Wilton Park

Report

Data for Retention: Addressing under-representation of LGBT+ minorities in STEM

Wednesday 01 - Friday 03 February 2023 | WP3105

In association with: the UK Science and Innovation Network (SIN)

Foreign, Commonwealth & Development Office

Wilton Park

Report

Data for Retention: Addressing under-representation of LGBT+ minorities in STEM

Wednesday 01 - Friday 03 February 2023 | WP3105

In association with: the UK Science and Innovation Network (SIN)

Introduction

LGBT+ minorities are under-represented in STEM disciplines in both the US and the UK. In both countries, however, policymakers are struggling to design evidence-based interventions to address this problem. Often, this is because there is very little data collected on this minority population, which prevents policymakers from targeting interventions. Policymakers cannot focus efforts on a particular discipline (e.g., the problem might be worse in Chemistry than in Physics) or on a particular timepoint in a research career trajectory (e.g., LGBT+ scientists might be dropping out in graduate school or at post-doctoral level) or even once they become established in their own labs.

Some consider this to be a 'leaky pipeline' problem, with insufficient information on where, when and why the 'leaks' occur. Others note that this does not fully highlight the problem, as active discrimination can also push people out of the system. In the absence of data to characterise the problem writ-large, there is obviously a limit to the efficacy of interventions.

In 2021, the UK Science and Innovation Network (SIN) established a partnership with the National Science Policy Network (NSPN) on a project to address LGBT+ attrition in STEM fields. To date, outputs from this project include the first bilateral report on the issue based on discussions with universities, NGOs, researchers, and funders in both the UK and the US.

The conference aims to build on work undertaken to date, in order to further understand and address under-representation of LGBT+ people in STEM. Discussion will explore what data is currently collected, gaps in existing data and ways in which to overcome barriers to future data collection.

The conference represented a key milestone toward collaboration on data sharing and the establishment of a US-UK repository of datasets. In particular, it aimed to progress the following objectives:

- **Establish a UK-US community** of experts working together on DEI in STEM from government, NGOs, university administrations, researchers, and funders.
- **Create the world's first repository of datasets** for researcher access.
- **Draft a UK-US open-source policy guide for universities** to reference in designing evidence-based interventions to stop the attrition of LGBT+ minorities in STEM.

The conference also supported the launch of new bilateral funding for research into LGBT+ attrition and retention in STEM.

Executive summary

The problem: A growing body of evidence shows that LGBT+ people are underrepresented in STEM fields and face high rates of exclusion, harassment, and career limitations. However, due to a dearth of data collected on these populations, policymakers in the UK and the US have struggled to design evidence-based policies which support retention of LGBT+ minorities in STEM. For example, policymakers don't know when attrition is likely to occur in the career trajectory of these scientists, nor which disciplines, geographies or identities are most likely to be affected. For a host of complex reasons, LGBT+ identities have not been tracked as broadly and consistently as other demographics and this results in policy interventions that are well-intentioned, but speculative.

The solution: In order to make the STEM ecosystem inclusive, attractive and welcoming for all, scientific communities in both the UK and the US are embarking on the enormous challenge of collection, collation and study of sexual orientation and gender identity (SOGI) data in STEM. Not only is this data critical to our understanding of the current status, but it will also be the foundation on which we must design policies to stem the attrition of LGBT+ and other minorities. We must do this if we want to create environments in which all talented individuals who want to pursue scientific careers can flourish.

The collaboration: Despite different legal structures, organisation of agencies, and extents of SOGI data collection in the two nations, the UK and the US share similar goals and have been having similar conversations separately. In fact, because they collect different types of data, it is clear that by working together the UK and the US can enrich each other's understanding of this complex issue. The meeting provided a unique opportunity to learn across sectors and from different perspectives, seeking to improve collaboration and information exchange.

The event: This meeting at Wilton Park brought together over 40 professionals from government agencies, funding bodies, university leadership, academia, and non-governmental organisations in the United Kingdom and United States. Outcomes included:

(1) Fostering bilateral collaboration on the topic of data for LGBT+ retention in STEM

(2) Developing a set of recommendations on data collection

(3) Compiling a list of existing datasets and literature

(4) Announcing new funding mechanisms to promote research on improving STEM for LGBT+ communities.

