

Everyday Voices as Big Data: A Call for the Secondary Analysis of Large-Scale Qualitative Interview Data

Sociology

1–18

© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/00380385251344472

journals.sagepub.com/home/soc



Jane Elliott 

London School of Economics, UK

Carrie Friese 

London School of Economics, UK

Gaby Harris 

Manchester Metropolitan University, UK

Elizabeth Mann 

London School of Economics, UK

Mike Savage 

London School of Economics, UK

Abstract

This article reflects on the paradox that although research using qualitative interviews has developed sophisticated repertoires of data collection, it has not fully embraced secondary analysis and has struggled to address questions of representativeness. This contrasts with quantitative social science where this is now routine. We discuss recent innovations associated with the 'big qual' approach to assembling data from existing qualitative interview studies but argue for a development that champions secondary analysis of qualitative interview data collected from larger, more representative samples. We reflect on precedents for this approach from the 1958 British Birth Cohort Study and, more recently, the American Voices Project. We draw out the unrealised possibilities of secondary analysis, enhanced by recent affordances of computational social science. We argue that the wider deployment of secondary analysis will expand the appeal of qualitative research for policy audiences and contest the hegemony of quantitative survey analysis.

Keywords

big data, qualitative interview data, research design, secondary analysis, transferability

Corresponding author:

Jane Elliott, International Inequalities Institute, London School of Economics, Houghton St, London, WC2A 2AE, UK.

Email: B.J.Elliott@LSE.ac.uk

Introduction

In the past 50 years, research based on in-depth or semi-structured interviews has been the wellspring of some of the most inspiring work in British sociology. This research lay behind the rapid expansion of sociology during the 1950s and 1960s, as marked by the iconic studies of Bott (1956), Young and Willmott (1958), Townsend (1962), Stacey (1970), and Oakley (1974). These methods continue to be popular because of their success in eliciting hidden and marginalised views, in critiquing mainstream perspectives and offering a vision for sociology as a ‘critical’ discipline (Savage, 2010).

However, this path-breaking tradition of qualitative research has sadly lost ground to the analysis of large-scale representative survey data, which has become dominant across social scientific and policy research communities (Jerrim and De Vries, 2017). This mainly results from quantitative researchers’ claims for the generalisability and transferability of their findings, which is underscored by investment in nationally representative survey data, the widespread adoption of secondary analysis and most recently the availability of digital administrative data for social science research (Connelly et al., 2016; Dale et al., 2008; Deluca, 2023; Payne and Williams, 2005). By contrast, the secondary analysis of data from qualitative interview studies remains relatively muted (Bishop and Kuula-Luumi, 2017), which holds back the vital need to bring less prominent voices to the fore.

This article seeks to rectify this by building momentum for the collection of qualitative interview data at scale and further advocates the value of secondary qualitative analysis. We suggest that advances in Machine Learning and Large Language Models (LLMs) might be able to *assist* in navigating these new ‘big data’ resources (Bonikowski and Nelson, 2022; Franzosi, 2021) – though we emphasise the need for care in these initiatives. The application of Natural Language Processing (NLP) and LLM in qualitative sociological research is a comparatively new development. In November 2020 a search for the term ‘language model’ in major US social scientific and sociological journals found no references related to probabilistic LLMs (Jensen et al., 2022); we expect this to change, offering exciting new possibilities for secondary qualitative analyses.

We place our argument in the wider context of the dramatic transformation of the data landscape in the past half-century, associated with the proliferation of digital infrastructures. We do not subscribe to epochal claims that the ‘information age’ sweeps all before it. But nor can it be ignored. So far, its impact has been uneven. In quantitative social science the secondary analysis of survey data has become utterly routinised and has been supplemented by easier access to administrative and transactional data, either alongside or matched to survey data. To offer one such example, the UK’s longitudinal survey Understanding Society was downloaded from the UK Data Service 8202 times in the year April 2022–March 2023, its associated COVID-19 study a further 2765 times that same year (UK Data Service, 2024: 63). Freed from the burden of collecting their own data, many quantitative researchers can focus full attention on data analysis.

By contrast, it is striking that the proliferation of digital infrastructures has hitherto not involved the radical transformation of qualitative interview studies. These typically remain centred around an expert researcher interviewing respondents chosen largely by some mix of purposive and convenience sampling. There is no doubting that qualitative

interviews continue to play a vital role in social research, as highlighted by Edwards and Holland (2020). However, we believe that qualitative research should have even greater reach. As we discuss below, the ‘bespoke’ nature of many qualitative studies has not driven significant efforts for secondary analysis of interview data, even with the increased focus on archiving and data availability (Bishop and Kuula-Luumi, 2017).

Meanwhile, one question often asked of qualitative research is whether evidence is sufficiently secure or ‘reproducible’ to influence policy and legislation (Guba and Lincoln, 2005); that is, whether it is transferable across contexts and has ‘*moderatum* generalisability’ (Payne and Williams, 2005)?¹ Debates on whether qualitative research should be assessed by quantitative criteria are extensive (Tracy, 2010). Nevertheless, the importance of transferability for impactful qualitative research is well established (Daniel, 2018; Deluca, 2023; Lincoln and Guba, 1985). Indeed, transferability and ‘requisite variety’ has been argued to ensure rigour and resonance (Weick, 2007), ontologically, conceptually and epistemologically shaping how data are generated and analysed (Collins et al., 2024; Tracy, 2010).

We believe the social and political power of qualitative interview studies needs to be better recognised. Moreover, the increased reuse of qualitative interview data can amplify the reach and impact of the original inquiries. We further argue that social scientific research could be advanced and supported by a large-scale qualitative interview resource explicitly designed for secondary analysis, providing reassurance of rigour and resonance both for secondary studies, but also as a resource to demonstrate the wider transferability of the findings of smaller-scale studies.

In this article we champion the exciting potential for such larger-scale qualitative interview analysis. We first draw out the contingent historical factors that made secondary analysis more common for survey rather than in-depth interview data, refuting suggestions that quantitative analysis is somehow necessarily more robust. Second, we consider lessons to be learnt from the Timescapes initiative, the most impressive effort to develop ‘big qual’, arguing that its focus on ‘assemblage’ needs to be pushed further to ensure the transferability of qualitative evidence. Third, we discuss the recent development of initiatives to collect more representative qualitative interview data, notably the American Voices Project (AVP). Finally, we sketch out the implications of advances in NLP and LLM for facilitating analysis of much larger-scale qualitative resources. Although our focus is predominantly on qualitative interview data within a UK context, our arguments have wider international resonances.

