A QUICK GUIDE TO DATA COLLECTION FOR LGBTQ+ CHARACTERISTICS

ALEXANDER L. BOND AND TYLER L. KELLY (LGBTQ+ STEM)

Abstract. We provide guidance on how to handle data collection with respect to LGBTQ+ characteristics and reasons one may look to use such data. We then outline model questions, various pitfalls one can introduce in questionnaires, and provide guidance on how to update one's questionnaire over time.

1. Introduction

In recent years, there has been an increase in interest from professional organisations, funding agencies, and academic institutions in inclusion of people with diverse sexual and gender identities (LGBTQ+ people) in science, technology, engineering and mathematics (STEM). Over the last 40 years, the language for how people have identified in this area has evolved and it will continue to evolve. Many organisations have adapted their own guidelines, or use those developed by others but which are now outdated and no longer reflect best practice.

Such data collection is essential to measure progress against organisational equity, diversity and inclusion (EDI) aspirations, evaluate change over time or in relation to initiatives, and to communicate the diversity of a group to others within or outwith the organisation [Stonewall19]. This latter reason is beneficial to other LGBTQ+ workers who will see themselves reflected in the organisation, which is an important aspect of professional belonging.

2. Why collect these data?

Historically, data on member or staff diversity was often limited to visible characteristics such as sex, race, and disability. Indeed, it is only since 2015 that there has been a more concerted effort to capture data on the LGBTQ+ identities of STEM workers [YM16]. Despite this, a lack of data is consistently highlighted as a shortcoming [APPG21, HOC22], and organisations hesitate in performing actions without having data that prompts a response. However, we note that without proper data, LGBTQ+ people are by definition unrepresented (and hence underrepresented) in one's organisation and this requires action.

LGBTQ+ workers in STEM face many systemic barriers not faced by their straight counterparts, including career limitations, professional devaluation, and harassment [CW21]. Often, these concerns are dismissed owing to a lack of data. Some professional organisations have started to include data collection on their membership [RSC22, RAE23], while others do not include any data on LGBTQ+ participation [RS21]. It is important to collect data with respect to gender identity, sexual identity and trans identity to ensure equality over protected characteristics. Without data, there is no representation.

In STEM, there are many ways in which inequity can present itself, and best practices involve constantly overseeing and evaluating a body's actions. For example, one can use data to evaluate:

- Whether certain groups in the LGBTQ+ umbrella are underrepresented in their community or organisation.
- Whether the people they are platforming in their event or organisation properly represent diverse genders and sexualities.
- If there is bias affecting their awarding of grants, prizes, honours, or peer-review processes.

- If research opportunities, leave, or resources are disproportionately allocated to non-LGBTQ+ researchers.
- Whether there is an attainment gap in one's industry or organisation with respect to LGBTQ+ identities in promotion.
- Establishing whether or not there is an LGBTQ+ pay gap among people in similar roles or across the entire organisation.
- If an organisation's LGBTQ+ employees are disproportionately represented in non-research roles, such as teaching, outreach, or administration.
- If LGBTQ+ people are disproportionately not applying to their job advertisements.
- If LGBTQ+ people have higher or lower success rates in job applications in senior academic or administrative positions.
- How many badges of each set of pronouns one needs to purchase before a large event.

These are important questions that need to be addressed to ensure an inclusive community in STEM and none of them can be answered without data.

Organisations that *do* collect data often do so in ways that exclude individuals based on their identities or conflate gender and sexual identity. Such questions can unintentionally introduce bad data. This is often not done maliciously, but a result of sometimes confusing and outdated guidance.

This document aims to provide a modern adaptable guide in order to help organisations ask questions with respect to gender, sexual and trans identities. In particular, we provide model questions for asking about these characteristics and highlight potential pitfalls and misinterpretations to be wary of. This document is primarily written for a UK audience, and we recognize that some terms may differ in their use in other places (for example, we suggest adding *two-spirit* as an option for sexual identity in North America). Data collection should always be done to reflect the audience in question, and further adaptations to these recommendations should be made with that in mind and in consultation with the groups about whom data collection is targeted.

Response rates can be lower for LGBTQ+ people who may not wish to volunteer information about themselves, especially to organisations who have historically been less than supportive or even actively harmful to LGBTQ+ people. Some may even consider signifying that they do not wish to disclose information to be divulging that they do not fit with the (likely) majority response. This is an issue of organisational trust rather of survey or question design. If organisations are concerned about the potential for significantly lower response rates from LGBTQ+ respondents, the root causes of that mistrust should be the primary focus.