Recurring themes of discussion included:

- We cannot change what we can't measure – data is key: Participants agreed that rich, reliable data is a crucial foundation for social change, and for policy progress. They gave deep consideration to the myriad of complex challenges facing institutions collecting SOGI data. These included issues spanning legality, privacy, trust, longevity, reliability and the need to appropriately balance data quality, respondent burden, confidentiality, and data user needs.
- Passion, power, and possibilities: The group aimed to inspire and imagine a world where LGBT+ people can thrive in the scientific milieu. The group carried this powerful vision to their deep consideration of the data questions at hand.
- Tackling multiple barriers: The LGBT+ community encompasses all other demographic categories, and efforts to understand and serve the community should give a voice to the most marginalised (particularly communities of colour or those tackling multiple barriers) a particular challenge when the data is limited given the small numbers.

- Collaboration is key - UK and the US synergies abound: The group relied on three questions to compare and contrast the UK and the US data infrastructure, asking: What data do we have? What data do we need? What are the barriers to collating this data? The answers to these core questions reveal complementary (rather than duplicative) data sets and myriad opportunities for collaboration.

Terminology and scope

- **"LGBT+"** is the official terminology used by the UK Foreign, Commonwealth & Development Office and is not meant to be an exhaustive representation of the community.
- **"SOGI data"** refers to demographic information about sexual orientation and gender identity. SOGI data are valid demographic data that should be handled similarly to other more commonly collected demographic items such as race, ethnicity, socioeconomic status, and ability. Collection of this data requires careful consideration of privacy, sensitivity, and mixed methods of analysis, especially in cases of disaggregation or small sample sizes.
- **"STEM"** refers broadly to the fields of science, technology, engineering, and mathematics, which may also include education, medicine, social sciences, or research and innovation generally, depending on context. For the purposes of this discussion, its scope has not been restricted.
- **"Retention"** in this report refers to individuals being able to stay in the STEM enterprise if they want to do so. Relatedly, we use the term "STEM pathways" as a framework for considering free movement within and in/out of STEM, rather than emphasising a single strict career pipeline which requires holding onto people who would otherwise leak out.
- "**Harmonisation"** refers to the creation of data standards that allow for semantic interoperability. Harmonised standards create a flexible but shared language of survey items that enables efficient processing and analysis across institutions so that they may share data effectively. Data harmonisation, in contrast to standardisation, does not intend to create a rigid question format, but instead focuses on a shared methodology to improve interoperability.

Key issues and recommendations

Participants agreed on four key areas of future collaboration:

1. **Sharing best practices on SOGI data collation at the institutional level**
   Participants considered prior work which demonstrates that higher education institutions in both the UK and the US face challenges which have deterred them from collecting SOGI data. These challenges include sensitivity of data, safety concerns, political sensitivities, and uncertainty of how to ask questions. However, because these challenges also exist to varying extents for other demographic categories, participants agreed that they can be overcome through thoughtful design. Participants heard from speakers who covered their experiences in these methodologies. Like all data collection, SOGI data collection at institutions should involve:
   (a) Purposeful design
   (b) Rigorous methodologies
   (c) Self-education on the issues in advance
   (d) An internal audit of available resources
   (e) Involvement of communities throughout the process

"much can be learned from working with datasets already collated by colleagues, especially given the scant availability of SOGI data, and the survey burden on relatively small groups"

(f) Attention to privacy and security

(g) Inclusive and flexible language in survey items

Data collection must also be accompanied by careful data analysis and data-informed interventions, though these topics generally fell outside the scope of this discussion.

2. **Sharing research datasets**
   Participants agreed that much can be learned from working with datasets already collated by colleagues, especially given the scant availability of SOGI data, and the survey burden on relatively small groups. Participants discussed safe, ethical ways in which data sharing might be possible. Participants also agreed about the utility of a high-level repository of datasets which would not allow access to individual data, rather would illuminate what data exists.

3. **Supporting social science and anthropology research which studies the mechanics of attrition and retention for SOGI communities**
   Participants agreed on the need for deep social science research into data for retention, including the study of survey design alongside experts, for example census professionals. The Royal Society of Chemistry will take this forward as part of a new funding mechanism for UK-US collaboration on Data for LGBT+ Retention in STEM, with the first cohort set to be announced by the Society in August 2023.

