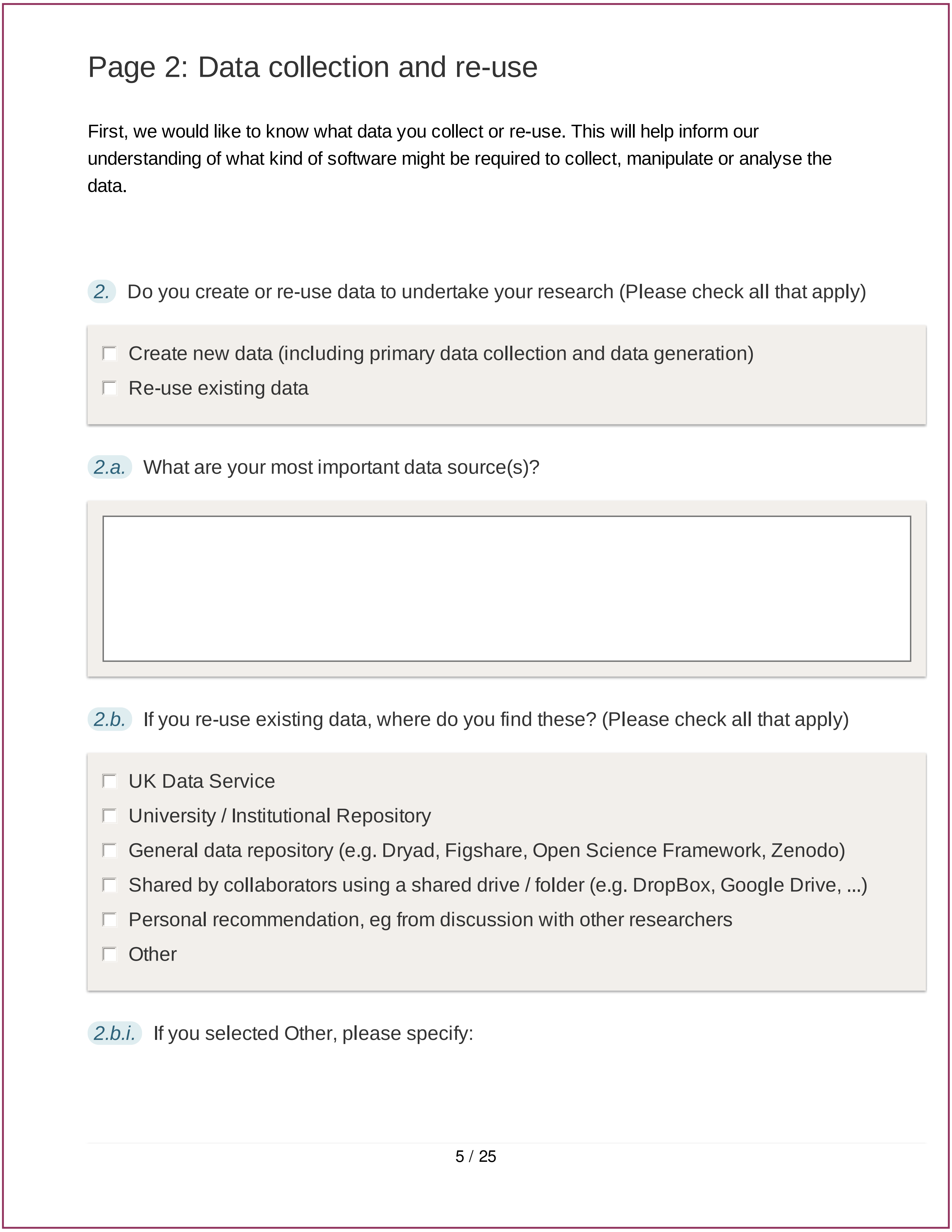
3 / 25



○ Yes

○ No

4 / 25



Page 2: Data collection and re-use

First, we would like to know what data you collect or re-use. This will help inform our understanding of what kind of software might be required to collect, manipulate or analyse the data.

2. Do you create or re-use data to undertake your research (Please check all that apply)

□ Create new data (including primary data collection and data generation)

□ Re-use existing data

2.a. What are your most important data source(s)?

2.b. If you re-use existing data, where do you find these? (Please check all that apply)

□ UK Data Service

□ University / Institutional Repository

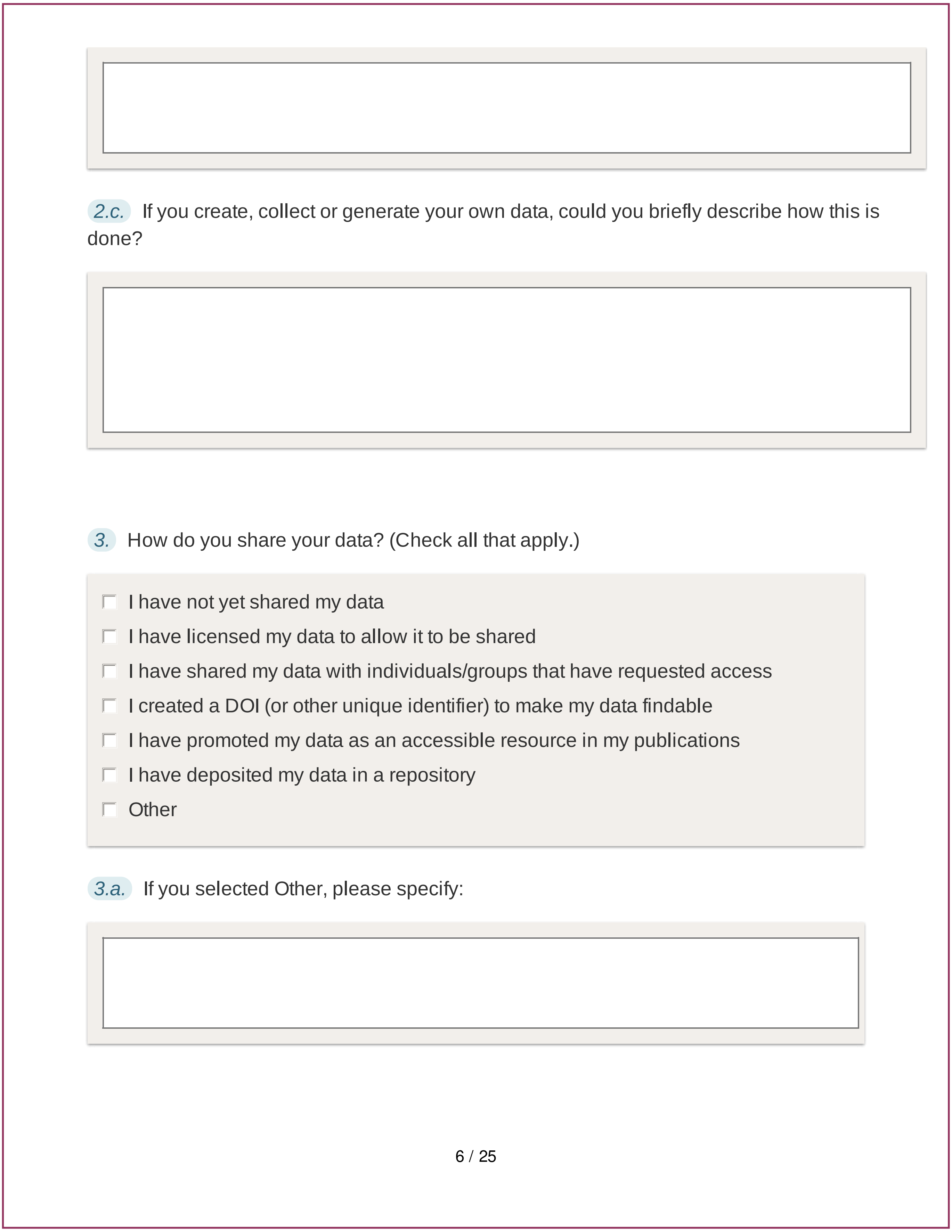
□ General data repository (e.g. Dryad, Figshare, Open Science Framework, Zenodo)

□ Shared by collaborators using a shared drive / folder (e.g. DropBox, Google Drive, ...)

□ Personal recommendation, eg from discussion with other researchers

□ Other

2.b.i. If you selected Other, please specify:

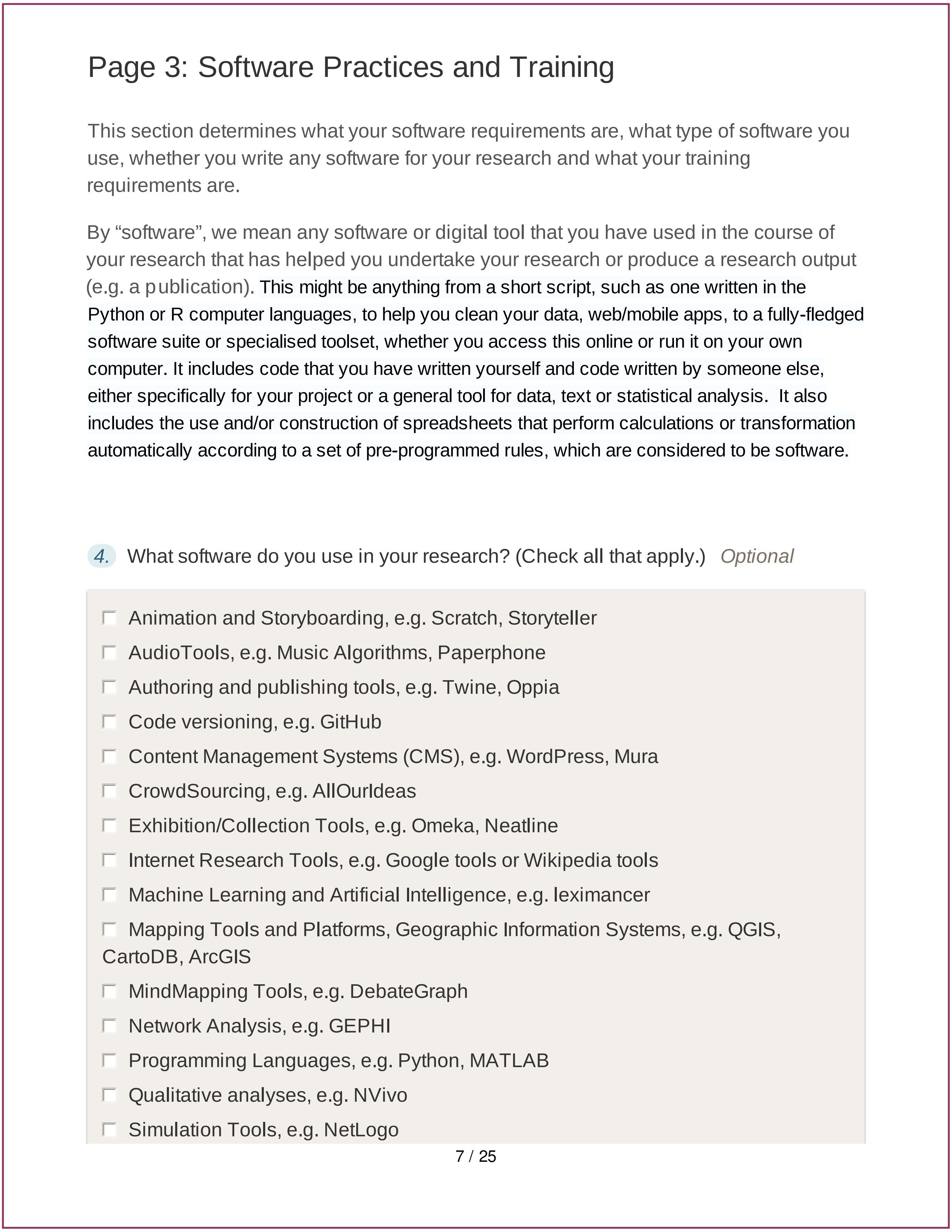


2.c. If you create, collect or generate your own data, could you briefly describe how this is done?

3. How do you share your data? (Check all that apply.)

□ I have not yet shared my data
□ I have licensed my data to allow it to be shared
□ I have shared my data with individuals/groups that have requested access
□ I created a DOI (or other unique identifier) to make my data findable
□ I have promoted my data as an accessible resource in my publications
□ I have deposited my data in a repository
□ Other

3.a. If you selected Other, please specify:



Page 3: Software Practices and Training

This section determines what your software requirements are, what type of software you use, whether you write any software for your research and what your training requirements are.

By “software”, we mean any software or digital tool that you have used in the course of your research that has helped you undertake your research or produce a research output (e.g. a publication). This might be anything from a short script, such as one written in the Python or R computer languages, to help you clean your data, web/mobile apps, to a fully-fledged software suite or specialised toolset, whether you access this online or run it on your own computer. It includes code that you have written yourself and code written by someone else, either specifically for your project or a general tool for data, text or statistical analysis. It also includes the use and/or construction of spreadsheets that perform calculations or transformation automatically according to a set of pre-programmed rules, which are considered to be software.

