How can we make conferences more inclusive? Lessons from the International Ethological Congress

Rebecca S. Chen[1*], Tuba Rizvi[¶2], Ane Liv Berthelsen[&1], Anneke J. Paijmans[& 1], Avery L. Maune[& 3], Barbara A. Caspers[& 3,4], Bernice Sepers[& 1], Isabel Damas-Moreira[& 3], Isabel Schnülle[& 3], Jana Könker[& 3,6], Joseph I. Hoffman[& 1,4,5,7], Joelyn de Lima[& 8], Jonas Tebbe[& 1,3], Kai-Philipp Gladow[& 9], Lisa de Vries[& 10], Marc Gilles[& 3], Nadine Schubert[& 3], Nayden Chakarov[& 4,9], Peter Korsten[& 11], Petroula Botsidou[& 1], Sabine Kraus[& 3,9], Stephen M. Salazar[& 9], Svenja Stoehr[& 9], Wolfgang Jockusch[& 12], Öncü Maraci[¶3]

[1] Department of Evolutionary Population Genetics, Bielefeld University, Bielefeld, Germany

[2] Department of Evolutionary Biology, Bielefeld University, Bielefeld, Germany

[3] Department of Behavioural Ecology, Bielefeld University, Bielefeld, Germany

[4] Joint Institute for Individualisation in a Changing Environment (JICE), Bielefeld University and University of Münster, Bielefeld, Germany

[5] British Antarctic Survey, High Cross, Madingley Road, Cambridge, United Kingdom

[6] Department of Biochemistry and Physiology of Plants, Bielefeld University, Bielefeld, Germany

[7] Center for Biotechnology (CeBiTec), Faculty of Biology, Bielefeld University, Bielefeld, Germany

[8] The Teaching Support Center & Center for Learning Sciences, The Swiss Federal Institute of Technology, Lausanne, Switzerland

[9] Department of Animal Behaviour, Bielefeld University, Bielefeld, Germany

[10] German Institute for Adult Education, Bonn, Germany

[11] Department of Life Sciences, Aberystwyth University, Aberystwyth, United Kingdom

[12] Berliner Akademie für Mediation und Interkulturelle Kommunikation (BAMIK GmbH), Berlin, Germany

* Corresponding author

E-mail: rebecca.chen@uni-bielefeld.de

[¶] These authors contributed equally

[&] These authors are listed in alphabetical order

Short title: Inclusivity at academic conferences

Abstract

Despite growing awareness of the importance of researcher diversity, barriers to inclusion and equity persist in science and at academic conferences. As hosts of the 37th International Ethological Congress, "Behaviour 2023", we studied equity, diversity and inclusivity (EDI) issues using observational and experimental behavioural data collected during question and answer (Q&A) sessions in addition to surveys conducted before and after the congress. Perceived women asked fewer questions than perceived men because they raised their hands less often to ask questions, and not because they were chosen less often by the session host. Self-reports indicated that women felt more comfortable asking questions when their own gender was represented (in the audience, by the speaker, and/or by the host) and when the setting was smaller. However, this pattern was not reflected in the observational data as perceived women asked fewer questions regardless of the situation. We report potential reasons why women asked fewer questions using survey data, and experimentally tested whether we could reduce gender disparity in question-asking. Our results indicate that session hosts cannot mitigate the gender disparity in question-asking by actively selecting women to start the Q&A session. We addressed further inclusivity barriers of underrepresented minorities beyond gender in a post-congress survey, which showed that underrepresented minorities did not have a more positive or negative congress experience but did perceive EDI issues as more severe. We conclude by providing recommendations for organising more inclusive scientific events, such as (i) ensuring that people who are less likely to ask questions do not miss out on academic opportunities, (ii) organising topic-, language-, and/or career-stage specific discussions and (iii) utilising technology to make presenting and listening smooth for everyone.

Introduction

Diversity within the scientific community is essential for advancing science because it facilitates the inclusion of a wide range of perspectives and contributions. There is growing evidence that increased gender and/or ethnic diversity can benefit science as a whole (1) by increasing productivity (2,3) , delivering higher quality science (4), and producing papers with higher scientific impact (5). Despite these known advantages of researcher diversity in academia, the persistent lack of underrepresented minorities (groups of people whose representation in academia is lower compared to their representation in the general population) and ongoing inequities remain ubiquitous, including in the biological sciences (6–10).

Unwelcoming, unsupportive and/or hostile working environments for certain groups of people, known as "chilly climates", and systemic biases can impede equity, reduce diversity and hinder inclusion in academia. Such groups of people include women, ethnic minorities, LGBTQ+ (lesbian, gay, bisexual, transsexual, queer) researchers, and people with a disability. Chilly climates can manifest in the form of discrimination (active or passive), harassment, microaggression and professional devaluation based on sexism (11–13), racism (14–16), queerphobia (17,18) and ableism (19–21), amongst others. These factors can prevent certain groups from entering, progressing or staying in academia (22–24), ultimately leading to the underrepresentation of these groups over time and to a reduction in researcher diversity. In recent years, solutions to address this "leaky pipeline" affecting the underrepresentation of senior women (25–27) and senior scientists from other underrepresented minorities (28–30) have been discussed. Additionally, several initiatives and guidelines have been formed to promote inclusive academic environments more broadly (e.g. 9,31–34).

Systemic biases or prejudices arise due to academic cultures or built-in systems (e.g. institutional policies) that disadvantage certain groups of scientists. Some groups might have different needs to ensure a healthy work-life balance (e.g. due to caretaking duties; 35,36), get access to mental health services (e.g. due to being more vulnerable to experiencing mental health issues which is the case for LGBTQ+ scientists (17)), or receive adequate mentorship (e.g. due to cultural differences (37)). These aspects are essential to progress and excel, yet the focus is often put on distributing resources equally rather than equitably (38,39). Moreover, systemic bias can lead to some groups experiencing feelings of exclusion, which can ultimately reduce academic performance (40,41). For example, financial constraints disproportionately restrict exposure to the field of ecology among ethnic minorities and groups with low socio-economic status, reducing the number of role models and decreasing the sense of belonging for these groups (29). Additionally, physical and social inaccessibility on campuses (42) and physical limitations during fieldwork (43) can contribute to the exclusion of researchers with a disability. Policies and practises subject to systemic bias need to be reformed to ensure an academic environment that accommodates all social identities, as they perpetuate unfairness and hinder the inclusivity of marginalised individuals. To effectively improve these policies and practices, we require a better understanding of existing barriers to inclusion and equity, for example by encouraging dialogue and observing what barriers unfold in natural settings.

Barriers to inclusion and equity also unfold at academic conferences. Conferences are crucial events for networking and gaining exposure as they provide a space to connect with researchers with similar interests, promote one’s own work (44) and collect information on jobs and funding opportunities, which are particularly important for early-career researchers (45). However, certain groups of people can face barriers to invitation, participation or recognition at scientific conferences. For example, women and ethnic minorities are underrepresented as invited speakers (46,47) and on average, women receive a lower turnout at talks than men (48,49). High registration fees and travel expenses also create obstacles for researchers from low-income countries, generating economic disparities. Additionally, factors such as a lack of proper accessibility to and within the conference venue, limited childcare options for caretakers, and the need for English proficiency can represent major barriers to ensuring an inclusive scientific conference.

Studies on equity, diversity and inclusivity (EDI) issues at conferences have increased in number over the last decade and awareness of common issues is growing, yet knowledge gaps persist. International conferences provide an excellent opportunity to understand EDI issues, as a diverse group of people comes together with regard to their educational background, (work) culture, and career stage, therefore not biasing observations towards a specific institute only. One frequently studied issue is the gender disparity in question-asking probability after oral presentations. Evidence from within and outside of the biological sciences show that women tend to ask fewer questions at Q&A sessions than their male peers (49–54), possibly due to a mix of factors like “not working up the nerve” or men asking the first question which has consequences for the rest of the session (50). However, the causes and consequences of this disparity remain unknown. Do women ask fewer questions due to lack of self-esteem, discouragement due to chilly climates, or direct discrimination against women that might hinder their active participation? And although session hosts are increasingly instructed to choose young researchers or women to ask questions, to what extent does this encourage these groups to ask more questions? In addition to gender disparities in question-asking behaviour, the exact barriers that certain social identities face must be identified, together with the more specific barriers that form when multiple identities intersect (e.g. barriers that are specific to women of colour or LGBTQ+ people with a disability). Barriers can arise from discrimination, prejudice and/or a tendency to dismiss specific contributions, which all play a role in forming a “chilly conference climate”, negatively impacting the experience of those affected. It is therefore crucial to improve our current understanding of contemporary EDI-related issues that occur at scientific conferences to be able to organise more inclusive events.

To address the knowledge gaps highlighted above, we conducted a comprehensive study during the 37th International Ethological Congress, 'Behaviour 2023', hosted at Bielefeld University, Germany. This congress focussed on animal behaviour and was attended by delegates from a range of backgrounds including ethology, behavioural genetics and anthropology. We used a combination of observational and experimental data collected from three different sources: (i) congress registration (quantitative self-reports regarding attendees' social identities; 727 responses), (ii) Q&A sessions (quantitative observational data regarding gender disparities in question-asking probability; 1278 questions asked in 67 sessions), and (iii) a post-congress survey (quantitative self-reports regarding congress experiences and perceptions of EDI issues as well as qualitative feedback; 391 responses). We experimentally tested whether session hosts can increase the probability that women ask questions by instructing them to either direct the first question after a talk to a male or female participant. Combining qualitative and quantitative (observational and experimental) data allowed us to gain a deeper understanding of key inclusivity issues related not only to gender identity, but also to nationality, sexual orientation, and disability.

Results

We investigated various aspects of inclusivity and inequality among social identities at a scientific event using a case study. The congress took place in August 2023 and was attended by more than 850 researchers. The language of the congress was English, and a total of 661 oral presentations were given, distributed across continuous parallel sessions, as well as eleven plenary talks presented by invited speakers.

The organising committee of the congress took a number of measures to boost inclusivity at the congress (see Methods for details). Briefly, (i) plenary speakers were invited to represent gender and ethnic diversity; (ii) all attendees were obliged to agree to an appropriate Code of Conduct (55); (iii) attendees were able to express concerns and report discrimination and harassment to an awareness team; (iv) help was offered to those with auditory, visual, mobility and/or dietary needs before and during registration; (v) limited travel grants were given to researchers based in the Global South; (vi) free childcare was offered and parent-children rooms were reserved for attendees and their families; (vii) a symposium on "Equality, diversity and equity in behaviour, ecology and evolution" was hosted and three EDI-related workshops were given by external facilitators; (viii) pronouns were optionally printed on name tags. We acknowledge that many more steps can be taken to foster inclusivity at academic conferences, such as hosting the conference in a hybrid format (56,57).

What social identities were present at the congress?

A total of 727 attendees took part in the pre-congress survey, which gathered data on the social identities of congress attendees. A total of 65% of the attendees who provided their pronouns used she/her (hereafter referred to as "women", but see S1 Methods 1) and 33% used he/him (hereafter referred to as "men", but see S1 Methods 1). Fifteen attendees (2%) used she/they, he/they or they/them pronouns. A total of 14.4% of attendees who responded to the question 'if they identified with the LGBTQ+ community' responded with "yes" (*n* = 92). A total of 59 nationalities were represented among the congress attendees. The majority of attendees were of European nationality (*n* = 481), followed by Asian (*n* = 85), North American (*n* = 48), Oceanic (*n* = 20), South American (*n* = 18), and African nationalities (*n* = 5). Most of the attendees with European, North American and Oceanic nationalities were female, but the majority of Asian and South American attendees were male (Fig 1). Lastly, four people acknowledged the need for some form of assistance during the congress due to either physical or mental disabilities.

Fig 1. Gender distribution of congress registrants across the continents of their nationality. Gender was inferred from people's self-reported pronouns during registration, and countries with a colour were those that were represented at the congress.

Is there a gender disparity in question-asking probability?

We tested for a gender disparity in question-asking probability using two lines of evidence based on (i) observational data on question-asking behaviour collected during Q&A sessions after oral presentations (388 questions asked after 134 unmanipulated talks that were not part of our experiment, see below) and (ii) self-reports on question-asking collected in the post-congress survey (373 complete responses).

To identify a gender disparity in question-asking probability using the observational data, we asked whether fewer questions are asked by women. We fitted a binomial generalised linear mixed effect model (GLMM), where the dependent variable indicates whether a question was asked by a perceived man (0) or a perceived woman (1), while accounting for the gender proportion of the audience and the non-independence of talks within a session (see Methods for details). Across all unmanipulated talks, 48% of questions were asked by perceived women without accounting the proportion of perceived women in the audience. The overall probability that a perceived woman asked a question, corrected for the proportion of perceived women in the audience, was 0.34 (GLMM intercept = -0.66, $p < 0.001$, Fig 2a; S2 Table 1), providing clear evidence that perceived women are less likely to ask questions compared to perceived men. When repeating the analysis with a more conservative dataset that excluded questions where the observer noted any source of uncertainty in the data collected, the results remained virtually identical (GLMM intercept = -0.67, $p < 0.001$; S2 Table 1). Moreover, we analysed whether age affects the gender disparity in question-asking (S3 Analysis 1), but because (i) we found a gender disparity regardless of age and (ii) this analysis was based on many assumptions that we do not think are always valid, we did not take age into account in further models on question-asking based on the observational data.

Fig 2. **Gender disparity in question-asking behaviour.** a) Intercepts and 95% confidence intervals (CI) for models QA.1, QA.2 and QA.3 that tested for gender disparities in asking questions, raising hands and being chosen respectively. Yellow points indicate statistically significant intercepts ($p < 0.05$). A negative intercept indicates that the probability that a woman asked a question was lower than expected from the number of women in the audience (male bias) whereas a positive intercept indicates that this probability was higher than expected (female bias); b) Model estimates and 95% CIs for the effect of gender on the probability of asking a question based on the survey data. Yellow points indicate statistically significant effects ($p < 0.05$). A negative estimate indicates that the gender in question is negatively associated with the probability that a woman asked a question (i.e. positively associated with the probability that a man asked a question; a male bias); c) Raw data with added jitter, null hypothesis and model estimates for model QA.2, which tested for a gender disparity in raising hands; d) Raw data with added jitter, null hypothesis (no gender disparity) and model estimates for model QA.3, which tested for a gender disparity in being chosen to ask a question.

In all of the models that used the observational data on question-asking, we perceived gender based on a person's self-reported pronouns when possible, or alternatively on the person's appearance. We acknowledge that perceiving someone's gender based on their appearance is predisposed to observer bias and assumes that gender identity can be visually assessed. We investigated how often we correctly perceived the gender of session hosts and speakers, and assessed how often one's pronouns differed from their gender (S1 Methods 1). Because the gender perceived by observers corresponded to the person's self-reported pronoun(s) in over 94% of observations of women and men, and over 97% of women and men use she/her and he/him pronouns respectively, we conclude that there is reasonably good agreement between observer perception and self-reported gender. In the rest of the Results, we refer to perceived women and perceived men as women and men respectively for simplicity purposes. Because observers had a low accuracy (27%) of correctly perceiving the gender of non-binary scientists, we focused on the question-asking behaviour of non-binary scientists using the post-congress survey only.

We further collected observational data on question-asking behaviour during plenary talks. We analysed these data separately because these sessions were held in larger lecture rooms attended by the vast majority of congress participants. Consequently, we corrected an estimated proportion of women in the audience by using the proportion of women registered at the congress as a whole, rather than correcting for audience counts. We did not correct for the proportion of women in the audience because it was unfeasible to count the audience by eye due to lack of visibility, size of the room, and difficulty keeping track of which people were and which people were not already counted. A total of 60 questions were asked during eleven plenary talk Q&A sessions, 17 (28%) of which were asked by women. Despite this relatively small sample size, the gender disparity in question-asking was even greater during plenary sessions as compared to regular oral presentations, with the probability of a woman asking a question being only 0.20 when correcting for the estimated proportion of women in the audience (GLMM intercept = -1.54, $p < 0.001$; S2 Table 1).

Similarly, we tested for a gender disparity in question-asking probability using self-reports from the post-congress survey, where we asked if fewer women asked questions. We fitted a binomial generalised linear model (GLM) using the binomial response to the question "Did you ask a question at the congress?" (1 = yes, 0 = no) as the dependent variable and the self-reported gender identity (woman, man, non-binary, other) as the independent variable. Note that this question therefore addresses the likelihood that a woman asked a question across the entire congress, as opposed to the observational data that addresses if questions were less likely to be asked by women based on each talk. Although including gender in the model barely improved the model fit (likelihood ratio test (LRT) $p = 0.05$), the results again showed that women were less likely to have asked a question during the congress compared to men (*beta* estimate female = -0.49, $p = 0.04$; Fig 2b; S2 Table 1), while non-binary people ($n = 7$) were not more or less likely to ask a question compared to men (*beta* estimate non-binary = 0.92, $p = 0.40$; Fig 2b; S2 Table 1).

Do women raise their hands less often or get chosen to ask a question less often?

We next tested whether women asked fewer questions than men did because they raised their hands less often to ask a question. We fitted a multivariate binomial GLMM where the dependent variable was the fraction of women who raised their hands over the total number of people who raised their hands, while accounting for the proportion of women in the audience and the non-independence of talks within a session (see Methods for details). The probability that a woman raised their hand was 0.36 (intercept = -0.58, $p < 0.001$; Fig 2a,c; S2 Table 1), indicating that women were less likely to raise their hand than men were.

Additionally, women might be chosen less often to ask their questions by the session hosts when both women and men raise their hands. We tested this hypothesis by fitting another binomial GLMM using the gender of the questioner as the dependent variable, but this time correcting for the proportion of the people who raised their hand that were women. In this model, we only included cases where at least one woman and one man raised their hand, so that the session host had to make a choice between assigning the question to one gender over the other. The probability that a woman was chosen to ask their question by the session host was 0.46, indicating that women were not chosen significantly less often by session hosts to ask their question compared to men (intercept = -0.14, p = 0.53; Fig 2a,d; S2 Table 1). We investigated this same question using the post-congress survey data, where we collected data on a person's gender and whether one of the reasons they did not ask a question was due to not being chosen despite raising their hand ("not being chosen" in short). We fitted a binomial GLM with the response to "not being chosen" as the dependent variable and self-reported gender as the independent variable. Including gender in the model did not significantly improve the model fit (LRT $\chi^2 = 1.49$, LRT $p = 0.47$) indicating that women were equally likely to be chosen to ask their question, in line with our results based on the observational data.

Why do women ask fewer questions than men do?

Next, we tested whether women and non-binary respondents were less comfortable asking a question using data collected in the post-congress survey. Respondents of the survey indicated their agreement to the statement "I feel comfortable asking questions during Q&A sessions" on a 7-point Likert scale (1 = "Strongly disagree", 7 = "Strongly agree"). We fitted an ordinal logistic regression model (OLR) for the response to this statement and included the self-reported gender as an independent variable while correcting for career stage (early, mid or late career). We found that both women (*beta* estimate = -1.26, SE = 0.21, t = -5.99, $p < 0.001$) and non-binary respondents (*beta* estimate = -1.60, SE = 0.68, t = -2.36, $p < 0.02$) felt less comfortable asking questions during Q&A sessions compared to men. Both mid-career (*beta* estimate = 1.04, SE = 0.21, t = 5.04, $p < 0.001$) and late-career researchers (*beta* estimate =2.48, SE = 0.33, t = 7.50, $p < 0.001$) were more comfortable asking questions compared to early-career researchers.

The post-congress survey additionally included questions on what aspect(s) motivated people to ask questions at the Behaviour 2023 conference (hereafter referred to as "motivations"), and what aspect(s) made people more hesitant to ask a question (hereafter referred to as "hesitations"). We tested in two steps if women asked fewer questions than men because they had different motivations and hesitations to ask questions than men did. The first step tested which motivations and hesitations were more often selected by women compared to men. We fitted multiple binomial GLMs, one per motivation and hesitation. In each case, the dependent variable was the binomial response whether the motivation or hesitation was ticked (1) or not (0), and the independent variables were self-reported gender and career stage (early, mid or late career). The second step tested which of the motivations and hesitations that were significantly affected by gender were significant predictors of the probability of a person asking a question during the congress, where we then examined which of the significant ones were also affected by gender. We fitted a second set of binomial GLMs, again one for each motivation and hesitation. The dependent variable in these models was the response to the question "Did you ask one or more questions during Q&A sessions?" (1 = yes, 0 = no) and the independent variable was the binomial response whether the motivation or hesitation was ticked (1) or not (0), while also including gender and career stage as covariates.

Including gender as an independent variable did not improve the fit of any of the models fitted to the motivations (FDR-corrected LRT $q < 0.05$; Fig 3a, S2 Table 2), indicating that women were not more likely to select any of the motivations compared to men. However, when looking at the hesitations, "afraid I would not be able to phrase/articulate my question well" was significantly affected by gender after correcting for multiple testing, where women were more likely to tick this hesitation compared to men (*beta* estimate for wome*n* = 0.90, *p* for wome*n* = 0.002, Fig 3c, S2 Table 3). Two more hesitations were affected by gender but were not statistically significant after correcting for multiple testing: "I did not have the confidence" (*beta* estimate for wome*n* = 0.78, $p = 0.01$, FDR-corrected LRT $q = 0.08$; Fig 3c, S2 Table 3) and "I felt intimidated by the audience" (*beta* estimate for wome*n* = 0.76, $p = 0.02$, FDR-corrected LRT $q = 0.13$; Fig 3c, S2 Table 3). For all of the other hesitations, the inclusion of gender did not improve the fit of the models (LRT FDR-corrected q-value < 0.05; Fig 3c, S2 Table 3). Early career researchers were more likely to tick almost all hesitations compared to mid- and late-career researchers (S2 Table 4).