4. **Enriching the data that exists by supporting new work to study LGBT+ communities in STEM**
   Participants were unanimous in their call for more data, and their desire to support policymakers, institutions and researchers who are painstakingly working to collate it.

Background: The nature of SOGI data

"there is a growing body of evidence that demonstrates that LGBT+ people face harassment, discrimination, and exclusion in STEM"

Quantitative and qualitative data are essential for identifying LGBT+ representation and understanding the community's experience in STEM. Large population-based demographics rely upon quantitative demographic collection which is necessary for benchmarking LGBT+ population sizes in the broader population and in STEM to accurately assess the amount of underrepresentation in STEM. However, even without representation data, there is a growing body of evidence that demonstrates that LGBT+ people face harassment, discrimination, and exclusion in STEM (e.g. Cech et al., 2021 – see Bibliography). To further identify the specific experiences of the community, qualitative data can offer insight into the underlying culture in STEM that leads to poor outcomes for LGBT+ people (e.g. Bilimoria & Stewart 2009; Formby 2017). The user of demographic data collection should assess whether quantitative or qualitative methods best suit the needs of the organisation and the questions that are being asked. Importantly, disaggregation of qualitative data, when safely possible, offers the most power in addressing the needs of the most marginalised members of the community that may be overlooked due to insufficient numbers for statistical analysis in traditional quantitative methods.

"institutions should make clear that they are not just collecting data, but willing to use it toward policies to improve STEM environments."

Demographic data can be used as evidence for diagnosing a problem, but in itself is only one piece of the puzzle in addressing LGBT+ retention in STEM. Interlocutors were clear that more data is needed. Not only is it crucial in diagnosing the mechanics of the problem and informing policy interventions, but can also, in some cases, even be the determinant for whether or not a minority is legally considered as such. Interlocutors noted that in some ways, the act of collecting data in and of itself increases visibility for an often invisible population in the STEM enterprise. However, participants cautioned against survey fatigue among the relevant populations. To avoid this, they advised institutions to articulate what demographic data collection will do prior to implementing an extensive (and expensive) collation effort; institutions should make clear that they are not just collecting data, but willing to use it toward policies to improve STEM environments.

The fluid and complex nature of demographic data often creates complexities with statistical analysis. These are surmountable, even given perceived complexities of SOGI data in particular.

An example subset discussed at Wilton Park includes:

"There is fluidity and change in how individuals may identify, and identities change over time"

- **SOGI data as relational.** Individuals have a deep and sometimes complicated relationship with their own identities that can change depending upon the context of the survey and their current personal circumstances. Clear understanding by those filling out the surveys of how SOGI data will be used, the purpose of its incorporation, and the safety of this data can help mitigate issues of data inaccuracies when filling out SOGI items.
- **SOGI data as temporal.** There is fluidity and change in how individuals may identify, and identities change over time. This can constitute analytical boundaries especially for longitudinal data collection. However, understanding that a singular survey is a snapshot of the community and analysing the flux in the population can also provide invaluable information to the user. Survey professionals discussed ways in which this can be used as a data asset rather than a detraction.
- **SOGI data as contextual.** Disclosure of SOGI identities will vary upon the context of the survey, the institution providing the survey, and the trust the individual has in these institutions. SOGI data should be decoupled from access to resources but should be used post-hoc to identify how these resources are being distributed to LGBT+ individuals. Emphasising communication and trust on the part of the data collector has been shown to help with accurate data collection (though some may still not feel safe/able to disclose their identities in particular data collation contexts.)
- **SOGI data as sensitive.** The safety of LGBT+ individuals should be at the forefront of any data collection. Institutions are rightly extremely wary of this issue, and conscious that discoverability and "outing" of individuals is not a risk they are prepared to take. They are also conscious that the format and wording of SOGI items has the possibility to inflict harm if improperly executed. These risks can be mitigated with careful survey design and rigorous privacy protocols. The mechanics of these issues were discussed in detail.

"The safety of LGBT+ individuals should be at the forefront of any data collection."

Examining the existing evidence base

Participants reviewed an expanding evidence base on LGBT+ representation in STEM, within the broader context of emerging LGBT+ demographic data.