The Differential Take-up of Secondary Analysis in Quantitative and Qualitative Methods – and Why This Matters

The development of qualitative interview methods in UK social science is now well known. Before 1945, interviews were generally used as an adjunct to ‘social surveys’, often as part of broader ethnographic community studies (Bulmer et al., 1991; see also Lee (2004) and Platt (2002) on the Chicago School). Examples include the poverty studies, famously conducted by Booth and Rowntree, where interviews took place with key

informants such as school attendance officers. These figures were seen to possess expert knowledge. Where qualitative interviews were conducted with various kinds of subaltern populations, their testimony was often treated sceptically, and sometimes deployed as part of a sensationalist ‘exposure’, as with Henry Mayhew’s famous journalistic investigations of poverty (Mayhew, 1985 (orig. 1865)).

It was only after the Second World War that the qualitative interview, as a singular method, detached from larger contextualising community research projects, became a major repertoire in British social science (see Savage, 2008, 2010). The idea of the ‘depth interview’ was initiated in Freudian psychotherapy during the inter-war years, and then deployed by anthropologists, notably those associated with the Tavistock Institute, in investigations of post-war social reconstruction. Elizabeth Bott recast a study of problems in marital relationships into an analysis of social network dynamics (Savage, 2008). Shortly afterwards, Michael Young’s Institute of Community Studies took up this approach to a fanfare of public interest. During the 1960s, these methods commanded huge attention for articulating ‘counter-cultural’ perspectives, revealing the voices of people who had previously been ignored or marginalised. A range of iconic qualitative projects from Young and Willmott’s (1958) *Family and Kinship in East London* to Ann Oakley’s (1974) *The Sociology of Housework* captured huge public and policy interest. From this point on, the bespoke qualitative interview project became a sociological staple.

In the past four decades there has been considerable innovation in the collection of in-depth qualitative interview data. In the 1980s and 1990s narrative-oriented approaches to interviewing grew in popularity (Elliott, 2005; Hollway and Jefferson, 2000; Mishler, 1986). This emphasised the need for open-ended questions – encouraging respondents to describe the concrete details of their lives. More recently there have been experiments with new ways of collecting qualitative data such as photo elicitation, creative and performative methods (Brown, 2019; Carabelli and Lyon, 2016; Pearce et al., 2020). The digital revolution and ubiquity of the smart phone has encouraged the collection of interview data online (Anderdal Bakken, 2023; Pearce et al., 2014) and the COVID-19 pandemic prompted debates on the advantages and challenges of conducting on-line interviews (Rahman et al., 2021; Thunberg and Arnell, 2022). Methodological innovation has focused more on the *collection* rather than *analysis* of data. Analysis has remained largely dependent on an interpretative close reading of text, albeit often with the aid of increasingly sophisticated software.² Additionally, questions rooted in the inter-subjectivity between researcher and interviewee often receive greater methodological attention than the challenge of how to achieve an appropriately rich, varied and more representative sample such that research results are transferrable to other contexts (Weick, 2007).

By contrast, the trajectory has been very different for the development of quantitative methods, especially those championing the use of nationally representative surveys. During a rather similar period of post-war trailblazing, survey research also centred on data collection. It was only from the 1980s that the turn towards secondary analysis became significant: before this, surveys had mainly been carried out on a one-off basis. The ability to share and re-analyse survey data began to improve as the capacity for computer storage and transmission expanded, and during the 1980s several books

championing secondary analysis were published in the UK (Dale et al., 1988; Hakim, 1982; Kielcolt and Nathan, 1985). A striking example of this late uptake of secondary quantitative analysis is that until the early 1990s, official government interpretations of social data were not subject to scrutiny from social scientists using independent secondary analysis (Römer, 2023).

Quantitative researchers were therefore not quick off the block in championing secondary analysis. However, rapid change took place in the 1990s, and the secondary analysis of quantitative data became routine. This was additionally driven by funding councils who demanded the archiving and accessibility of survey data as a precondition for research funding.³ By the 2000s the Economic and Social Research Council (ESRC) was explicitly asking applicants for survey funding to demonstrate that relevant data were absent from existing data sets. In addition, by the 2000s funding councils supported resource centres to document and archive large-scale quantitative studies such as the British Election Study, the British Household Panel Study and the British Birth Cohort studies. The aim was to facilitate efficient data use by a wide range of researchers (Pearson, 2016). More recently, the Administrative Data Research UK partnership has been established by the ESRC to ensure that the wealth of quantitative data collected routinely by government departments is more accessible for secondary analysis to produce policy-relevant insights (Gordon, 2020).

This differential trajectory was not inevitable. Precedents for the archiving and re-analysis of qualitative social science data go back almost as far as for survey data. In fact, UK social science led other nations in championing *qualitative* archiving and secondary analysis from a relatively early period. Almost three decades ago, in 1996, the Qualidata archive at Essex issued the following statement:

The QUALIDATA Resource Centre located in the Department of Sociology at the University of Essex has now been in existence for almost two years. Its aims are: locating, assessing and documenting qualitative data and arranging for their deposit in suitable public archive repositories; disseminating information about such data; and raising archival consciousness among the social science research community.

The UK Data Archive now includes over 1600 deposits of ‘qualitative and mixed methods’ studies out of 9655 (5067 of which are surveys).⁴ Yet, most qualitative archived studies have a small number of cases (i.e. fewer than 100), include multiple types of data and are focused on specific geographic area(s). Only a small minority include a sufficient number and range of qualitative cases to suggest that issues of representativeness have been considered.⁵

The provision of a dedicated qualitative data archive encouraged intellectual momentum around the archiving and secondary analysis of qualitative interview data, much of which was promoted by Heaton (1998, 2008; see also Hammersley, 1997). However, compared with the enthusiastic endorsement of secondary analysis in the quantitative community (e.g. Goldthorpe, 2016), resistance to secondary analysis among qualitative researchers remained strong. Mauthner et al. (1998) argued against what they saw as ‘naïve realism’ (p. 733) and insisted on a reflexivity such that qualitative evidence could only be understood in the context that it was collected and could therefore not readily be

re-analysed. Doubts have persisted. Although she ultimately endorses the potential for re-analysis, Irwin (2013: 297) acknowledges the ‘contextual embeddedness of data (which) engenders ethical and epistemological challenges to analysts’.