Fig 3. The results of models that tested for gender disparities in question-asking due to different motivations and hesitations between men and women. Model estimates of the effect of female gender on (a) six motivations and (b) the effects of the motivation on the probability that the person asked a question during the congress, as well as the effect of female gender on (c) twelve hesitations, (d) the effects of the hesitation on the probability that the person asked a question during the congress. Yellow points indicate including the variable in the model significantly improved the model fit compared to the null model after correcting for multiple-testing.

We found that most of the motivations and hesitations that were predictive of question-asking probability were not influenced by gender (Fig 3b, Fig 3d, S2 Table 2, S2 Table 3). The only hesitation that varied significantly by gender (fear of the inability to phrase/articulate a question well) did not influence the probability of asking a question during the congress (*beta* estimate = -0.34, $p = 0.18$; Fig 3d, S2 Table 3). However, the two hesitations that were associated with gender only before applying a multiple-testing correction (lack of confidence and feeling intimidated by the audience) were significant predictors of the probability of asking a question (lack of confidence: *beta* estimate = -0.70, $p = 0.008$, FDR-corrected LRT $q = 0.02$; feeling intimidated by the audience *beta* estimate = -0.77, $p = 0.07$, FDR-corrected LRT $q = 0.02$; Fig 3d, S2 Table 3). Taken together, these results suggest that women are more likely to indicate that they are hesitant to ask a question because of a lack of confidence and/or feeling intimidated by the audience compared to men, which may make them less likely to ask a question, although insignificant after multiple-testing correction.

What conditions might encourage women to ask questions?

We investigated which conditions might reduce the gender disparity in question-asking probability. First, we tested which of the following five variables significantly affected the probability of a woman asking a question based on the observational data: (i) speaker's gender, (ii) gender proportion of the audience, (iii) host's gender, (iv) total audience size, and (v) room size. We fitted five binomial GLMMs for the probability that a woman asked a question with one of the five variables as an independent variable, while correcting for the gender of the audience and the non-independence of talks within a session. None of the five factors significantly improved the fit of the models, indicating that they did not significantly affect the probability that a woman asked a question (LRT $p > 0.05$ for all five GLMMs, Fig 4a; S2 Table 5).

Fig 4. The results of models that evaluated what conditions can encourage women to ask questions. a) Model estimates and 95% CI for the effect of five variables on the probability of a woman asking a question based on the behavioural data. A negative estimate indicates that the variable in question was negatively associated with the probability that a woman asked a question (i.e. it was positively associated with the probability that a man asked a question; a male bias). Yellow points indicate that including the variable in the model significantly improved the model fit compared to the null model; b) Model estimates and 95% CIs for the effect of female gender on the Likert-scale response of four statements asked in the post-congress survey. Yellow points indicate a statistically significant effect of female gender ($p < 0.05$).

Next, we addressed the same question using data collected in the post-congress survey, addressing whether women but also non-binary participants (despite low sample size, $n = 7$) were more or less comfortable asking questions in particular situations compared to men. We asked respondents to indicate on a 7-point Likert scale to what extent they agree with the following five statements: "I feel more comfortable asking a question if…" (i) "… the presenter is of my own gender", (ii) "… there is representation of my gender in the audience", (iii) "… the host is of my own gender", (iv) "… the audience size is smaller", and (v) "... if I know the speaker", partially reflecting the variables described above. We fitted four OLR models, with the Likert-scale response to each of the five questions as the dependent variable and self-reported gender identity and career stage as independent variables.

Including gender in the model improved the fit of almost all models (Fig 4b; S2 Table 6), where women and non-binary participants felt more comfortable asking questions compared to men when: the speaker was of their own gender (women: *beta* estimate = 1.23, $p < 0.001$; non-binary: *beta* estimate = 2.44, $p < 0.001$), their own gender was represented in the audience (women: *beta* estimate = 1.34, $p < 0.001$; non-binary: *beta* estimate = 1.90, $p < 0.01$), the host was of their own gender (women: *beta* estimate = 0.93, $p < 0.001$; non-binary: *beta* estimate = 1.58, $p = 0.02$). Only women felt more comfortable than men asking questions when the audience size was smaller (women: *beta* estimate = 0.79, $p < 0.001$; non-binary: *beta* estimate = -0.07, $p = 0.91$). Compared to men, neither women nor non-binary people felt more or less comfortable asking questions when they knew the speaker (women: *beta* estimate = 0.92, $p = 0.12$; non-binary: *beta* estimate = -0.19, $p = 0.78$).

Can session hosts mitigate the gender disparity in question-asking?

Previous research has shown that women can be encouraged to ask questions if a woman asks the first question in a Q&A session (50). We used observational data to test for this pattern in our data by quantifying the effect of the gender of the first questioner on gender disparities in question-asking in the rest of that session. More specifically, we fitted three binomial GLMMs to test for an effect of the gender of the person who started the Q&A on the probability that: (i) a question was asked by a woman, corrected for the proportion of women in the audience;

(ii) a woman raised their hand, corrected for the proportion of women in the audience; and (iii) a woman was chosen by the session host to ask their question, corrected for the proportion of people who raised their hand who were women. The models had a near identical structure to the three models presented in Methods Section ii and iii, but included an additional fixed effect of the gender of the first questioner, and we removed the intercept for easier interpretation of the model output.

The gender of the first questioner significantly affected the probability of women asking a question (LRT $p = 0.01$, S2 Table 7). Indeed, women were less likely than men to ask a question after a woman started the Q&A (*beta* estimate = -1.04, $p < 0.001$; Figure 5; S2 Table 7), but not after a man started the Q&A (*beta* estimate = -0.33, $p = 0.12$; Figure 5; S2 Table 7). Similarly, the gender of the first questioner significantly affected the probability of women raising their hands (LRT $p = 0.03$), as women were less likely to raise their hand than men after a woman started the Q&A (*beta* estimate = -0.90, $p < 0.001$; Figure 5; S2 Table 7), but not when a man started the Q&A (*beta* estimate = -0.31, $p = 0.16$; Figure 5; S2 Table 7). The gender of the first questioner did not significantly affect the probability of a woman being chosen to ask a question (LRT $p = 0.74$) as women were not significantly more or less likely to get chosen than men, regardless of whether a woman (*beta* estimate = -0.13, $p = 0.72$; S2 Table 7) or a man (*beta* estimate = -0.33, $p = 0.48$; Figure 5; S2 Table 7) started the Q&A. Similar results were obtained for all three models when testing for the effect of the gender of the first questioner on the probability of a woman asking the second question only (S2 Table 8).

Fig 5. Model results showing the effect of the gender of the first questioner on question-asking probability. Points indicate the probability that a woman asked a question, raised their hand, and was chosen to ask a question (left to right) for the unmanipulated and manipulated sessions. Yellow points indicate statistical significance ($p < 0.05$).

We sought to find causal insights into the effect of the gender of the first questioner by conducting an experiment in which we manipulated host behaviour. In the experiment, session hosts were instructed to either give the first question in the Q&A session to a woman or to a man. This manipulation allowed us to directly evaluate whether the gender of the first questioner affected the probability of women asking questions subsequently, regardless of the dynamics between the audience's behaviour and session host's choice. The same models as described above were fitted using data collected from the successfully manipulated talks.

The gender of the first questioner did not significantly affect the probability of a woman asking a question, raising their hand or being chosen to ask a question in the sessions where the host choice was manipulated (LRT all $p > 0.13$, S2 Table 7). Indeed, women were always less likely to ask a question than men, although this difference was only significant after a woman started the Q&A (*beta* estimate = -0.66, $p = 0.001$; Figure 5; S2 Table 7) but not after a man started the Q&A (*beta* estimate = -0.25, $p = 0.18$; Figure 5; S2 Table 7). Women always raised their hands significantly less often than men, regardless of whether a woman (*beta* estimate = -0.92, $p < 0.001$; Figure 5; S2 Table 7) or a man started the Q&A (*beta* estimate = -0.62, $p < 0.001$; Figure 5; S2 Table 7). Finally, women were not chosen to ask their question more or less often than men were, regardless of whether a woman (*beta* estimate = 0.61, $p = 0.17$; Figure 5; S2 Table 7) or a man started the Q&A (*beta* estimate = 0.68, $p = 0.10$; Figure 5; S2 Table 7). Interestingly, if we only selected the second question in each session, we found that women were significantly less likely to raise their hand than men after a woman started the Q&A (*beta* estimate = -0.93, $p = 0.003$; S2 Table 8) but not after a man started the Q&A (*beta* estimate = -0.32, $p = 0.28$; S2 Table 8).

How did people with different social identities experience the congress?

In the post-congress survey, we asked respondents to indicate their agreement with the following three statements on a 7-point Likert scale:

1. "I felt heard during the conversations I had, both during Q&A sessions and social activities" ("feeling heard" in short)
2. "I felt comfortable being myself" ("comfortable being myself" in short)
3. "Attending the Behaviour 2023 congress helped me feel like I belong in my research field" ("sense of belonging" in short)

We tested which of the following social identity variables were associated with the response to each of the three statements: gender, LGBTQ+, nationality (continent), affiliation (continent), and expatriate status ("expat" in short, defined as a person whose country of affiliation was different from the country of their nationality). Expatriate status was included because research has shown that expatriation for work helps the development of cultural intelligence (58), which is "the capability for success in new cultural settings" (59), which we would expect to play an important role at international scientific events. Additionally, we tested for the effects of the level of comfort a person had speaking English ("English comfort") which reflects a combination of factors including social environments, culture, and socio-economic status that affect one's English language proficiency, as well as fear and anxiety to use the language (60,61). We further tested for a person's self-reported level of expertise ("expertise rating"), which is highly correlated with age (*beta* estimate for ages 35-50 = 2.02, $p < 0.001$; *beta* estimate for ages > 50 = 3.43, $p < 0.001$) and career stage (*beta* estimate for mid-career stage = 2.21, $p < 0.001$; *beta* estimate for late-career stage = 4.16, $p < 0.001$) but also captures variation in confidence.

First, we fitted one univariate OLR model per statement and per social identity. If including the social identity in the univariate model significantly improved model fit, assessed with an LRT, we included the variable in the final model for that statement. We found that people with higher agreement to the "feeling heard" statement also felt more comfortable speaking English (*beta* estimate = 0.28, $p = 0.006$; Fig 6a; S2 Table 9) and rated themselves as having a higher level of expertise in their field (*beta* estimate = 0.24, $p < 0.001$; Fig 6a; S2 Table 9). Similarly, people with higher agreement to the "comfortable being myself" statement also felt more comfortable speaking English (*beta* estimate = 0.28, *p* = 0.01; Fig 6b; S2 Table 9) and rated themselves as having a higher level of expertise in their field (*beta* estimate = 0.22, *p* < 0.001 =; Fig 6b; S2 Table 9). Moreover, women and non-binary people felt less comfortable being themselves (*beta* estimate wome*n* = -0.48, *p* wome*n* = 0.03; *beta* estimate non-binary = -2.26, *p* non-binary = 0.001; Fig 6b; S2 Table 9) compared to men. Lastly, people with higher agreement to the "sense of belonging" statement also felt more comfortable speaking English (*beta* estimate = 0.31, *p* = 0.002; Fig 6c; S2 Table 9) and rated themselves as having a higher level of expertise in their field (*beta* estimate = 0.36, *p* < 0.001; Fig 6c; S2 Table 9). People with a North American affiliation had higher agreement to "sense of belonging" compared to those with a European affiliation (*beta* estimate = 1.16, *p* = 0.03); however, thwe interpret any effects of affiliation with care due to variation in sample sizes, as only 19 North American affiliates filled in the post-congress survey as opposed to 334 European affiliates.

Fig 6. The results of models evaluating which social identities were significantly associated with variation in congress experiences. a) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on feeling heard at the congress; b) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on feeling comfortable being yourself; c) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on congress attendance increasing ones feeling of belonging in the research field. The reference continent for affiliation to which the other continents were compared to was Europe. The estimates and 95% CIs for African and South American affiliations on statement c) were excluded due to small sample sizes (S2 Table 9). Yellow points indicate a statistically significant effect of the social identity variable in the final models.

Respondents to the post-congress survey were also asked if they experienced discrimination and/or harassment (of any sort) at the congress and whether they reported it to the awareness team, or if they witnessed someone else experiencing this. A total of eleven respondents reported experiencing some form of discrimination or harassment, of which two cases were reported to the awareness team. Eight of the eleven cases were reported by women, two by men, six by LGBTQ+ and/or non-binary attendees. A total of three survey respondents witnessed somebody else experiencing some form of discrimination or harassment, of which one case was reported to the awareness team.

How do perceptions of the severity of EDI issues differ among people with different social identities?

To test for differences among social identities in their perceptions of EDI issues, we asked post-congress respondents to indicate their agreement with the following three statements on a 7-point Likert scale:

1. "I think the Congress attendees represented the diversity of researchers in our field" ("attendee diversity" in short)
2. "Our research field experiences equity, diversity and inclusion-related issues (e.g. racism, homophobia, harassment, bullying etc.)" ("EDI issues" in short)
3. "I think the questions asked after the talks were equally divided across genders" ("no QA gender disparity" in short).

We used the same analytical approach as described above for the congress experience models. However, instead of fitting "expertise rating" as an independent variable, we fitted age category, as we expected that older researchers would be more likely to have experienced different research environments as well as cultural diversity and consequently, they might potentially have experienced more EDI issues independent of their level of expertise.

Women agreed less with the "attendee diversity" statement compared to men (*beta* estimate = -0.53, *p* = 0.01; Fig 7a; S2 Table 10), and LGBTQ+ people agreed less to this statement compared to non-LGBTQ+ people (*beta* estimate = -0.60, *p* = 0.03; Fig 7a; S2 Table 10).

Similarly, women agreed more with the "EDI issues" statement compared to men (ordinal *beta* estimate = 0.48, p = 0.03; Fig 7b; S2 Table 10), and LGBTQ+ identities agreed more to this statement compared to non-LGBTQ+ identities (*beta* estimate = 0.73, p = 0.009; Fig 7b; S2 Table 10). Moreover, expats agreed more with the statement on EDI issues compared to non-expats (*beta* estimate = 0.55, p = 0.006; Fig 7b; S2 Table 10). Furthermore, compared to people of European nationalities, people with North American nationalities (*beta* estimate = 0.77, p = 0.03; Fig 7b; S2 Table 10) agreed more with the EDI issue statement. Lastly, people of South American nationalities agreed more to the "no QA gender disparity" statement (ordinal *beta* estimate = 2.64, p = 0.04; S2 Table 10) compared to people with European nationalities, although those with South American affiliations agreed less compared to those with European affiliations (ordinal *beta* estimate = -5.39, p = 0.006; S2 Table 10), a contradicting result which could have arisen due to low sample size. People who are more comfortable speaking English agreed less with the statement about no QA gender disparity (ordinal *beta* estimate = -0.23, p = 0.03, Fig 7c; S2 Table 10). Although including gender, LGBTQ+ identity and nationality significantly improved model fit in the univariate regression models for no QA gender disparity, they did not explain significant variation in the final model that included all significant covariates (S2 Table 10).

Fig 7. The results of models evaluating which social identities were significantly associated with variation in EDI issue perception. a) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on congress attendees showing good representation of the diversity of the field; b) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on our field experiencing EDI-related issues. The reference continent for nationality to which the other continents were compared to was Europe; c) Model estimates and 95% CIs of the final model that tested for the effect of social identity variables on the Likert-scale response to the statement on there being no gender disparity in question-asking after talks. The reference continent for affiliation to which the other continents were compared to was Europe. The estimates and 95% CIs for African and

South American nationalities and affiliations on statements b) and c) were excluded for easier visual presentation, because the confidence intervals were large which made visual interpretation of the other confidence intervals difficult (S2 Table 10). Yellow points indicate a statistically significant effect of the social identity variable in the final models.

What can be done to promote inclusivity at scientific conferences?

The organising committee took a number of measures to make the International Ethological Congress 2023, "Behaviour 2023" more inclusive. We asked participants to respond to an open-ended question in the post-congress survey to obtain qualitative feedback from the participants on the conference, for example on the various inclusivity initiatives taken, their overall experience, as well as suggestions for improvement. Of the 391 total respondents, 48% (*n* = 191) provided a response to this question, of which 185 could be assigned to a particular topic (i.e. a "code", for details, see the Methods).

Most of the open-ended responses in the post-congress survey consisted of a combination of three sentiments (positive, suggestions, negative; S2 Table 11), however 51 responses contained only positive feedback, 22 contained only negative feedback, and 4 contained only suggestions. We coded 691 elements across 24 codes. Among these were 112 general compliments on the conference (e.g. "Great conference, thank you.") that will not be included in the further descriptions and analyses. Of the remaining 579 elements, 50% (*n* = 288) were positive, 34% (*n* = 197) were negative, and 16% (*n* = 94) were suggestions (Fig 8). While the participants offered feedback on a number of different topics, multiple responses included feedback about one or more specific EDI-related measures taken during the congress, which we elaborate on below. Although such feedback was relatively infrequent, we argue that this is as expected as these measures are often only perceived by the ones who need them the most.

We report these numbers as well as direct quotes from respondents to illustrate the positive impact that these measures can have.

Fig 8. Frequency of ideas expressed in each category for the three sentiments (positive, negative, suggestion).

1) *Plenary speaker diversity.* Three participants mentioned their appreciation for gender and/or ethnic diversity in plenary speakers, with one person indicating why this was appreciated, e.g. "It makes a huge difference to see gender and ethnic diversity represented in these head-line names, so well done on selecting this set of speakers. It sets a positive tone for the whole meeting."
2) *Pronouns on name tags.* Three people thanked us for allowing the option to print pronouns on their nametags (of which not all were non-binary), where one person commented that they appreciated the option as they "care about making sure everyone can feel more included just by default".
3) *Code of Conduct and awareness team.* The official Behaviour 2023 website contained a webpage on "Inclusivity and Accessibility" which included the Code of Conduct and additional information on who to contact about special needs. The responses from the post-congress survey indicated that 43% of respondents read this webpage. Out of those that read the page, 25.6% of respondents indicated that it played a role in their decision to attend the congress. A total of 19 people mentioned in the open text that they appreciated our general push for inclusivity at the congress, with four people specifically mentioning the Code of Conduct and/or awareness team and some highlighting how the presence of the awareness team helped them feel safe, e.g. "I was very grateful that the awareness team existed, which really helped me feel safe during this conference".
4) *Childcare.* A total of eleven people who filled out the survey used the free childcare service offered during the congress, seven of which stated that they would not have been able to attend the congress without this service. Seven respondents also indicated

that they would be able to attend more conferences if (free) childcare was available as a standard. The responses to the open-ended question in the post-congress survey included five positive mentions of the free childcare provided, where one person highlighted the difference this makes in the conference experience of parents, e.g. "After becoming a parent this was the first conference I could really enjoy fully and focus on the lectures and talking with colleagues".

5) *Accessibility/disability.* A total of 18 people indicated in the post-congress survey that they have some form of a disability, although eleven did not inform us about this prior to the conference. Out of those that did, three indicated that we were able to accommodate their disability, five indicated that the accommodation could have been better, and one person said that we were not able to accommodate their disability. The qualitative feedback included comments and suggestions for event organisers in general to make scientific conferences more accessible and inclusive, especially for researchers with a disability. The common themes of these comments included: (i) the difficulty of moving around the conference venue for people with mobility issues (in our case, mostly related to distances and stairs in the lecture rooms), (ii) the distraction caused by using (animal) sounds to indicate time limits to speakers, (iii) the appreciation of a quiet room for everyone who needs a space to "recharge and reflect", (iv) the overwhelming experience during poster sessions that was non-inclusive to people sensitive to sound and/or prone to anxiety in large crowds, and (v) the importance of ensuring the availability of presentation programs' notes that can be seen by only the presenter during the talk.

6) *EDI-related activities.* A total of 66 people that responded to the post-congress survey attended the EDI symposium and 21 attended one of the EDI workshops (one on unconscious bias, one on inclusive teaching). Reasons for attending the symposium and/or workshop included the participants being motivated to (i) learn about EDI issues (61% and 66% respectively), (ii) improve the way they do research (61% and 67%), and (iii) talk about their own (10% and 24% respectively) or others' (18% and 62% respectively) EDI-related issues. Out of the reported symposium/workshop

attendees, many respondents stated that attending will influence their practice, with some being sure about the changes they would make (41% and 29% respectively), and others seeing the potential but being less sure (20% and 62% respectively). Suggestions for EDI-related workshops in general, that were not specific to the content and facilitators of the workshops we hosted in particular, mostly focused on the need to shift from theoretical work to practical implications.

Discussion

Barriers to inclusion and equity persist in science, including at academic conferences. Our aim was to identify and address equity, diversity and inclusivity issues present at the 37th International Ethological Congress that stretch beyond gender, using a number of different approaches. We identified barriers that unfold during Q&A sessions, as well as barriers that affect the congress experience of attendees not only when presenting or discussing science, but also when simply attending the activities that are part of the conference programme. A summary of all results can be found in Fig 9.

Fig 9. Summary of our results based on both the behavioural and survey data. The single asterisk (*) refers to non-binary researchers. The double asterisk (**) refers to a marginally significant result (not significant after applying a multiple-testing correction). The identification of the ten barriers is based on the results presented in the following Results sections: (i) section ii; (ii) section iii; (iii) section iv; (iv) section iv; (v) section iv; (vi) section v; (vii) section vi based on both the unmanipulated and manipulated data; (viii) section vii; (ix) section vii; (x) section vii.