The majority of LGBT+ representation data has come from large national surveys from national governments. However, this isn't STEM-specific. Other essential quantitative and qualitative data sets have arisen from non-governmental organisations, academic researchers, and academic institutions. While harmonisation of data collection methods has remained a key hurdle in collection of SOGI data, there has been a growing body of evidence from multiple sectors on how to adapt SOGI items for the context of the survey. The analysis of the current evidence base also provides insights into the key barriers and gaps in data collection which have highlighted the methodological and ethical challenges to collecting this data.

- **National statistics agencies such as the UK Office for National Statistics and the US Census Bureau have recently begun implementation of SOGI items on large national surveys.**

  National census surveys offer the potential of significant data on LGBT+ representation in the general public, which can in turn be used to analyse representation in STEM and other fields. The UK is well ahead of its peers on this issue. The English and Welsh census included SOGI items on the 2021 census[1-3], an extremely innovative and novel accomplishment. The National Records of Scotland also included SOGI items on the 2022 Scottish census[4-6]. The Northern Ireland census incorporated a sexual orientation question but not a gender identity question on its census in 2021[7-9].

  By contrast, the US Census Bureau has yet to incorporate SOGI items on the US census or the American Community Survey (ACS). It does, however, allow for analysis of individuals sharing a household with a same-sex partner[10, 11] Incorporation of SOGI items is yet to be tested and implemented on further iterations of the ACS. (The census bureau did incorporate SOGI items on the Household Pulse Survey which measured household experiences during the COVID-19 pandemic. This study found that LGBT+ respondents were more likely than their non-LGBT+ counterparts to face food insecurity and economic insecurity[12].)

  While these data sets are limited by their ability to only provide quantitative data, information on benchmark population sizes and cross-sectional analysis is critical to our understanding of various populations in the LGBT+ community and can be foundational in associating SOGI census items to economic items in order to identify LGBT+ representation in STEM fields at scale.

- **Funding agencies often are the first line of large-scale data collection on diversity and inclusion in STEM fields but largely have yet to include SOGI in their surveys.**

  The largest source of governmental STEM funding in the UK is provided by UK Research and Innovation (UKRI). While UKRI has yet to include SOGI items on their demographic data collection, SOGI item inclusion in demographics is included in their equality, diversity, and inclusion strategic plan which will hopefully lead to SOGI data collection soon[13-15].

In the US, governmental funding is decentralised but relies upon the National Science Foundation (NSF) for demographic data collection and classification of marginalised identities as underrepresented in STEM[16]. NSF conducts a wide variety of surveys to collect demographics on the STEM workforce and education throughout different career stages including the National Survey of College Graduates, the Survey of Earned Doctorates, the Early Career Doctorates Survey, the Survey of Doctorate Recipients, and the Survey of Graduate Students and Postdoctorates in Science and Engineering[17]. Although NSF agreed to pilot SOGI items on their surveys, they ultimately decided not to include these items. Participants discussed the significance of this decision and researcher engagement with it. The National Science and Technology Council announced the first-ever Federal Evidence Agenda on LGBTQI+ Equity which urges that "data collection (on LGBT+ communities) must start immediately"[18]. The report follows the White House Summit on STEMM Equity and Excellence and its white paper urging improved data collection to achieve equity outcomes.

"Education agencies have the statistical power and capability to analyse a large section of the STEM population"

- **SOGI data collection has been rapidly developing in education agencies and provides a gold mine of opportunity for SOGI data analysis in both the US and the UK.**

  Education agencies have the statistical power and capability to analyse a large section of the STEM population. These agencies also provide a unique opportunity to provide longitudinal data collection and analysis to follow individuals throughout their participation in the academic enterprise.

  The US Department of Education was one of the first national agencies to begin collecting SOGI data when it incorporated SOGI items on the 2016 follow up survey to the High School Longitudinal survey of 2009 using the Office of Management and Budget working group on SOGI data collection's recommendations[19, 20]. This study follows more than 23,000 students beginning in high school and through to postsecondary education, and beyond. The Baccalaureate and Beyond Longitudinal Study, which follows students for ten years after completion of a Baccalaureate degree, also incorporated SOGI metrics as of 2018[21]. The Beginning Postsecondary Students Longitudinal Study, which follows students through postsecondary education, incorporated SOGI metrics as of 2020[22].