3. Best practice for questionnaires

When asking about sexuality or gender in order to ensure equality or diversity in an organisation and event, best practice gives respondents agency to identify as they see fit. While this may present more challenges in terms of data analysis, it is the most inclusive approach and the least likely to become outdated as user responses can vary freely. All questions should be optional, as individuals may not wish to disclose certain aspects of their identity to the organisation.

Organisations can provide example terminology, but should not restrict responses. Organisational, national, or other guidance may require infrequent responses to be pooled to ensure individuals are not identifiable. Such decisions should be made in consultation with LGBTQ+ community members, and organisations must recognize that such decisions may not be clear-cut, and may vary over time.

It is most useful to divide questions into three broad categories: gender identity, sexual identity, and trans identity. Often the terms used for these three areas are conflated, confused, and intermixed, leading to lower data quality, and feelings of exclusion. For example, "cis gay man" combines the identity "cis" (where one's gender identity matches that assigned at birth), the sexual

identity "gay" with the gender identity "man", and therefore artificially reduces the subset of individuals who would select this option. Similarly, questions that ask if individuals "are members of the LGBTQ+ community" or "identify as LGBTQ+", though well-meaning, conflate these three concepts, and result in poor quality data that is not fit for purpose.

It is important to inform respondents the reasons for which the organisation is collecting these data and how it will be accessed or processed. For example, in some cases, aggregate data may be appropriate, while in other circumstances, individual-level data are required. Organisations should have an objectively justifiable reason for collecting data, and articulate that at the outset of the survey.

Finally, if a defined vocabulary of terms *must* be used, these should be presented as multiple choice options, allowing respondents to select more than one option. It is important to list the options in a non-prejudicial way—typically putting options in alphabetical order is a common solution. Moreover, the options of "*I do not know*" and "*prefer not to say*" must be included so that individuals are not forced to choose between responding and potentially putting themselves in a harmful situation. Lastly, there must be a free-text option that allows respondents freedom to identify how they see fit.

3.1. **Gender identity.** Gender identity refers to a personal sense of one's own gender and concept of self. The first question we recommend is:

Question 3.1. *What term best describes your gender identity?*

If a set vocabulary of options is required, we recommend the following:

Please check all that apply:
- ☐ *agender / no gender*
- ☐ *female*
- ☐ *genderqueer*
- ☐ *male*
- ☐ *non-binary*
- ☐ *I do not know / questioning*
- ☐ *prefer not to say*
- ☐ *I am:* _____

Cautions. A variety of approaches have been used in the scientific community to collect data on gender identity. Below, we outline why we discourage some relatively common approaches.

- **Do not use perceived gender identities.** While a person may express behaviours, attitudes and appearances that are aligned with a particular gender in your interactions with them, there are multiple reasons that this may differ from their gender identity. It is important to accurately capture an individual's own gender identity rather than how another may perceive their gender identity. We thus highly discourage trying to deduce this for yourself rather than asking a question, as then one is just creating bad data.
- **Do not assign automated gender.** Automated assignment of gender based on databases of first names has been a common tool in examining gender bias, for examine in journal authors or editorial boards, grant or award recipients, or other groups of individuals. These approaches use large volumes of data (typically birth records) to assign a probability of a name referring to a female or male, and then assign that probability to the individual [Wai16, Mul20, Elm13]. In these cases, individuals' inability to self-identify, the geographically-restricted set of names used (often North American or western European), and use of a binary gender are not inclusive and actively misgender individuals. Though there are often disclaimers included both in the processes themselves and the resulting publications,

these are most often dismissive and there is little or no attempt to address the biases and shortcomings [FBM16, TS16].

- **Do not use sex unless truly required for your purposes.** Many organisations have aimed to apply the UK's census questions about gender directly. The UK census used the following questions around sex and gender: "What is your sex?" and "Is the gender you identify with the same as your sex registered at birth?" The census' needs probably differ from those of scientific or research organisations (e.g., a learned society is probably not aiming to understand how many urologists or gynaecologists are needed to serve a geographic area effectively). This would mean that your organisation probably does not need information about individuals' biological sex. In most situations, organisations are truly seeking information on respondents' gender.
- **Make it possible to let a trans person to provide their gender identity without disclosing their trans identity.** Another issue with the UK census questions is that it is impossible for a trans person to provide their gender identity without also providing their trans identity. For example, this effectively does not allow trans women to identify as women. This does not give your respondents full agency when answering your questions. There are many reasons that a trans person may not want to divulge their trans identity in your questionnaire. This issue can be resolved by separating the questions about trans and gender identity as we outline here.