We show that women tend to ask fewer questions than men despite the fact that they do not appear to be actively discriminated against. Although we find clear evidence that a question is less likely to be asked by a woman compared to a man based on the behavioural data, women only appear to be slightly less likely to have asked a question across the entire congress. This pattern may arise if men on average asked more questions per individual (e.g. 3 questions during the congress) compared to women (e.g. 1 question during the congress), which does not affect the probability that a woman asked a question in the survey but does affect the probability that a question was asked by a woman. Alternatively, the pattern may arise if there are certain men that ask a lot of questions across different sessions, or if women who did not ask any questions during the congress were also less likely to fill in our post-congress survey. We further found that women likely ask fewer questions due to a lack of self-confidence and because they feel intimidated by the audience (although only significant before applying a multiple-testing correction). Indeed, the gender gap in confidence (62,63), as well as the inaccuracy of women's self-perception (64) have previously been proposed to play a role in various gender disparities, including the underrepresentation of women in senior leadership positions (62). The reasons why women tend to have lower self-confidence and belief in their own abilities are however complex and difficult to generalise, as they could be rooted in both internal and external processes that take place within and outside of the academic environment (e.g. family environment (65), gender stereotypes (66), and a lack of role models (67)).

Women's representation could potentially improve women's confidence as it has been shown to boost female engagement (50–52,68), yet our findings only partially support this. Whereas the data collected in the post-congress survey suggests that women are more comfortable asking questions when their gender is represented (in the audience, by the presenter or session host), the data collected during the Q&A show that women were less likely to raise their hand and ask questions than men, regardless of the situation. Moreover, women appeared to be less inclined to raise their hand to ask questions, specifically after a woman started the Q&A. We speculate this could be caused by (i) a lower feeling of competitiveness of women (69) towards the opposite gender compared to men leading to a lower motivation to ask the second question, and/or (ii) women stop feeling motivated to represent their gender among questioners after another woman asked a question instead of themselves. While our results suggest that session hosts cannot mitigate the gender disparity in question-asking by actively selecting women to start the Q&A, we found different results for the manipulated talks compared to the unmanipulated ones when a man started the Q&A. When host behaviour was not manipulated, we found no gender disparity when a man started the Q&A, as women were equally as likely to raise their hands. Yet, in our experiment, we did find a gender disparity in raising hands when a man started the Q&A. These results indicate that either the deliberate choice of a man over a woman (as happened in our manipulated talks) or the (conscious or unconscious) change in behaviour of the session host due to higher awareness of their choices might have discouraged women from asking questions during the rest of the session. Testing what exact perceived behaviours from session hosts affect the probability that women raise their hands to ask questions would require further research, yet the effects of female representation among questioners are evidently complex and appear to not always be positive.

While gaining a deeper understanding of the causes and consequences of gender disparities in question-asking probability is important, we argue that it is more critical to ensure that women do not miss out on academic opportunities as a consequence of this disparity. The same accounts for non-binary participants who also appeared to be uncomfortable asking questions. Questioners might gain academic benefits by (i) expressing their interest and participating in the scientific discussion, (ii) increasing their likability by showing responsiveness (70), (iii) growing their visibility, which can help them connect with people working on similar topics, and (iv) facilitating collaborations and/or exchanging ideas that can improve the quality of their research. To our knowledge, there is no empirical evidence of the academic benefits of question-asking during Q&A sessions.

Assuming that there are benefits of question-asking at conferences, we expect that similar outcomes could be achieved in alternative ways that might be more likely to be adopted by people who are less likely to ask questions, including but not limited to women. For example, conference organisers could plan topic-focused discussion rounds, provide an online platform where attendees can connect based on mutual interests, and/or schedule more time after presentations for the audience members to engage in one-on-one discussions with the speaker. Such activities would benefit not only women, but also introverted people and non-native English speakers, who are less inclined to ask questions in Q&A sessions, as revealed by our quantitative and qualitative data. We thus urge for a shift in focus towards addressing those potentially missed academic opportunities for people who are less inclined to ask questions during Q&A's, which disproportionately include women, and ensuring equity by providing alternative pathways to reclaim those opportunities.

Moreover, our results have important implications with regard to differences in congress experiences. People who do not feel like an expert in the field appear to have a less positive congress experience. Expertise is undoubtedly, but not exclusively, related to age, and therefore it does not come as a surprise that people who feel like they have less expertise do not feel heard as much, are not as comfortable being themselves, and do not feel like the congress contributed to their sense of belonging as much compared to those who rated themselves higher in their expertise. We also found that older attendees are less likely to appraise oral presenters compared to younger attendees, yet are also more likely to ask a critical question (results only presented in S1 Results 1). Some qualitative responses mentioned the huge negative impact a critical comment from a senior researcher can have on the experience of early-career researchers. Although we have no data indicating that these findings on congress experience and presenter feedback are directly connected, we suspect that the opposite is also true: senior researchers can have a positive influence on the experience of early career researchers through their feedback on oral presentations as well as during scientific discussions. Therefore, we encourage senior researchers to give positive appraisal to presenters when they see fit, which we expect to boost the congress experience by "warming up" the "chilly conference climate" that early-career researchers might experience. In addition, we encourage future research into activities that can help empower early-career researchers and improve their congress experience, such as (i) organising Q&A sessions between (PhD) students and senior scientists, (ii) hosting events tailored towards early-career researchers specifically, or (iii) setting up a buddy network that connects (PhD) students that work on similar topics.

Similarly, people who feel less comfortable speaking English also had a less positive congress experience. The dominance of the English language at international academic events causes a systemic bias. Indeed, recent work has started to uncover the many disadvantages faced by non-native English speakers, such as spending more time on scientific activities compared to native speakers (71). We encourage critical thinking about initiatives that can improve the inclusivity of people who are less comfortable speaking English, such as (i) hosting social events that accommodate foreign languages, for example language-specific discussion rounds (also previously suggested by (72)), (ii) utilising AI-assisted translation services during talks and/or Q&A sessions, similar to AI-assisted academic writing (73,74) and (iii) emphasising the importance of teaching English proficiency during early and higher education. Such activities have the potential to make people feel more like they are heard, especially in early stages of their academic career, which can increase a person's sense of professional worth and belonging.

Our results further show that different social identities have dissimilar perceptions of equity, diversity and inclusivity issues. Evidently, historically underrepresented minorities, including women and LGBTQ+ identities, seem to better recognize EDI issues. Previous research has also shown that men are less likely to notice gender disparities in question-asking probability (49). We expect that minorities are more likely to notice EDI issues either because these groups experience more EDI issues themselves, or because they are more aware of issues that other people face, or a combination of the above. Interestingly, expat scientists agree more with the statement that our field (behavioural, ecological and evolutionary sciences) experiences EDI issues, which could be attributed to the link between expatriation and cultural intelligence (58). This finding emphasises the importance of active listening (75), especially to those with a cultural background or social identity different from one's own, which can increase awareness of issues both inside and outside of academia. The importance and value of listening is directly reflected by the comprehensive constructive feedback that we received in the post-congress survey, where many congress attendees took the opportunity to provide suggestions for making conferences more inclusive and raised both minor and major points for improvement that would not have been brought to our attention if we had not specifically asked for this feedback. We therefore encourage every research group to provide the opportunity for members to express their concerns, and to foster an environment where dialogue about EDI issues is encouraged (76).

The responses to the open-ended questions in the post-congress survey revealed that participants had an overall positive experience during the conference. Nonetheless, there were also critiques and suggestions that were not only specific to this event but could be relevant to scientific conferences in general. Although we are well aware of the many logistic, financial and time-related limitations that event organisers face, we would like to emphasise a number of aspects that have been suggested by respondents to foster more inclusive conferences. These think these aspects can be addressed to improve the experience of the minority without sacrificing the experience of the majority, by making small tweaks or implementing small additions to accommodate to everyone. First, giving an oral presentation can itself be stressful regardless of a person's social identity and abilities. Attention to a few simple details can help mitigate some of this stress. For example, ensuring that the presentation program's notes are available to the presenter can especially benefit neurodivergent and non-native language speakers. Stress can additionally be lowered by limiting the scope for distractions, such as auditory cues indicating the presentation time remaining. Although these cues can be helpful for the majority of people, if they are played too loud, they can be distracting to neurodivergent speakers with heightened auditory sensitivity. So, we encourage event organisers to ensure such sounds are played at an appropriate volume for everyone.

Secondly, although international conferences in theory provide an excellent opportunity to host workshops on EDI-related themes, we believe that such workshops are likely to be more effective if they are organised as satellite events. This way, the workshops can be longer in duration allowing the discussion of both theoretical and practical aspects, attendees do not have to choose between attending workshops or scientific talks, and having these satellite events during the year can help increase interactions and build community. Lastly, poster sessions held in loud, crowded venues can be overwhelming, especially for people sensitive to large crowds and/or auditory overstimulation. Alternatives to poster sessions have previously been proposed (e.g. virtual posters: (77,78), and we encourage future event organisers to critically think about the setup, size and location of the poster sessions and/or alternative modes for more inclusive and equitable ways of presenting science. This does not necessarily have to go at the expense of traditional posters sessions which are effective for the majority of attendees, but we encourage to have alternative options available. We summarise all our general recommendations for inclusive scientific events aimed at future organisers, based on our data and personal experiences, in Fig 10.

Fig 10. Summary of our recommendations for more inclusive scientific events, based on the data we collected as well as our personal experience.

Several inferences about certain groups of social identities made in our study are based on relatively low sample sizes. We acknowledge the statistical limitations of these inferences; nevertheless, we argue that these inferences address barriers experienced by social minorities that have rarely been researched. For example, we find a clear signal that non-binary respondents felt uncomfortable being themselves in the post-congress survey even though there were only seven non-binary respondents. Including this small group of people in our analysis helps to illuminate the social barriers faced by certain minorities, which by definition are represented in small numbers. We further argue that, as opposed to quantitative analyses, qualitative data can be more insightful in identifying and addressing barriers experienced by minorities, as shown by the comprehensive feedback given by the handful of respondents on mobility- and neurodiversity-related issues.

Our case study investigated equity, diversity and inclusivity issues at an academic conference. We expect that many of the inferences that we draw from our data can be generalised to settings outside of conferences. For example, our conclusions on question-asking behaviour are likely to be applicable to Q&A sessions not only at conferences, but also within the setting of seminars given at academic institutes. We also expect that our findings on differences in congress experiences between people of different genders, with different levels of comfort in speaking English, and with different perceived levels of expertise will be applicable to many different academic social settings, such as lab meetings and collaborative projects. Our study therefore does not only have implications for the way we host and attend scientific events, including conferences, but also for conducting science overall. Removing barriers that are present across different academic settings requires acknowledgement of those barriers, especially by those in leadership positions, identifying the causes and mechanisms by which these barriers are established and maintained, understanding how they affect researchers, and developing effective strategies to tackle them through open, accepting and respectful dialogue.

Methods

Conference description

Bielefeld University, located in Germany, hosted a seven-day International Ethological Congress, "Behaviour 2023" in August 2023 which was attended by more than 850 people. The official language of the congress was English. Six of the days consisted of scientific talks, including eleven plenary talks given by invited international speakers, which lasted 60 minutes each including a 10–15-minute question-and-answer (Q&A) session. After each plenary talk (except on the last day), oral sessions took place, which consisted of 1-7 seven talks. In total, there were 56 general oral sessions, as well as 42 oral sessions that were part of symposia on a specific theme. General oral sessions and symposia were moderated by internal and/or external session hosts. Each talk slot lasted fifteen minutes, with the speakers being instructed to limit their speaking time to a duration of twelve minutes, leaving three minutes for the Q&A. Each day (except the last day) consisted of parallel morning and afternoon sessions, and each session included a coffee break.

Various initiatives were taken to promote inclusivity at Behaviour 2023. First, all of the congress attendees were obliged to agree to a Code of Conduct when registering for the congress. The Code of Conduct outlined expected and unacceptable behaviours and clearly stated the consequences of non-compliance. During the congress, attendees were able to inform an awareness team about any concerns and cases of discrimination or harassment. The awareness team was a group of organising committee members who had received harassment training from an external organisation (Frauen Notruf Bielefeld e.V.) who could be contacted by email, phone, via social media, or directly in person during the congress. Recognition of awareness team members was facilitated by them wearing a recognizable badge.

Moreover, the programme of plenary talks was curated in a way that ensured a balanced representation of gender and ethnic diversity among plenary speakers, ensuring that at least half of the plenary speakers were female and that each continent was represented at least once. Prior to the congress, we offered information and help to people with auditory, visual, mobility and/or dietary needs through the website and during congress registration. We offered a number of full travel grants to researchers based in the Global South. During the congress, we offered free childcare provided by an external company, which was funded by the Bielefeld Equal Opportunities Committee. We additionally offered parent-children offices, breastfeeding rooms and free congress attendance to the partners of attendees that were only there to provide childcare. We further offered quiet rooms that were open between at least the first and last talk of each day. Moreover, we convened a symposium on "Equality, diversity and equity in behaviour, ecology and evolution" with talks given by three invited speakers, and organised three half-day workshops given by external moderators in an attempt to foster engagement and critical dialogue on EDI issues among congress attendees. We organised workshops on two different topics: one on unconscious bias and one on inclusive teaching in higher education. The former workshop was given two times on the same day, independently from each other with different groups of workshop attendees. Lastly, we offered the option to congress attendees to print their pronouns on their nametags, in an attempt to avoid misgendering among congress attendees and to build an inclusive culture for non-binary people.

Pre-congress survey

Congress attendees were asked to fill in a voluntary online survey on their social identity when registering for the congress. The survey included questions on: (i) their pronouns, (ii) if they identified as lesbian, gay, bisexual, transgender, queer, intersex or any other non-heterosexual, non-heteroromantic, or non-cisgender identity (LGBTQ+), (iii) their nationality, and (iv) if they have any dis-/para-bilities.

Question-asking study

We collected data on question-asking behaviour during Q&A sessions at the congress. Although it is important to understand disparities in question-asking behaviour among multiple social identities as well as the intersections of those identities, we focused only on gender disparities due to logistical and practical reasons, as this was the most conspicuous identity that could be perceived in a real-life setting. We observed the question-asking behaviour of the participants of 67 oral sessions at the congress.

A total of 25 observers (organising committee members, students and/or colleagues) collected data on question-asking behaviour across the five days of talks. Observers were randomly allocated to collect data in oral sessions within the timeframe of their availability. When collecting data, observers conducting the study were seated in the back corner(s) of the lecture hall to obtain a better overview of the audience and to reduce our visibility when counting the number of people in the audience (see below). In 32 of the 67 sampled sessions (48%), data were gathered by multiple observers to evaluate inter-observer reliability (hereafter referred to as "double-sampled sessions"). Sessions were held in lecture halls of three different sizes: small (63-77 seats), medium (102-132 seats) and large (308-404 seats). Because it is difficult to observe people in large lecture rooms while remaining stationary, sessions held in large rooms were always sampled by two observers, where some variables were collected by one observer but not the other and vice versa (see below). Therefore, data collection in a double-sampled session in a large room was done by four people.

We collected data on the perceived gender (female, male, other) and perceived age class (< 35, 35-50, or > 50 years) of session hosts, speakers and questioners (see below). We acknowledge that inferring someone's gender and/or age based on their appearance is subjective and prone to error. We therefore elaborate on our methods used to infer gender and age at the end of this section. Data were collected at three different levels: per session, per talk and per question as described below.

Data collected per oral session

For each oral session, we noted down the gender, career stage and age class of the session host, as well as three meta-data variables including the day of the congress (day 1-5), lecture hall (1-9), and whether the session was part of a general oral session or symposium. Although general oral sessions were hosted by just one person, a symposium could be hosted by up to three session hosts. If a symposium was hosted by more than one person, we focused on the host that led the Q&A session. If multiple hosts led the Q&A session, or if the hosts swapped roles, this was noted down and accounted for in the relevant analyses as described below.

Data collected per talk

At the start of each talk, the total audience size was counted, as well as the total number of men in the audience. Because more women than men registered for the congress, we counted only the men in order to accelerate the counting process. The session hosts, speaker, observers and technical assistants were excluded from these counts. We noted down if there was any uncertainty in the number of people counted due to, for example, the view of the observer being partially blocked, people sitting in areas out of sight to the observer, or limited light in the room. Similar to above, the gender, career stage and age class of the speaker were recorded. In addition, the duration of the Q&A session was recorded in minutes and we also noted occasions when the speaker talked for longer than their allocated time slot.

Data collected per question

For each question asked after each talk, we counted the total number of people and the total number of men who raised their hands to ask a question. Because it was more difficult to reliably count all of the people who raised their hands in large rooms, two observers were always present in the large rooms (and four people in double-sampled large rooms). One of the two observers counted the total number of people raising their hands and the other observer counted only the number of men who raised their hands. For each person who asked a question, the following data were collected: the gender of the person asking the question, age class of the questioner, if they showed appreciation towards the speaker (e.g. "Thank you for the interesting talk") and whether the question contained criticism and/or a counterargument. Lastly, the observers noted down if one of the following situations occurred: a person asked a question without raising their hand ("jumper"), the session host asked the question, the speaker chose who asked the question instead of the session host, an observer asked a question, a person asked multiple questions in one turn, or a person asked multiple questions in one Q&A but not consecutively.

Data collected during plenary talks

Plenary talks were held in a different building with a large lecture hall containing 638 seats and were not run in parallel with any of the other congress activities. Due to the difficulty of counting the number of people sitting down and raising their hands in this large lecture room, we only collected data on the gender of the people asking questions. At least two observers collected data during plenary talks, and the gender and number of questions for plenary talks were manually cross-checked based on the notes taken by each observer.

Inferring gender

The gender of session hosts, speakers and questioners was inferred from the pronouns printed on their nametags as well as mentions of their pronouns (e.g. shown on a speaker's title slide). If the name tag could not be read from a distance, but if we did know the person's name (which was the case for session hosts and speakers) and if they had consented to print their pronouns on their name tags during congress registration, we confirmed a person's pronouns based on the registration sheets. For questioners and hosts and speakers who did not opt to print their pronouns on their name tags, we inferred gender from visual appearance (e.g. hair length, clothing, voice pitch, body size, name if stated when asking the question). We acknowledge, however, that (i) inferring a person's gender based on their appearance is flawed (and we address our accuracy of inferring gender in S1 Methods 1) and that (ii) gender identity and pronouns can be independent of each other, as not every woman uses she/her pronouns, not every man uses he/him pronouns, and not every non-binary person uses they/them pronouns, which is a topic that we also address in S1 Methods 1.

Inferring career stage and age

The information used to characterise the career stage of a session host and speaker was their title and/or academic position, which speakers regularly mentioned at the start or end of a talk, and session hosts when introducing themselves. If there was no mention of the session host or speaker's career stage, we attempted to find this information after the congress based on publicly available data (e.g. using Twitter/X, ResearchGate, and university websites).

To estimate the career stage of a person without any such confirmation, we estimated the age of session hosts, speakers and questioners. We classified people into three age categories: under 35, between 35 and 50, and above 50 years of age based on their appearance (facial features, hair colour, voice, clothing). We instructed observers to be careful not to bias their age estimation by a person's career stage if this was known, as age and career stage are not always directly linked to each other.

Experimental manipulation of session host choice

We investigated if the session host's choice of questioner can help overcome gender disparity in question-asking probability. For a subset of sessions (40 sessions, 62.5%), we manipulated the behaviour of the session host. We used stratified random assignment of session hosts to either an unmanipulated or manipulated session. If the session host was part of the organising committee, they were automatically assigned to a manipulated session because they were aware of the study and its purposes, and consequently they might be biased if assigned to an unmanipulated session. The hosts of unmanipulated sessions were unaware of our study and were not contacted prior to the congress about the study. Two weeks prior to the congress, the hosts of manipulated sessions were asked by email if they wanted to participate in our study, without mentioning the exact goal or describing the tasks in detail. If the session host agreed, they were given instructions specific to their session. If the session host declined to participate ($n = 2$), we did not sample that session and swapped data collection with a session whose host agreed to participate.

In manipulated sessions, the host was instructed for each talk within that session to assign the first question of the Q&A to either a woman or a man, resulting in two possible conditions. The conditions were randomly assigned across all of the talks in all of the manipulated sessions, ensuring an overall equal distribution of the two conditions over all sampled talks but not necessarily an equal distribution of the two conditions within a manipulated session. If the raising of hands did not meet the experimental condition (e.g. the condition was the first question given to a woman, but no women raised their hands), the hosts were instructed to select a person as they normally would.

Hosts successfully assigned the first question to the assigned gender in 102 talks (48 to a woman, 54 to a man). The manipulation was unsuccessful in 106 talks either because nobody of the assigned gender raised their hand (*n* = 63) or because of other unknown reasons (*n* = 43).

Data curation and validation

A number of steps were taken to curate the collected data on question-asking into the final dataset used for analyses, which are described in detail in the S1 Methods 2. Briefly, we checked whether data collected in double-sampled sessions had a good inter-observer reliability. Indeed, agreement between observers was "good" to "almost perfect" for all of the variables except for age which had "moderate" to "substantial" agreement (S1 Methods 3). Due to the low reliability of our age estimates, we did not investigate the effect of age on question-asking probability.

Because there were slight differences in how certain situations were noted down by observers of double-sampled sessions, we manually checked and corrected the data when the observers appeared to disagree over the number of questions that were asked (9 talks). After manual correction, data from different observers of the same session were combined using a conditional workflow dependent on the variable as described in the S1 Methods 4. Briefly, (i) if observers disagreed on the inferred gender of a person, we discarded the data; (ii) we took the mean of audience number estimations; (iii) we used the maximum of the number of hands raised, and (iv) we assumed that disagreement on the variables that recorded whether something was or was not done or said (e.g. a questioner appreciating the speaker) was due to one observer having missed it or forgetting to not it down rather than the other observing taking note of something that did not happen or was not said.