Therefore, despite the initial enthusiasm for secondary analysis of qualitative interviews, interest has not blossomed. Annual reports from the UK Data Archive suggest that the reuse of *qualitative* data increased substantially over the first decade of the 21st century, but remained infrequent compared with reuse of *quantitative* data. In 2003/2004 there were just 56 qualitative data sets provided to users compared with 17,779 *quantitative* data sets; by 2009/2010 this had increased to 1187 qualitative data sets provided out of a total of 56,777 data sets (Economic and Social Data Service, 2004: 20, 2010: 24). Analysis of qualitative data downloads and published papers that mention secondary analysis of qualitative data, indicates that the substantial increase in the reuse of UK qualitative data between 2002 and 2012 has not been maintained (Bishop and Kuula-Luumi, 2017).⁶

The lack of momentum in reusing qualitative data could reflect growing sensitivity to research ethics, particularly concerning data from in-depth, bespoke research projects. Crucial ethical considerations can be more manageable when a large-scale resource of diverse qualitative interviews collected from a broad geographic area is designed for secondary analysis. Some researchers argue that anonymising qualitative interview transcripts is challenging, cautioning against archiving or secondary analysis. For example, Parry and Mauthner (2004) note that the detailed nature of qualitative interviews makes de-identification difficult, as the richness of individual accounts can reveal respondents’ identities. Removing such material diminishes the data’s quality and utility. However, advances in information technology and data linkage capabilities have made it increasingly challenging to ensure true anonymity, even in quantitative studies collecting detailed personal information (ter Meulen et al., 2011). While complete anonymity is difficult to guarantee in any research, there is no intrinsic reason why this issue cannot be mitigated via data management and appropriate access and licensing arrangements.

Informed consent regarding the possible secondary analysis of interviews is crucial (Enriquez, 2024; Murphy et al., 2021; Parry and Mauthner, 2004). For example, in any original study, researchers will have gained entree into the lives of individuals and at times a community in a dyadic relationship (Bishop, 2007; Irwin and Winterton, 2012). While people may consent to discuss private matters with a specific researcher whom they have come to know, a person may not be comfortable having their interviews shared with other, unknown, researchers who are asking different questions (Murphy et al., 2021; Ruggiano and Perry, 2019). The ethics of secondary qualitative analysis can thus emphasise that the secondary study should have aims that match those of the original study (e.g. Etkind et al., 2017). These legitimate ethical concerns help to explain why studies conducting secondary qualitative analysis often include the researchers on the original study (see Ruggiano and Perry, 2019 for evidence of this pattern generally; for an example see Chew-Graham et al., 2012). However, as transparency and replicability crises are occurring across qualitative and quantitative research, the ethics of secondary data analysis are also being rethought – with some arguing that the ethical concerns

around anonymity and research positionality are ‘overblown’ (Freese et al., 2022; Murphy et al., 2021) although others argue that making qualitative data publicly available will decrease data quality (Khan et al., 2024).

In summary, and as discussed further below, the impetus in the 1990s and early 2000s to promote the secondary analysis of qualitative interview data seems to have stalled. Even though major research funders may still demand that such data are documented and archived, in practice the small-scale, bespoke nature of many studies results in limited demand for their reuse. This subdued interest in secondary qualitative interview analysis makes the lessons to be learned from the recent major British intervention ‘big qual’, associated with the Timescapes initiative, of strategic importance.

The Timescapes Initiative and ‘Big Qual’

The Timescapes study originated from a 2003 ESRC scoping study to establish a multi-purpose qualitative data resource, paralleling the quantitative British Birth Cohort Studies and British Household Panel Study (for a fuller discussion see Davidson et al., 2019). The ESRC subsequently launched a competitive call for a new, largely interview-based, qualitative resource, awarded to a team led by Professor Bren Neale at the University of Leeds. Initial funding covered five years (2007–2012), extended to document and archive the data.

The project’s network of longitudinal empirical studies aimed to deepen understandings of personal relationships and family life dynamics. It produced an archive of qualitative longitudinal data, including interview transcripts and multimedia materials including essays and drawings from participants. The programme drew in researchers from five UK universities across sociology, social policy, psycho-social research, oral history and the sociology of health. Seven projects ranged from studying 100 children’s sibling relationships to eight longitudinal case studies of grandparents. Alongside some fresh projects, existing studies were also extended, adding new sweeps of longitudinal data.

Collectively, these seven projects followed over 300 individuals, complementing each other by focusing on different family transitions. Theoretically sophisticated, the programme emphasised the interconnections between biographical time, generational time and historical time (Adam, 1998; Neale et al., 2012). Much of the material is now documented and archived as nine rich data sets at the Timescapes Archive at the University of Leeds, a satellite of the UK Data Archive.

Timescapes is therefore exactly the kind of ambitious initiative that is needed. The team have written multiple methodological publications, built capacity and generated an impressive literature on how synergies can be built between a set of studies (Davidson et al., 2019; Irwin and Winterton, 2012). An especially arresting intervention is the development of the ‘big qual’ approach, spearheaded by Ros Edwards and her team (Edwards et al., 2021), which aims to link cases from multiple archived qualitative studies.⁷ Davidson et al. (2019) describe this method as creating assemblages that allow new research questions to be addressed through comparative attention to differences between studies. Metadata detailing a study’s focus, sample

characteristics, geographic and temporal details are vital to structure comparative designs (Davidson et al., 2019).

The use of archaeological metaphors invokes how evidence is to be linked, by offering a broad and detailed mapping of diverse data. It is fully acknowledged that combining small, unrepresentative samples does not yield a representative sample, thus limiting claims to generalisability (Davidson et al., 2019). In promoting 'qualitative integrity' within 'big qual' initiatives, Timescapes advocates for a contextual and nuanced understanding of time, temporality and context. This perspective leads Timescapes to question the scalability of qualitative research and to refrain from claiming that findings are representative. This distinguishes the 'big qual' approach from the secondary analysis of large quantitative data sets, which aim for representativeness and often use weighting factors for more accurate population estimates (Dale et al., 2008). Consequently, the 'big qual' approach operates in parallel to quantitative methods and does not challenge their claims to offer more transferable findings.