3.2. **Sexual identity.** Sexual identity refers to one's personal sense of romantic and/or sexual attraction to others. We recommend the following question:

Question 3.2. *What is your sexual identity?*

Recall that best practice provides an open text box for respondents to best describe their identity. However, if a set vocabulary of options must be used, they should allow multiple choices to be selected, and we recommend the following:

Please check all that apply:
☐ *asexual*
☐ *bisexual*
☐ *gay*
☐ *lesbian*
☐ *pansexual*
☐ *queer*
☐ *straight or heterosexual*
☐ *I do not know / questioning*
☐ *prefer not to say*
☐ *I am:* _____

Cautions.

- **Question the standard usage of *gay man* and *gay woman* for sexual identity.** Many standard templates use these terms to distinguish between gay men and women. The issue with using these options is that renders a nonbinary person unable to identify as gay without misgendering themselves, and conflates two concepts: a sexual identity (gay) and gender identity (woman or man). The questions above are made so that they decouple these two aspects, while enabling data to be cross-tabulated to obtain intersectional information if required.
- **Don't be afraid of the word queer when used appropriately.** While historically the term *queer* has been used as a slur, many people in the LGBTQ+ community have since reclaimed it as a term of identity and empowerment. It is often used as an umbrella term for

people that identify not in the majority with respect to sexual identity. Grammatically, this reclaiming has repurposed the word to be an adjective or a verb (e.g., "I am a queer woman"; "queering the curriculum"), rather than that using it as a noun which is typically regarded as offensive.[1] Despite the fact that many use it as an umbrella term, many members of the LGBTQ+ community do not identify with the term and would prefer to identify with more well-established terminology.

3.3. **Trans identity.** Broadly speaking, a transgender person (or trans person) is an individual whose gender identity does not match the one that they were assigned at birth. There are multiple ways to ask one about their trans identity, depending on the purposes the organisation is trying to accomplish. We recommend giving multiple choices and clarification on what the purpose of asking the question. An example of such a question is as follows:

Question 3.3. *Is your gender or gender identity different to that assigned at birth?*

☐ *no*
☐ *yes*
☐ *prefer not to say*
☐ *custom response:* _____

Cautions.

- **Do not mix gender identity and trans identity questions.** For a variety of reasons, one may want to disclose their gender identity but not their trans identity or vice versa. For this reason, mixing the questions to have options such as "cis man" and "trans man" can lead to less data than what one could obtain with two questions.
- **Consider if you need a more bespoke question.** Our guidance typically relates to questions concerning representation or climate in a professional body or research community. If an organisation is looking to understand if their support for trans colleagues during transition are serving this community appropriately, employers may want to consider alternative questions such as "Is your gender identity that which you were assigned at birth?" Such questions should be developed with the trans community's input.

3.4. **Pronouns.** Organisations may want to ask about individuals' pronouns. Pronouns often relate to one's gender identity, but they do not uniquely map onto various gender identities. Individuals can have multiple pronouns, which may reflect diverse aspects of their identity. This should be a free text field for respondents to complete, but if one cannot do this, we provide options below.

Question 3.4. *What pronouns should be used to refer to you?*

Check all that apply
☐ *he/him*
☐ *she/her*
☐ *they/them*
☐ *any pronouns*
☐ *no pronouns*
☐ *prefer not to say*
☐ *I use:* _____

As with all questions, it should only be included if capturing this data is required and is not covered by other questions in the survey. We remind the reader this question should be optional, as is the case for all of the above.

Cautions.

[1] We leave an example as an exercise for that controversial uncle you try to avoid during holidays.

- **Reminder: Be clear with respondents how this information will be used.** It is good practice to include pronouns on conference badges at events so that correct pronouns are used and gender-diverse individuals feel that they are in a safe space. However, it is important when asking respondents for their pronouns that its usage in such ways is clear. Many people are not ready for their pronouns to be known broadly or may use different pronouns in different contexts, and this may put event attendees in uncomfortable situations, so it is important this question stays optional.
- **Do not ask respondents for their *preferred* pronouns.** In the past, personal pronouns were referred to as "preferred pronouns," implying that other pronouns were acceptable even if not preferred. This is an outdated term that is not used and can cause offense.