Statistical analyses of behavioural data on question-asking

To test whether there was a gender disparity in question-asking probability, we built a series of generalised linear mixed effect models (GLMMs) using the R package lme4 v1.1 (79). Unless indicated otherwise, the data used to construct the models below excluded sessions where we manipulated session host behaviour, as well as questions that were follow-up questions by the same person, questions asked by the session host, or questions asked by people who did not raise their hands (jumpers). For clarity, a summary of the models that use the observational data can be found in S2 Table 12, which includes a clarification of the subset of the data used, the research question it addresses, and the formula written in lme4 syntax (79).

The first model (QA.1) tested whether women ask fewer questions than men do in regular oral sessions. We fitted a binomial GLMM to the perceived gender of the questioner (1 = female, 0 = male). Under the null hypothesis, we would expect that the proportion of questions asked by women is equal to the proportion of women in the audience. This would therefore mean that the audience consists of 60% women, the null hypothesis is that 60% of questions are asked by women. Therefore, we corrected for the gender proportion of the audience by specifying the *offset* argument in the GLMM as the logit of the proportion of women in the audience. We corrected for the non-independence of talks within a session by including the random effect of talk ID nested within session ID. If the resulting intercept was significantly negative, this would indicate that women asked fewer questions than men did. We repeated this analysis with a conservative subset of the data that excluded any questions where there was uncertainty in the data, for example because the observer could not count the audience reliably (QA.1c).

We also tested for gender disparity in question-asking probability in the plenary sessions only. A similar GLMM was fitted as described above (QA.1) using the observational data collected during plenary talks, where the dependent variable was the inferred gender of the questioner and a random effect was included for plenary ID (QA.1p). Because of the large audience and room size, it was not possible to accurately count the number of women and men in the audience. Therefore, instead of correcting for the proportion of the women in the audience, we corrected for the gender proportion by using the proportion of women who registered for the congress, assuming that the vast majority of registrants attended the plenary sessions.

Next, we used a similar model structure to model QA.1 to address what conditions can encourage women to ask questions. Specifically, we tested for the effects of the following five variables on the gender disparity in question-asking probability: (a) the gender of the speaker (male, female or non-binary), (b) the gender proportion of the audience (where 1 would theoretically indicate a 100% female audience), (c) the gender of the session host (male, female or non-binary), (d) the total size of the audience, and (e) the size of the room (small, medium or large), further referred to as models QA.1a – QA.1e respectively. We constructed five binomial GLMMs using the inferred gender of the questioner as the dependent variable and one of the five variables as an independent variable. We again corrected for the gender proportion of the audience using the *offset* function as described above and included the random effect of talk ID nested within session ID. For the model that tests for the gender of the session host (QA.1c), we excluded sessions where there were multiple session hosts who alternated leading the Q&A. We determined whether a variable was a significant predictor of the likelihood that a woman asked a question by conducting a likelihood-ratio test (LRT) using the *anova* function from the stats R package v4.3.2 (80), which compared the model in question with the null model that only included the intercept (QA.1).

How does a gender bias in question-asking arise?

Women might ask fewer questions than men do due to two different reasons: women raise their hands less often than men do, or women are chosen less often to ask their question by session hosts when they do raise their hands. We tested which reason was the most probable cause for the gender disparity in question-asking probability by fitting two GLMMs.

The first GLMM (QA.2) evaluated whether women raised their hands less often than men did by fitting the number of hands raised by women and men as the response variable using the *cbind* function. Similar to above, we corrected for the gender proportion of the audience by specifying the *offset* argument as the logit of the proportion of women in the audience. Again, we used a binomial error distribution and corrected for the non-independence of talks within a session by including the random effect of talk ID within session ID. Under the null hypothesis, we expected that the number of hands raised by women and men would be proportional to the number of female and male audience members respectively. If the resulting intercept was significantly negative, this would indicate that women raised their hands less often than men did.

The second GLMM (QA.3) evaluated whether women were chosen less often by session hosts than men were by fitting the gender of the questioner as the response variable, but instead of correcting for the gender proportion of the audience, we corrected for the proportion of women out of those people who raised their hands. Under the null hypothesis, we expected that the number of questions asked by women would be proportional to the number of women who raised their hand. We therefore specified the *offset* argument as the logit of the proportion of women out of the people who raised their hands. For this analysis, we only used a subset of the data where the session host could make a choice between allocating the question to a man or women, meaning that the subset only included situations where at least one woman and one man raised their hand. We again used a binomial error distribution and corrected for the non-independence of talks within a session by including the random effect of talk ID within session ID. If the resulting intercept was significantly negative, this would indicate that women were chosen less often to ask their question than men were.

Do women ask more questions if other women have asked questions previously in the Q&A?

Session hosts can potentially help to reduce the gender disparity in question-asking probability by selecting women to ask the first question, and/or by encouraging other women to raise their hands and ask questions. We tested whether the gender of the first questioner affected the probability of (i) a woman asking a question compared to proportion of women in the audience, (ii) a woman raising their hand and (iii) a woman being chosen to ask their question compared to the proportion of people raising their hand who are women by fitting three different binomial GLMMs to unmanipulated talks only. We used similar models to QA.1 (the response was the gender of the questioner, corrected for the gender proportion of the audience), QA.2 (the response was the gender of the people who raised their hands, corrected for the gender proportion of the audience), and QA.3 (the response was the gender of the questioner, corrected for the proportion of women out of the people who raised their hands), respectively. Additionally, we excluded the first question asked in each Q&A session from the dataset and used the gender of the first questioner as a fixed effect instead, as the gender of this first questioner was our variable of interest. We removed the intercept (by adding -1 to the formula) to allow for an easier interpretation of the output. For clarity, a summary of the models that address the effect of the gender of the first questioner can be found in S2 Table 13, which includes a clarification of the subset of the data used, the research question it addresses, and the formula written lme4 syntax (79).

The three models were fitted using two separate datasets, first using the data collected in unmanipulated sessions only (QA.4u-QA6u respectively) and second using data collected in manipulated sessions where the first question was successfully assigned according to the condition of the manipulation (i.e. a woman or man asked the first question as instructed, QA.4m-6m respectively). We repeated all six GLMMs with a subset of the data that only included the second question asked in each session (QA.4u.2 - QA.46u.2 and QA.4m.2 - QA.46m.2 respectively) rather than all questions asked after the first one. These models helped us determine whether the gender of the first questioner only affected the probability that only the next question was asked by a woman, rather than all questions in the remainder of the session. To test whether the gender of the first questioner had a significant effect on the response variable, we compared the fit of the model to a null model that only included the random factors using an LRT.

All of the models described above excluded follow-up questions by the same questioner, cases where the speaker assigned the question rather than the host, questions asked by the session host, questions asked by jumpers, and questions where the gender of the questioner or the proportion of women in the audience was unknown. The models using the number of hands raised (QA.2, QA.3, QA.5u, QA.5m, QA.6u, QA.6m) also excluded cases where the number of women and/or men raising their hands was unrecorded (e.g. because the observer did not see it) or when no hands were raised. The models where the probability of being chosen to ask a question was investigated (QA.3, QA.6m, QA.6u) excluded cases where only men or only women raised their hands, as here the host could not choose whether a woman or man got to ask their question. The model estimates (predicted log-odds) were obtained from Wald tests using the *summary* function in lme4 v1.1-35.5 and back-transformed to probabilities (inverse logit) using the *plogis* function in stats v4.4. We additionally obtained profiled confidence intervals using the *confint* function in stats v4.4. A probability was considered to be significantly different from the expected probability under the null hypothesis (no gender disparity, probability = 0.5) if the *p*-value of the Wald test was lower than 0.05 and if the corresponding 95% confidence intervals did not overlap zero.

Other gender disparities in oral sessions

We further investigated whether men or women have a higher probability to: (i) ask a question without being chosen to (i.e. being a "jumper"), (ii) speak for longer than their allocated time, (iii) give and/or receive a compliment after an oral presentation and (iv) ask and/or receive critical questions. We investigated which variables of interest (e.g. gender, career stage dependent on the dependent variable) were significantly associated with the probability that one of the four mentioned cases occurred by constructing a binomial GLMM for each of the dependent variables of interest (S2 Table 14). Statistical significance of the variable was inferred from an LRT which determined whether including the variable significantly improved the fit of the model compared to the null model that did not include the variable. Only statistically significant predictors (LRT p-value < 0.05) were retained in the final model. In all of these models, we included the random effect of talk ID nested in session ID. The results of these models are described in S1 Results 1.

Post-congress survey

Three days after the end of the congress, we advertised a post-congress survey on the congress website, Twitter/X and e-mailed this to people that registered for the congress or signed up for the newsletter. The survey was filled in by 391 people (approx. 45% of all attendees) and included sections with questions on (a) social identity (gender, pronouns, age, career stage, LGBTQ+, nationality, affiliation), (b) congress-related questions on attendance, (c) self-assessment of one's expertise and comfort speaking English, (d) conference experience, (e) question-asking, (f) attendance of and feedback on EDI-related activities such as the symposium and workshops, (g) perceived equality at the congress and in the field of behaviour, ecology and evolution in general, (h) childcare (was childcare used and how important was the offer for free childcare to the attendee), (i) dis/para-ability (do you have a dis/para-ability and was this adequately accommodated for) and (j) qualitative feedback. People that did not attend the congress were also able to fill in a shortened version of the survey that only asked for their social identity variables and reasons why they did not attend. As very few non-attendees filled out the survey (n = 3), we do not report these results. At the start of the survey, respondents were asked to consent to their data being used for research, and answering the questions was optional.

Prior to the statistical analyses, we simplified and processed a number of variables obtained from the personal details section of the post congress survey (section a). First, we condensed the career stages into three categories: early-career (BSc students, MSc students, post-graduates, and PhD students), mid-career (postdocs, lecturers, and researchers), late-career (associate professors, assistant professors, and full professors) and "other" (applied scientists, non-academics, retired scientists, technicians, etc.). Second, we added a variable expatriate status ("expat"), which indicated whether the country of affiliation is the same as the country of nationality (same = no expat, not the same = expat). We acknowledge that this variable is imperfect and only provides a contemporary snapshot of someone's expat status, yet it serves as an indicator of cultural exposure. Third, we categorised all countries (nationalities and affiliations) into the continents for simplification purposes and due to unequal and sometimes small sample sizes per country. People who indicated multiple countries of nationality ($n = 6$) were excluded from all analyses as the countries were often located in different continents. From here onwards, we collectively refer to gender, career stage, sexual and gender identity (LGBTQ+), nationality, affiliation and expat status as the "social identity variables". We also tested for the effect of expertise rating (Likert-scale response to "I am an expert in my field") and the effect of English comfort (Likert-scale response to "I feel comfortable speaking in English") and collectively refer to these variables as the "controlling" variables. Both of these responses were measured on a 7-point Likert scale which indicated to what extent people agreed, ranging from "Strongly disagree" (1) to "Strongly agree" (7), where 4 would indicate a neutral attitude. For clarity, a summary of the models that use the data collected in the post-congress survey can be found in S2 Table 15, which includes the research question it addresses, and the formula written lme4 syntax (79).

Gender effects on question-asking motivation and hesitation

In section e) of the post-congress survey, we collected data on question-asking behaviour. First, we asked whether participants asked one or multiple questions during the Q&A sessions at the congress (yes/no). We tested whether gender was predictive of a person having asked a question during the congress by fitting a binomial GLM to the response to this question as the dependent variable and using self-reported gender as the independent variable. We used an LRT to evaluate whether gender was a significant predictor of the probability that a person asked a question.

Second, we asked which factor(s) motivated attendees to ask a question:

1) "Interest in the topic"
2) "Making my voice heard"
3) "Appraising the speaker's work"
4) "Deeper understanding"
5) "Showing the audience and speaker my understanding of the topic"
6) "Relevance for my own research".

Next, we asked which factor(s) contributed to their hesitation to ask a question during the Q&A sessions:

1) "Not feeling clever enough"
2) "Afraid I misunderstood the content of the presentation"
3) "I felt intimidated by the speaker"
4) "I felt intimidated by the audience"
5) "I felt intimidated by the setting (e.g. size of the room)"
6) "I felt intimidated by the session chair"
7) "I did not think my question was relevant/important"
8) "Afraid I would not be able to phrase/articulate my question well"
9) "I did raise my hand but was not chosen to ask a question"
10) "There was no time left to ask my question"
11) "I am too much of an introvert"
12) "I would rather ask my question after the session one-to-one with the speaker"
13) "I did not have the confidence"

Note that hesitation number 9 is presented separately from the other hesitations in the results, as the response to this hesitation was used in combination with the observational data to understand whether women ask less questions because they were chosen less often by the session hosts than men.

Lastly, we presented a series of statements to identify which conditions might make people feel more comfortable to ask a question:

1) "I feel comfortable asking questions during Q&A sessions"
2) "I feel more comfortable asking questions to a speaker who is of my own gender"
3) "I feel more comfortable asking question when my own gender is represented in the audience"
4) "I feel more comfortable asking questions when the audience is smaller"
5) "I feel more comfortable asking questions when the session host is of my own gender"
6) "I feel more comfortable asking questions when I know the speaker".

Similar to above, survey participants indicated on a 7-point Likert scale to what extent they agreed with the six statements, where the scale ranged from "Strongly disagree" (1) to "Strongly agree" (7), where 4 would indicate a neutral attitude.

We built two sets of models to identify what motivations, hesitations and conditions were more important for some gender identities than for others, and consequently which motivations, hesitations and conditions were the best predictors of whether a person asked a question at the congress or not. First, we only selected motivations and hesitations that were ticked at least 15 times in general. Next, we identified which factors out of the selected motivations, hesitations and conditions were significantly affected by gender. Separately for each motivation and hesitation, we then built binomial generalised linear models (PCS.1, S2 Table 15) using the lme4 R package v1.1.35.3 (Bates et al., 2015). The binary response of whether this motivation or hesitation was applicable or not (1 = yes, 0 = no) was used as the dependent variable, and gender was used as an independent variable (female, male or non-binary) as well as career stage (early-, mid-, late-career stage). For the ordinal condition responses, we built one ordered logistic regression (OLR) model for each one of the conditions with the R package MASS (81). To investigate whether gender had a significant effect on the response variable, we compared the fit of the model to a null model that only included the intercept using an LRT. We applied a multiple testing correction to all motivation, hesitation and condition LRTs collectively using the false discovery rate (FDR, Benjamini and Hochberg, 1995).

Next, we asked which of the motivations, hesitations and conditions affected the probability that the person asked a question during the congress. For this, we built 24 separate binomial linear models (PCS.2, S2 Table 15) using lme4, where the binary response whether the person asked a question during the congress (1 = one or multiple questions asked, 0 = no questions asked) was used as a dependent variable and the response of the motivation/hesitation/condition as the independent variable. We further included both gender and career stage in the models to account for potential direct effects of these variables on the probability that a person asked a question independent from the motivation/hesitation/condition. Again, we evaluated whether the motivation/hesitation/condition had a significant effect on question-asking probability using an LRT which compared the fit of the model to a null model that only included the intercept and applied an FDR correction to the LRT outputs of all 24 models collectively.

How did different social identities and people with different levels of expertise and English comfortability experience the conference?

We next identified which social identity and/or controlling variable(s) explained variation in congress experience. Post-congress survey participants indicated on a 7-point Likert scale (similar to above) to what extent they agreed with the following three statements about their congress experience:

1) “I felt heard during the conversations I had, both during Q&A sessions and social activities”

2) “I felt comfortable being myself”

3) “Attending the Behaviour 2023 congress helped me feel like I belong in my research field”

We built ordinal logistic regression (OLR) models to the responses to each of the three statements (PCS.3, PCS.4 and PCS.5 respectively, S2 Table 15) using the *polr* function from the R package MASS (81). First, we identified which of the social identity variables significantly improved the fit of the models by fitting six separate models for each statement, with one of the social identity variables included as an independent variable. A significant social identity was identified using an LRT which compared the model that included the social identity variable to a null model that only included the intercept. In addition to identifying significant social identity variables, we also fitted expertise rating and English comfort rating as potential confounding variables and assessed if they improved the fit of the models using an LRT. Only variables that significantly improved the fit of the model (i.e. the *p-value* of the LRT was less than 0.05) were included in the final model for that conference experience statement. We conducted a Wald test using the *coeftest* function from the R package lmtest v0.9-40 (82) to generate coefficients, standard errors and p-values, and the *confint* function from the same package to generate the corresponding confidence intervals.

Perceptions of equity, diversity and inclusivity among congress attendees (statistical analyses)

Similar to the analysis of congress experience, we investigated which social identity and/or controlling variable(s) explained variation in how attendees perceived EDI issues in the context of the congress and the broader research field. Survey participants indicated on a 7-point Likert scale to what extent they agreed with three statements about perceived EDI issues:

1) "I think the Congress attendees represented the diversity of researchers in our field"

2) "Our research field experiences equity, diversity and inclusion related issues (e.g. racism, homophobia, harassment, bullying etc)"

3) "I think the questions asked after the talks were equally divided across genders"

We took a similar approach as described above: (i) we fitted OLR models to the responses of each of the three EDI issue perception statements (PCS.6, PCS.7 and PCS.8 respectively, S2 Table 15), (ii) we identified which of the social identity variable(s) were significantly associated with the response to the statement by conducting LRTs that compared the model for that social identity or controlling variable against a null model that did not include the variable, (iii) we built the final model to include only social identity variables that significantly improved the fit of the model. In addition to identifying significant social identity variables, we also fitted age and English comfort rating as potential confounding variables and assessed if they improved the fit of the models using an LRT.

Qualitative analysis of open-ended questions

In the post-congress survey, participants were asked to respond to an open-ended question with their feedback or opinions on the congress. Of the 391 number of total respondents, 48% (*n* = 191) provided a response to this question, of which 185 could be coded into their respective sentiments.

We used Qualitative Content Analysis methodology (83) to code the open-ended responses. Codes were assigned to the main elements (distinct pieces of information that convey a particular idea; e.g. organisation, provision for accessibility, etc.) in the responses. These elements were further tagged with the sentiments expressed as being 'Positive' (e.g. *well* organised, *good* focus on EDI), 'Negative' (e.g. *tight* schedule/*inadequate* scheduling, *inadequate* provisions for accessibility) or providing a 'Suggestion' (e.g. *alternative* scheduling, search function in abstracts). Since multiple respondents provided extensive responses to the question, each response could therefore have more than one code and/or sentiment expressed in it. This preliminary coding was done by two independent people (both members of the research team) who coded all of the responses. The coders then discussed misalignments in coding until a consensus was achieved for all of the responses. At the end of this phase, we had 824 coded elements across 78 codes. These codes were then aggregated based on their similarity. At the end of this phase, we had 24 codes (8 in each sentiment).

All statistical analyses were implemented in R v.4.3.2 (80) using RStudio v. 2023.09.1. Data were visualised using the packages ggplot2 v3.5.1 (84), cowplot v1.1.3 (85) and viridis v0.6.5 (86).

Data and code availability

All anonymized data for the pre-congress survey, question-asking behaviour and post-congress survey can be found on https://github.com/rshuhuachen/ms_edi_behaviour23 and Zenodo https://zenodo.org/records/13825175 with DOI 10.5281/zenodo.13825175. These repositories also include all code used to analyse the data and additional documents shared to increase transparency and reproducibility, such as the Code of Conduct and the protocol used for collecting data on question-asking behaviour. Although all respondents of the post-congress survey consented to their data being used for research anonymously, did not publish the qualitative feedback that was part of the survey as anonymity cannot be guaranteed. A summary of the entire workflow, including the code and results, can be found on https://rshuhuachen.github.io/ms_edi_behaviour23/.

Author contributions

Conceptualization and funding acquisition: Ö.M., R.S.C., T.R.,

Project administration: N.S., R.S.C.

Supervision and data curation: R.S.C.

Investigation: A.L.B., A.J.P., A.L.M., B.S., I.D., J.K., J.I.H, J.T., K.P.G., M.G., N.C., Ö.M., P.B., P.K., R.S.C., S.K., S.Sa., S.St, T.R.

Methodology: A.J.P., B.A.C., J.d.L., J.I.H., J.T., P.K., K.P.G., M.G., N.C., P.B., S.K., S.Sa., T.R., R.S.C (lead)

Formal analysis and visualisation registration data: A.L.B., A.L.M., R.S.C.

Formal analysis and visualisation behavioural data: B.S., J.T., M.G., P.B., R.S.C.

Formal analysis and visualisation survey data: J.d.L., R.S.C., S.K., T.R.

Writing – original draft preparation: B.S., T.R., R.S.C. (lead)

Writing – review and editing: all

Science communication: I.D., I.S., R.S.C., S.St.

Ethical statement

This study has been approved by the Ethics Committee of Bielefeld University under application number 2023-140.

Funding Disclosure

This study was funded by the Behavioural Ecology Department, the Gender Equality Committee of Bielefeld University and the Equal Opportunities Initiative Fund awarded by the European Society of Evolutionary Biology. Part of the authors have been funded through the CRC-TRR 212 project number 316099922.

Competing Interests

The authors declare no competing interests.

Acknowledgements

This work would not have been possible without the following volunteers who helped us to collect data on question-asking during the congress: Athina Giannakara, Ettore Camerlenghi, Eva Rousselle, Magdalena Fus, and Pierre Guiart. We would also like to give a special thanks to the following student assistants that helped us digitise the collected data: Kiên Si Ho, Arne Kalinowski, Carolin Stober, Pascal Tebler, Philip Wolf, Berkan Boran, and Katharina Vockel. We acknowledge and thank the support for all our EDI initiatives, including conducting our study, by the three main responsible scientists of the congress: Oliver Krüger, Barbara Caspers and Klaus Reinhold. Next, we would like to give special thanks to Maria Fernanda Mastrocesare Araujo for her support, feedback and insights from a different STEM discipline. We also thank Tjomme van Mastrigt for his insights and contributions to the statistical analyses and visualisation of the model outputs. Furthermore, we would like to acknowledge the help in creating figures for science communication purposes by Gabe Winters. We are also grateful to Irene B. Jacobs for giving constructive feedback on the manuscript based on her expertise in the lack of female acknowledgement in maritime history. Lastly, we thank all of the session chairs who agreed to take part in our experiment during the congress.