We therefore seek to open up a 'second front' that complements Timescapes by supporting high-quality, large-scale qualitative interview studies with transparent and systematic sampling methods. These could be widely used for secondary analysis potentially in tandem with quantitative studies. This approach ensures a comprehensive understanding by integrating qualitative insights with quantitative findings, maintaining the strengths of both methodologies.

A New Path? The Secondary Analysis of Large-Scale Qualitative Data

We take our cues from important precedents for this approach. For example, between 2008 and 2010 an ESRC-funded project, the 'Social Participation and Identity project' (SPI) conducted qualitative biographical interviews with 220 members of the 1958 British Birth Cohort Study. This project, leveraged a sub-sample from an ongoing longitudinal survey, matching qualitative data to extensive longitudinal information collected since birth. Purposive stratified sampling ensured the diversity of interviewees (Elliott et al., 2010).

The SPI was a smaller-scale intervention than Timescapes. Publications by the immediate research team (e.g. Elliott, 2013; Miles and Leguina, 2018) showed how qualitative data could be used to make stronger claims, for instance about the significance of racist and nationalist attitudes (Flemmen and Savage, 2017). However, the wider research community has not extensively used the qualitative data; the study has been downloaded over 770 times, but Google Scholar suggests that Elliott et al.'s (2010) account of the qualitative SPI study has been cited only 43 times as of January 2025.

The SPI was successful in establishing the feasibility of collecting qualitative material from a subset of a large quantitative longitudinal study and indeed the study design was replicated as part of the interdisciplinary 'Halcyon' programme on healthy ageing (Kuh et al., 2013). This has not yet energised wider interest in creating large-scale qualitative resources based on samples reflecting the diversity of the population in the UK. However, we might reflect on a major uptick in qualitative interview secondary analysis in the USA, which could have major repercussions across the globe.

Of particular importance here is the American Voices Project (AVP), which provides qualitative evidence on everyday life in the USA (see Edin et al., 2024). The AVP aims to generate large-scale qualitative evidence on respondents' own perspectives and the narratives of their lives, with 2700 interviews completed to date. The long-term aim is to provide transferable evidence for social research, policy audiences, public interest and journalism.

The AVP involves a dramatic re-tooling to allow qualitative research to be more nationally representative, large scale, well documented and accessible for re-analysis by other researchers. Data collection was based on three-stage cluster sampling, starting with a stratified sample of counties and oversampling low-income groups to ensure these voices are included (Alexander et al., 2017). This approach closely resembles that taken to sampling for quantitative longitudinal studies such as the Millennium Cohort Study or Understanding Society in the UK, both of which were designed for use by secondary analysts (Buck and McFall, 2012; Plewis, 2007).

The AVP interview guide covers topics such as early childhood development and parenting, residential segregation, poverty and deprivation, policing and criminal justice, health disparities, immigration and ethnicity, educational inequality, the labour market, housing and eviction, public surveillance, populism and the radical right, and science and genetics (Edin et al., 2024). Interviews concluded with short, structured questionnaires, which allowed for cross-comparison between the qualitative and quantitative data. Field work was completed between 2019 and 2021 with adjustments in mode due to the pandemic. The AVP was therefore able to capture the everyday voices and experiences of Americans not only during a global pandemic, but also in the context of the Black Lives Matter campaign following the murder of George Floyd in May 2020; documented in the resulting 'Monitoring the crisis' report series.⁸

By drawing on the strengths of existing qualitative interview methods, and applying them within a large-scale initiative, the AVP offers remarkable possibilities for reimagining data sharing practices and for qualitative research to command more authority in social science research. It also shifts the ethics from the problems of reusing bespoke research data to the ethics of creating a shareable and open qualitative data set, with a particular focus on informed consent (Enriquez, 2024; Freese et al., 2022; Murphy et al., 2021). Far from supplanting smaller-scale qualitative interview studies, the proponents of the AVP believe they offer a complementary data source (Edin et al., 2024). Such a resource could be used to motivate the need for further in-depth research.

Many high-profile US social scientists have responded with enthusiasm to the potential of large-scale open-access qualitative data. When the AVP worked with the Russell Sage Foundation (RSF) to invite researchers to access these qualitative data, it recorded the second highest number of applications ever to an RSF call (Edin et al., 2024). Although it is too early to judge the impact of the 20 articles appearing in the RSF issue 'Building an open qualitative science' (Volumes I & II, 2024), together with the seven crisis monitoring reports, the prospects seem better than at any previous time. The AVP has established an important beachhead demonstrating the viability of large-scale qualitative interview data from a diverse sample of the population that allows secondary analysis. We believe that UK social science needs similar, ambitious, thinking.

Natural Language Processing and New Possibilities for the Analysis of Large-Scale Qualitative Studies

Cross-fertilising the AVP model, there is also the real potential that large-n qualitative analysis can be further enhanced by a new generation of NLP algorithms. These can assist researchers with the analysis of much larger amounts of qualitative interview data than has previously been possible and permits the selection of subsets of cases with specific characteristics embedded within national samples.

Interpretation is crucial in qualitative analysis, and we do not advocate for AI replacing human researchers in analysing interview data. Nonetheless, we need to engage with this rapidly moving field and computational social science offers new ways to interrogate large volumes of qualitative data. While early work focused on the technical aspects of NLP with unstructured data, there is now increasing use of computational text analysis to support researchers from a range of disciplines in addressing substantive questions (Baumer et al., 2017; Bonikowski and Nelson, 2022).

Ethical considerations are important when using AI to assist with analysing biographical textual data. The use of closed-source LLMs can help ensure interview data are not used outside the research project (e.g. are not used as a broader resource to 'train' the LLM), protecting confidentiality. Additionally, the role of the researcher in interpretive analysis should remain central, with AI serving to manage, and sift through, large corpora of text, allowing for more efficient identification of relevant material, but with the human (reflexive) researcher crucial for nuanced hermeneutic analysis that takes account of the power dynamics within society. This approach can help assuage the ethical concerns that have been raised by some scholars, fearful that analysis based on generative AI could 'perpetuate or exacerbate the colonisation or marginalisation of other modes of knowledge, cultures, or values, by privileging a certain perspective on the data analysis process, for instance, one reflecting Western cultures, because of training data prevalently collected online' (Davison et al., 2024: 1436).