4. What software do you use in your research? (Check all that apply.) *Optional*

☐ Animation and Storyboarding, e.g. Scratch, Storyteller

☐ AudioTools, e.g. Music Algorithms, Paperphone

☐ Authoring and publishing tools, e.g. Twine, Oppia

☐ Code versioning, e.g. GitHub

☐ Content Management Systems (CMS), e.g. WordPress, Mura

☐ CrowdSourcing, e.g. AllOurIdeas

☐ Exhibition/Collection Tools, e.g. Omeka, Neatline

☐ Internet Research Tools, e.g. Google tools or Wikipedia tools

☐ Machine Learning and Artificial Intelligence, e.g. leximancer

☐ Mapping Tools and Platforms, Geographic Information Systems, e.g. QGIS, CartoDB, ArcGIS

☐ MindMapping Tools, e.g. DebateGraph

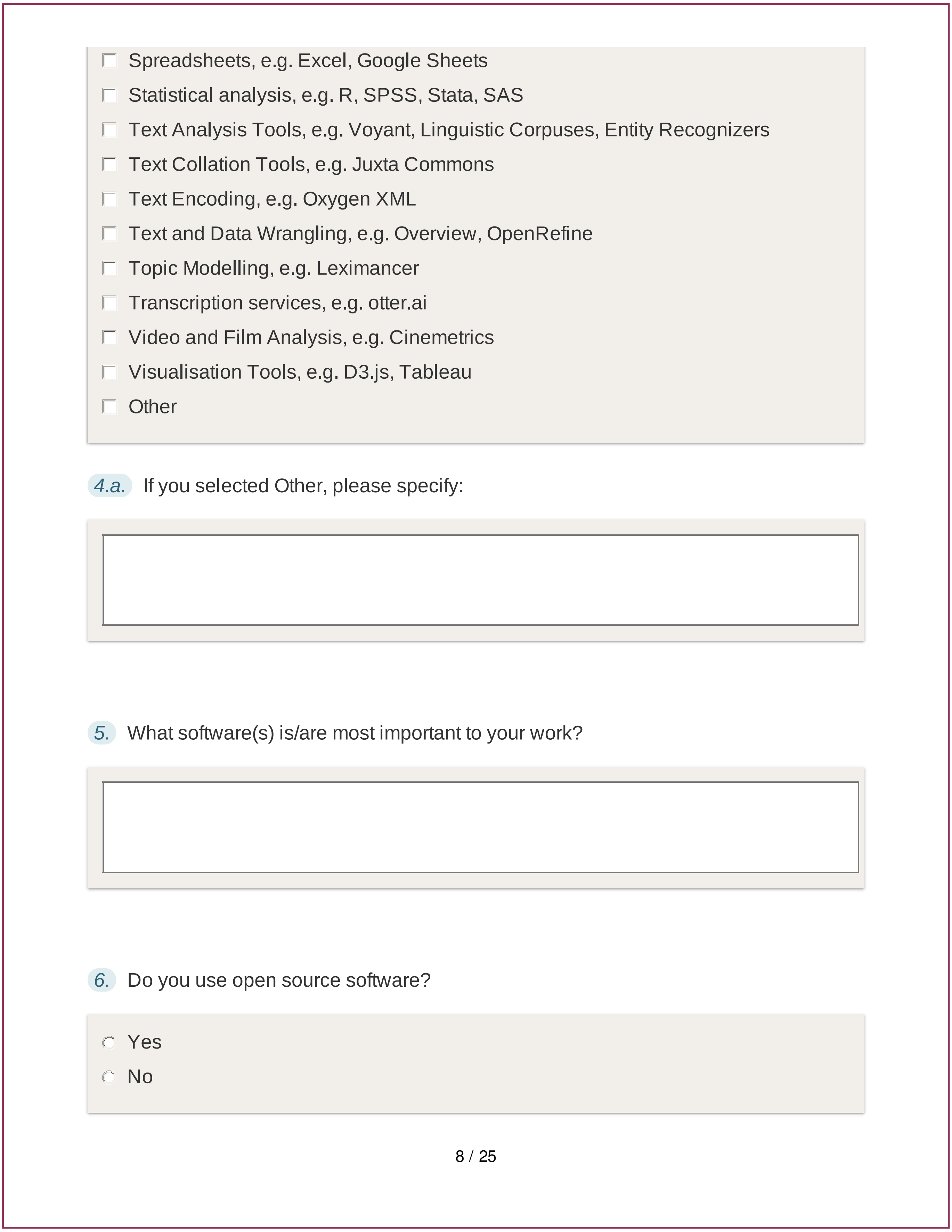
☐ Network Analysis, e.g. GEPHI

☐ Programming Languages, e.g. Python, MATLAB

☐ Qualitative analyses, e.g. NVivo

☐ Simulation Tools, e.g. NetLogo

7 / 25



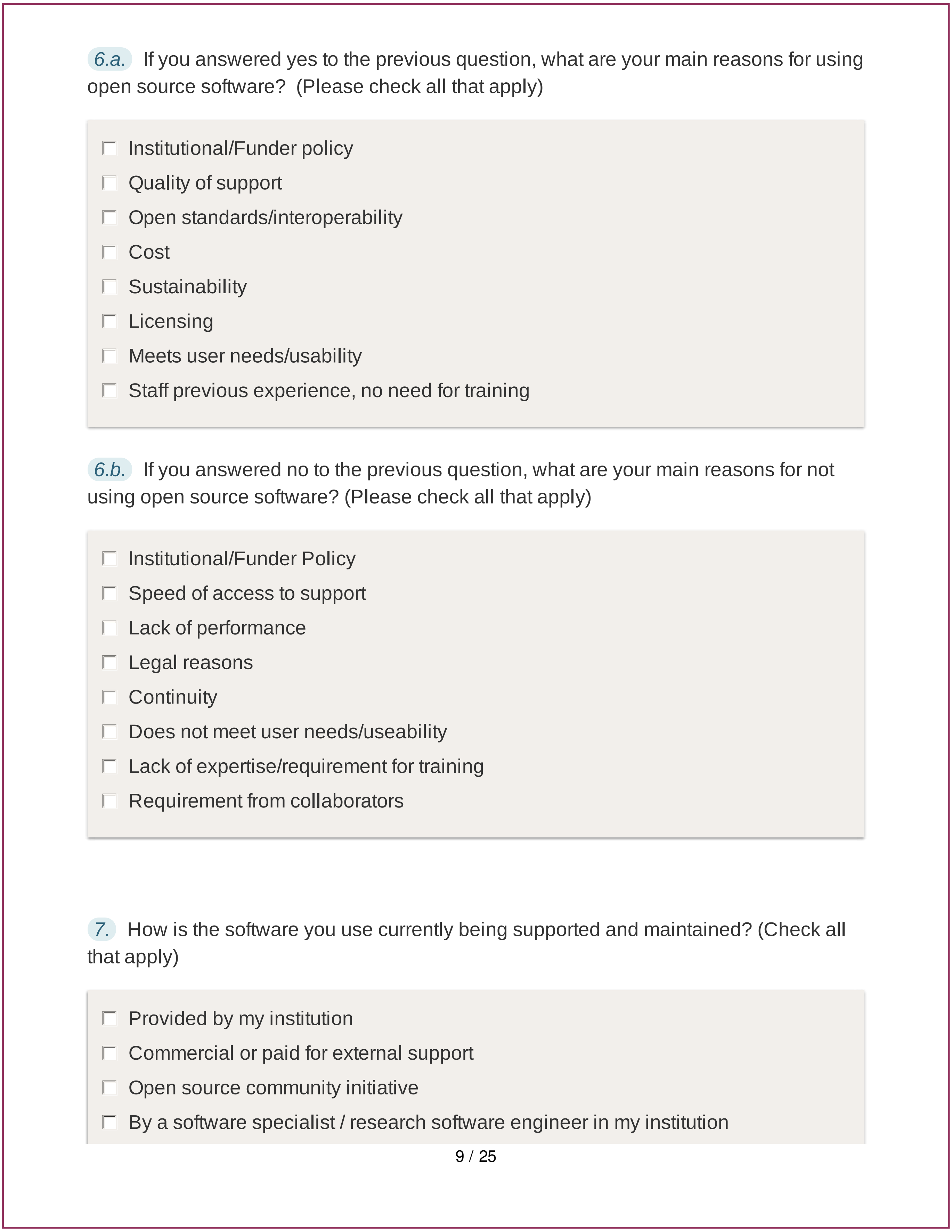
- ☐ Spreadsheets, e.g. Excel, Google Sheets
- ☐ Statistical analysis, e.g. R, SPSS, Stata, SAS
- ☐ Text Analysis Tools, e.g. Voyant, Linguistic Corpuses, Entity Recognizers
- ☐ Text Collation Tools, e.g. Juxta Commons
- ☐ Text Encoding, e.g. Oxygen XML
- ☐ Text and Data Wrangling, e.g. Overview, OpenRefine
- ☐ Topic Modelling, e.g. Leximancer
- ☐ Transcription services, e.g. otter.ai
- ☐ Video and Film Analysis, e.g. Cinemetrics
- ☐ Visualisation Tools, e.g. D3.js, Tableau
- ☐ Other

4.a. If you selected Other, please specify:

5. What software(s) is/are most important to your work?

6. Do you use open source software?

- ○ Yes
- ○ No



6.a. If you answered yes to the previous question, what are your main reasons for using open source software? (Please check all that apply)

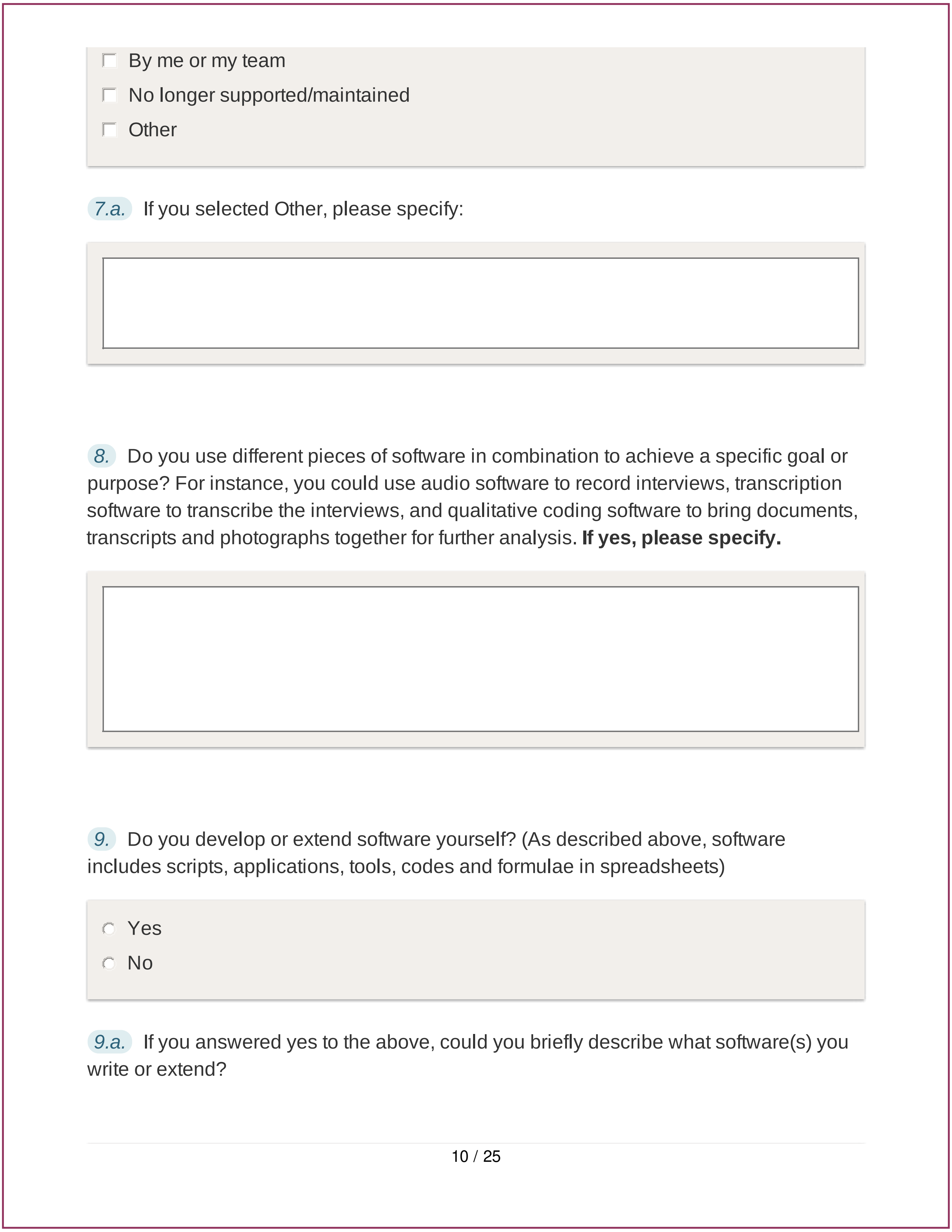
□ Institutional/Funder policy
□ Quality of support
□ Open standards/interoperability
□ Cost
□ Sustainability
□ Licensing
□ Meets user needs/usability
□ Staff previous experience, no need for training