"Professional societies and NGOs also collect SOGI data and often have high levels of trust from their membership leading to excellent survey-response data."

  In the UK, the Higher Education Statistics Agency (HESA), has collected SOGI items on their administrative census of all graduates of universities. HESA is a non-governmental organisation which adheres rigorously to UK statistical standards, allowing data harmonisation. SOGI data is incorporated in the Graduate Outcomes Survey, or its predecessors the DLHE and Longitudinal DLHE (Destinations of Leavers from Higher Education) records, which can both be linked back to the student record data, where the special category (EDI) personal information is held.[1*]

[1*] HESA publishes extracts of these datasets in a disclosure controlled manner as Open Data on their website: https://www.hesa.ac.uk/data-and-analysis. They supplement these data with a range of services, which are described here: https://www.hesa.ac.uk/services

"SOGI minorities were found to be 7% less likely to stay in a STEM degree than their peers, despite higher rates of participation in undergraduate research"

- **Universities, professional societies, non-governmental organisations, and academic researchers have also led efforts to illuminate attrition and the challenges facing LGBT+ people in STEM.**
  Universities are in a strategically important position to alter the landscape of retention by increasing data collection and collation across institutions and sharing best practices to encourage a welcoming STEM environment. Professional societies and NGOs also collect SOGI data and often have high levels of trust from their membership leading to excellent survey-response data. Participants examined some examples of these, including university-led climate and perceptions studies which have identified important factors in poor LGBT+ retention in STEM at both the undergraduate and faculty levels.[24, 25] One such study determined that SOGI minorities were found to be 7% less likely to stay in a STEM degree than their peers, despite higher rates of participation in undergraduate research, a key predictor of STEM retention.[23] In the UK, a joint report from the Universities and Colleges Admissions Service and Stonewall did important work to identify the percentage of university applicants who are LGBT+. [26] At the professional level, a few key studies and surveys have examined climate in STEM workplaces. [29,30] Notably, the American Association for the Advancement of Science (AAAS) has launched a new SOGI data collection initiative, in addition to its SEA change programme.[27, 28] Participants encouraged transatlantic collaboration in ongoing phases of all of these to enrich data and share best practice.

"For the first time in history, SOGI data collection is gaining momentum"

"Harmonisation of SOGI items will support data analysis and allow for systems to be nimble with respect to context and needs of the organisation /community"

"participants recommended that the harmonised standards be continually reviewed in collaboration with the LGBT+ community to keep pace with the evolution of queer and trans identities over time"

Participants welcomed this trajectory of increased data collection. In order to translate this momentum into datasets which can support effective policymaking to increase retention, they recommended:

- **Increased analytical capability and policy intentionality**
  For the first time in history, SOGI data collection is gaining momentum. Participants cautioned against viewing this as the 'end' rather than the 'means' to increased retention for SOGI minorities in STEM. They recommended a focus on downstream analytical capabilities to ensure that data collected adequately answers user needs. Additionally, participants recommended clarifying policy intentionality at all stages of collection.[2#]

- **Data harmonisation at a national and international level**
  Harmonisation of SOGI items will support data analysis and allow for systems to be nimble with respect to context and needs of the organisation /community**.** Harmonised standards have been created by the UK Office for National Statistics, the US Office of Management and Budget (OMB) working group on SOGI data collection, and the US Federal Committee on Statistical methodology. These standards, built in concert with LGBT+ communities, offer a framework for data collection but do not intend to be a rigid protocol on data collection. Additionally, participants recommended that the harmonised standards be continually reviewed in collaboration with the LGBT+ community to keep pace with the evolution of queer and trans identities over time. Participants emphasised harmonisation to enhance interoperability of data sets, noting that these remain flexible to respond to the specific questions being asked at individual institutions.

[2#] Participants considered the potentially circular issue of SOGI data in many institutions whereby sexual orientation and gender identity are not classified as underrepresented and therefore are not allocated resources for data collection. Of course, this leads to insufficient data collection, and the inability to show definitively that LGBT+ individuals are marginalised in STEM. This cycle must be broken to continue moving forward with allocation of resources and programming which addresses this issue.