It is beyond the scope of this article to provide an overview of all the new developments in qualitative analysis afforded by NLP, but two distinct strands of work are of relevance. First, NLP can aid the analysis of qualitative data such that much larger samples can be subjected to detailed textual analysis, attending to the form of the data as much as its content (Benoit et al., 2016; Franzosi, 2021; Mohr et al., 2013; Tebaldi et al., 2019). As Franzosi (2021) carefully demonstrated, this does not imply settling for a 'distant' or shallow reading of text. The quantitative identification of patterns in text, that would have been impossibly time consuming by hand, can now be automated such that they provide another lens through which material can be viewed. In turn this can raise new research questions and insights that can be pursued using close reading methods and hermeneutic analysis (Franzosi, 2021; Tebaldi et al., 2019). Examples include the ability to create open-source NLP algorithms that will quantify sentence length, sentence complexity, noun and verb analysis, including the gender of individuals spoken about, and the use of singular and plural terms. However, Franzosi also acknowledges these new tools are indeed only tools and can never replace the social scientist; and even though many are freeware and open source they are not necessarily easy to use. However, in

Franzosi's (2021: 1537) words: 'We either embrace the "new science" and use its tools to our advantage or risk being left behind.'

Second, recent advances in LLMs, such as each new instantiation of the generative AI ChatGPT, allow researchers automatically to select relevant interviews (or interview sections) from a much larger corpus for in-depth analysis. Importantly, the speed and capabilities of these LLMs are such that the whole text of each interview can be automatically interrogated for the characteristics, experiences or discourse of interest, without the need for pre-defined metadata. Analysis at scale no longer requires a team of researchers to code large swaths of data. The implications for qualitative research are considerable. Qualitative scholars would no longer necessarily need to design data collection processes that would target specific (and often hidden) groups of individuals (e.g. those in pain; those who are political activists; those who are infertile). Instead, a large-scale qualitative omnibus study, such as the AVP, described above, can provide a diverse and well-documented sample from which specific subgroups could be selected based on prompts offered to a chatbot. This in turn could remedy one of the key weaknesses of much current qualitative work where convenience or snowball samples risk basing conclusions on a very select group of respondents (Payne and Williams, 2005).⁹

The ability to use large language models to identify sections of relevant text within a qualitative interview also allows for analysis that moves beyond consideration of variation *between* individuals to focus instead on variation *within* individuals' accounts of their experiences. This endorses the ethos of much qualitative work that strives to allow for ambiguity and ambivalence in the way that individuals make sense of their lives (Watson, 2006).

These approaches could boost computational social science and bring it into closer connection with empirical data collection. This is ethically important, as sociological research would assuredly be based on interviews with people rather than composites of online personas that are biased towards white, English-speaking people who engage in online activity that results in digital data (Gallegos et al., 2024). It could become possible for qualitative social scientists to use the new tools offered by NLP to provide more sociologically inflected perspectives, which question the default engineering and naturalistic framings that might otherwise dominate these initiatives. This is surely a project of vital sociological urgency.

Conclusions

We are at a turning point. During recent decades quantitative research has gained increasing academic and policy traction by emphasising its superiority around secondary analysis and thereby replication and testing. There is no intrinsic reason for qualitative researchers to concede this ground. Especially in the context of 21st-century 'polycrises', which standard survey-based methods have not proven adept in anticipating, large-scale qualitative interview research has the potential to provide an understanding of how individuals cope with the structural challenges they face and could even provide an 'early warning' system to detect emerging societal threats. Social scientists can more effectively 'listen' (Back, 2007) to the perspectives of lay people and gain a greater

understanding of how diverse individuals are struggling to make sense of the changing world around them. Neither traditional survey methods, nor bespoke qualitative studies, including those which may be amalgamated into wider assemblages, are fully equipped for this task. Our approach will allow the vital strengths of qualitative inquiry to be harnessed so that they can reinvigorate the ‘sociological imagination’ at the heart of social science methods.

Of course, some research questions can only be answered by recruiting very specific groups of individuals and more familiar modes of bespoke, small-n qualitative research must continue. Alongside these studies, we advocate for much larger samples of *qualitative* interviews, created to map the diversity and variability within the population. These could be designed for secondary analysis and allow a focus on more generic and perennial research questions, with a depth not currently offered by representative surveys and at a scale and coverage not normally achieved in qualitative research.

There are important exemplars of this emerging work. The UK Data Archive has already demonstrated how qualitative material can be shared without compromising the anonymity and confidentiality of respondents. Timescapes has made important advances but positions qualitative secondary analysis in parallel to, and apart from, quantitative secondary analysis. More recently, the AVP provides a model for representative, large-scale qualitative interviewing. The rapid improvement of AI and machine learning that has resulted in sophisticated LLMs provides an important opportunity. The use of large and varied samples does not preclude detailed interpretative analysis of a selected sub-sample, but can locate that sub-sample more precisely within the wider population. In turn, this would be powerful in lending greater rigour to qualitative research and greater confidence in its insights and conclusions for policy makers. Although we have focused on interview-based research, we encourage a broader relationship between qualitative resources and secondary analysis.

The stakes are high. We are passionate about the potential of interview-based research to address the scale and nature of multiple social crises evident in the 21st century, and we hope that this article will be a helpful provocation to encourage investment in the creation and use of more representative, large-scale, qualitative data sets for sociologists, social policy researchers and others to use.

Authors' note

Jane Elliott was affiliated to University of Exeter, UK when the first draft of this paper was written and has now moved to the LSE.

Acknowledgements

We are very grateful to Bren Neale and Libby Bishop for their extensive and helpful comments on a first draft of this article. We would also like to thank the anonymous reviewers for their very careful reading of our submission and their constructive comments.


Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: this work was supported by the Economic and Social Research Council (Grant number: UKRI/ES/B000147/1).