6.b. If you answered no to the previous question, what are your main reasons for not using open source software? (Please check all that apply)

□ Institutional/Funder Policy
□ Speed of access to support
□ Lack of performance
□ Legal reasons
□ Continuity
□ Does not meet user needs/useability
□ Lack of expertise/requirement for training
□ Requirement from collaborators

7. How is the software you use currently being supported and maintained? (Check all that apply)

□ Provided by my institution
□ Commercial or paid for external support
□ Open source community initiative
□ By a software specialist / research software engineer in my institution



□ By me or my team

□ No longer supported/maintained

□ Other

7.a. If you selected Other, please specify:

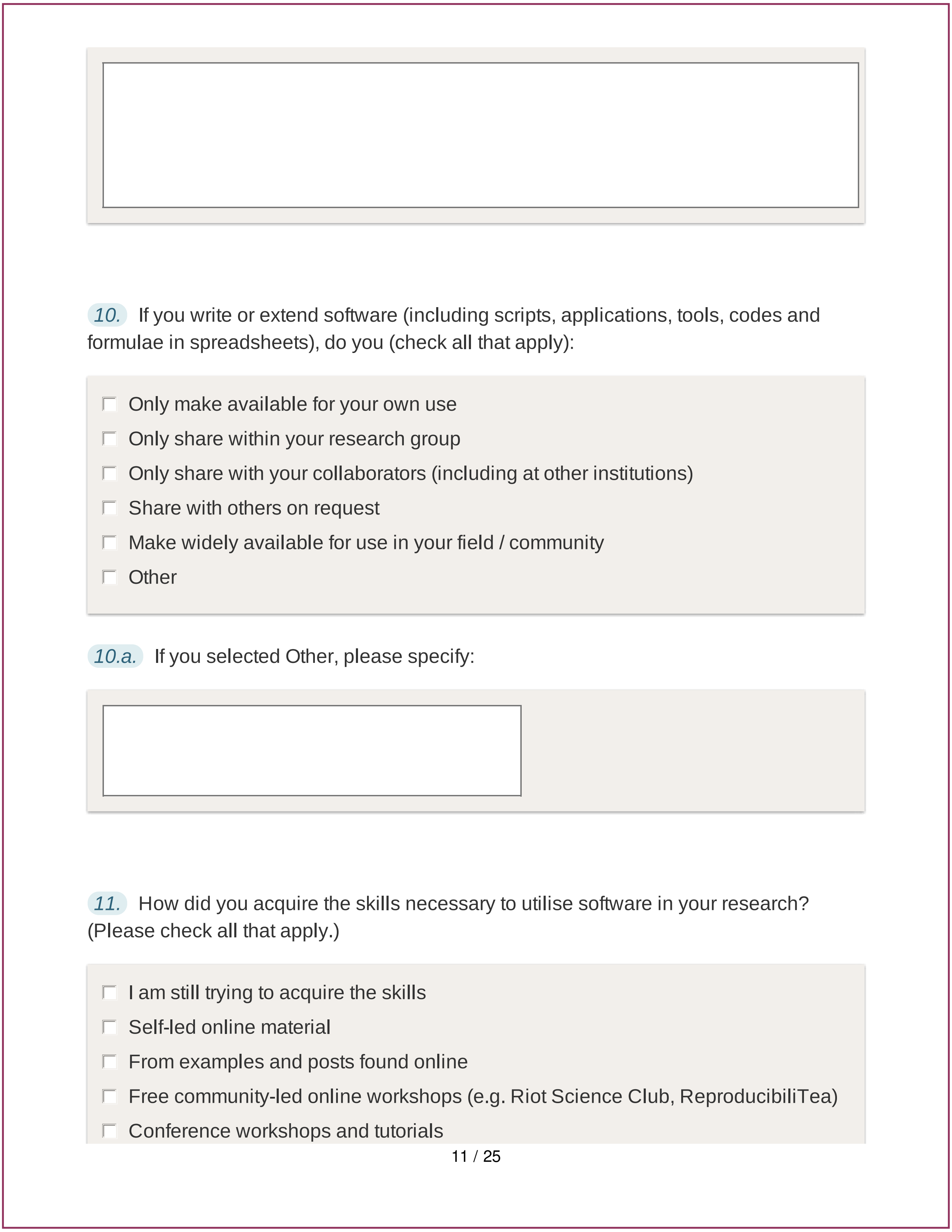
8. Do you use different pieces of software in combination to achieve a specific goal or purpose? For instance, you could use audio software to record interviews, transcription software to transcribe the interviews, and qualitative coding software to bring documents, transcripts and photographs together for further analysis. **If yes, please specify.**

9. Do you develop or extend software yourself? (As described above, software includes scripts, applications, tools, codes and formulae in spreadsheets)

○ Yes

○ No

9.a. If you answered yes to the above, could you briefly describe what software(s) you write or extend?



10. If you write or extend software (including scripts, applications, tools, codes and formulae in spreadsheets), do you (check all that apply):

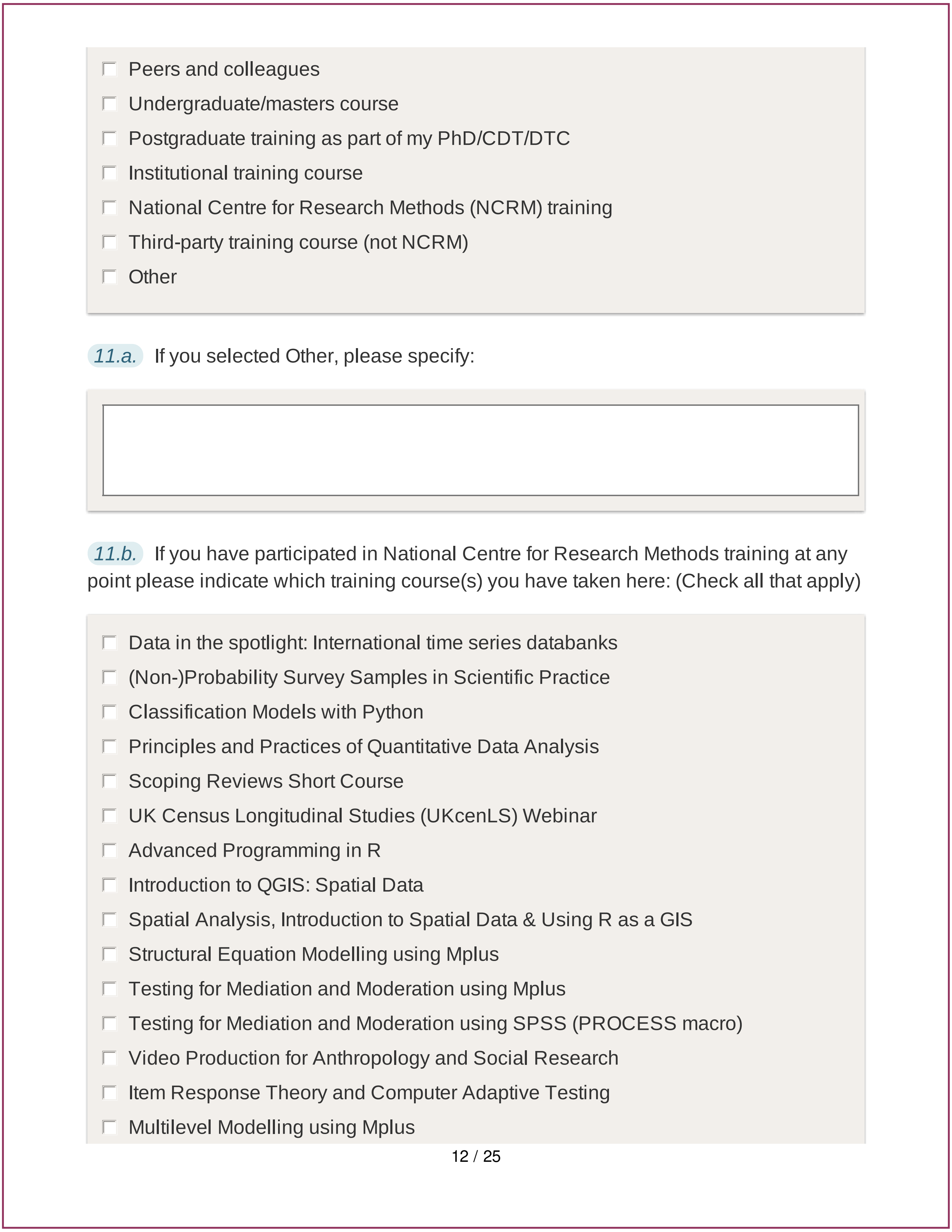
- □ Only make available for your own use
- □ Only share within your research group
- □ Only share with your collaborators (including at other institutions)
- □ Share with others on request
- □ Make widely available for use in your field / community
- □ Other

10.a. If you selected Other, please specify:

11. How did you acquire the skills necessary to utilise software in your research? (Please check all that apply.)

- □ I am still trying to acquire the skills
- □ Self-led online material
- □ From examples and posts found online
- □ Free community-led online workshops (e.g. Riot Science Club, ReproducibiliTea)
- □ Conference workshops and tutorials

11 / 25

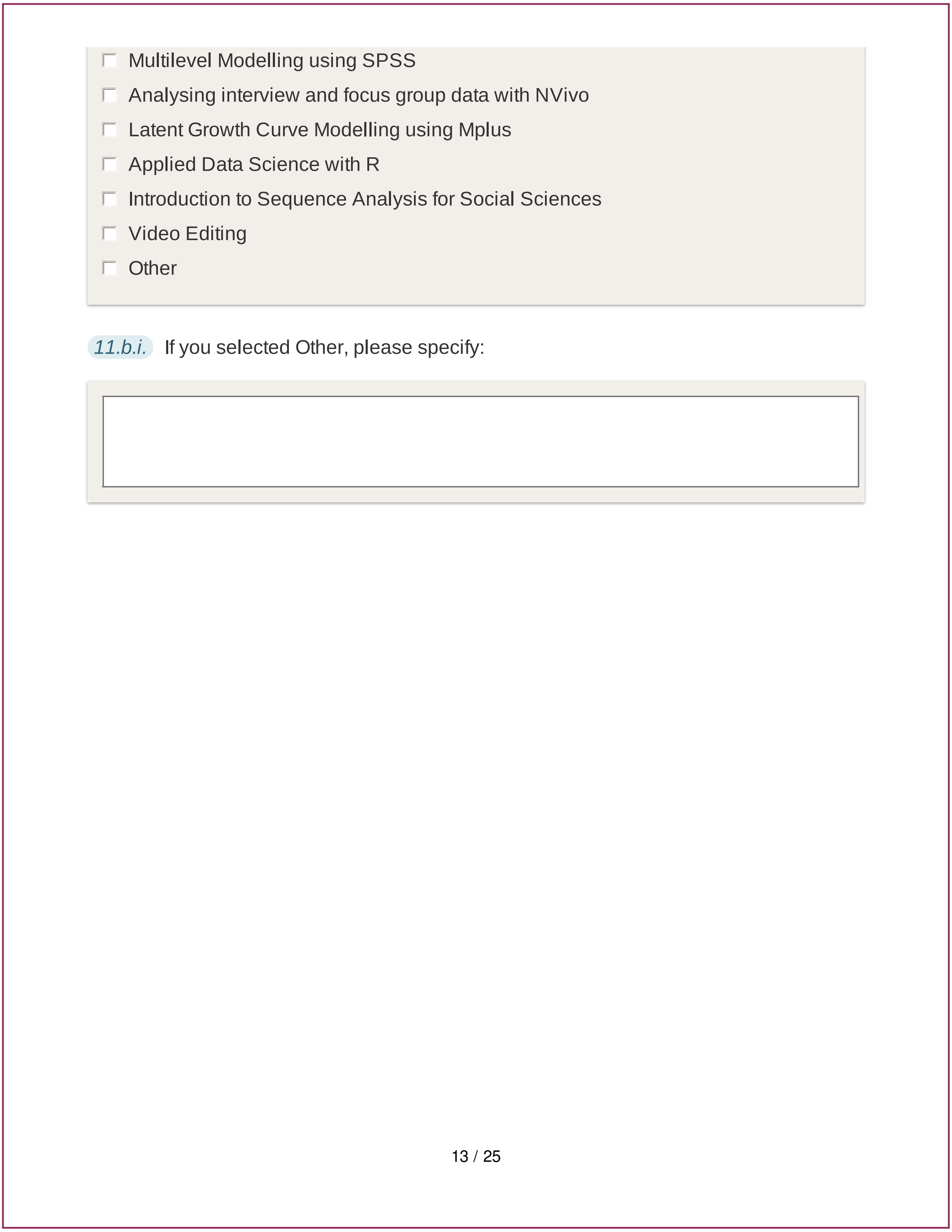


- □ Peers and colleagues
- □ Undergraduate/masters course
- □ Postgraduate training as part of my PhD/CDT/DTC
- □ Institutional training course
- □ National Centre for Research Methods (NCRM) training
- □ Third-party training course (not NCRM)
- □ Other