References

1. Nielsen MW, Alegria S, Börjeson L, Etzkowitz H, Falk-Krzesinski HJ, Joshi A, et al. Gender diversity leads to better science. Proc Natl Acad Sci. 2017 Feb 21;114(8):1740–2.

2. Martin GC. The Effects Of Cultural Diversity In The Workplace. J Divers Manag JDM. 2014 Nov 21;9(2):89–92.

3. Saxena A. Workforce Diversity: A Key to Improve Productivity. Procedia Econ Finance. 2014;11:76–85.

4. Campbell LG, Mehtani S, Dozier ME, Rinehart J. Gender-Heterogeneous Working Groups Produce Higher Quality Science. Larivière V, editor. PLoS ONE. 2013 Oct 30;8(10):e79147.

5. AlShebli BK, Rahwan T, Woon WL. The preeminence of ethnic diversity in scientific collaboration. Nat Commun. 2018 Dec 4;9(1):5163.

6. Cronin MR, Alonzo SH, Adamczak SK, Baker DN, Beltran RS, Borker AL, et al. Anti-racist interventions to transform ecology, evolution and conservation biology departments. Nat Ecol Evol. 2021 Aug 9;5(9):1213–23.

7. Lagisz M, Aich U, Amin B, Rutkowska J, Sánchez-Mercado A, Lara C, et al. Little transparency and equity in scientific awards for early and mid-career researchers in ecology and evolution [Internet]. Life Sciences; 2022 Sep [cited 2023 Jan 16]. Available from: https://ecoevorxiv.org/repository/view/3655

8. Lee DN. Diversity and inclusion activisms in animal behaviour and the ABS: a historical view from the U.S.A. Anim Behav. 2020 Jun;164:273–80.

9. McGill BM, Foster MJ, Pruitt AN, Thomas SG, Arsenault ER, Hanschu J, et al. You are welcome here: A practical guide to diversity, equity, and inclusion for undergraduates embarking on an ecological research experience. Ecol Evol. 2021 Apr;11(8):3636–45.

10. Tulloch AIT. Improving sex and gender identity equity and inclusion at conservation and ecology conferences. Nat Ecol Evol. 2020 Aug 3;4(10):1311–20.

11. Blithe SJ, Elliott M. Gender inequality in the academy: microaggressions, work-life conflict, and academic rank. J Gend Stud. 2020 Oct 2;29(7):751–64.

12. Casad BJ, Franks JE, Garasky CE, Kittleman MM, Roesler AC, Hall DY, et al. Gender inequality in academia: Problems and solutions for women faculty in STEM. J Neurosci Res. 2021 Jan;99(1):13–23.

13. Clancy KBH, Nelson RG, Rutherford JN, Hinde K. Survey of Academic Field Experiences (SAFE): Trainees Report Harassment and Assault. Apicella CL, editor. PLoS ONE. 2014 Jul 16;9(7):e102172.

14. Martínez-Blancas A, Bender A, Zepeda V, McGuire R, Tabares O, Amarasekare P, et al. Surviving racism and sexism in academia: Sharing experiences, insights, and perspectives. Bull Ecol Soc Am. 2023;104(1):1–8.

15. McGee EO. Interrogating Structural Racism in STEM Higher Education. Educ Res. 2020 Dec;49(9):633–44.

16. Settles IH, Jones MK, Buchanan NT, Dotson K. Epistemic exclusion: Scholar (ly) devaluation that marginalizes faculty of color. J Divers High Educ. 2021;14(4):493.

17. Cech EA, Waidzunas TJ. Systemic inequalities for LGBTQ professionals in STEM. Sci Adv. 2021 Jan 15;7(3):eabe0933.

18. Marosi NKM, Avraamidou L, López López M. Queer individuals' experiences in STEM learning and working environments. Stud Sci Educ. 2024 Feb 29;1–39.

19. Brown N, Leigh J. Ableism in academia: where are the disabled and ill academics? Disabil Soc. 2018 Jul 3;33(6):985–9.

20. Cech EA. Engineering ableism: The exclusion and devaluation of engineering students and professionals with physical disabilities and chronic and mental illness. J Eng Educ. 2023 Apr;112(2):462–87.

21. Crabtree A, Neikirk K, Marshall A, Barongan T, Beasley HK, Lopez EG, et al. Strategies for change: thriving as an individual with a disabilty in STEMM. Pathog Dis. 2023 Jan 17;81:ftac045.

22. Dorenkamp I, Weiß EE. What makes them leave? A path model of postdocs' intentions to leave academia. High Educ. 2018;75:747–67.

23. Douglas HM, Settles IH, Spence Cheruvelil K, Montgomery GM, Elliott KC, Cech EA, et al. The importance of inclusive climate within the research group, department, and profession for marginalized science scholars' career outcomes. J Divers High Educ. 2024;

24. White-Lewis DK, O'Meara K, Mathews K, Havey N. Leaving the institution or leaving the academy? Analyzing the factors that faculty weigh in actual departure decisions. Res High Educ. 2023;64(3):473–94.

25. Almukhambetova A, Torrano DH, Nam A. Fixing the Leaky Pipeline for Talented Women in STEM. Int J Sci Math Educ. 2023 Jan;21(1):305–24.

26. Pell AN. Fixing the leaky pipeline: women scientists in academia. J Anim Sci. 1996;74(11):2843.

27. Resmini M. The 'Leaky Pipeline'. Chem – Eur J. 2016;22(11):3533–4.

28. Figueiredo D. Walking through the Leaky Academic Pipeline in STEM: Equity Not Equality Needed for Women and under Represented Minorities (URMs). In: Bhatti F, Taheri E, editors. Sustainable Development [Internet]. IntechOpen; 2024 [cited 2024 Jul 26]. Available from: https://www.intechopen.com/chapters/87062

29. O'Brien LT, Bart HL, Garcia DM. Why are there so few ethnic minorities in ecology and evolutionary biology? Challenges to inclusion and the role of sense of belonging. Soc Psychol Educ. 2020 Apr;23(2):449–77.

30. Sarraju A, Ngo S, Rodriguez F. The leaky pipeline of diverse race and ethnicity representation in academic science and technology training in the United States, 2003–2019. Rees CA, editor. PLOS ONE. 2023 Apr 26;18(4):e0284945.

31. Sardelis S, Oester S, Liboiron M. Ten Strategies to Reduce Gender Inequality at Scientific Conferences. Front Mar Sci. 2017 Jul 25;4:231.

32. Schell CJ, Guy C, Shelton DS, Campbell-Staton SC, Sealey BA, Lee DN, et al. Recreating Wakanda by promoting Black excellence in ecology and evolution. Nat Ecol Evol. 2020 Jul 24;4(10):1285–7.

33. Tilghman S, Alberts B, Colón-Ramos D, Dzirasa K, Kimble J, Varmus H. Concrete steps to diversify the scientific workforce. Science. 2021 Apr 9;372(6538):133–5.

34. Tseng M, El-Sabaawi RW, Kantar MB, Pantel JH, Srivastava DS, Ware JL. Strategies and support for Black, Indigenous, and people of colour in ecology and evolutionary biology. Nat Ecol Evol. 2020 Jul 7;4(10):1288–90.

35. Hooker SK, Simmons SE, Stimpert AK, McDonald BI. Equity and career-life balance in marine mammal science? Mar Mammal Sci. 2017 Jul;33(3):955–65.

36. Michailidis MP, Morphitou RN, Theophylatou I. Women at workequality versus inequality: barriers for advancing in the workplace. Int J Hum Resour Manag. 2012 Nov;23(20):4231–45.

37. Womack VY, Wood CV, House SC, Quinn SC, Thomas SB, McGee R, et al. Culturally aware mentorship: Lasting impacts of a novel intervention on academic administrators and faculty. Mughal MAZ, editor. PLOS ONE. 2020 Aug 7;15(8):e0236983.

38. Espinoza O. Solving the equity–equality conceptual dilemma: a new model for analysis of the educational process. Educ Res. 2007 Dec;49(4):343–63.

39. Secada WG. Educational equity versus equality of education: An alternative conception. Equity Educ. 1989;68–88.

40. Edwards JD, Barthelemy RS, Frey RF. Relationship between Course-Level Social Belonging (Sense of Belonging and Belonging Uncertainty) and Academic Performance in General Chemistry 1. J Chem Educ. 2022 Jan 11;99(1):71–82.

41. Li Y, Singh C. Sense of belonging is an important predictor of introductory physics students' academic performance. Phys Rev Phys Educ Res. 2023 Oct 3;19(2):020137.

42. Miller RA, Downey M. Examining the STEM climate for queer students with disabilities. J Postsecond Educ Disabil. 2020;33(2):169–81.

43. Morales N, Bisbee O'Connell K, McNulty S, Berkowitz A, Bowser G, Giamellaro M, et al. Promoting inclusion in ecological field experiences: Examining and overcoming barriers to a professional rite of passage. Bull Ecol Soc Am. 2020 Oct;101(4):e01742.

44. De Leon FLL, McQuillin B. The Role of Conferences on the Pathway to Academic Impact: Evidence from a Natural Experiment. J Hum Resour. 2020;55(1):164–93.

45. Hauss K. What are the social and scientific benefits of participating at academic conferences? Insights from a survey among doctoral students and postdocs in Germany. Res Eval. 2021 Oct 7;30(1):1–12.

46. Ford HL, Brick C, Blaufuss K, Dekens PS. Gender inequity in speaking opportunities at the American Geophysical Union Fall Meeting. Nat Commun. 2018 Apr 24;9(1):1358.

47. Bhayankaram KP, Prathivadi Bhayankaram N. Conference panels: do they reflect the diversity of the NHS workforce? BMJ Lead. 2022 Mar;6(1):57–9.

48. Barreto J, Romitelli I, Santana P, Assis AP, Pardini R, Leite M. Is the audience gender-blind? Smaller audience in female talks highlights prestige differences in academia [Internet]. 2024 [cited 2024 Jul 25]. Available from: https://ecoevorxiv.org/repository/view/7175/

49. Lupon A, Rodríguez-Lozano P, Bartrons M, Anadon-Rosell A, Batalla M, Bernal S, et al. Towards women-inclusive ecology: Representation, behavior, and perception of women at an international conference. Risse-Buhl U, editor. PLOS ONE. 2021 Dec 10;16(12):e0260163.

50. Carter AJ, Croft A, Lukas D, Sandstrom GM. Women's visibility in academic seminars: Women ask fewer questions than men. Sugimoto CR, editor. PLOS ONE. 2018 Sep 27;13(9):e0202743.

51. Davenport JRA, Fouesneau M, Grand E, Hagen A, Poppenhaeger K, Watkins LL. Studying Gender in Conference Talks -- data from the 223rd meeting of the American Astronomical Society [Internet]. arXiv; 2014 [cited 2023 May 10]. Available from: http://arxiv.org/abs/1403.3091

52. Hinsley A, Sutherland WJ, Johnston A. Men ask more questions than women at a scientific conference. Pavlova MA, editor. PLOS ONE. 2017 Oct 16;12(10):e0185534.

53. Käfer J, Betancourt A, Villain AS, Fernandez M, Vignal C, Marais GAB, et al. Progress and Prospects in Gender Visibility at SMBE Annual Meetings. Genome Biol Evol. 2018 Mar 1;10(3):901–8.

54. Pritchard J, Masters K, Allen J, Contenta F, Huckvale L, Wilkins S, et al. Asking gender questions. Astron Geophys. 2014 Dec 1;55(6):6.8-6.12.

55. Favaro B, Oester S, Cigliano JA, Cornick LA, Hind EJ, Parsons ECM, et al. Your Science Conference Should Have a Code of Conduct. Front Mar Sci [Internet]. 2016 Jun 22 [cited

2023 Feb 23];3. Available from: http://journal.frontiersin.org/Article/10.3389/fmars.2016.00103/abstract

56. Niner HJ, Johri S, Meyer J, Wassermann SN. The pandemic push: can COVID-19 reinvent conferences to models rooted in sustainability, equitability and inclusion? Socio-Ecol Pract Res. 2020 Sep;2(3):253–6.

57. Raby CL, Madden JR. Moving academic conferences online: Aids and barriers to delegate participation. Ecol Evol. 2021 Apr;11(8):3646–55.

58. Morin G, Talbot D. Cultural intelligence of expatriate workers: a systematic review. Manag Rev Q. 2023 Feb;73(1):413–54.

59. Earley PC, Ang S. Cultural intelligence: Individual interactions across cultures. 2003;

60. Khan SD, Mohammed LA, Khan S. Influence of socio-economic status on english speaking and writing anxiety: a comprehensive literature review. Educ Adm Theory Pract. 2024;30(5):14703–9.

61. Janardhan M. The quest for fluency: English language challenges for non-native learners. Int J Engl Lit Soc Sci. 2024;9(3):469–72.

62. Carlin BA, Gelb BD, Belinne JK, Ramchand L. Bridging the gender gap in confidence. Bus Horiz. 2018 Sep;61(5):765–74.

63. Vajapey SP, Weber KL, Samora JB. Confidence gap between men and women in medicine: a systematic review. Curr Orthop Pract. 2020 Sep;31(5):494–502.

64. Herbst THH. Gender differences in self-perception accuracy: The confidence gap and women leaders' underrepresentation in academia. SA J Ind Psychol [Internet]. 2020 Feb 24 [cited 2024 Jul 25];46. Available from: http://www.sajip.co.za/index.php/SAJIP/article/view/1704

65. Krauss S, Orth U, Robins RW. Family environment and self-esteem development: A longitudinal study from age 10 to 16. J Pers Soc Psychol. 2020;119(2):457.

66. Master A. Gender Stereotypes Influence Children's STEM Motivation. Child Dev Perspect. 2021 Sep;15(3):203–10.

67. González-Pérez S, Mateos De Cabo R, Sáinz M. Girls in STEM: Is It a Female Role-Model Thing? Front Psychol. 2020 Sep 10;11:2204.

68. Bailey EG, Greenall RF, Baek DM, Morris C, Nelson N, Quirante TM, et al. Female In-Class Participation and Performance Increase with More Female Peers and/or a Female Instructor in Life Sciences Courses. Eddy SL, editor. CBE—Life Sci Educ. 2020 Sep;19(3):ar30.

69. Sutter M, Glätzle-Rützler D. Gender Differences in the Willingness to Compete Emerge Early in Life and Persist. Manag Sci. 2015 Oct;61(10):2339–54.

70. Huang K, Yeomans M, Brooks AW, Minson J, Gino F. It doesn't hurt to ask: Question-asking increases liking. J Pers Soc Psychol. 2017;113(3):430.

71. Amano T, Ramírez-Castañeda V, Berdejo-Espinola V, Borokini I, Chowdhury S, Golivets M, et al. The manifold costs of being a non-native English speaker in science. Dirnagl U, editor. PLOS Biol. 2023 Jul 18;21(7):e3002184.

72. Stefanoudis PV, Biancani LM, Cambronero-Solano S, Clark MR, Copley JT, Easton E, et al. Moving conferences online: lessons learned from an international virtual meeting. Proc R Soc B Biol Sci. 2021 Oct 27;288(1961):20211769.

73. Giglio AD, Costa MUPD. The use of artificial intelligence to improve the scientific writing of non-native english speakers. Rev Assoc Médica Bras. 2023;69(9):e20230560.

74. Hwang SI, Lim JS, Lee RW, Matsui Y, Iguchi T, Hiraki T, et al. Is ChatGPT a “Fire of Prometheus” for Non-Native English-Speaking Researchers in Academic Writing? Korean J Radiol. 2023;24(10):952.

75. Decady Guijarro R, Bourgeault IL. Supporting diverse health leadership requires active listening, observing, learning and bystanding. Equal Divers Incl Int J. 2023 Mar 28;42(3):346–63.

76. Holmes MH, Jackson JK, Stoiko R. Departmental Dialogues: Facilitating Positive Academic Climates to Improve Equity in STEM Disciplines. Innov High Educ. 2016 Nov;41(5):381–94.

77. Arcila Hernández L, Chodkowski N, Treibergs K. A guide to implementing inclusive and accessible virtual poster sessions. J Microbiol Biol Educ. 2022;23(1):e00237-21.

78. Holt EA, Heim AB, Tessens E, Walker R. Thanks for inviting me to the party: Virtual poster sessions as a way to connect in a time of disconnection. Ecol Evol. 2020 Nov;10(22):12423–30.

79. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw [Internet]. 2015;67(1). Available from: http://www.jstatsoft.org/v67/i01/

80. Team RC. R: A language and environment for statistical computing. Vienna; 2021.

81. Venables WN, Ripley BD. Modern applied statistics with S [Internet]. 4th ed. New York: Springer; 2002. Available from: https://www.stats.ox.ac.uk/pub/MASS4/

82. Zeileis A, Hothorn T. Diagnostic checking in regression relationships. R News. 2002;2(3):7–10.

83. Schreier M. Qualitative Content Analysis. In: The SAGE Handbook of Qualitative Data Analysis [Internet]. 1 Oliver’s Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Inc.; 2014 [cited 2024 Sep 10]. p. 170–83. Available from:

https://sk.sagepub.com/reference/the-sage-handbook-of-qualitative-data-analysis/i1108.xml

84. Wickham H. ggplot2: Elegant graphics for data analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org

85. Wilke CO. cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Internet]. 2024. Available from: https://wilkelab.org/cowplot/

86. Garnier, Simon, Ross, Noam, Rudis, Robert, et al. viridis(Lite)- colorblind-friendly color maps for R [Internet]. 2024. Available from: https://sjmgarnier.github.io/viridis/

87. Cohen J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20.

88. Gamer M, Lemon J, Singh FP. irr: Various Coefficients of Interrater Reliability and Agreement [Internet]. 2005 [cited 2024 Sep 16]. p. 0.84.1. Available from: https://CRAN.R-project.org/package=irr

89. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63.

Supporting information

S1 File: Methods and Results

S1 Methods 1: Gender and pronouns

S1 Methods 2: Manual correction and curation of data on question-asking behaviour

S1 Methods 3: Inter-observer reliability

S1 Methods 4: Combining data collected from multiple observers of the same session

S1 Results 1: Can we identify other gender disparities in oral sessions?

S2 File. Supplementary Tables and Figures

S2 Table 1. Model output for question-asking models based on both behavioural and survey data. The models investigated gender disparities in question-asking (QA), raising hands (RH) or being chosen to ask a question (GC), where some QA models were given a name in the Methods and are indicated in the table in parenthesis. For the models using behavioural data (BD), the intercepts are indicative of a gender disparity. For the models using survey data (SD), we tested whether including gender improved model fit using an LRT and additionally computed a Wald test. Bold numbers indicate statistical significance ($p < 0.05$). One observation in the BD represents a single question, whereas one observation in the SD represents a single response. Abbreviations: O = observations, T = talks, IC = intercept.

S2 Table 2. Model output for gender differences in question-asking motivation and probability based on post-congress survey data. We first tested which motivations were significantly influenced by gender using LRTs and consequently conducted Wald tests (left). We next tested which motivations were predictive of the probability that a person asked a question during the congress, also using LRTs and Wald tests (right). Bold numbers indicate statistical significance ($p < 0.05$).

S2 Table 3. Model output for gender differences in question-asking hesitation and probability based on post-congress survey data. We first tested which hesitations were significantly influenced by gender using LRTs and consequently conducted Wald tests (left). We next tested which hesitations were predictive of the probability that a person asked a question during the congress, also using LRTs and Wald tests (right). Bold numbers indicate statistical significance ($p < 0.05$).

S2 Table 4. Model output for age differences in question-asking motivation and hesitation based on post-congress survey data. This output is based on the same models presented in S2 Table 2 and S2 Table 3, but here we report the estimates of the career stages. The reference level was the early-career stage, against which both mid- and late-career stages were compared. We only conducted a Wald test and did not correct for multiple testing, as career stage was not our main variable of interest (but gender was). Bold numbers indicate statistical significance ($p < 0.05$).

S2 Table 5. Model output for variables affecting female question-asking probability using the behavioural data. We tested whether including each variable significantly improves the model fit using a likelihood ratio test (LRT) and additionally report model output of the Wald test.

S2 Table 6. Model output for gender effects on feeling comfortable asking questions using the post-congress survey data. We tested whether including gender significantly improved model fits using LRTs and additionally report model output for the estimates of female gender compared to male gender using Wald tests. Bold numbers indicate statistical significance ($p < 0.05$).

S2 Table 7. Model output testing gender effects of the first questioner on the probability that a woman asks a question in the rest of the Q&A using all questions (except the first one, Q1). We first tested whether including the condition significantly improved the model fit using LRTs and additionally report model output of Wald tests. Bold numbers indicate statistical significance (p < 0.05). Abbreviations: T = talks, Q = questions, W = woman, M = man, P = probability.

S2 Table 8. Model output testing gender effects of the first questioner on the probability that a woman asks the second question (Q2). We first tested whether including condition (first question to a woman or first question to a man) significantly improved the model fit using LRTs and additionally report model output of Wald tests. Bold numbers indicate statistical significance (p < 0.05). Abbreviations: T = talks, Q = questions, W = woman, M = man, P = probability.

S2 Table 9. Model output for the three statements on congress experience. Univariate models tested for the significance of each variable using LRTs and only variables that significantly improved the model fit (indicated in bold) were included in the final model. Bold numbers indicate statistical significance (p < 0.05).

S2 Table 10. Model output for the three statements on EDI issue perception. Univariate models tested for the significance of each variable using LRTs and only variables that significantly improved the model fit (indicated in bold) were included in the final model. Bold numbers indicate statistical significance (p < 0.05).