ORCID iDs

Jane Elliott  <https://orcid.org/0000-0003-2683-0099>

Carrie Friese  <https://orcid.org/0000-0001-7144-8046>

Gaby Harris  <https://orcid.org/0009-0002-1827-8767>

Elizabeth Mann  <https://orcid.org/0000-0003-4267-1614>

Mike Savage  <https://orcid.org/0000-0003-4563-9564>

Notes

1. There are broader debates on qualitative research and generalisability. Payne and Williams critique a lack of rigour in how generalisability is used, and advocate for a ‘*moderatum* generalisability’. They lamented that ‘avoiding the question [of generalisability] is apparently a legitimate practice under contemporary canons of academic publishing in sociology’ (Payne and Williams, 2005: 299). It is beyond the scope of the article to engage fully with these debates. Our focus instead is on the transferability of findings from qualitative research beyond the immediate context in which it was conducted.
2. Frequently used software packages such as ATLAS.ti and NVivo have developed from facilitating manual coding to incorporating AI that allows for more automatic thematic coding (D Mortelmans, Ch 19 ‘NVivo and AI (semi) automatic coding’ in Mortelmans D. (2024).
3. The ESRC data policy states that: ‘Publicly-funded research data are a public good, produced in the public interest, which shall be made openly available and accessible with as few restrictions as possible in a timely and responsible manner that meets a high ethical standard.’ <https://www.ukri.org/wp-content/uploads/2021/07/ESRC-200721-ResearchDataPolicy.pdf>.
4. Browsing the UK data Archive on 15 July 2024.
5. For example, the ‘Qualitative Election Study of Great Britain’ (2015, SN 8117) contrasts with ‘Young Women, Agency and Intimacy in Sexual Relationships’ (2008, SN6928) – both downloaded over 200 times. The Qualitative Election Study consists of 23 focus groups (including 94 eligible voters) carried out during and after the 2015 General Election and held across Great Britain. In contrast, the Young Women, Agency and Intimacy Study employed multiple methods all within a single school in south-east England.
6. There are examples of the secondary analysis of qualitative data (Anderson and Roy, 2013; Chew-Graham et al., 2012; Etkind et al., 2017; Reeves, 2015). However, these remain the exceptions to the rule. Whereas in quantitative research it is rare for researchers to conduct their own survey, in qualitative research it is rare for a researcher *not* to collect their own data.
7. Although there are previous examples of researchers combining qualitative data from separate projects before conducting new analysis of the data (e.g. Etkind et al., 2017), the ‘big qual’ approach has significantly developed this method.
8. Crisis monitoring reports are available at: <https://inequality.stanford.edu/covid/american-voices-project>.
9. A shortcoming of much contemporary qualitative work in the social sciences is its focus on the relationship between the researcher and the respondent at the expense of providing a clear account of sample recruitment. Although the focus of qualitative work is on how people understand themselves and the social world, if insufficient attention is paid to a well-justified sample design rigour will be diminished (Deluca, 2023).

References

- Adam B (1998) *Timescapes of Modernity: The Environment & Invisible Hazards*. New York, NY: Routledge.

- Alexander JT, Andersen R, Cookson PW, et al. (2017) A qualitative census of rural and urban poverty. *The Annals of the American Academy of Political and Social Science* 672(1): 143–161.
- Anderdal Bakken S (2023) App-based textual interviews: Interacting with younger generations in a digitalized social reality. *International Journal of Social Research Methodology* 26(6): 631–644.
- Anderson C and Roy T (2013) Patient experiences of taking antidepressants for depression: A secondary qualitative analysis. *Research in Social and Administrative Pharmacy* 9(6): 884–902.
- Back L (2007) *The Art of Listening*. London: Berg.
- Baumer EPS, Mimno D, Guha S, et al. (2017) Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68: 1397–1410.
- Benoit K, Conway D, Lauderdale BE, et al. (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *The American Political Science Review* 110(2): 278–295.
- Bishop L (2007) A reflexive account of reusing qualitative data: Beyond primary/secondary dualism. *Sociological Research Online* 12(3): 43–56.
- Bishop L and Kuula-Luumi A (2017) Revisiting qualitative data reuse: A decade on. *Sage Open* 7(1): 215824401668513.
- Bonikowski B and Nelson LK (2022) From ends to means: The promise of computational text analysis for theoretically driven sociological research. *Sociological Methods & Research* 51(4): 1469–1483.
- Bott E (1956) Urban families: The norms of conjugal roles. *Human Relations* 9(3): 325–342.
- Brown N (2019) Identity boxes: Using materials and metaphors to elicit experiences. *International Journal of Social Research Methodology* 22(5): 487–501.
- Buck N and McFall S (2012) Understanding Society: Design overview. *Longitudinal and Life Course Studies* 3(1): 5–17.
- Bulmer M, Bales K and Sklar KK (eds) (1991) *The Social Survey in Historical Perspective, 1880–1940*. Cambridge: Cambridge University Press.
- Carabelli G and Lyon D (2016) Young people's orientations to the future: Navigating the present and imagining the future. *Journal of Youth Studies* 19(8): 1110–1127.
- Chew-Graham C, Kovandžić M, Gask L, et al. (2012) Why may older people with depression not present to primary care? Messages from secondary analysis of qualitative data: Depression in older people. *Health & Social Care in the Community* 20(1): 52–60.
- Collins C, Neely MT and Khan S (2024) 'Which cases do I need?': Constructing cases and observations in qualitative research. *Annual Review of Sociology* 50: 21–40.
- Connelly R, Playford CJ, Gayle V, et al. (2016) The role of administrative data in the big data revolution in social science research. *Social Science Research* 59: 1–12.
- Dale A, Arbor S and Procter M (1988) *Doing Secondary Analysis*. London, UK: Unwin Hyman.
- Dale A, Wathan J and Wiggins V (2008) Secondary analysis of quantitative data sources. In: Alasuutari P, Bickman L and Brannen J (eds) *The Sage Handbook of Social Research Methods*. London: Sage, 520–535.
- Daniel BK (2018) Empirical verification of the 'TACT' framework for teaching rigour in qualitative research methodology. *Qualitative Research Journal* 18(3): 262–275.
- Davidson E, Edwards R, Jamieson L, et al. (2019) Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality & Quantity* 53(1): 363–376.
- Davison RM, Chughtai H, Nielsen P, et al. (2024) The ethics of using generative AI for qualitative data analysis. *Information Systems Journal* 34: 1433–1439.
- DeLuca S (2023) Sample selection matters: Moving toward empirically sound qualitative research. *Sociological Methods & Research* 52(2): 1073–1085.