11.a. If you selected Other, please specify:

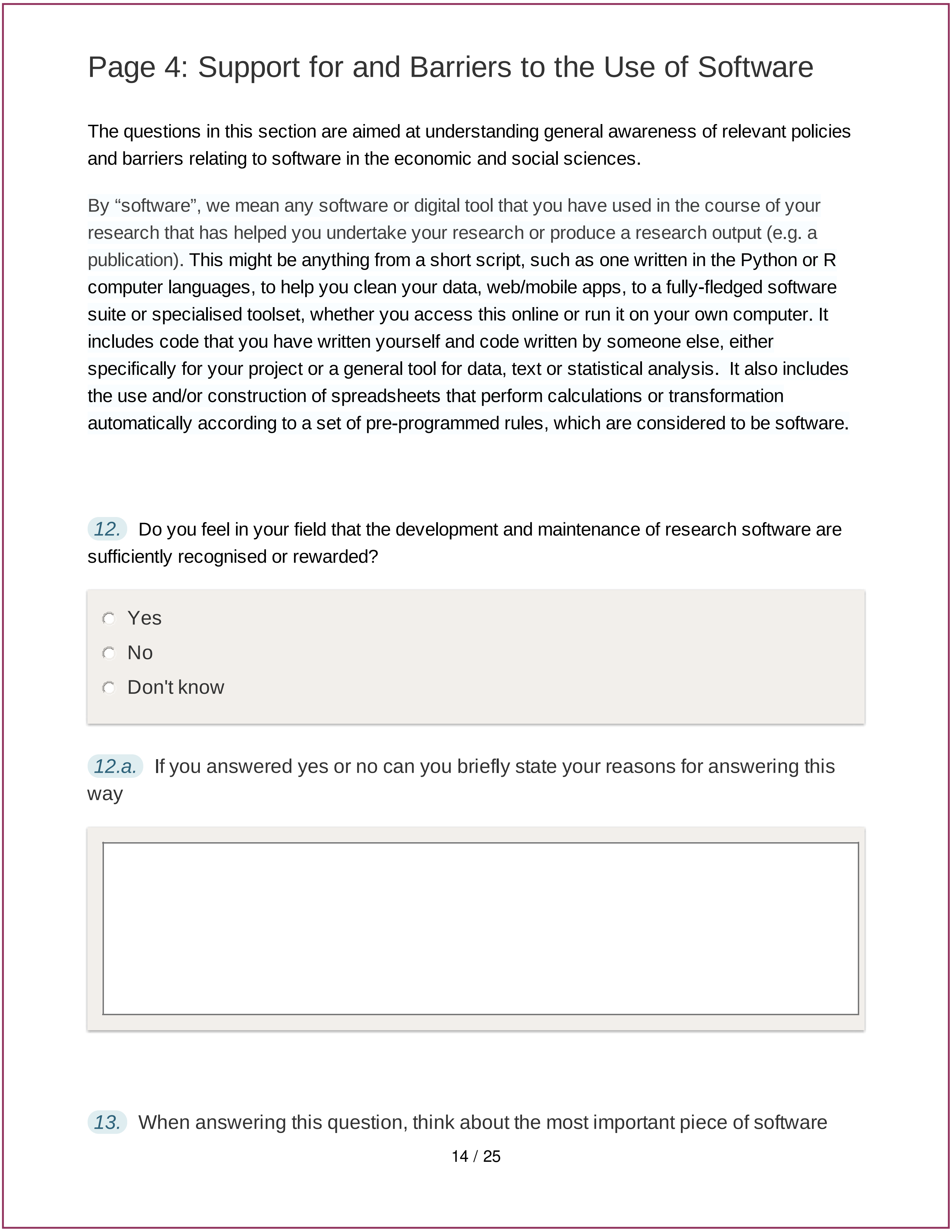
11.b. If you have participated in National Centre for Research Methods training at any point please indicate which training course(s) you have taken here: (Check all that apply)

- □ Data in the spotlight: International time series databanks
- □ (Non-)Probability Survey Samples in Scientific Practice
- □ Classification Models with Python
- □ Principles and Practices of Quantitative Data Analysis
- □ Scoping Reviews Short Course
- □ UK Census Longitudinal Studies (UKcenLS) Webinar
- □ Advanced Programming in R
- □ Introduction to QGIS: Spatial Data
- □ Spatial Analysis, Introduction to Spatial Data & Using R as a GIS
- □ Structural Equation Modelling using Mplus
- □ Testing for Mediation and Moderation using Mplus
- □ Testing for Mediation and Moderation using SPSS (PROCESS macro)
- □ Video Production for Anthropology and Social Research
- □ Item Response Theory and Computer Adaptive Testing
- □ Multilevel Modelling using Mplus



- □ Multilevel Modelling using SPSS
- □ Analysing interview and focus group data with NVivo
- □ Latent Growth Curve Modelling using Mplus
- □ Applied Data Science with R
- □ Introduction to Sequence Analysis for Social Sciences
- □ Video Editing
- □ Other

11.b.i. If you selected Other, please specify:



Page 4: Support for and Barriers to the Use of Software

The questions in this section are aimed at understanding general awareness of relevant policies and barriers relating to software in the economic and social sciences.

By “software”, we mean any software or digital tool that you have used in the course of your research that has helped you undertake your research or produce a research output (e.g. a publication). This might be anything from a short script, such as one written in the Python or R computer languages, to help you clean your data, web/mobile apps, to a fully-fledged software suite or specialised toolset, whether you access this online or run it on your own computer. It includes code that you have written yourself and code written by someone else, either specifically for your project or a general tool for data, text or statistical analysis. It also includes the use and/or construction of spreadsheets that perform calculations or transformation automatically according to a set of pre-programmed rules, which are considered to be software.

12. Do you feel in your field that the development and maintenance of research software are sufficiently recognised or rewarded?

○ Yes

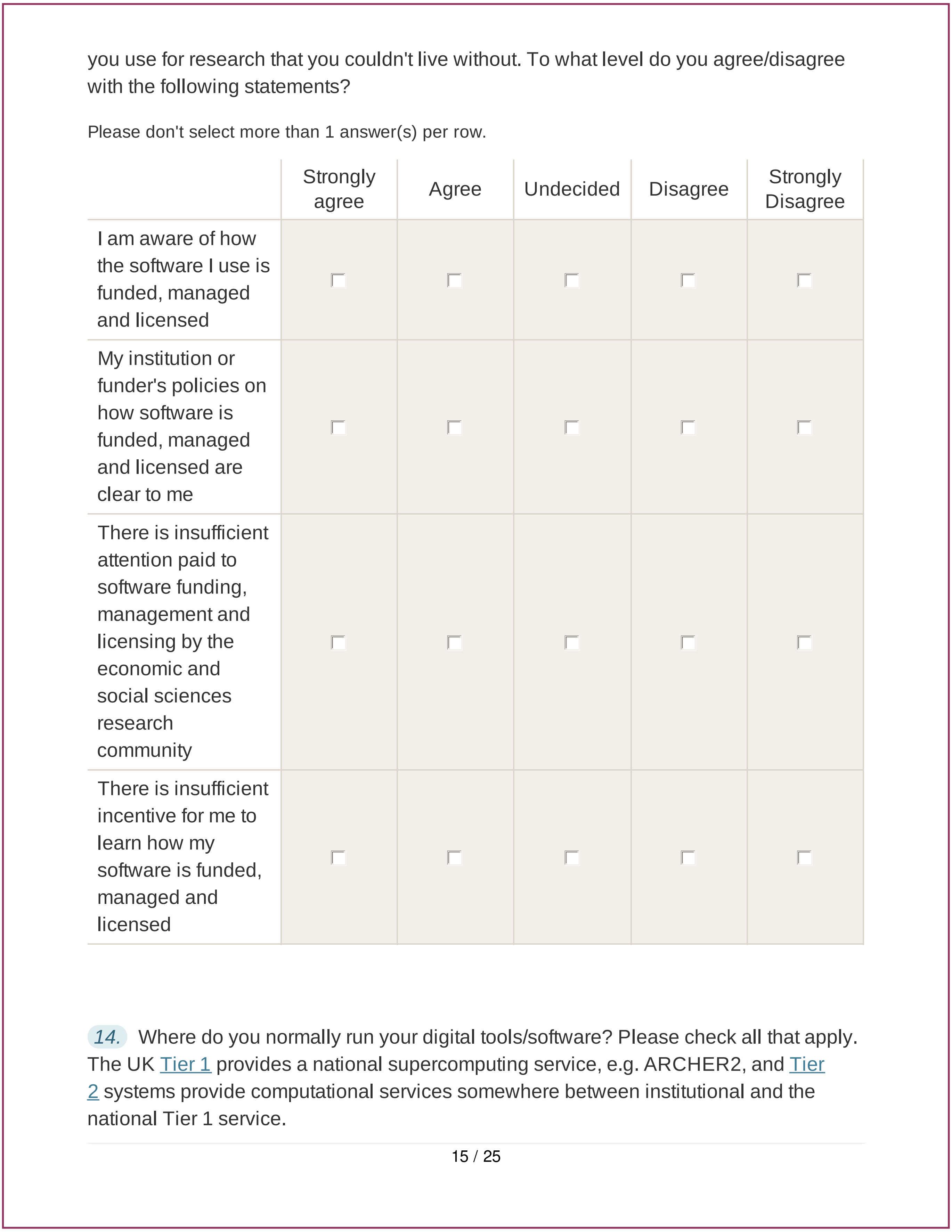
○ No

○ Don't know

12.a. If you answered yes or no can you briefly state your reasons for answering this way

13. When answering this question, think about the most important piece of software

14 / 25

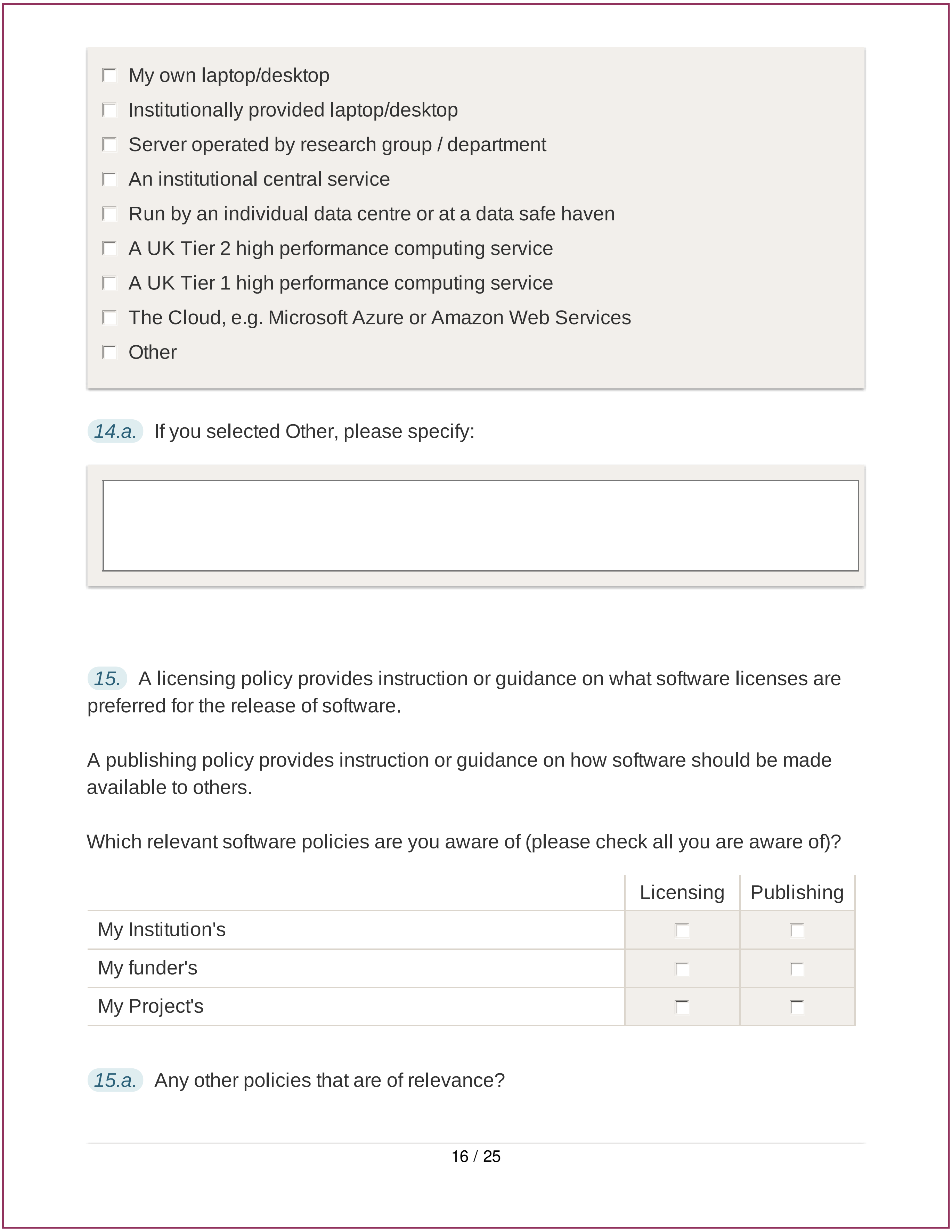


you use for research that you couldn't live without. To what level do you agree/disagree with the following statements?