"Identities, especially intersectional identities, represent small subsets of the data and are therefore often left out of analysis which can lead to the exclusion of the needs of the most marginalised"

- **Disaggregation and qualitative methods**
  Identities, especially intersectional identities, represent small subsets of the data and are therefore often left out of analysis which can lead to the exclusion of the needs of the most marginalised. Participants recommended that further work be done to address who is left out and how their experiences can be assessed through disaggregation or qualitative methods.
- **The UK and the US collect different types of data but taken together are more than the sum of their parts**
  The cultural, political, and structural contexts of the UK and US differ and have thereby led to differences in data collection on SOGI items, as described above. Attitudes towards the LGBT+ community, intrinsic community differences, education and funding structures, and the politicisation of SOGI data collection are just a few of the key differences when assessing the current landscapes in the respective countries. The combination of both countries' strengths in data collection can help cover some of the gaps that either country faces independently to create a more in-depth picture of LGBT+ representation and their experiences in STEM.

Recommendations for data collection at institutions

Participants agreed on the following eight tenets which might assist institutions in their first forays into collecting SOGI data. The Appendix includes supporting materials.

1. **Share best practices – Universities might be unique but they're not alone**
   Crafting sensitive questions, building community trust, overcoming legal barriers, ensuring robust privacy policy: All of these are issues that almost every institution faces in first-generation efforts at gathering SOGI data.[3‡] Participants agreed to help institutions informally connect with one another as they undertake these processes, and began this in earnest at Wilton Park. Participants formed continued informal working groups.

"Institutions should recognise the cultural and political milieus which contextualise their collection efforts."

2. **Self-educate on the issues in advance**
   Institutions should recognise the cultural and political milieus which contextualise their collection efforts. Unfortunately, these still sometimes include active violence and attempts to delegitimise or leverage those with LGBT+ identities. Sensitivity to these issues in data collection and how this relates to other intersectional identities is key. Similarly, to avoid survey fatigue (particularly prevalent in campus climate surveys), the existing literature and data on LGBT+ experiences in STEM should be considered. Some resources are provided in the Bibliography.

"Thoughtful co-production is key to developing trust and building relationships - both those being consulted and those being surveyed."

3. **Involve LGBT+ communities in the entire process**
   Thoughtful co-production is key to developing trust and building relationships - both those being consulted and those being surveyed.[4§] Ensure a clear plan for communicating the data uncovered about these communities back to the communities themselves. Participants recommended that institutions take careful note of the possibility of a small number of LGBT+ individuals taking on considerable burdens to assist universities in this work (often at personal cost). Institutions can

[3‡] Participants commended, for example, the open blog of Chief Statistician of Scotland sharing challenges to SOGI data collection candidly here: Sex, gender identity, trans status - data collection and publication: guidance - gov.scot (www.gov.scot)

[4§] Some groups emphasised that addressing LGBT+ retention in STEM must prioritise voices from the set of characteristics referenced in the Equality Act of 2010, including people of colour, women, trans, intersex, and asexual individuals, people with disabilities, and a mix of socioeconomic backgrounds. Doing so can help ensure that even those with the smallest numbers have a voice, and the experiences of the community can be defined by the most marginalised within the community.

"an institution should have the resources to collect sensitive personal information, analyse it, distribute the results, and consider the downstream policy interventions accurately and safely."

mitigate against this by rewarding the consultation work. In general, clear communication and agreement upon the expectations, outcomes, and accountability during the co-production process can ensure trust with the communities involved.

4. **Conduct an internal audit of available resources and capacity for change.**
   Prior to collecting SOGI data, an institution should have the resources to collect sensitive personal information, analyse it, distribute the results, and consider the downstream policy interventions accurately and safely. The process requires a myriad of professionals and resources including IT and analytics staff, communications departments, survey design professionals and others. Participants cautioned against the harms that can result in under-resourced or hasty data collection efforts, and highlighted the need for community involvement in survey development to mitigate some of the risks and harmful practices noted above.

5. **Design data collection with purpose.**
   It is valuable to define and communicate the institution's motivation for collecting the data as part of survey communications. As more institutions begin collecting data, there is a risk of collection for collection's sake and while more data might be helpful to researchers, institutions are unlikely to get buy-in from the communities they are trying to serve. When leadership is clear about data uses (in this case the improvement of the STEM system for all marginalised groups) those groups are more likely to respond. Harm reduction and the empowerment for change should be at the forefront of these data collection efforts.