S2 Table 11. Codes used for the qualitative analysis of open text responses. Both condensed and expended codes are presented as well as their frequency the codes were expressed in the responses.

S2 Table 12. Models for the observational behavioural data. This table includes both the research question each model addressed expressed verbally and in lme4 model syntax.

S2 Table 13. Models for the effect of the gender of the first questioner. This table includes both the research question each model addressed expressed verbally and in lme4 model syntax.

S2 Table 14. Dependent variables and predictors used to identify other gender disparities in oral sessions. The results of this analysis are only presented in the Supplementary Materials.

S2 Table 15. Models for the post-congress survey data. This table includes both the research question each model addressed expressed verbally and in lme4 model syntax.

S2 Table 16. Conditional workflow used to combine data collected by different observers of the same session.

S2 Figure 1. Inter-observer reliability statistics for each variable collected on the three different levels. Cohen's Kappa and ICC statistics calculated for variables collected per session (a), talk (b) and question (c). Vertical purple and red dotted lines indicate commonly accepted thresholds for Cohen's kappa (Cohen's kappa > 0.8 = "near perfect"; Cohen, 1968) and ICC (ICC > 0.75 = "good"; Koo & Li, 2016) respectively.

S3 File. Supplementary analysis on age

Fig 1.

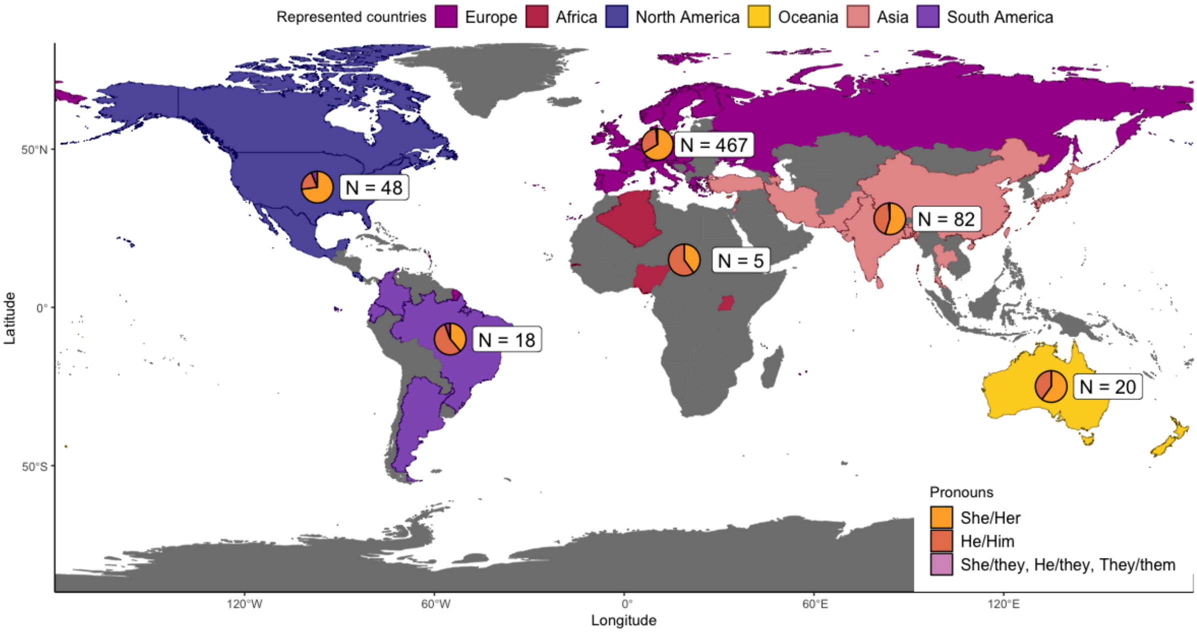
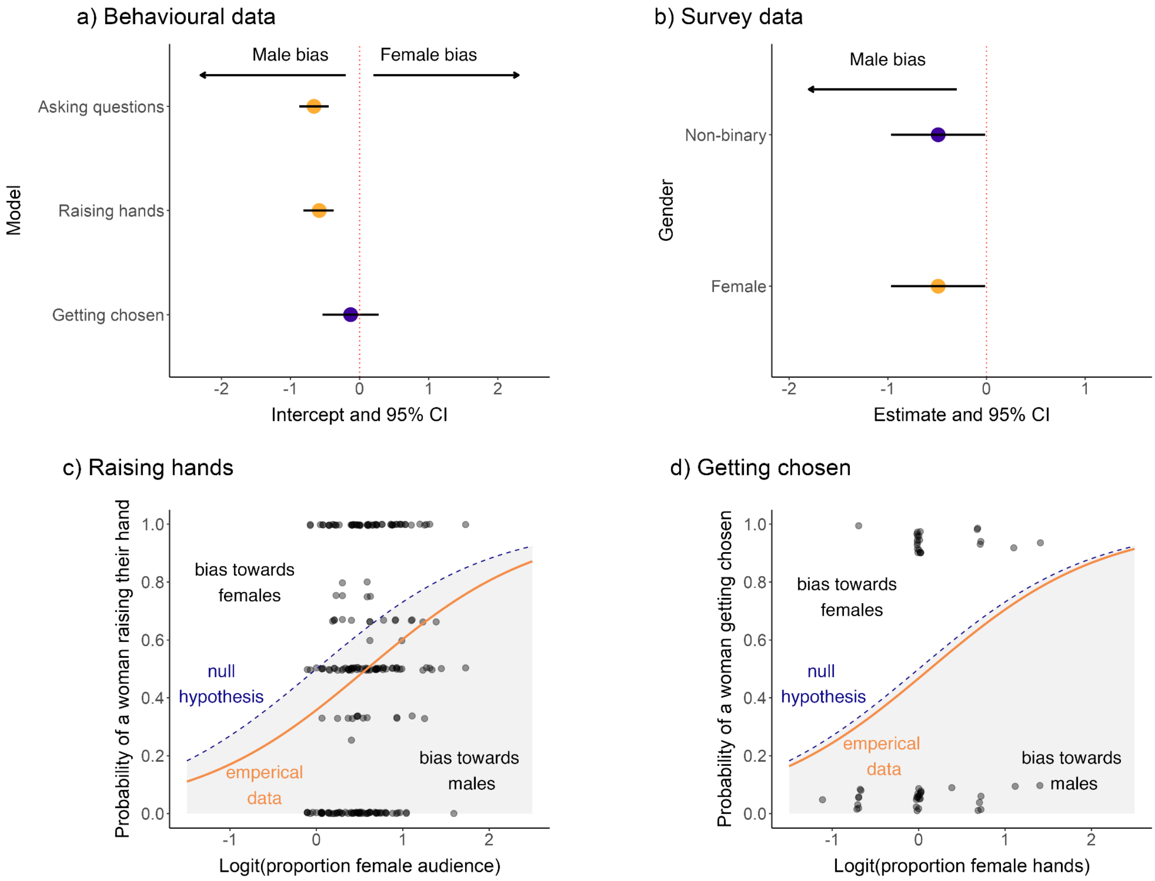
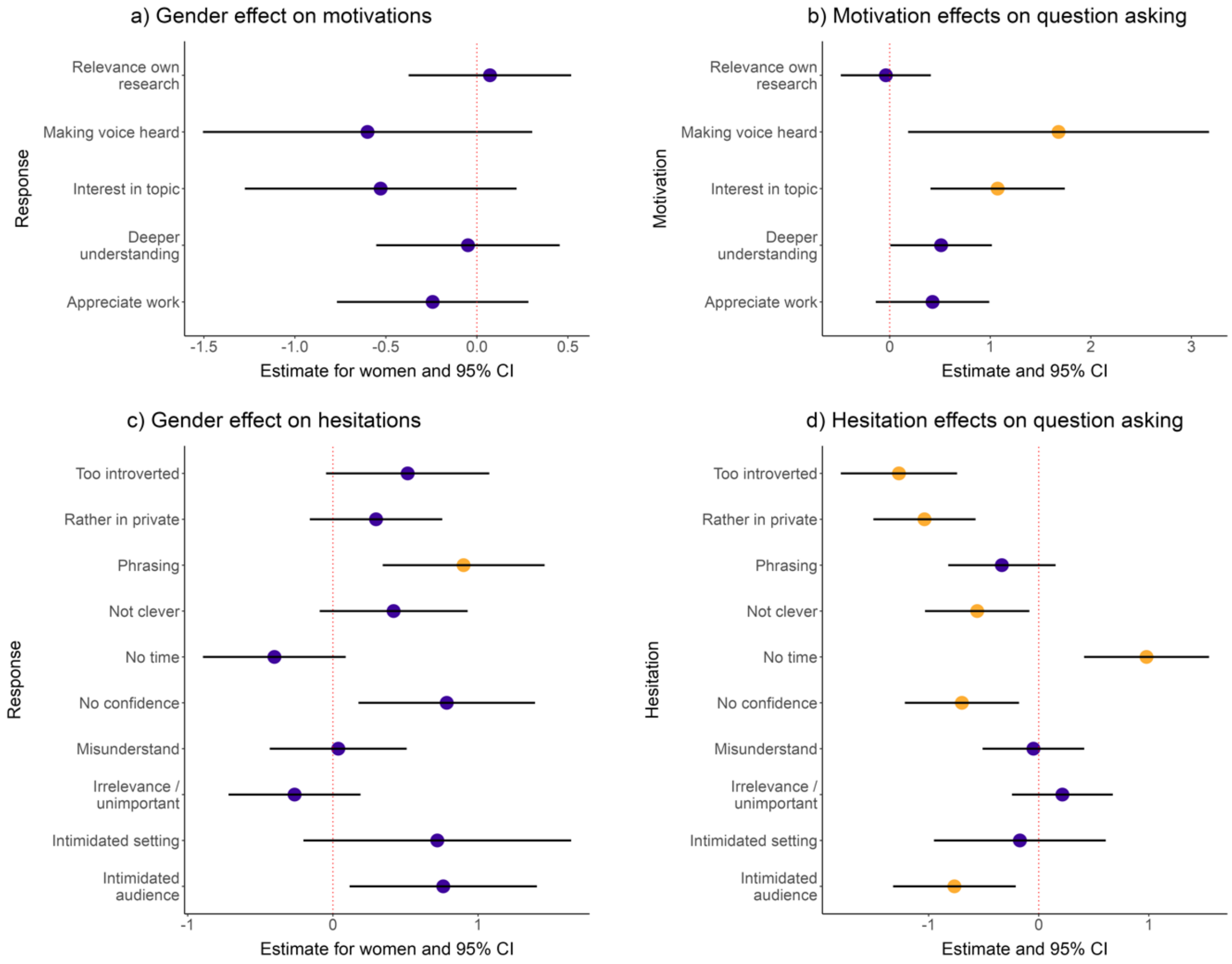

Represented countries
Europe
Africa
North America
Oceania
Asia
South America
N = 467
N = 48
N = 82
N = 5
N = 18
N = 20
Pronouns
She/Her
He/Him
She/they, He/they, They/them
50°N
0°
50°S
Latitude
120°W
60°W
0°
60°E
120°E
Longitude

Fig 2.



a) Behavioural data
Male bias
Female bias
Asking questions
Raising hands
Getting chosen
Model
-2
-1
0
1
2
Intercept and 95% CI
b) Survey data
Male bias
Non-binary
Female
Gender
-2
-1
0
1
Estimate and 95% CI
c) Raising hands
Probability of a woman raising their hand
1.0
0.8
0.6
0.4
0.2
0.0
bias towards females
null hypothesis
emperical data
bias towards males
-1
0
1
2
Logit(proportion female audience)
d) Getting chosen
Probability of a woman getting chosen
1.0
0.8
0.6
0.4
0.2
0.0
bias towards females
null hypothesis
emperical data
bias towards males
-1
0
1
2
Logit(proportion female hands)

Fig 3.



a) Gender effect on motivations
Response
Relevance own research
Making voice heard
Interest in topic
Deeper understanding
Appreciate work
-1.5
-1.0
-0.5
0.0
0.5
Estimate for women and 95% CI
b) Motivation effects on question asking
Motivation
Relevance own research
Making voice heard
Interest in topic
Deeper understanding
Appreciate work
0
1
2
3
Estimate and 95% CI
c) Gender effect on hesitations
Response
Too introverted
Rather in private
Phrasing
Not clever
No time
No confidence
Misunderstand
Irrelevance / unimportant
Intimidated setting
Intimidated audience
-1
0
1
Estimate for women and 95% CI
d) Hesitation effects on question asking
Hesitation
Too introverted
Rather in private
Phrasing
Not clever
No time
No confidence
Misunderstand
Irrelevance / unimportant
Intimidated setting
Intimidated audience
-1
0
1
Estimate and 95% CI

Fig 4.

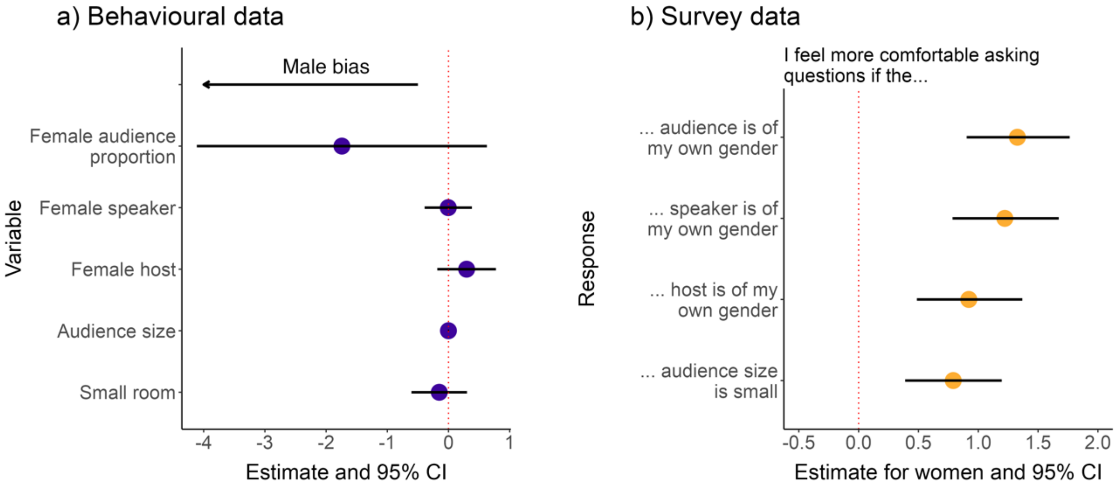
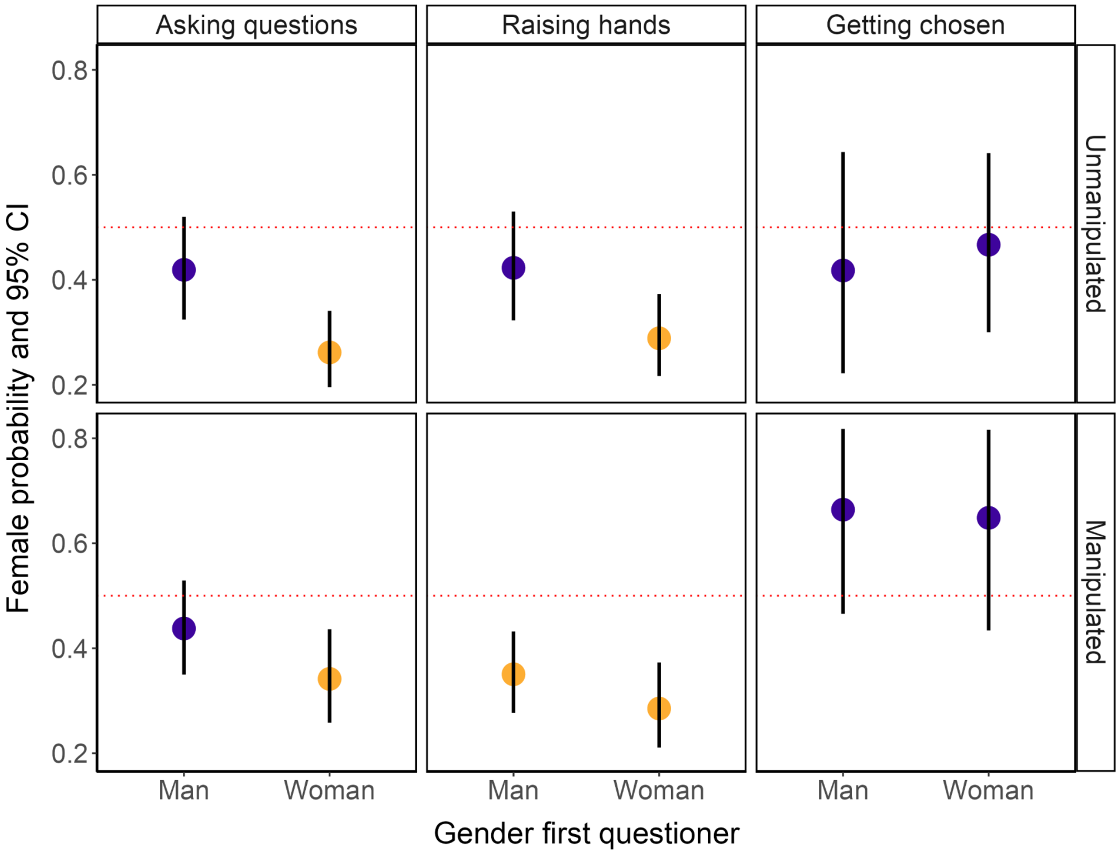
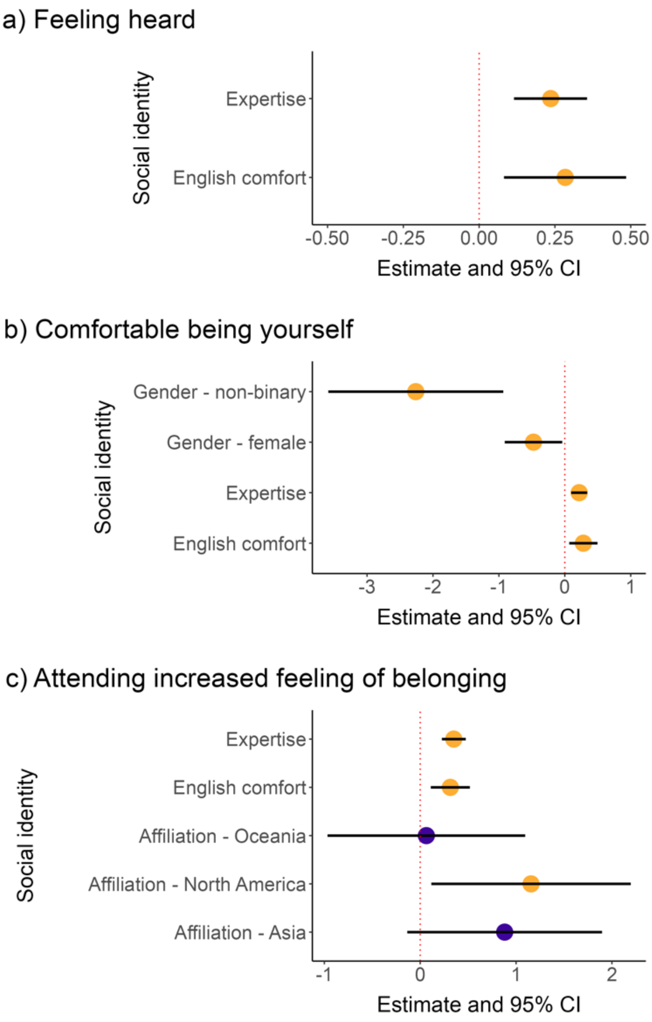

a) Behavioural data
Male bias
Female audience proportion
Female speaker
Female host
Audience size
Small room
Variable
-4
-3
-2
-1
0
1
Estimate and 95% CI
b) Survey data
I feel more comfortable asking questions if the...
... audience is of my own gender
... speaker is of my own gender
... host is of my own gender
... audience size is small
Response
-0.5
0.0
0.5
1.0
1.5
2.0
Estimate for women and 95% CI

Fig 5.