Please don't select more than 1 answer(s) per row.

	Strongly agree	Agree	Undecided	Disagree	Strongly Disagree
I am aware of how the software I use is funded, managed and licensed	□	□	□	□	□
My institution or funder's policies on how software is funded, managed and licensed are clear to me	□	□	□	□	□
There is insufficient attention paid to software funding, management and licensing by the economic and social sciences research community	□	□	□	□	□
There is insufficient incentive for me to learn how my software is funded, managed and licensed	□	□	□	□	□

14. Where do you normally run your digital tools/software? Please check all that apply. The UK Tier 1 provides a national supercomputing service, e.g. ARCHER2, and Tier 2 systems provide computational services somewhere between institutional and the national Tier 1 service.



- □ My own laptop/desktop
- □ Institutionally provided laptop/desktop
- □ Server operated by research group / department
- □ An institutional central service
- □ Run by an individual data centre or at a data safe haven
- □ A UK Tier 2 high performance computing service
- □ A UK Tier 1 high performance computing service
- □ The Cloud, e.g. Microsoft Azure or Amazon Web Services
- □ Other

14.a. If you selected Other, please specify:

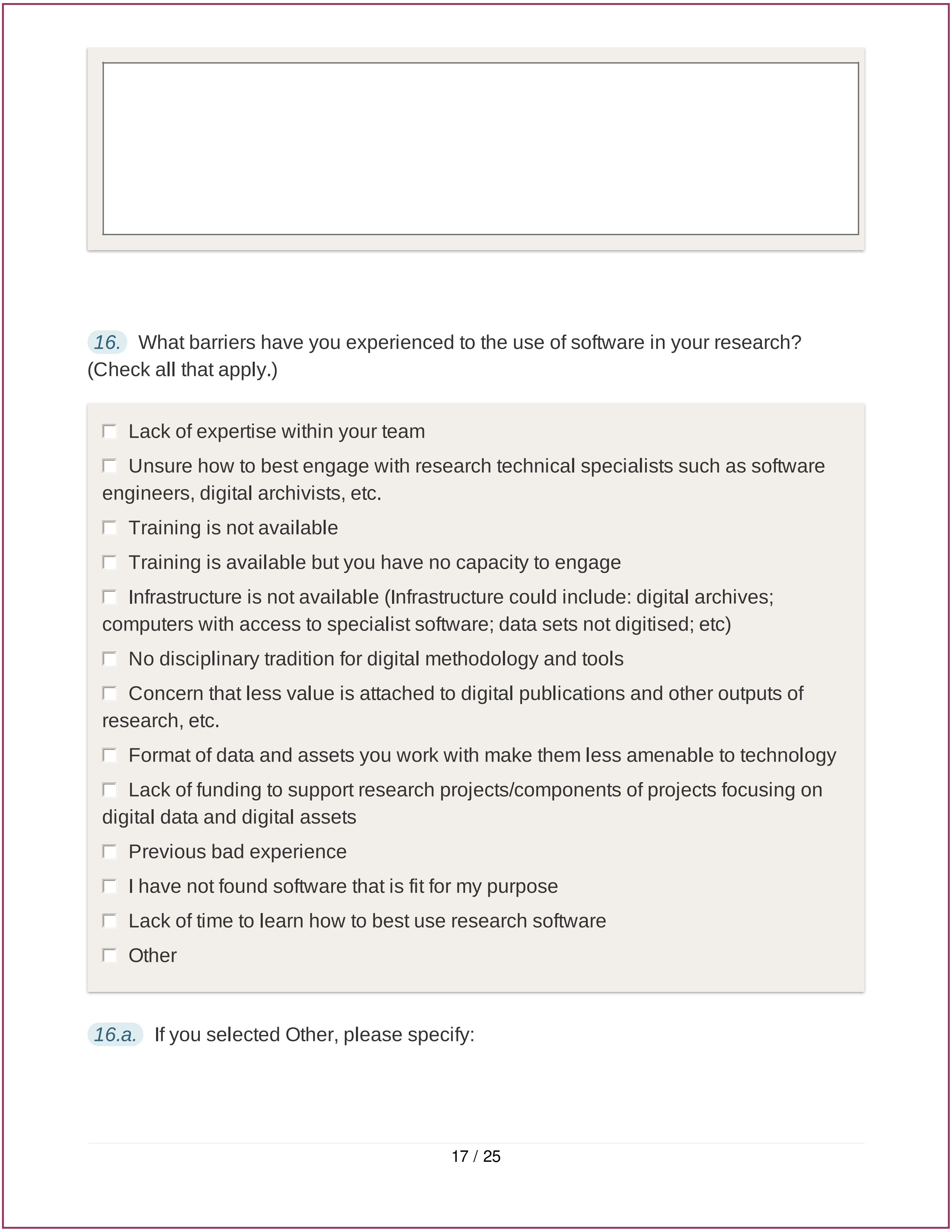
15. A licensing policy provides instruction or guidance on what software licenses are preferred for the release of software.

A publishing policy provides instruction or guidance on how software should be made available to others.

Which relevant software policies are you aware of (please check all you are aware of)?

	Licensing	Publishing
My Institution's	□	□
My funder's	□	□
My Project's	□	□

15.a. Any other policies that are of relevance?



16. What barriers have you experienced to the use of software in your research? (Check all that apply.)

□ Lack of expertise within your team

□ Unsure how to best engage with research technical specialists such as software engineers, digital archivists, etc.

□ Training is not available

□ Training is available but you have no capacity to engage

□ Infrastructure is not available (Infrastructure could include: digital archives; computers with access to specialist software; data sets not digitised; etc)

□ No disciplinary tradition for digital methodology and tools

□ Concern that less value is attached to digital publications and other outputs of research, etc.

□ Format of data and assets you work with make them less amenable to technology

□ Lack of funding to support research projects/components of projects focusing on digital data and digital assets

□ Previous bad experience

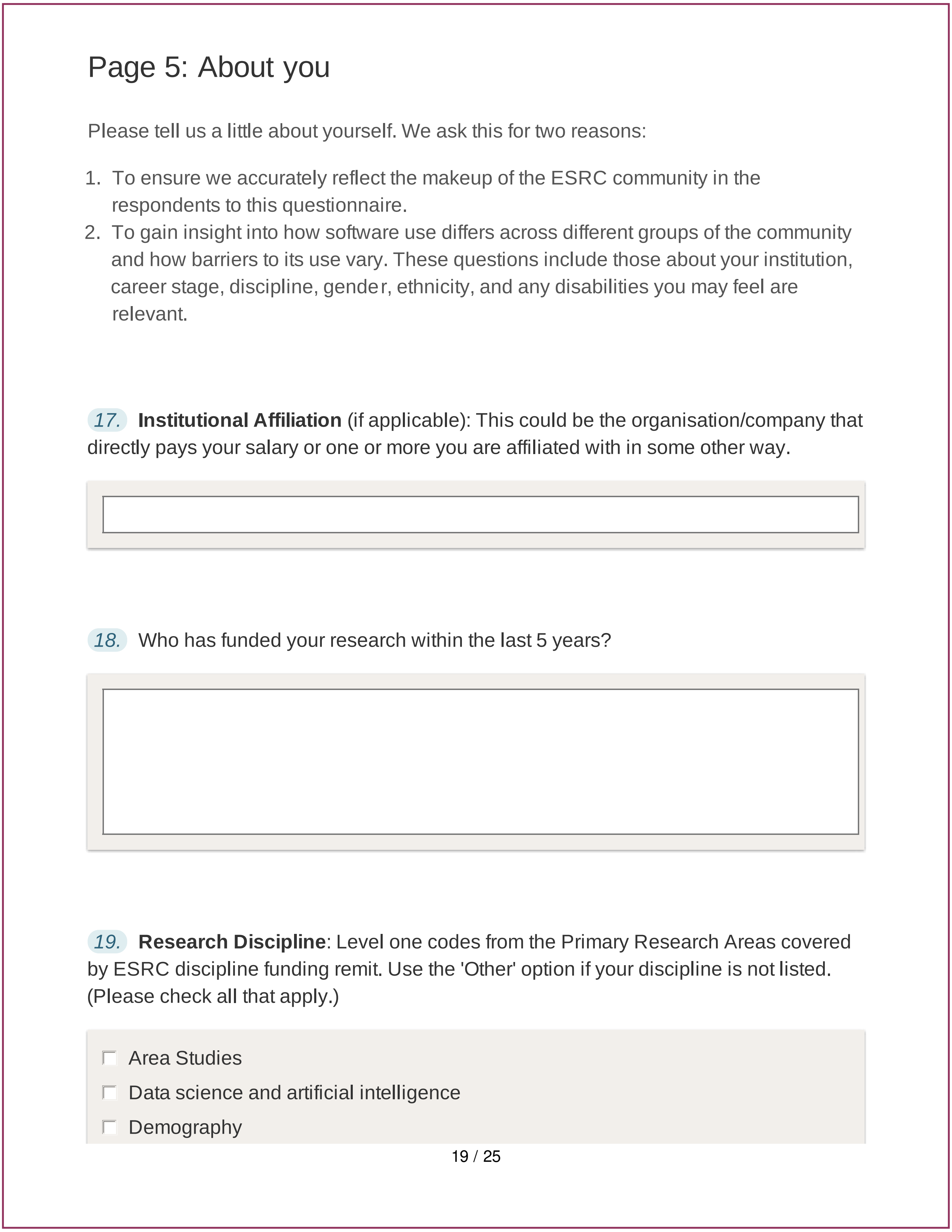
□ I have not found software that is fit for my purpose

□ Lack of time to learn how to best use research software

□ Other

16.a. If you selected Other, please specify:

17 / 25



Page 5: About you

Please tell us a little about yourself. We ask this for two reasons:

1. To ensure we accurately reflect the makeup of the ESRC community in the respondents to this questionnaire.
2. To gain insight into how software use differs across different groups of the community and how barriers to its use vary. These questions include those about your institution, career stage, discipline, gender, ethnicity, and any disabilities you may feel are relevant.

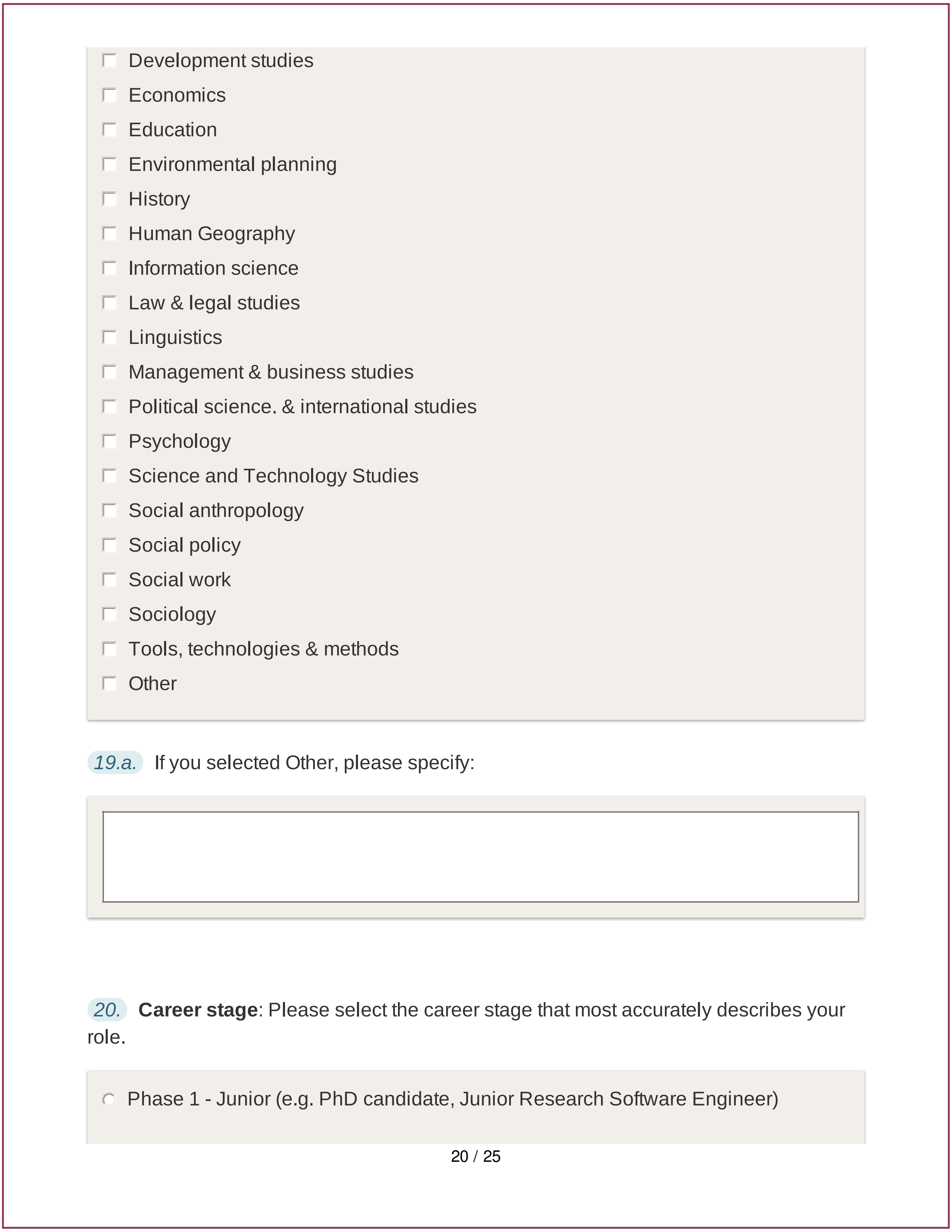
17. **Institutional Affiliation** (if applicable): This could be the organisation/company that directly pays your salary or one or more you are affiliated with in some other way.