"Harm reduction and the empowerment for change should be at the forefront of these data collection efforts"

   A considered approach to data collection will also help refine which survey questions are needed. For example, the question of "sex assigned at birth" is a topic of disagreement within the LGBT+ community and may not be necessary in circumstances when gender is the primary demographic of interest. Additionally, the need for quantitative versus qualitative data should be assessed. If quantitative data is desired, is there sufficient sample size and statistical power to answer the question being asked? If sample sizes are insufficient, can the data be safely disaggregated to identify the needs of smaller subsets of the community?

6. **Employ rigorous methodologies.**
   Designing SOGI data collection and analysis requires balancing scientific and methodological rigour with flexibility and evolution of language. LGBT+ identities do not often fall neatly into a predefined set of categories, and the language used by the community is subject to change over time. Consideration must be given by professional teams to (1) data aggregation and analysis (2) use of mixed qualitative and quantitative methods, (3) reproducibility, (4) data interoperability, (5) sufficient statistical power in the questions to provide insight from the collected data, and (6) incorporation of relevant techniques for multidimensional demographic data such as relational data analysis and cluster analysis. Data analysts should have both technical and cultural competence.

"Data analysts should have both technical and cultural competence."

7. **Ensure privacy and security.**
   Small sample sizes and the sensitive nature of SOGI data require particular attention to individuals' safety and protections against discoverability. Practitioners should remain vigilant and note that many social, cultural, and political contexts involve risks to the LGBT+ community. The collection of SOGI data requires investments in secure data systems and in IT staff, as well as accountability against data misuse.

"many social, cultural, and political contexts involve risks to the LGBT+ community."

8. **Use inclusive and flexible language in survey items**.
   The topic of sexual orientation and gender identity is complex, with evolving language and disagreement even within the LGBT+ community as to the best ways to ask questions. Institutions collecting SOGI data should therefore be open to feedback and willing to adapt based on their needs. General guidelines include:

   - Organise response options alphabetically rather than putting "straight" or "male" first by default.

"Poorly designed questions can cause harm or retraumatise respondents."

- Avoid "other". Instead, use "another identity not listed" or "write-in". Resources exist for dealing with and coding free text responses.
- Make all questions optional or include "decline to answer" options.
- Include checkboxes to enable select-all-that-apply functionality for response items.
- Avoid outdated or offensive terms. Poorly designed questions can cause harm or retraumatise respondents.
- Recognise the fluidity of identities and language over time. Data should be harmonised for interoperability but not necessarily standardised.

Conclusions and next steps

Data is powerful, not just in helping understand the status quo in the STEM enterprise, but also in helping the scientific community design and deliver evidence-based actions to ameliorate it. Until now, a dearth of SOGI data has created challenges for understanding when and why LGBT+ people leave STEM and for designing data-driven policy interventions. The collection of this data has been fraught with challenges of technical, legal, cultural, operational natures. Social science and survey science have helped to overcome some of these challenges with rigorous and sensitive methodology. These factors have led many, including the US National Science and Technology Council, to state that "data collection must start immediately."

The UK and the US are powerful partners on this issue on many fronts. Wilton Park participants highlight the following as key areas of continued work:

- sharing best practice on data collection and design
- highlighting synergistic datasets and working to share and learn from these
- continuing bilateral dialogue and collaboration on data-driven policy interventions to support retention
- establishing bilateral instruments to support UK-US leading social science research on this issue.

Acknowledgments

Kolin Clark and Shane Coffield served as rapporteurs for the Wilton Park conference and for this report. Wilton Park thanks Ronit Prawer and the UK Science and Innovation Network for their leadership of this bilateral initiative. Colbie Chinowski, Anna Dye, and Jacob O'Connor also contributed to an initial report on US/UK SOGI data collection in collaboration with the National Science Policy Network, which led to this event. Workshop participants and Garrett Dunlap contributed feedback and edits to the report. We also acknowledge the previous work of the Royal Society of Chemistry, Institute for Physics, and Royal Astronomical Society in their efforts to coalesce discussion around these topics.

Kolin Clark and Shane Coffield
Wilton Park | August 2023