- Economic and Social Data Service (2004) *Economic and social data service annual report 2003–2004*. UK Data Archive, Essex. Available at: <https://ukdataservice.ac.uk/app/uploads/esds-annualreport20032004.pdf> (accessed 29 May 2025).
- Economic and Social Data Service (2010) *Economic and social data service annual report August 2009 – July 2010*. UK Data Archive, Essex. Available at: <https://ukdataservice.ac.uk/app/uploads/esds-annualreport20092010.pdf> (accessed 29 May 2025).
- Edin KJ, Fields CD, Grusky DB, et al. (2024) Listening to the voices of America. *RSF: The Russell Sage Foundation Journal of the Social Sciences* 10(4): 1–31.
- Edwards R, Davidson E, Jamieson L, et al. (2021) Theory and the breadth-and-depth method of analysing large amounts of qualitative data: A research note. *Quality & Quantity* 55: 1275–1280.
- Edwards R and Holland J (2020) Reviewing challenges and the future for qualitative interviewing. *International Journal of Social Research Methodology* 23(5): 581–592.
- Elliott J (2005) *Using Narrative in Social Research: Qualitative and Quantitative Approaches*. London: Sage.
- Elliott J (2013) Talkin’‘bout my generation’: Perceptions of generational belonging among the 1958 cohort. *Sociological Research Online* 18(4): 122–137.
- Elliott J, Miles A, Parsons S, et al. (2010) *The Design and Content of the ‘Social Participation’ Study. A Qualitative Sub-Study Conducted as Part of the Age 50 (2008) Sweep of the National Child Development Study*. London: CLS.
- Enriquez D (2024) Publishing publicly available interview data: An empirical example of the experience of publishing interview data. *Frontiers in Sociology* 9: 1157514.
- Etkind SN, Bristowe K, Bailey K, et al. (2017) How does uncertainty shape patient experience in advanced illness? A secondary analysis of qualitative data. *Palliative Medicine* 31(2): 171–180.
- Flemmen M and Savage M (2017) The politics of nationalism and white racism in the UK. *The British Journal of Sociology* 68(S1): S233–S264.
- Franzosi R (2021) What’s in a text? Bridging the gap between quality and quantity in the digital era. *Quality & Quantity* 55(4): 1513–1540.
- Freese J, Rauf T and Voelkel JG (2022) Advances in transparency and reproducibility in the social sciences. *Social Science Research* 107: 102770.
- Gallegos IO, Rossi RA, Barrow J, et al. (2024) Bias and fairness in large language models: A survey. *Computational Linguistics* 50(3): 1097–1179.
- Goldthorpe JH (2016) *Sociology as a Population Science*. Cambridge: Cambridge University Press.
- Gordon E (2020) Administrative data research UK. *Patterns (N Y)* 1(1): 100010.
- Guba EG and Lincoln YS (2005) Paradigmatic controversies, contradictions, and emerging confluences. In: Denzin NK and Lincoln YS (eds) *The Sage Handbook of Qualitative Research*, 3rd edn. Los Angeles, CA: Sage, 191–216.
- Hakim C (1982) *Secondary Analysis in Social Research: A Guide to Data Sources and Methods with Examples*. London: Unwin Hyman.
- Hammersley M (1997) Qualitative data archiving: Some reflections on its prospects and problems. *Sociology* 31(1): 131–142.
- Heaton J (1998) Secondary analysis of qualitative data. *Social Research Update*. Available at: <https://sru.soc.surrey.ac.uk/SRU22.html> (accessed 30 May 2025).
- Heaton J (2008) Secondary analysis of qualitative data: An overview. *Historical Social Research* 33(3 (125)): 33–45.
- Hollway W and Jefferson T (2000) *Doing Qualitative Research Differently: Free Association, Narrative and the Interview Method*. London: Sage.

- Irwin S (2013) Qualitative secondary data analysis: Ethics, epistemology and context. *Progress in Development Studies* 13(4): 295–306.
- Irwin S and Winterton M (2012) Qualitative secondary analysis and social explanation. *Sociological Research Online* 17(2): 1–12.
- Jensen JL, Karell D, Tanigawa-Lau C, et al. (2022) Language models in sociological research: An application to classifying large administrative data and measuring religiosity. *Sociological Methodology* 52(1): 30–52.
- Jerrim J and De Vries R (2017) The limitations of quantitative social science for informing public policy. *Evidence & Policy* 13(1): 117–133.
- Khan S, Hirsch JS and Zeltzer-Zubida O (2024) A dataset without a code book: Ethnography and open science. *Frontiers in Sociology* 9: 1308029.
- Kiecolt KJ and Nathan LE (1985) *Secondary analysis of survey data. Sage University Paper Series on Quantitative Applications in the Social Sciences*. Vol. 53. London: Sage publications Ltd.
- Kuh D, Cooper R, Hardy R, et al. (2013) *A Life Course Approach to Healthy Ageing*. Oxford: Oxford University Press.
- Lee RM (2004) Recording technologies and the interview in sociology, 1920–2000. *Sociology* 38(5): 869–889.
- Lincoln YS and Guba EG (1985) *Naturalistic Inquiry*. Newberry Park, CA: Sage.
- Mauthner NS, Parry O and Backett-Milburn K (1998) The data are out there or are they? Implications for archiving and revisiting qualitative data. *Sociology* 32(4): 733–745.
- Mayhew H (1985) *London Labour and the London Poor*. Milton Keynes: Penguin UK. [Originally published 1865]
- Miles A and Leguina A (2018) Socio-spatial mobilities and narratives of class identity in Britain. *The British Journal of Sociology* 69(4): 1063–1095.
- Mishler EG (1986) *Research Interviewing: Context and Narrative*. Cambridge, MA: Harvard University Press.
- Mohr JW, Wagner-Pacifi R, Breiger RL, et al. (2013) Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics. *Poetics* 41(6): 670–700.
- Mortelmans D (2024) NVivo and AI: (Semi)-Automatic Coding. In: *Doing Qualitative Data Analysis with NVivo*. Cham: Springer Nature Switzerland, 229–250.
- Murphy AK, Jerolmack C and Smith D (2021) Ethnography, data transparency, and the information age. *Annual Review of Sociology* 47(1): 41–61.
- Neale B, Henwood K and Holland J (2012) Researching lives through time: An introduction to the Timescapes approach. *Qualitative Research* 12(1): 4–15.
- Oakley A (1974) *The Sociology of Housework*. London: Martin Robertson.
- Parry O and Mauthner NS (2004) Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. *Sociology* 38(1): 139–152.
- Payne G and Williams M (2005) Generalization in qualitative research. *Sociology* 39(2): 295–314.
- Pearce G, Thøgersen-Ntoumani C and Duda JL (2014) The development of synchronous text-based instant messaging as an online interviewing tool. *International Journal of Social Research Methodology* 17(6): 677–692.
- Pearce S, Gibson F, Whelan J, et al. (2020) Untellable tales and uncertain futures: The unfolding narratives of young adults with cancer. *International Journal of Social Research Methodology* 23(4): 377–390.
- Pearson H (2016) *The Life Project: The Extraordinary Story of Our Ordinary Lives*. London, UK: Penguin.
- Platt J (2002) The history of the interview. In: Gubrium JF and Holstein JA (eds) *Handbook of Interview Research: Context and Method*. Los Angeles, CA: Sage, 35–54.