18. Who has funded your research within the last 5 years?

19. **Research Discipline**: Level one codes from the Primary Research Areas covered by ESRC discipline funding remit. Use the 'Other' option if your discipline is not listed. (Please check all that apply.)

□ Area Studies
□ Data science and artificial intelligence
□ Demography

19 / 25



□ Development studies

□ Economics

□ Education

□ Environmental planning

□ History

□ Human Geography

□ Information science

□ Law & legal studies

□ Linguistics

□ Management & business studies

□ Political science. & international studies

□ Psychology

□ Science and Technology Studies

□ Social anthropology

□ Social policy

□ Social work

□ Sociology

□ Tools, technologies & methods

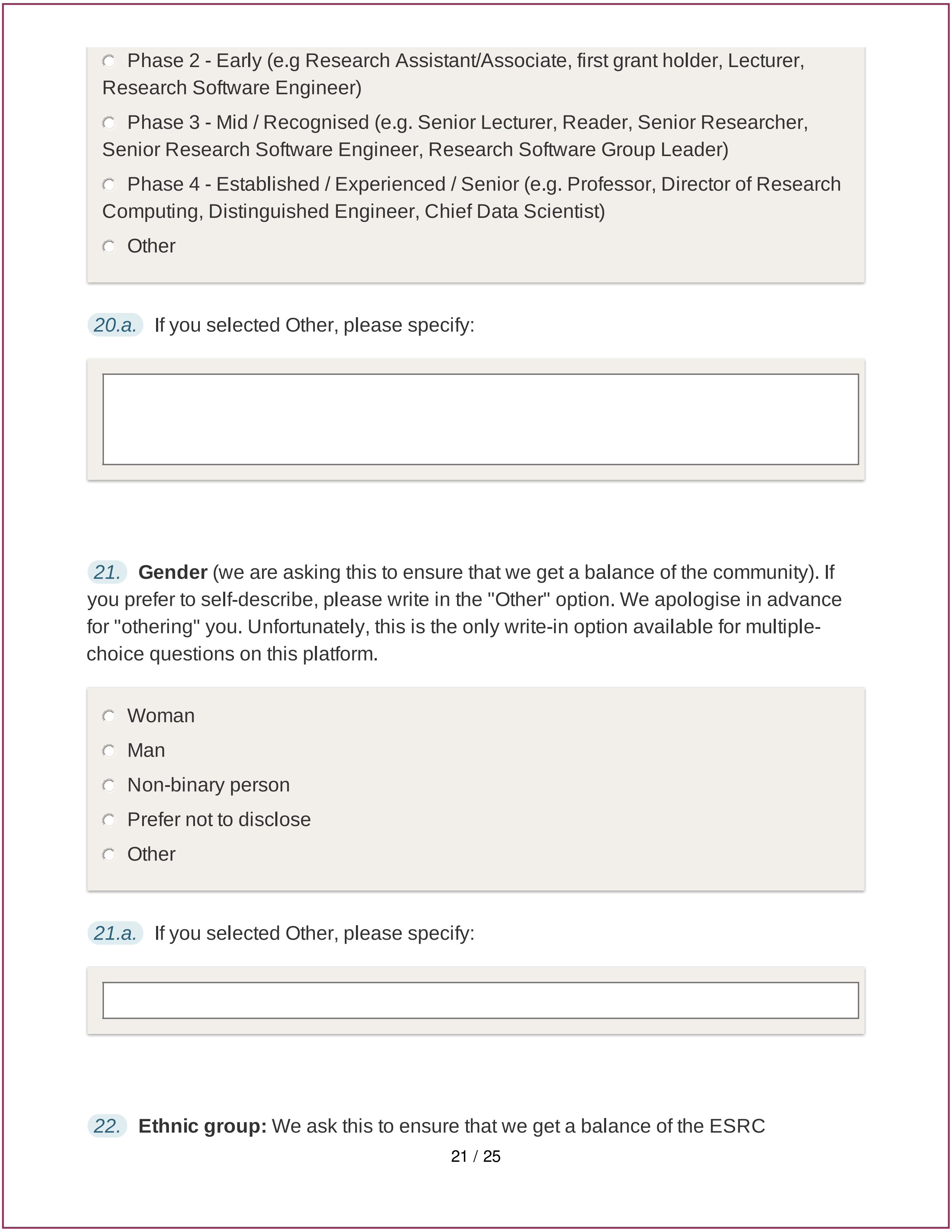
□ Other

19.a. If you selected Other, please specify:

20. **Career stage**: Please select the career stage that most accurately describes your role.

○ Phase 1 - Junior (e.g. PhD candidate, Junior Research Software Engineer)

20 / 25



○ Phase 2 - Early (e.g Research Assistant/Associate, first grant holder, Lecturer, Research Software Engineer)

○ Phase 3 - Mid / Recognised (e.g. Senior Lecturer, Reader, Senior Researcher, Senior Research Software Engineer, Research Software Group Leader)

○ Phase 4 - Established / Experienced / Senior (e.g. Professor, Director of Research Computing, Distinguished Engineer, Chief Data Scientist)

○ Other

20.a. If you selected Other, please specify:

21. **Gender** (we are asking this to ensure that we get a balance of the community). If you prefer to self-describe, please write in the "Other" option. We apologise in advance for "othering" you. Unfortunately, this is the only write-in option available for multiple-choice questions on this platform.

○ Woman

○ Man

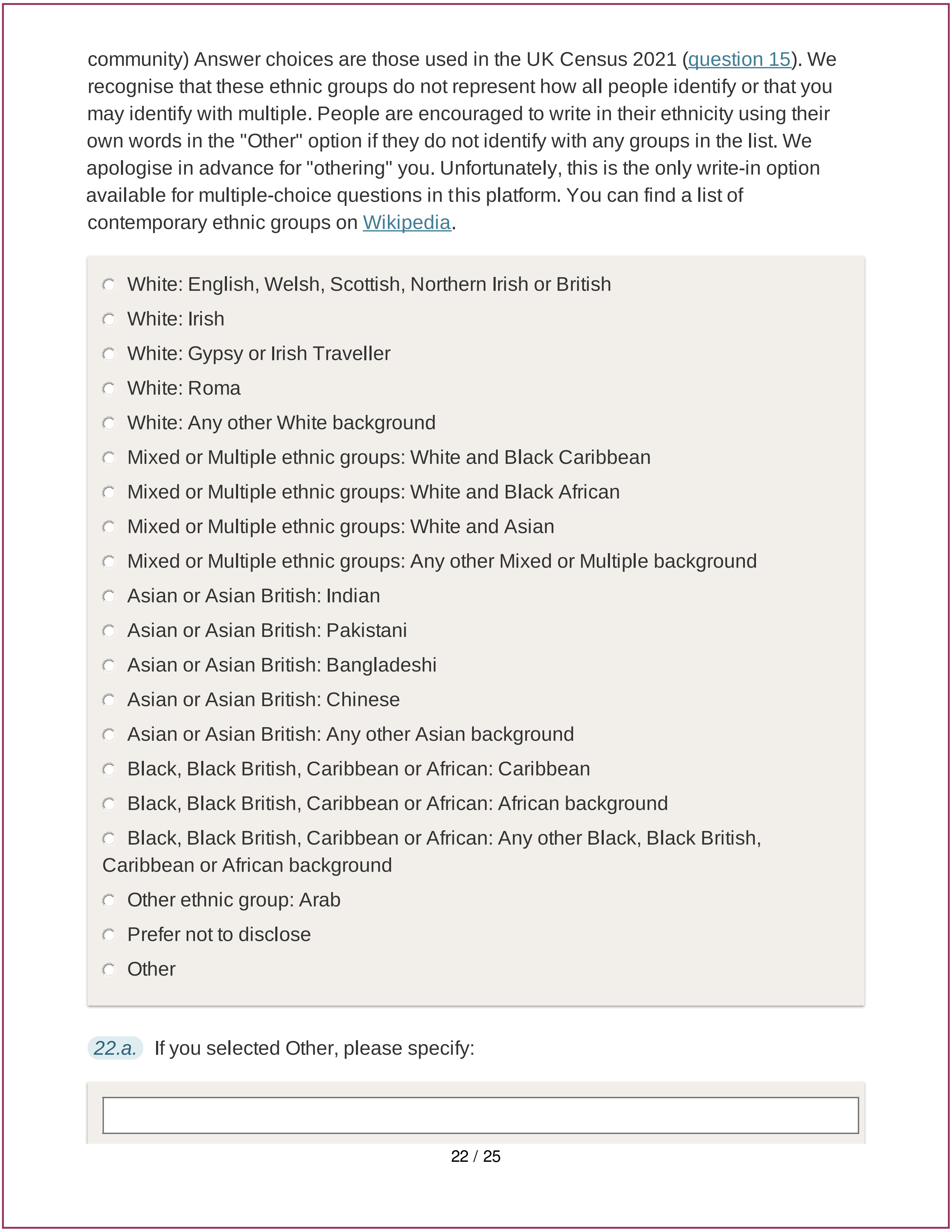
○ Non-binary person

○ Prefer not to disclose

○ Other

21.a. If you selected Other, please specify:

22. **Ethnic group:** We ask this to ensure that we get a balance of the ESRC



community) Answer choices are those used in the UK Census 2021 (question 15). We recognise that these ethnic groups do not represent how all people identify or that you may identify with multiple. People are encouraged to write in their ethnicity using their own words in the "Other" option if they do not identify with any groups in the list. We apologise in advance for "othering" you. Unfortunately, this is the only write-in option available for multiple-choice questions in this platform. You can find a list of contemporary ethnic groups on Wikipedia.