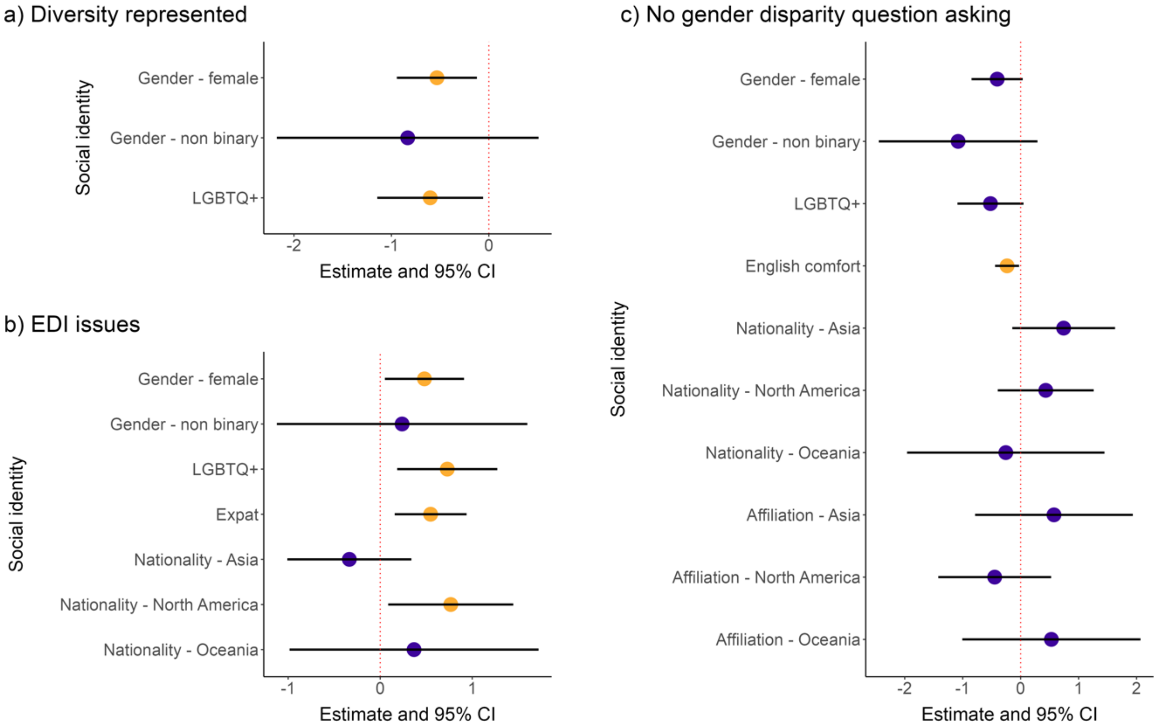
Asking questions
Raising hands
Getting chosen
Unmanipulated
Manipulated
Female probability and 95% CI
0.8
0.6
0.4
0.2
Man
Woman
Gender first questioner

Fig 6.



a) Feeling heard
Social identity
Expertise
English comfort
-0.50
-0.25
0.00
0.25
0.50
Estimate and 95% CI
b) Comfortable being yourself
Social identity
Gender - non-binary
Gender - female
Expertise
English comfort
-3
-2
-1
0
1
Estimate and 95% CI
c) Attending increased feeling of belonging
Social identity
Expertise
English comfort
Affiliation - Oceania
Affiliation - North America
Affiliation - Asia
-1
0
1
2
Estimate and 95% CI

Fig 7.



a) Diversity represented
Social identity
Gender - female
Gender - non binary
LGBTQ+
-2
-1
0
Estimate and 95% CI
b) EDI issues
Social identity
Gender - female
Gender - non binary
LGBTQ+
Expat
Nationality - Asia
Nationality - North America
Nationality - Oceania
-1
0
1
Estimate and 95% CI
c) No gender disparity question asking
Social identity
Gender - female
Gender - non binary
LGBTQ+
English comfort
Nationality - Asia
Nationality - North America
Nationality - Oceania
Affiliation - Asia
Affiliation - North America
Affiliation - Oceania
-2
-1
0
1
2
Estimate and 95% CI

Fig 8.

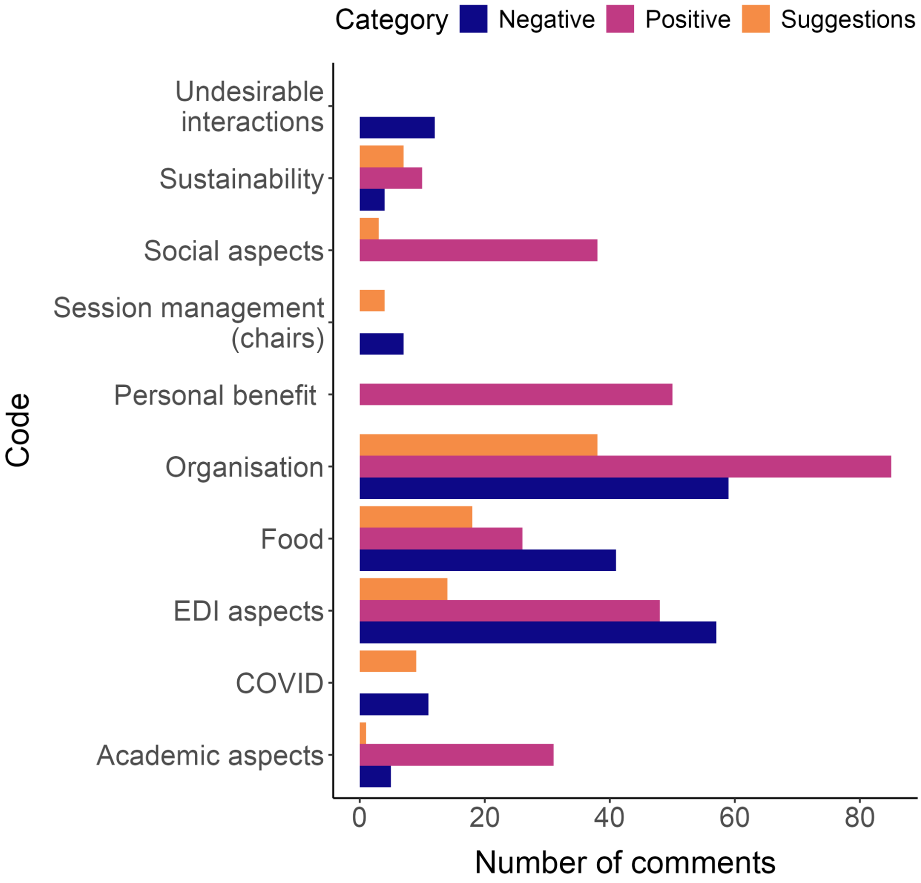
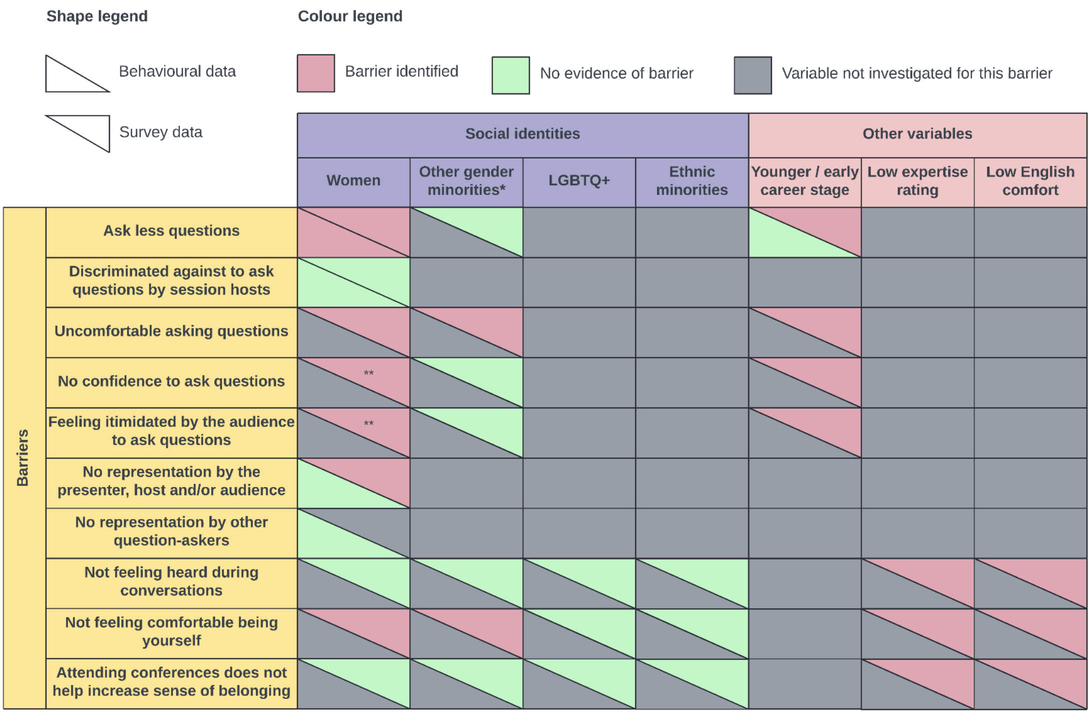

Category
Negative
Positive
Suggestions
Code
Undesirable interactions
Sustainability
Social aspects
Session management (chairs)
Personal benefit
Organisation
Food
EDI aspects
COVID
Academic aspects
0
20
40
60
80
Number of comments

Fig 9.



Shape legend

◺ Behavioural data

◹ Survey data

Colour legend

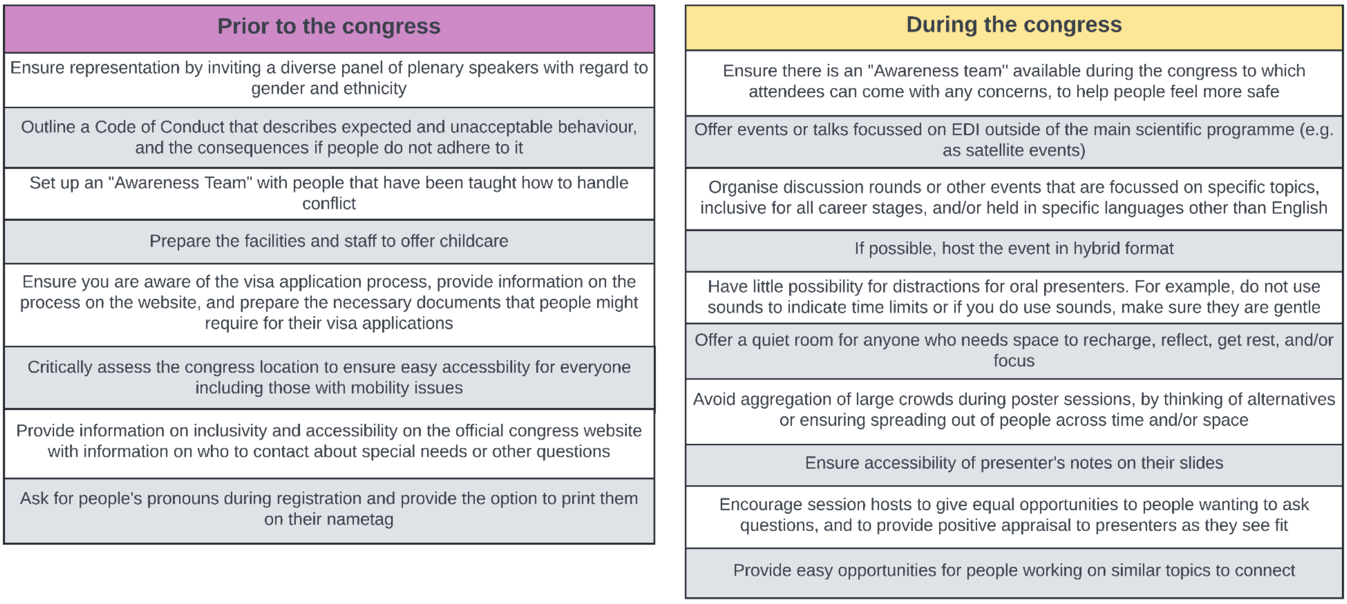
🟥 Barrier identified

🟩 No evidence of barrier

⬛ Variable not investigated for this barrier

		Social identities				Other variables		
		Women	Other gender minorities*	LGBTQ+	Ethnic minorities	Younger / early career stage	Low expertise rating	Low English comfort
Barriers	Ask less questions	◺🟥 ◹🟥	◺⬛ ◹🟩	⬛	⬛	◺🟩 ◹🟥	⬛	⬛
	Discriminated against to ask questions by session hosts	◺🟩 ◹🟩	⬛	⬛	⬛	⬛	⬛	⬛
	Uncomfortable asking questions	◺⬛ ◹🟥	◺⬛ ◹🟥	⬛	⬛	◺⬛ ◹🟥	⬛	⬛
	No confidence to ask questions	◺⬛ ◹🟥 **	◺⬛ ◹🟩	⬛	⬛	◺⬛ ◹🟥	⬛	⬛
	Feeling itimidated by the audience to ask questions	◺⬛ ◹🟥 **	◺⬛ ◹🟩	⬛	⬛	◺⬛ ◹🟥	⬛	⬛
	No representation by the presenter, host and/or audience	◺🟩 ◹🟥	⬛	⬛	⬛	⬛	⬛	⬛
	No representation by other question-askers	◺🟩 ◹⬛	⬛	⬛	⬛	⬛	⬛	⬛
	Not feeling heard during conversations	◺⬛ ◹🟩	◺⬛ ◹🟩	◺⬛ ◹🟩	◺⬛ ◹🟩	⬛	◺⬛ ◹🟥	◺⬛ ◹🟥
	Not feeling comfortable being yourself	◺⬛ ◹🟥	◺⬛ ◹🟥	◺⬛ ◹🟩	◺⬛ ◹🟩	⬛	◺⬛ ◹🟥	◺⬛ ◹🟥
	Attending conferences does not help increase sense of belonging	◺⬛ ◹🟩	◺⬛ ◹🟩	◺⬛ ◹🟩	◺⬛ ◹🟩	⬛	◺⬛ ◹🟥	◺⬛ ◹🟥

Fig 10.



Prior to the congress
Ensure representation by inviting a diverse panel of plenary speakers with regard to gender and ethnicity
Outline a Code of Conduct that describes expected and unacceptable behaviour, and the consequences if people do not adhere to it
Set up an "Awareness Team" with people that have been taught how to handle conflict
Prepare the facilities and staff to offer childcare
Ensure you are aware of the visa application process, provide information on the process on the website, and prepare the necessary documents that people might require for their visa applications
Critically assess the congress location to ensure easy accessbility for everyone including those with mobility issues
Provide information on inclusivity and accessibility on the official congress website with information on who to contact about special needs or other questions
Ask for people's pronouns during registration and provide the option to print them on their nametag

During the congress
Ensure there is an "Awareness team" available during the congress to which attendees can come with any concerns, to help people feel more safe
Offer events or talks focussed on EDI outside of the main scientific programme (e.g. as satellite events)
Organise discussion rounds or other events that are focussed on specific topics, inclusive for all career stages, and/or held in specific languages other than English
If possible, host the event in hybrid format
Have little possibility for distractions for oral presenters. For example, do not use sounds to indicate time limits or if you do use sounds, make sure they are gentle
Offer a quiet room for anyone who needs space to recharge, reflect, get rest, and/or focus
Avoid aggregation of large crowds during poster sessions, by thinking of alternatives or ensuring spreading out of people across time and/or space
Ensure accessibility of presenter's notes on their slides
Encourage session hosts to give equal opportunities to people wanting to ask questions, and to provide positive appraisal to presenters as they see fit
Provide easy opportunities for people working on similar topics to connect

Supporting Materials

S1 File: Methods and Results

S1 Methods 1

Inferring gender

Previous studies on gender disparities in question asking perceived gender through appearance only and have highlighted the limitations of this approach (50,52). We acknowledge that assuming a person's gender identity based on their appearance is imperfect as i) observers might be biased in assessing another person's gender identity due to cultural and personal differences, ii) a person's gender expression, in terms of clothing and appearance, is not necessarily related to their gender identity, and iii) gender is non-binary and can be fluid, and making assumptions can wrongly categorise a person into a binary gender. Therefore, we hoped to prevent misgendering by offering the option for attendees to print their preferred pronouns on their nametags.

However, not every person is comfortable having pronouns publicly shown, or does not want this printed because of other reasons. In practice the names were also difficult or impossible to read from a distance. Therefore, we did have to perceive the gender of people asking questions through their appearance. To evaluate how consistent the gender perceived by observers was with the preferred pronouns provided by congress attendees, we used the pre-congress survey (i.e. the registration form) to perform a cross-check based on the gender data collected on session hosts and speakers for whom we knew the name and who consented to print their pronouns on their name tags and consented to us using their data for our study. In 94% of these observations (305/325), the observers correctly inferred the gender of women and in 99% of observations this was correct for men (139/141), although the success rate was much lower for non-binary participants (27%, 3/11). If we were more likely to misgender men than women, we likely underestimated the gender disparity in question asking. If we were more likely to misgender women than men, we might have overestimated this gender disparity. So, if our ability to correctly perceive the gender of speakers and hosts is similar to our ability to perceive the gender of questions, this implies we are more likely to have underestimated the gender disparity. Nevertheless, we doubt that the occurrence of misgendering was high enough to have biased our conclusions.

Pronouns versus gender

Moreover, we acknowledge that gender identity and preferred pronouns are often interchangeable although there are subtle differences. Therefore, we used the post-congress survey to quantify this potential discrepancy. In the post-congress survey, we asked each person for their gender identity as well as their preferred pronoun(s). In 98.8% of cases, self-identified women preferred she/her pronouns and 97.2% of self-identified men preferred he/him pronouns. Self-identified non-binary attendees used he/them, she/them or they/them pronouns.

S1 Methods 2

Data collection sheets were digitised by a team of nine student assistants. Despite the training of observers, complex situations occurred that were not anticipated, or situations were interpreted differently by the observers that sampled the same session. Consequently, inconsistencies between two observers sometimes occurred, for example in records of the number of questions that were asked during a Q&A. Because inconsistencies in the number of questions asked in a session made it difficult if not impossible to match up the data collected by two observers in the same session, we manually resolved these inconsistencies based on the notes taken. To ensure this manual curation was reliable and did not introduce mistakes based on subjective interpretations by single people, two data curators assessed the inconsistencies independently. The most common reason for disagreements in the number of questions asked was due to one questioner asking multiple questions, which was noted down inconsistently by observers. We manually corrected for this by adding a note on whether a question was a follow-up question (defined as a question that was asked by the same person consecutively, i.e. without a question being asked by another person in between) and excluded these follow-up questions in our analyses.

In addition, there were certain sessions where collecting data was more difficult, for example when the room was very busy, making it difficult to estimate the audience size, or when the lighting in the room was suboptimal, making it difficult to estimate a person's age or infer their gender. We added an additional binomial parameter to each datapoint to indicate whether there was any kind of uncertainty in the data collected, based on the notes taken during that session/talk/question. This allowed us to implement a conservative analytical approach in which we compared the results of models that included and excluded these ‘unreliable’ data that included potential biases. The dataset excluding unreliable data of any kind is hereafter referred to as the “conservative dataset”.

S1 Methods 3

In 32 out of 67 sessions, multiple observers collected data on question-asking behaviour to quantify the reliability of our observations and consequently, the credibility of our data. To evaluate inter-observer reliability (IOR), we calculated the unweighted Cohen's kappa (87) for nominal variables and the intraclass correlation coefficient (ICC) for numeric variables using a two-way agreement model implemented in the R package irr v.0.84.1 (88). The sessions that were double-sampled and took place in large lecture rooms were sampled by four observers, with the role of counting the number of men and people in total that raised their hands to ask a question being split between a pair of two observers. Thus, in double-sampled sessions in large lecture rooms, all of the variables other than the number of hands raised were recorded by two pairs of observers rather than a single pair. Because the IOR statistic is calculated for a given number of observers and we aimed to calculate this statistic across all sessions regardless of room size, we treated the two pairs in large lecture rooms as two independent double-samples and thus only tested for the agreement within pairs and not between pairs.

A Cohen's kappa value between 0.40 and 0.60 is interpreted as "moderate" agreement, a value between 0.61 and 0.80 is interpreted as "substantial" agreement, and a value over 0.80 is interpreted as "near perfect" (87). An ICC between 0.50 and 0.75 indicates "moderate" reliability, whereas a value between 0.75 and 0.90 indicates "good" reliability, and above 0.90 "excellent" (89). Observers had an "almost perfect" agreement on gender (host gender Cohen's kappa = 0.94, $p < 0.001$; speaker gender Cohen's kappa = 0.96, $p < 0.001$; questioner gender Cohen's kappa = 0.96, $p < 0.001$) and audience size (total audience size ICC = 0.96, $p < 0.001$; men in audience ICC 0.89, $p < 0.001$) and a "good" agreement on the duration of the Q&A (ICC = 0.83, $p < 0.001$). There was a "good" agreement on the number of hands raised in total (ICC = 0.77, $p < 0.001$) and by men only (ICC = 0.78, $p < 0.001$). However, observers had only a "substantial" agreement on host age (Cohen's kappa = 0.69, $p < 0.001$) and speaker age (Cohen's kappa = 0.64, $p < 0.001$), and only "moderate" agreement on the age of the questioner (questioner age Cohen's kappa = 0.40, $p < 0.001$).

S1 Methods 4

We combined the different observations of each parameter recorded in double-sampled sessions based on the conditions noted down in S2 Table 16. Due to the importance of gender and age for our analyses and the sensitivity of these data, we excluded any data points where the observers disagreed on these variables. If there was inconsistency on the noted number of hands raised by different observers, the most plausible explanation is that one of the two observers did not see one of the hands raised, and therefore we took the maximum number of raised hands. Similarly, if one observer noted down that a compliment was given to the speaker or that the speaker talked for longer than instructed, and the other observer did not, the most likely cause is that the other observer did not notice this or forgot to note it down. Lastly, interpreting a question as a challenge to the speaker might depend on the observer's expertise on the subject and/or conscious and unconscious bias in interpreting the questioner, which might lead to two observers interpreting the question differently. However, if one of the two observers interpreted the question as challenging, it is likely that at least part of the audience as well as the speaker also 'felt' this. Therefore, we applied the same logic as above and only one of the two observers had to note the question down as challenging for us to include this in the curated dataset.

S1 Results 1

In addition to identifying a gender disparity in asking questions, we asked if there were gender disparities in other aspects of the oral sessions that were related to the content of the question, waiting for your turn to ask a question, and accurately timing your talk. First, we found that older questioners were less likely to compliment the speaker (e.g. "Thank you for your interesting talk") compared to researchers estimated to be under 35 years old (estimate age category 35-50 = -0.61, $p = 0.001$; estimate age category > 50 = -0.49, $p = 0.052$). The probability that a person gave a compliment was highest at the start of the Q&A (estimate question number = -0.37, $p < 0.001$). Older questioners were also more likely to ask a critical question compared to questioners estimated to be under 35 years old (estimate age category 35-50 = 2.19, $p < 0.001$, estimate age category > 50 = 3.07, $p < 0.001$). Next, we found that jumping a question (i.e. asking a question without being chosen to do so) did not occur frequently (n = 18), but we did observe a non-significant but suggestive tendency for jumpers to be more likely male (estimate male questioner = 0.87, p = 0.10). Lastly, mid-career researchers were less likely to speak for longer than their allocated time slot (estimate mid-career = -0.57, $p < 0.001$), whereas late-career researchers were more likely to speak overtime (estimate late-career = 1.10, $p < 0.001$) compared to early-career researchers. The probability of a speaker receiving a compliment or a critical comment was not affected by speaker gender nor career stage (LRT p-value < 0.05).