- Plewis I (2007) Non-response in a birth cohort study: The case of the Millennium Cohort Study. *International Journal of Social Research Methodology* 10(5): 325–334.
- Rahman SA, Tuckerman L, Vorley T, et al. (2021) Resilient research in the field: Insights and lessons from adapting qualitative research projects during the COVID-19 pandemic. *International Journal of Qualitative Methods* 20: 160940692110161.
- Reeves A (2015) ‘Music’s a family thing’: Cultural socialisation and parental transference. *Cultural Sociology* 9(4): 493–514.
- Römer F (2023) *Inequality Knowledge: The Making of the Numbers about the Gap between Rich and Poor in Contemporary Britain*, vol. 89. Berlin: Walter de Gruyter.
- Ruggiano N and Perry TE (2019) Conducting secondary analysis of qualitative data: Should we, can we, and how? *Qualitative Social Work: Research and Practice* 18(1): 81–97.
- Savage M (2008) Elizabeth Bott and the formation of modern British sociology. *The Sociological Review* 56(4): 579–605.
- Savage M (2010) *Identities and Social Change in Britain Since 1940: The Politics of Method*. Oxford: Oxford University Press.
- Stacey M (1970) *Tradition and Change: A Study of Banbury*, 2nd edn. Oxford: Oxford University Press.
- Tebaldi M, Calaresu M and Purpura A (2019) The power of the president: A quantitative narrative analysis of the Diary of an Italian head of state (2006–2013). *Quality & Quantity* 53(6): 3063–3095.
- ter Meulen RH, Newson AJ, Kennedy MR, et al. (2011) *Genomics, health records, database linkage and privacy*. Background Paper. London: Nuffield Council on Bioethics.
- Thunberg S and Arnell L (2022) Pioneering the use of technologies in qualitative research - a research review of the use of digital interviews. *International Journal of Social Research Methodology* 25(6): 757–768.
- Townsend P (1962) *The Last Refuge: A Survey of Residential Institutions and Homes for the Aged in England and Wales*. London, England: Routledge and Kegan Paul.
- Tracy SJ (2010) Qualitative quality: Eight ‘big-tent’ criteria for excellent qualitative research. *Qualitative Inquiry* 16(10): 837–851.
- UK Data Service (2024) Trusted Data for Research: Annual report. Available at: <https://ukdata-service.ac.uk/app/uploads/UK-Data-Service-annual-report-2022-2023.pdf> (accessed 30 May 2025).
- Watson C (2006) Unreliable narrators? ‘Inconsistency’ (and some inconstancy) in interviews. *Qualitative Research* 6(3): 367–384.
- Weick KE (2007) The generative properties of richness. *Academy of Management Journal* 50(1): 14–19.
- Young M and Willmott P (1958) *Family and Kinship in East London*. London: Routledge.

Jane Elliott is now Professorial Research Fellow at the LSE’s International Inequalities Institute, where she leads the ESRC-funded project ‘UK Voices’. She has a longstanding interest and experience in combining qualitative and quantitative approaches to research, with a focus on using a narrative approach to understanding individual identities. Her publications include *Using Narrative in Social Research* (Sage, 2005).

Carrie Friese is an Associate Professor of Sociology at the LSE. Her research and teaching are in the sociology of health and illness, science and technology studies and qualitative research methods. Her books include *A Mouse in a Cage* (NYUP, 2025) and *Situational Analysis* (with Adele Clarke and Rachel Washburn, Sage, 2017).

Gaby Harris is a lecturer at Manchester Metropolitan University. Her ESRC-funded doctoral research examines how teenage girls navigate different social relationships through their wardrobe and consumption practices. Her research uses qualitative analysis to explore what the study of fashion can illuminate about individuals' social worlds, and how this can inform broader sociological concerns.

Elizabeth Mann is ESRC Postdoctoral Fellow at the Centre for Analysis of Social Exclusion, at the LSE. Liz's research interests centre on wealth inequality, particularly the intrahousehold allocation of wealth and the gender wealth gap in the UK context. Previous publications include Loutzenhiser G and Mann E (2021) Liquidity issues: Solutions for the asset rich, cash poor. *Fiscal Studies* 42(3–4): 651–675 and Summers K, Accominotti F, Burchardt T, et al. (2022) Deliberating inequality: A blueprint for studying the social formation of beliefs about economic inequality. *Social Justice Research* 35(4): 379–400.

Mike Savage is Professorial Research Fellow at the LSE's International Inequalities Institute where he convenes the research theme on 'wealth, elites and tax justice'. He has longstanding interests in using methodological innovations, both quantitative and qualitative to shed light on dynamics of social inequality. His books include the best-selling co-authored, *Social Class in the 21st Century* (Penguin, 2015) and *The Return of Inequality: Social Change and the Weight of the Past* (Harvard UP, 2021).

Date submitted January 2024

Date accepted April 2025