○ White: English, Welsh, Scottish, Northern Irish or British

○ White: Irish

○ White: Gypsy or Irish Traveller

○ White: Roma

○ White: Any other White background

○ Mixed or Multiple ethnic groups: White and Black Caribbean

○ Mixed or Multiple ethnic groups: White and Black African

○ Mixed or Multiple ethnic groups: White and Asian

○ Mixed or Multiple ethnic groups: Any other Mixed or Multiple background

○ Asian or Asian British: Indian

○ Asian or Asian British: Pakistani

○ Asian or Asian British: Bangladeshi

○ Asian or Asian British: Chinese

○ Asian or Asian British: Any other Asian background

○ Black, Black British, Caribbean or African: Caribbean

○ Black, Black British, Caribbean or African: African background

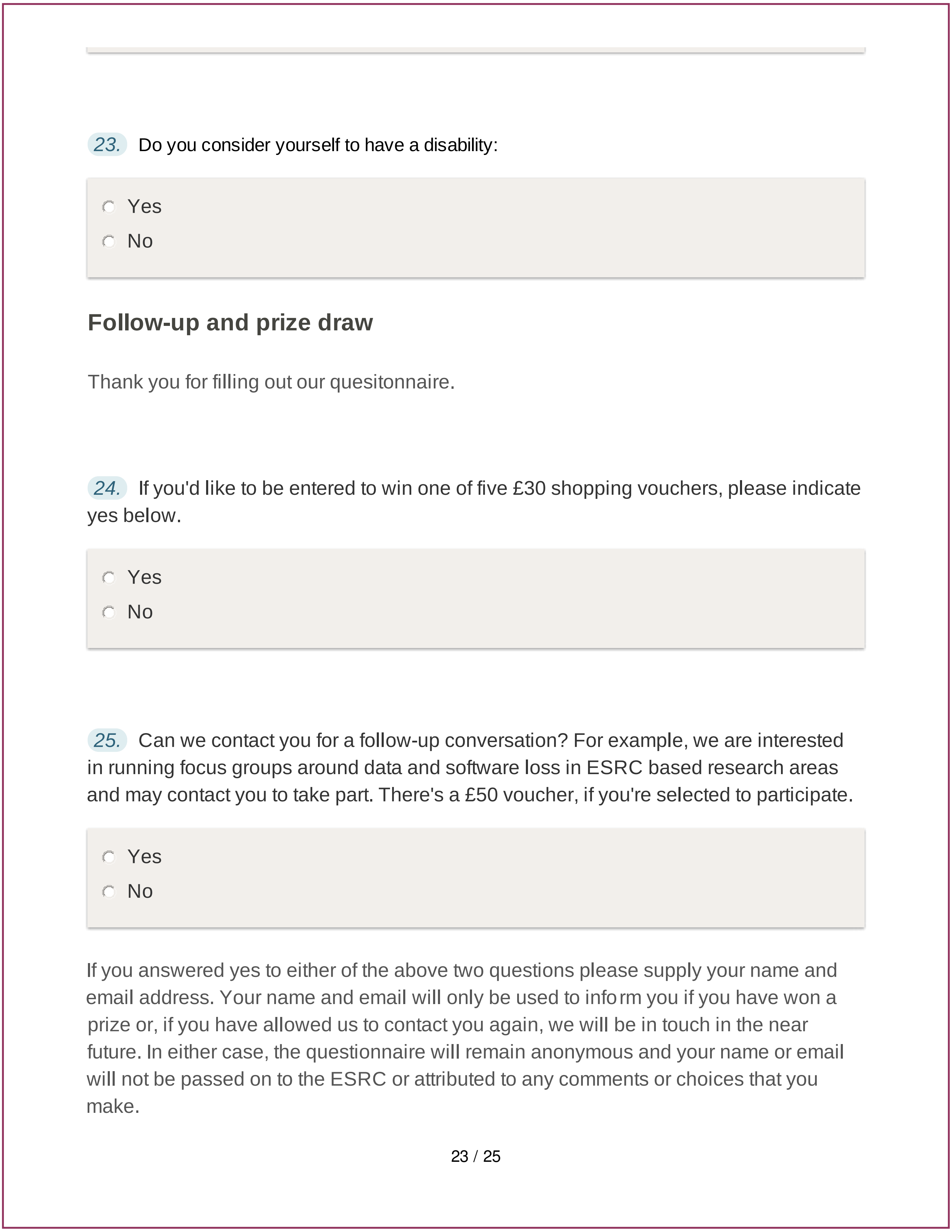
○ Black, Black British, Caribbean or African: Any other Black, Black British, Caribbean or African background

○ Other ethnic group: Arab

○ Prefer not to disclose

○ Other

22.a. If you selected Other, please specify:



23. Do you consider yourself to have a disability:

○ Yes
○ No

Follow-up and prize draw

Thank you for filling out our quesitonnaire.

24. If you'd like to be entered to win one of five £30 shopping vouchers, please indicate yes below.

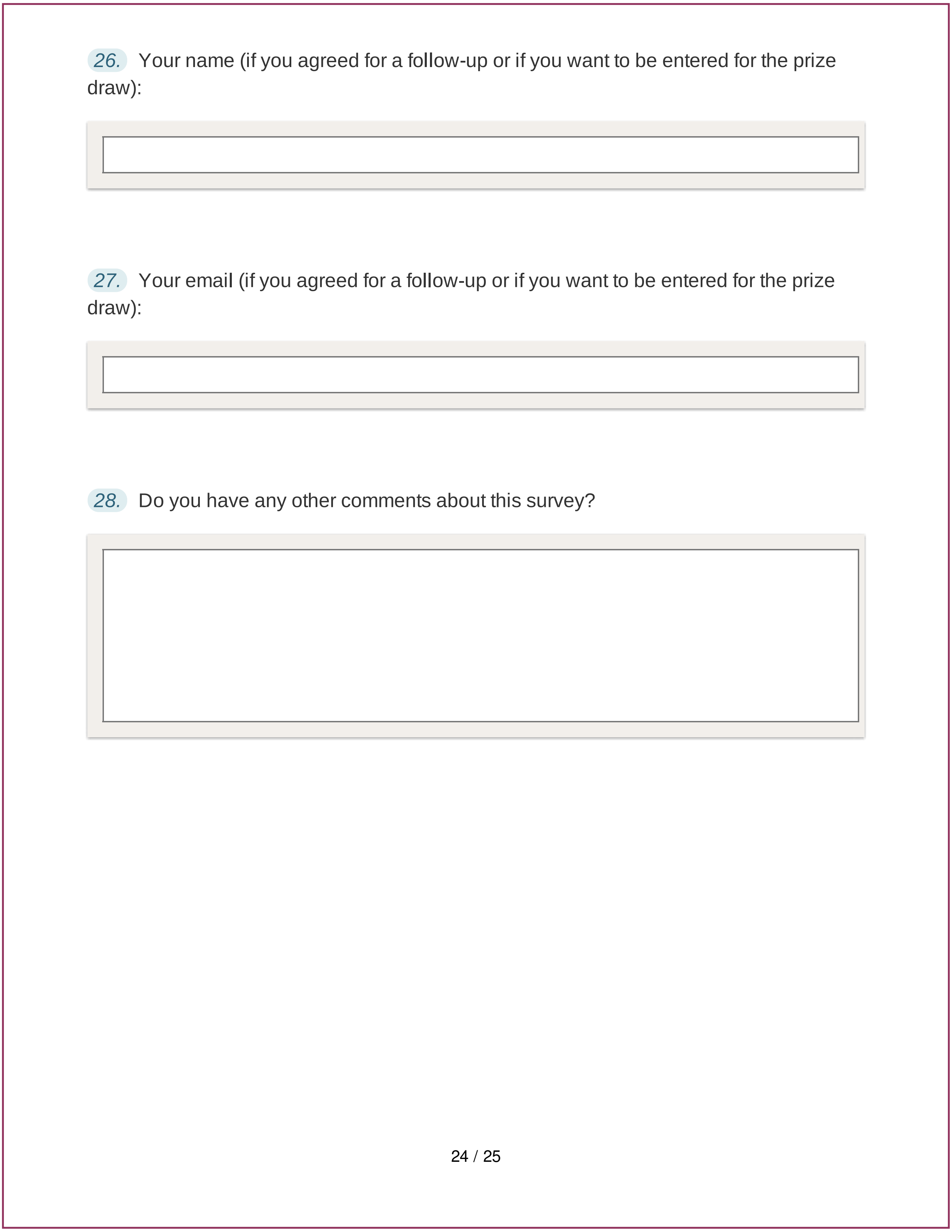
○ Yes
○ No

25. Can we contact you for a follow-up conversation? For example, we are interested in running focus groups around data and software loss in ESRC based research areas and may contact you to take part. There's a £50 voucher, if you're selected to participate.

○ Yes
○ No

If you answered yes to either of the above two questions please supply your name and email address. Your name and email will only be used to inform you if you have won a prize or, if you have allowed us to contact you again, we will be in touch in the near future. In either case, the questionnaire will remain anonymous and your name or email will not be passed on to the ESRC or attributed to any comments or choices that you make.

23 / 25

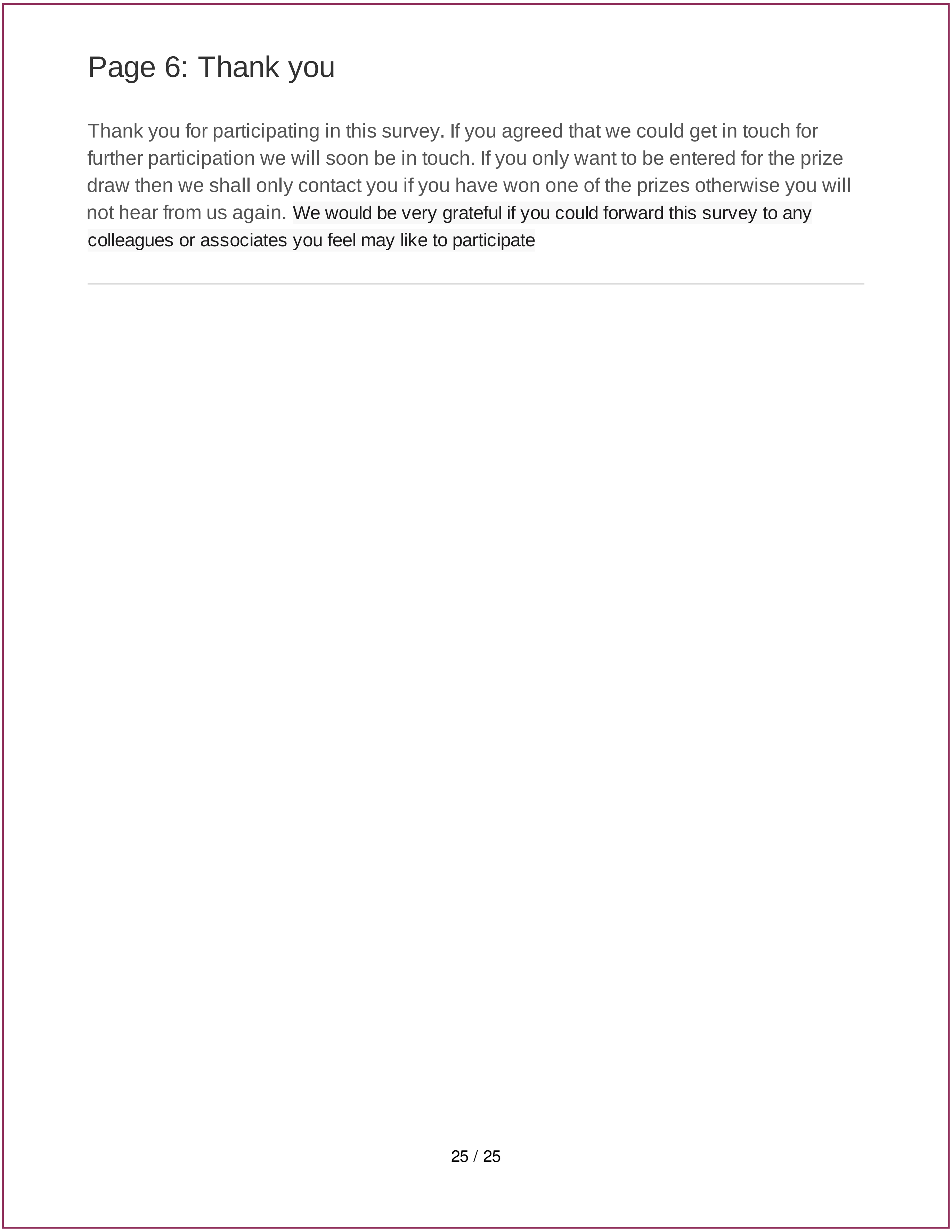


26. Your name (if you agreed for a follow-up or if you want to be entered for the prize draw):

27. Your email (if you agreed for a follow-up or if you want to be entered for the prize draw):

28. Do you have any other comments about this survey?

24 / 25



Page 6: Thank you

Thank you for participating in this survey. If you agreed that we could get in touch for further participation we will soon be in touch. If you only want to be entered for the prize draw then we shall only contact you if you have won one of the prizes otherwise you will not hear from us again. We would be very grateful if you could forward this survey to any colleagues or associates you feel may like to participate

25 / 25

APPENDIX 2. INTERVIEW TOPICS

The script below was used to guide the interview through the process.

RESEARCHER

Thank you for joining us today. Can you hear me? We are going to record this session, and as we need to record your verbal consent, **ask for consent to record and then start the recording**. The recording will be used to create a pseudonymised transcript.

So today, we're going to have a 45 minute conversation, which will take us through to [time now + 45mins]. **Ask if the participant has any time constraints.**

Hopefully, you have some idea of what we will cover from our previous communication, but just to go over it: first of all, we will ensure we have your consent to collect data; then we will dive straight into the questions about data and software - what you are using, what kind of workflows you are creating, how you are acquiring training and what the key barriers to software use are, and what support you receive from your institution and ESRC, as appropriate. OK with that?

Obviously, we would like to thank you for your participation today, and we'll email you a Love2Shop voucher for £50 in the next couple of weeks.

Consent

We're going to collect your consent verbally today. I'm going to read some statements and if you are happy to give your consent, please just say yes after each statement. If you're not happy at any point, you can say no and withdraw from the interview. Please do ask if you're not clear on any of the consent statements. We have six statements, because we need you to consent to both current and future use.

1. You should have received a copy of the participant information sheet and a consent form with your original email. The first statement is that **you have read and understood the participant information sheet and have had the opportunity to ask questions about the study**.

   Can you indicate verbally that you agree or identify if you don't

2. The next statement is that **you agree to take part in this research project and that you agree for your data to be used for the purpose of this study**.

- Can you indicate verbally that you agree or identify if you don't

3. Now we ask you to state that **you give permission for the pseudonymised transcript of this interview to be deposited in the researcher's data space (UoE SharePoint) and repository (DataShare) as described in the Participant Information Sheet so it can be used for future ethically approved research and learning on the Economics and Social Sciences research community. This is so there is explicit consent for the data to be reused**.

- Can you indicate verbally that you agree or identify if you don't

4. The next statement is, **you understand your participation is voluntary and you can withdraw (at any time) for any reason without my participation rights being affected**.

- Can you indicate verbally that they agree or identify if they don't

5. The next statement is related, which is that **you understand that you can withdraw from this study up to a week after the date of the interview. You understand that should you withdraw from the study after this date, then the information collected about you up to this point may still be used for the purposes of achieving the objectives of the study only**.