3.5. **Titles.** Titles, such as the gendered "Mr" or "Mrs", or gender neutral "Dr" or "Prof" are often collected alongside personal information, such as names, though should not be used in any analysis. Titles may not reflect the gender of respondents. Given the breadth of potential titles, it is simplest to ask respondents to fill in a free text field. This ensures those who may prefer titles such as "Mx" can include them, while not requiring an exhaustive list for data that is primarily administrative.

4. Discussion

Data collection about individuals' identity must be done in a secure, thoughtful way, and each question asked with a specific reason in mind. The autonomy of respondents must trump concerns about comparability, and data analysis methods must be flexible enough to accommodate changes in terminology over time. It is important to note that LGBTQ+ identities can be fluid and change, and people should have the agency to update how they identify in data over time.

Moreover, language changes over time. If your organisation is using a multiple choice option and is seeing multiple respondents provide the same custom response, then consider adding it to the list of options.

Many organisations have institutional bodies or groups that represent their LGBTQ+ staff, membership, or constituents. Working with those groups in survey design and implementation will improve the quality and quantity of the data collected. This can include trialling responses with a subset of individuals, refining questions to reflect cultural or geographic differences in terminology, or providing a critical eye into the purpose of data collection in the first place.

5. Acknowledgements

We acknowledge Beth Montague-Hellen, who has written a predecessor of this document for LGBTQ+ STEM [MH18]. We also thank Izzy Jayasinghe for conversations that have improved the quality of this paper.

References

[APPG21] APPG on Diversity and Inclusion in STEM. 2021. Inquiry into equity in the STEM workforce, final report. British Science Association, London.

[CW21] Cech, EA, Waidzunas, TJ. 2021. Systemic inequalities for LGBTQ professionals in STEM. Science Advances 7: eabe0933. doi: https://doi.org/10.1126/sciadv.abe0933

[Elm13] Elmas F. (2013) *SexMachine 0.1.1.* https://pypi.org/project/SexMachine/

[FBM16] Fox, C.W., Burns, C.S. and Meyer, J.A. 2016. *Editor and reviewer gender influence the peer review process but not peer review outcomes at an ecology journal.* Functional Ecology, 30: 140-153. doi: https://doi.org/10.1111/1365-2435.12529

[HOC22] House of Commons Science and Technology Committee Oral evidence: Diversity and inclusion in STEM, HC 903. 27 April 2022. https://committees.parliament.uk/oralevidence/10150/html/

[MH18] Montague-Hellen B. *Asking about Gender & Sexual Orientation in your Questionnaire.* doi: https://doi.org/10.6084/m9.figshare.6550277

[Mul20] Mullen, L. (2020). gender: Predict Gender from Names Using Historical Data. https://github.com/ropensci/gender

[RAE23] Royal Academy of Engineering. 2023. Inclusive cultures in engineering 2023. Commentary. Royal Academy of Engineering, London.

[RS21] Royal Society. 2021. Diversity data report. DES7487. Royal Society, London.

[RSC22] Royal Society of Chemistry. 2022. Diversity data report. Royal Society of Chemistry, London.

[Stonewall19] Stonewall UK. 2019. Understanding LGBT experiences: a guide for equalities monitoring for the UK. Stonewall UK, London.

[TS16] Topaz CM, Sen S (2016) Gender Representation on Journal Editorial Boards in the Mathematical Sciences. PLoS ONE 11(8): e0161357. doi: https://doi.org/10.1371/journal.pone.0161357

[Wai16] Wais K. *Gender Prediction Methods Based on First Names with genderizeR*, The R Journal,Vol. 8/1, Aug. 2016, https://journal.r-project.org/archive/2016-1/wais.pdf

[YM16] Yoder JB, Mattheis A. 2016. *Queer in STEM: Workplace Experiences Reported in a National Survey of LGBTQA Individuals in Science, Technology, Engineering, and Mathematics Careers*. Journal of Homosexuality 63(1):1-27. doi: https://doi.org/10.1080/00918369.2015.1078632.

Alexander L. Bond
Natural History Museum, Akeman Street,
Tring, Hertfordshire, United Kingdom HP23 6AP
Email: **a.bond@nhm.ac.uk**

Tyler L. Kelly
University of Birmingham, School of Mathematics
Birmingham, United Kingdom B15 2TT
Email: **t.kelly.1@bham.ac.uk**