- Can you indicate verbally that you agree or identify if you don't

6. Next, **you understand that you may be quoted directly in reports of the research but that you will not be directly identified (e.g. that your name will not be used and identifiable details will be changed)**.

- Can you indicate verbally that you agree or identify if you don't

INTERVIEW

Can you briefly state your background, your discipline and your (or your group's) research interests in particular what data sets and software are used.

Establish the diversity of community

First of all, we're interested in whether there are barriers that particularly affect people with certain experiences and who identify as having certain characteristics. We know, for instance, that disability, ethnicity and gender can create barriers. Do you feel any of these are relevant to you?

What data sets are being collected/re-used?

- How do you understand the data set term?
  - What does it mean to you in your research?
  - What form is it in when used in your research?

  This applies to all research projects you undertake, not just ESRC-funded ones.
- Our working definition of data is [Data as we mean it is anything digital (including non-digital data when stored in a digital format) that is collected and/or used for research, and the outputs of that research, including publications].
- Do you use any of the major ESRC datasets [specify here]
- How do you collect it? Share it? Store it? during or/and at the end of your research projects?
- Have you ever deposited data?
  - Where?
  - When and what terms do you use (archives repositories?)
- Have they had to write a Data Management Plan?
  - Use/devise any other data use protocols?
- What have been the challenges for data management in recent projects?

What software is being used in ESRC projects

- Do you use any software in your research?
  - What types?
- Look at this list [of software/data tools] (included at the end of this appendix) which of these types do you use in your research? The examples are just for illustration, you might use a different actual software product to those we've suggested.
- If not, what are the factors that stop you from using software
- How do you find the people who you collaborate with to create/work with software?
- How is it used, shared, and stored? How do you access it? (Software as a service, on a laptop?) How did you first find it?
- Where do you run your software? E.g.
  - My own laptop/desktop
  - Institutionally provided laptop/desktop
  - Server operated by research group / department
  - An institutional central service
  - Run by an individual data centre or at a data safe haven
  - A UK Tier 2 high performance computing service
  - A UK Tier 1 high performance computing service
  - The Cloud, e.g. Microsoft Azure or Amazon Web Services
  - Other
- What are the benefits to your work of moving to a different platform to run your software?
  - Are there any barriers to you doing that?

How software is currently being supported and maintained

- If yes, is it off the shelf or bespoke –
  - Who developed it? Someone in your project?

    - Someone elsewhere in your institution?
- What happens to it after the project?
    - Have you included it in a Software Management Plan or equivalent?
    - Have you published it for use by other people?
    - If so, where did you publish it (what repository, for eg) and what license did you use?
    - What are the challenges of developing technological tools/software/apps in your research? (eg lack of dedicated funding, deprecation of components, no time to support)

Which software training is being taken up [make this more general about training?]

- Thinking about the software you have just spoken about, how and when have you learned how to use these tools?
    - Follow up-to understand where they pick up skills, e.g. self-led learning, courses, peers and colleagues, workshops and conferences, etc [There's a list in the survey]
    - Drill down into how self-led learning actually works
- Have you attended any NCRM (National Centre for Research Methods) software training?
    - If so, you can ask the interviewee for details.
- How do you think training in the future should be organised?
    - Prompt: Which are the most effective?
    - Prompt: Do you have any ideas for improving software training that you'd like to share? Ask them to describe, e.g. how should it be financed, resources, what format?

Identifying priority barriers

- Which do you think are the most important barriers to the use of software? Here is our list of prompts (also at the end of this appendix) -
    - Lack of expertise within your team
    - Unsure how to best engage with research technical specialists such as software engineers, digital archivists, etc.
    - Training is not available
    - Training is available but you have no capacity to engage
    - Infrastructure is not available (Infrastructure could include: digital archives; computers with access to specialist software; data sets not digitised; etc)
    - No disciplinary tradition for digital methodology and tools
    - Concern that less value is attached to digital publications and other outputs of research, etc.
    - Format of data and assets you work with make them less amenable to technology
    - Lack of funding to support research projects/components of projects focusing on digital data and digital assets
    - Previous bad experience
    - I have not found software that is fit for my purpose
    - Lack of time to learn how to best use research software

Identifying work patterns/workflows

Ask people to walk us through a day and identify what data and software they use. Prompt questions; where did you turn for help; why did you make that choice (institutional/personal recommendation); where do you encounter barriers and what did you do to resolve them (or not)

Identify key possible interventions

- What is your view on ESRC and support available via its funding for research infrastructure and digital and data skills, including software management and development?
- What about your institution? What structures are in place to support data and software management? Is there support provided for writing DMPs?
- Where do you find other sources of support?
- If you require any bespoke software, would you know how to access or request this?
- If you were putting in a bid in a month's time, where would you go for support regarding the software aspects and how would you understand what was being asked of you?
- Another question is about what training ESRC should be providing,

- What is missing, what do you need, what training would be useful?
- How and when should that training be delivered?
- What would you like to see the ESRC do to facilitate good research practice/innovation around software?

Table 17.

MIght want to ask at the end, has this interview made you think about your own practice in any way?

Are there any things you think the ESRC should be made aware of regarding software?

List of the types of software

These were generally read out if the interviewee required prompting.

- Animation and Storyboarding, e.g. Scratch, Storyteller
- AudioTools, e.g. Music Algorithms, Paperphone
- Authoring and publishing tools, e.g. Twine, Oppia
- Code versioning, e.g. GitHub
- Content Management Systems (CMS), e.g. WordPress, Mura
- CrowdSourcing, e.g. AllOurIdeas
- Exhibition/Collection Tools, e.g. Omeka, Neatline
- Internet Research Tools, e.g. Google tools or Wikipedia tools
- Machine Learning and Artificial Intelligence, e.g. leximancer
- Mapping Tools and Platforms, Geographic Information Systems, e.g. QGIS, CartoDB, ArcGIS
- MindMapping Tools, e.g. DebateGraph
- Network Analysis, e.g. GEPHI
- Programming Languages, e.g. Python, MATLAB
- Qualitative analyses, e.g. NVivo
- Simulation Tools, e.g. NetLogo
- Spreadsheets, e.g. Excel, Google Sheets
- Statistical analysis, e.g. R, SPSS, Stata, SAS
- Text Analysis Tools, e.g. Voyant, Linguistic Corpuses, Entity Recognizers
- Text Collation Tools, e.g. Juxta Commons
- Text Encoding, e.g. Oxygen XML
- Text and Data Wrangling, e.g. Overview, OpenRefine
- Topic Modelling, e.g. Leximancer
- Transcription services, e.g. otter.ai
- Video and Film Analysis, e.g. Cinemetrics
- Visualisation Tools, e.g. D3.js, Tableau
- Other

Barriers to the use of software in your research

Like in the previous question these were read out if the interviewee required some suggestions:

- Lack of expertise within your team
- Unsure how to best engage with research technical specialists such as software engineers, digital archivists, etc.
- Training is not available
- Training is available but you have no capacity to engage
- Infrastructure is not available (Infrastructure could include: digital archives; computers with access to specialist software; data sets not digitised; etc)
- No disciplinary tradition for digital methodology and tools
- Concern that less value is attached to digital publications and other outputs of research, etc.
- Format of data and assets you work with make them less amenable to technology
- Lack of funding to support research projects/components of projects focusing on digital data and digital assets
- Previous bad experience
- I have not found software that is fit for my purpose
- Lack of time to learn how to best use research software

APPENDIX 3. LIST OF SOFTWARE TOOLS MENTIONED IN THE SURVEY

SOFTWARE MENTIONED IN THE SURVEY.

Access

- Website: https://www.microsoft.com/en-gb/microsoft-365/access
- Purpose: Database management system for storing and manipulating data.
- Licensing: Commercial, part of the Microsoft Office365 suite.
- Funding method:
- Issues:

Atlas.ti

- Website: https://atlasti.com/
- Purpose: Qualitative analysis
- Licensing: Commercial
- Funding method:
- Issues:

Audacity

- Website: https://www.audacityteam.org/
- Purpose: Manipulation and creation of audio files
- Licensing: GPL-2 (or later)
- Funding method:
- Issues:

CLAN (Computerized Language ANalysis)

- Website: https://dali.talkbank.org/clan
- Purpose: creating and analyzing transcripts in the Child Language Exchange System
- Licensing:GPL-2
- Funding method:
- Issues:

Elan

- Website: https://archive.mpi.nl/tla/elan
- Purpose: Annotate & Transcribe audio & video
- Licensing:
- Funding method: Free
- Issues:

Endnote

- Website: https://endnote.com/
- Purpose: Reference management
- Licensing: Commercial
- Funding method:
- Issues:

Excel

- Website: https://www.microsoft.com/en-gb/microsoft-365/excel
- Purpose: Spreadsheets
- Licensing: Commercial, part of the Microsoft Office365 suite.
- Funding method:
- Issues:

Express Scribe

- Website: https://www.nch.com.au/scribe
- Purpose: Transcription
- Licensing: Commercial
- Funding method:
- Issues:

Google Docs

- Website:https://docs.google.com/
- Purpose: word processing, collaborative document writing
- Licensing:
- Funding method: Free but requires a google account
- Issues:

Happy Scribe

- Website: https://www.happyscribe.com
- Purpose: Transcription and subtitles
- Licensing: Commercial
- Funding method:
- Issues:

HotGlue

- Website: https://hotglue.org/
- Purpose: a Content Manipulation System which allows to construct websites directly in a web-browser
- Licensing: GPL-3
- Funding method:
- Issues:

Leximancer

- Website: https://www.leximancer.com/
- Purpose: quantitative content analysis using machine learning
- Licensing: Commercial
- Funding method:
- Issues:

MATLAB

- Website: https://uk.mathworks.com/products/matlab.html
- Purpose: Programming platform
- Licensing: Commercial
- Funding method:
- Issues:

Microsoft Teams

- Website: https://www.microsoft.com/en-gb/microsoft-teams
- Purpose: Communications
- Licensing: Commercial
- Funding method:
- Issues:

MindGenius

- Website: https://www.mindgenius.com/

- Purpose: mind mapping and project management
- Licensing: Commercial
- Funding method:
- Issues:

Miro boards

- Website: https://miro.com
- Purpose: Collaboration / White boards
- Licensing: Commercial
- Funding method:
- Issues:

NVivo

- Website: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software
- Purpose: Qualitative data analysis
- Licensing: Commercial
- Funding method:
- Issues:

Otter.ai

- Website: https://otter.ai
- Purpose: Transcription
- Licensing: Commercial
- Funding method:
- Issues:

Praat

- Website: https://www.fon.hum.uva.nl/praat/
- Purpose: Speech analysis
- Licensing: GPL-3
- Funding method: Free
- Issues:

PsychoPy

- Website: https://www.psychopy.org/
- Purpose: Experiments in Behavioural Science
- Licensing: GPL-3
- Funding method: Free
- Issues:

Python

- Website: https://www.python.org/
- Purpose: Programming
- Licensing: Various (GPL compatible)
- Funding method: Free
- Issues:

QGIS

- Website: https://www.qgis.org/
- Purpose: Open Source Geographic Information System
- Licensing: Free

- Funding method:
- Issues:

Qualcoder

- Website: https://qualcoder.wordpress.com/
- Purpose: Qualitative analysis tool written in Python
- Licensing: MIT
- Funding method: Free
- Issues:

Qualtrics

- Website: https://www.qualtrics.com/uk/
- Purpose: surveys
- Licensing: Commercial
- Funding method:
- Issues:

Quirkos

- Website: https://www.quirkos.com/
- Purpose: Qualitative analysis
- Licensing: Commercial
- Funding method:
- Issues:

R

- Website: https://cran.r-project.org/
- Purpose: Programming
- Licensing: GPL-2 and GPL-3
- Funding method: Free
- Issues: Learning curve

Rayyan

- Website: https://www.rayyan.ai/
- Purpose: Organisation of collaborative literature reviews.
- Licensing: Commercial
- Funding method:
- Issues:

Recogito

- Website: https://recogito.pelagios.org/
- Purpose: an annotation platform for places.
- Licensing: Apache-2.0
- Funding method: Free
- Issues:

SPSS

- Website: https://www.ibm.com/uk-en/products/spss-statistics
- Purpose: Statistical software suite
- Licensing: Commercial
- Funding method:
- Issues:

Stata

- Website: https://www.stata-uk.com/software/stata.html
- Purpose: Programming platform
- Licensing: Commercial
- Funding method:
- Issues:

Trint

- Website: https://trint.com/
- Purpose: Transcription
- Licensing: Commercial
- Funding method:
- Issues:

Word

- Website: https://www.microsoft.com/en-gb/microsoft-365/word
- Purpose: word processing
- Licensing: Commercial, part of the Microsoft Office365 suite.
- Funding method:
- Issues:

Zoom

- Website: https://zoom.us/
- Purpose: Communications
- Licensing: Commercial
- Funding method:
- Issues:

Zotero

- Website: https://www.zotero.org/
- Purpose: Reference management
- Licensing:
- Funding method: Free
- Issues:

Software
Sustainability
Institute

www.software.ac.uk