S2 File. Supplementary Tables and Figures

S2 Table 1.

Model	**Data**	**# O**	**LRT χ^2**	**LRT *p***	**Reference level**	**Term**	**Estimate ± SE**	***z***	**Wald test *p***
QA (QA.1)	BD	350 (127 T)	N/A	N/A	N/A	IC	-0.66 ± 0.11	-6.07	< 0.001
QA **(QA.1c)**	BD – conser-vative	60 (124 T)	N/A	N/A	N/A	IC	-0.67 ± 0.11	-6.12	< 0.001
QA **(QA.1p)**	BD – plenary	342 (10 T)	N/A	N/A	N/A	IC	-1.54 ± 0.31	-4.95	< 0.001
QA	SD	373	5.96	0.05	Male	Female	-0.49 ± 0.24	-2.01	0.04
						Non-binary	0.92 ± 1.10	0.84	0.40
RH (QA.2)	BD	349 (127 T)	N/A	N/A	N/A	IC	-0.58 ± 0.11	-5.45	< 0.001
GC (QA.3)	BD	99 (67 T)	N/A	N/A	N/A	IC	-0.14 ± 0.23	-0.62	0.53
GC	SD	375	1.49	0.48	Male	Female	0.26 ± 0.33	0.78	0.44
						Non-binary	1.03 ± 0.88	1.17	0.24

S2 Table 2.

Motivation	**Gender effect on motivation**					**Motivation effect on probability of asking a question**				
	LRT χ^2	LRT FDR-*q*	Female estimate ± SE	*z*	Wald test *p*	LRT χ^2	LRT FDR-*q*	Estimate ± SE	*z*	Wald test *p*
Relevance own research	1.29	0.61	0.07 ± 0.23	0.31	0.75	0.03	0.91	-0.04 ± 0.23	-0.17	0.87
Making voice heard	5.95	0.13	-0.60 ± 0.46	-1.30	0.19	7.03	0.02	1.68 ± 0.76	2.20	0.02
Interest in topic	4.05	0.23	-0.53 ± 0.38	-1.39	0.17	10.32	0.00	1.07 ± 0.34	3.15	0.00
Deeper under-standing	0.84	0.69	-0.05 ± 0.26	-0.19	0.85	3.94	0.09	0.51 ± 0.26	1.98	0.08
Appreciate work	0.98	0.68	-0.24 ± 0.27	-0.91	0.37	2.25	0.23	0.43 ± 0.29	1.48	0.19

S2 Table 3.

Hesitation	Gender effect on hesitation					Hesitation effect on probability of asking a question				
	LRT χ^2	LRT FDR-*q*	Female estimate ± SE	*z*	Wald test *p*	LRT χ^2	LRT FDR-*q*	Estimate ± SE	*z*	Wald test *p*
Too introverted	4.04	0.23	0.52 ± 0.29	1.80	0.07	23.12	0.00	-1.27 ± 0.27	-4.72	< 0.001
Rather in private	5.37	0.16	0.30 ± 0.23	1.28	0.20	19.84	0.00	-1.04 ± 0.24	-4.38	< 0.001
Phrasing	11.19	0.02	0.90 ± 0.29	3.16	0.00	1.80	0.24	-0.34 ± 0.25	-1.35	0.18
Not clever	2.74	0.33	0.42 ± 0.26	1.61	0.11	5.32	0.04	-0.56 ± 0.24	-2.31	0.02
No time	3.78	0.20	-0.40 ± 0.25	-1.61	0.11	12.41	0.00	0.98 ± 0.29	3.38	< 0.001
No confidence	7.64	0.08	0.78 ± 0.31	2.53	0.01	6.80	0.02	-0.70 ± 0.27	-2.64	0.01
Misunderstand	0.36	0.83	0.04 ± 0.24	0.15	0.88	0.04	0.87	-0.05 ± 0.24	-0.21	0.84
Irrelevance/ unimportant	3.89	0.23	-0.26 ± 0.23	-1.14	0.26	0.85	0.44	0.22 ± 0.23	0.92	0.36
Intimidation setting	4.87	0.18	0.72 ± 0.47	1.53	0.13	0.18	0.76	-0.17 ± 0.40	-0.43	0.67
Intimidation audience	6.27	0.13	0.76 ± 0.33	2.31	0.02	7.33	0.02	-0.77 ± 0.28	-2.70	0.01

S2 Table 4.

	Motivation or hesitation	Career stage	Estimate ± SE	*z*	Wald test *p*
Motivations	Relevance own research	Mid-career	0.10 ± 0.23	0.42	0.67
		Late-career	-0.05 ± 0.33	-0.16	0.88
	Making voice heard	Mid-career	0.86 ± 0.51	1.67	0.09
		Late-career	0.7 ± 0.73	0.97	0.33
	Interest in topic	Mid-career	-0.53 ± 0.33	-1.61	0.11
		Late-career	1.73 ± 1.04	1.67	0.10
	Deeper understanding	Mid-career	0.21 ± 0.26	0.80	0.43
		Late-career	-0.45 ± 0.35	-1.27	0.20
	Appreciate work	Mid-career	0.14 ± 0.28	0.51	0.61
		Late-career	0.5 ± 0.38	1.30	0.19
Hesitations	Too introverted	Mid-career	-0.35 ± 0.27	-1.30	0.19
		Late-career	-0.96 ± 0.47	-2.04	0.04
	Rather in private	Mid-career	-0.05 ± 0.23	-0.21	0.84
		Late-career	-1.11 ± 0.38	-2.92	< 0.001
	Phrasing	Mid-career	-0.57 ± 0.25	-2.25	0.02
		Late-career	-1.59 ± 0.5	-3.16	< 0.001
	Not clever	Mid-career	-0.82 ± 0.25	-3.26	< 0.001
		Late-career	-1.79 ± 0.5	-3.58	< 0.001
	No time	Mid-career	0.55 ± 0.26	2.12	0.03
		Late-career	1.20 ± 0.35	3.41	< 0.001
	No confidence	Mid-career	-0.42 ± 0.27	-1.56	0.12
		Late-career	-2.91 ± 1.03	-2.84	0.01
	Misunderstand	Mid-career	-0.96 ± 0.24	-4.00	< 0.001
		Late-career	-1.50 ± 0.4	-3.72	0.00
	Irrelevance/un-important	Mid-career	-0.33 ± 0.23	-1.40	0.16
		Late-career	-0.71 ± 0.36	-1.96	0.05
	Intimidation setting	Mid-career	-0.97 ± 0.48	-2.02	0.04
		Late-career	0.29 ± 0.51	0.58	0.56
	Intimidation audience	Mid-career	-1.21 ± 0.32	-3.76	0.00
		Late-career	-1.2 ± 0.51	-2.37	0.02

S2 Table 5.

Variable	**LRT χ^2**	**LRT FDR-*q***	**Estimate ± SE**	***z***	**Wald test *p***
Female speaker (QA.1a)	0.00	0.98	-0.00 ± 0.20	-0.02	0.98
Proportion of audience that's female (QA.1b)	2.03	0.15	-1.74 ± 1.20	-1.44	0.15
Female host (QA.1c)	1.52	0.22	0.30 ± 0.24	1.21	0.23
Audience size (QA.1d)	0.06	0.81	-0.00 ± 0.00	-0.23	0.81
Small room (compared to large room) (QA.1e)	1.43	0.49	-0.15 ± 0.23	-0.65	0.51

S2 Table 6.

Response (Likert-scale)	LRT χ^2	LRT FDR-*q*	Level	Estimate ± SE	*t*	Wald test *p*
Audience is of own gender	41.06	< 0.001	Women	1.33 ± 0.22	6.05	< 0.001
			Non-binary	1.90 ± 0.66	2.88	0.004
Speaker is of own gender	36.30	< 0.001	Women	1.22 ± 0.23	5.40	< 0.001
			Non-binary	2.44 ± 0.68	3.58	< 0.001
Host is of own gender	19.64	< 0.001	Women	0.92 ± 0.22	4.11	< 0.001
			Non-binary	1.58 ± 0.67	2.34	0.02
Audience size is smaller	15.81	< 0.001	Women	0.79 ± 0.21	3.84	< 0.001
			Non-binary	-0.07 ± 0.67	-0.10	0.91

S2 Table 7.

Data	Model	#T, #Q	LRT χ^2	LRT *p*	Condition	Estimate ± SE	P	*z*	Wald test *p*
Unmanipulated, all Q minus Q1	Question-asking (QA.4.u)	96, 212	6.34	0.01	W first	-1.04 ± 0.19	0.26	-5.38	< 0.001
					M first	-0.33 ± 0.21	0.42	-1.57	0.12
	Raising hands (QA.5.u)	96, 209	4.90	0.03	W first	-0.90 ± 0.20	0.29	-4.62	< 0.001
					M first	-0.31 ± 0.22	0.42	-1.42	0.16
	Getting chosen (QA.6.u)	37, 51	0.11	0.74	W first	-0.13 ± 0.36	0.47	-0.36	0.72
					M first	-0.33 ± 0.47	0.42	-0.71	0.48
Manipulated, all Q minus Q1	Question-asking (QA.4.m)	90, 220	2.14	0.14	W first	-0.66 ± 0.20	0.34	-3.22	0.001
					M first	-0.25 ± 0.19	0.44	-1.34	0.18
	Raising hands (QA.5.m)	85, 204	1.32	0.25	W first	-0.92 ± 0.20	0.29	-4.51	< 0.001
					M first	-0.62 ± 0.18	0.35	-3.52	< 0.001
	Getting chosen (QA.6.m)	32, 49	0.01	0.91	W first	0.61 ± 0.45	0.65	1.37	0.17
					M first	0.68 ± 0.42	0.66	1.63	0.10

S2 Table 8.

Data	Model	# T, # Q	LRT χ^2	LRT p	Condition	Estimate ± SE	P	z	Wald test p
Unmanipulated, Q2	Question-asking (QA.4.u.2)	76, 76	5.68	0.02	W first	-1.30 ± 0.36	0.21	-3.59	< 0.001
					M first	-0.15 ± 0.35	0.46	-0.43	0.67
	Raising hands (QA.5.u.2)	75, 75	7.01	0.008	W first	-1.08 ± 0.23	0.25	-4.78	< 0.001
					M first	-0.14 ± 0.28	0.46	-0.52	0.60
	Getting chosen (QA.6.u.2)	26, 26	0.02	0.90	W first	-0.25 ± 0.56	0.44	-0.45	0.65
					M first	-0.36 ± 0.66	0.41	-0.55	0.58
Manipulated, Q2	Question - asking (QA.4.m.2)	75, 75 Model failed to converge	N/A	N/A	W first	N/A	N/A	N/A	N/A
					M first	N/A	N/A	N/A	N/A
	Raising hands (QA.5.m.2)	71, 71	2.21	0.14	W first	-0.93 ± 0.31	0.28	-3.00	0.003
					M first	-0.32 ± 0.29	0.42	-1.08	0.28
	Getting chosen (QA.6.m.2)	20, 20	2.33	0.13	W first	0.04 ± 0.69	0.51	0.05	0.96
					M first	1.62 ± 1.05	0.84	1.54	0.12

S2 Table 9.

	Variable	Univariate models		Final models			
		LRT χ^2	LRT *p*	Level	Estimate ± SE	*t*	Wald test *p*
Feeling heard (PCS.3)	Gender	4.38	0.11	N/A			
	LGBTQ+	3.57	0.06				
	Nationality	9.24	0.06				
	Affiliation	8.70	0.12				
	Expat	1.66	0.20				
	English comfort	14.38	< 0.001	N/A	0.28 ± 0.10	2.77	0.006
	Expertise	21.91	< 0.001	N/A	0.24 ± 0.06	3.85	< 0.001
Comfortable being myself (PCS.4)	Gender (relative to male)	13.30	0.001	Female	-0.48 ± 0.22	-2.14	0.03
				Non-binary	-2.26 ± 0.68	-3.35	0.001
	LGBTQ+	3.30	0.07	N/A			
	Nationality	2.01	0.74				
	Affiliation	6.17	0.29				
	Expat	0.01	0.95				
	English comfort	10.60	0.001	N/A	0.28 ± 0.11	2.56	0.01
	Expertise	17.77	< 0.001	N/A	0.22 ± 0.06	3.53	< 0.001
Sense of belonging (PCS.5)	Gender	4.48	0.11	N/A			
	LGBTQ+	0.82	0.37				
	Nationality	7.50	0.11				
	Affiliation (relative to Europe)	14.46	0.01	Asia	0.88 ± 0.52	1.71	0.09
				Africa	-1.02 ± 1.48	-0.69	0.49
				North America	1.16 ± 0.53	2.19	0.03
				Oceania	0.06 ± 0.52	0.12	0.90
				South America	15.76 ± 0.0	Inf	< 0.001
	Expat	0.52	0.47	N/A			
	English comfort	19.43	< 0.001	N/A	0.31 ± 0.10	3.03	0.003
	Expertise	45.30	< 0.001	N/A	0.35 ± 0.06	0.06	< 0.001

S2 Table 10.

	Variable	Univariate models		Final models			
		LRT χ^2	LRT p	Level	Estimate ± SE	t	Wald test p
Attendee diversity (PCS.6)	Gender (relative to male)	9.05	0.01	Female	-0.53 ± 0.21	-2.56	0.01
				Non-binary	-0.83 ± 0.68	-1.22	0.22
	LGBTQ+	6.95	0.01	LGBTQ+	-0.60 ± 0.28	-2.18	0.03
	Nationality	4.02	0.40	N/A			
	Affiliation	3.96	0.55				
	Expat	1.31	0.25				
	English comfort	0.77	0.38				
	Age	1.23	0.54				
EDI issues (PCS.7)	Gender (relative to male)	10.92	< 0.01	Female	0.48 ± 0.22	2.20	0.03
				Non-binary	0.24 ± 0.69	0.34	0.73
	LGBTQ+	10.40	0.001	LGBTQ+	0.73 ± 0.28	2.64	< 0.01
	Nationality (relative to Europe)	12.39	0.02	Asia	-0.34 ± 0.34	-0.98	0.33
				North America	0.77 ± 0.35	2.22	0.03
				Oceania	0.37 ± 0.69	0.54	0.59
				South America	1.27 ± 0.80	1.59	0.11
	Affiliation	6.78	0.24	N/A			
	Expat	8.88	< 0.01	Expat	0.55 ± 0.20	2.76	0.01
	English comfort	0.30	0.58	N/A			
	Age	2.52	0.28				
No QA gender disparity (PCS.8)	Gender (relative to male)	8.58	0.01	Female	-0.41 ± 0.22	-1.81	0.07
				Non-binary	-1.08 ± 0.70	-1.55	0.12
	LGBTQ+	7.60	< 0.01	LGBTQ+	-0.52 ± 0.29	-1.80	0.07
	Nationality (relative to Europe)	13.09	0.01	Asia	0.74 ± 0.45	1.64	0.10
				North America	0.43 ± 0.42	1.03	0.30
				Oceania	-0.26 ± 0.87	-0.30	0.77
				South America	2.64 ± 1.30	2.04	0.04
	Affiliation (relative to Europe)	15.32	< 0.01	Asia	0.58 ± 0.70	0.83	0.41
				Africa	1.74 ± 1.51	1.16	0.25
				North America	-0.45 ± 0.50	-0.91	0.37
				Oceania	0.53 ± 0.78	0.68	0.50
				South America	-5.39 ± 1.95	-2.76	0.01
	Expat	0.06	0.80	N/A			
	English comfort	5.80	0.01	N/A	-0.23 ± 0.11	-2.23	0.03
	Age	3.51	0.17	N/A			

S2 Table 11.

Category	Condensed code	Condensed code frequency	Expanded code	Expanded code frequency
Positive	Compliment	112		
	Organisation	85	well organised	76
			timekeeping in sessions	15
			problem solving by organisers	7
			venue	5
			good swag	4
			technical support	2
	Personal benefit	50	Personal benefit	48
			learnt a lot	6
			will return	6
	EDI aspects	48	focus on EDI	38
			transport pass	8
			childcare	5
			cost	5
			Trained Awareness Team	4
			signage	3
			grants	4
			quiet room	3
	Social aspects	38	good atmosphere	17
			good activities (social program)	16
			good participants	11
	Academic aspects	31	good topics / academic diversity	11
			good talks	10
			plenary talks	9
			good sessions	11
	Food	26		
	Sustainability	10		
Negative	Organisation	59	tight schedule / inadequate scheduling	25
			inadequate space in room	20
			long days/conference	19

			too many parallel sessions / talks	16
			inadequate communication	8
			inadequate tech	4
			missed printed program	3
Negative	Organisation	59	Problematic sponsor	1
	EDI aspects	57	inadequate provisions for accessibility	33
			lack of diversity	10
			high costs	9
			inadequate integration / networking of new/alone	4
			issues with travel / venue	3
			inaccessible conference materials	2
			personal pronouns not visible on badges	2
			Visa issues	2
			quiet room	2
			inadequate level of childcare	1
	Food	41		
	Undesirable interactions	12	disrespectful / sexist interactions	8
			unproductive mean questions	4
			intolerance to other ideas	2
	COVID	11	covid cases	9
			inadequate covid preventative measures	9
	Session management (chairs)	7		
	Academic aspects	5	inadequate academic rigour in talks	7
			homophobic ideas in talks	1
			ideological motivations	2
	Sustainability	4		
Suggestions	Organisation	38	alternative scheduling	21
			plan rooms according to expected audience	6
			better communication	5
			hybrid conference	4

			search function in abstracts	3
			better tech	2
	Food	18		
	DEI aspects	14	focus on DEI	8
Suggestions	DEI aspects	14	font/ options on nametag	3
			support for VISAs	2
			registration for part of the conference	1
	COVID	9		
	Sustainability	7	choice of swag	5
			sustainability	2
	Session management (chairs)	4		
	Social aspects	3	themed networking	3
	Academic aspects	1		

S2 Table 12.

Model name	Data subset	Research question	Model formula in lme4 syntax
QA.1	Unmanipulated oral sessions	Do women ask less questions than men do relative to the proportion of the audience who are women?	gender_questioner_female ~ 1 + (1|session_id / talk_id), offset = logit(audience_women_prop)
QA.1c	Conservative unmanipulated oral sessions		
QA.1p	Plenary sessions		gender_questioner_female ~ 1 + (1|plenary_id), offset= logit(registration_women_prop),
QA.1a-QA.1e	Unmanipulated oral sessions	What conditions can encourage women to ask questions? a) Gender of the speaker b) Gender proportion of the audience c) Gender of the session host d) Total size of audience e) Size of room	gender_questioner_female ~ condition + (1|session_id/talk_id), offset=logit(audience_women_prop)
QA.2		Do women raise their hands less often relative to the proportion of the audience who are women?	cbind(hands_women, hands_men) ~ 1 + (1|session_id/talk_id), offset = logit(audience_women_prop)
QA.3	Unmanipulated oral sessions where at least one woman and one man raised their hand	Do women get chosen less often than men relative to the proportion of people who raised their hand who are women?	gender_questioner_female ~ 1 + (1|talk_id), offset = logit(hands_prop_women)

S2 Table 13.

Model name	Data subset	Research question	Model formula in lme4 syntax
QA.4.u	Unmanipulated oral sessions minus question 1	Do women ask less questions than men do relative to the proportion of the audience who are women?	gender_questioner_female ~ - 1 + gender_first_questioner + (1|session_id / talk_id), offset = logit(audience_women_prop)
QA.4.u.2	Unmanipulated oral sessions only question 2		
QA.5.u	Unmanipulated oral sessions minus question 1	Do women raise their hands less often relative to the proportion of the audience who are women?	cbind(hands_women, hands_men) ~ - 1 + gender_first_questioner + (1|session_id/talk_id), offset = logit(audience_women_prop)
QA.5.u.2	Unmanipulated oral sessions only question 2		
QA.6.u	Unmanipulated oral sessions where at least one woman and one man raised their hand minus question 1	Do women get chosen less often than men relative to the proportion of people who raised their hand who are women?	gender_questioner_female ~ - 1 + gender_first_questioner + (1|talk_id), offset = logit(hands_prop_women)
QA.6.u.2	Unmanipulated oral sessions where at least one woman and one man raised their hand only question 2		
QA.4.m	Manipulated oral sessions minus question 1	Do women ask less questions than men do relative to the proportion of the audience who are women?	gender_questioner_female ~ - 1 + condition + (1|session_id / talk_id), offset = logit(audience_women_prop)
QA.4.m.2	Manipulated oral sessions only question 2		
QA.5.m	Manipulated oral sessions minus question 1		

QA.5.m.2	Manipulated oral sessions only question 2	Do women raise their hands less often relative to the proportion of the audience who are women?	cbind(hands_women, hands_men) ~ - 1 + condition + (1|session_id/talk_id), offset = logit(audience_women_prop)
QA.6.m	Manipulated oral sessions where at least one woman and one man raised their hand minus question 1	Do women get chosen less often than men relative to the proportion of people who raised their hand who are women?	gender_questioner_female ~ - 1 + condition + (1|talk_id), offset = logit(hands_prop_women)
QA.6.m.2	Manipulated oral sessions where at least one woman and one man raised their hand only question 2		

S2 Table 14.

Dependent variable	**Predictors**
Jumping a question	Question number, questioner gender, host gender, host age
Speaking longer than your allocated time	Speaker gender, speaker career stage
Giving a compliment	Question number, questioner gender, questioner age
Receiving a compliment	Speaker gender, speaker career stage
Asking a critical question	Questioner gender, questioner age
Receiving a critical question	Speaker gender, speaker career stage

S2 Table 15.

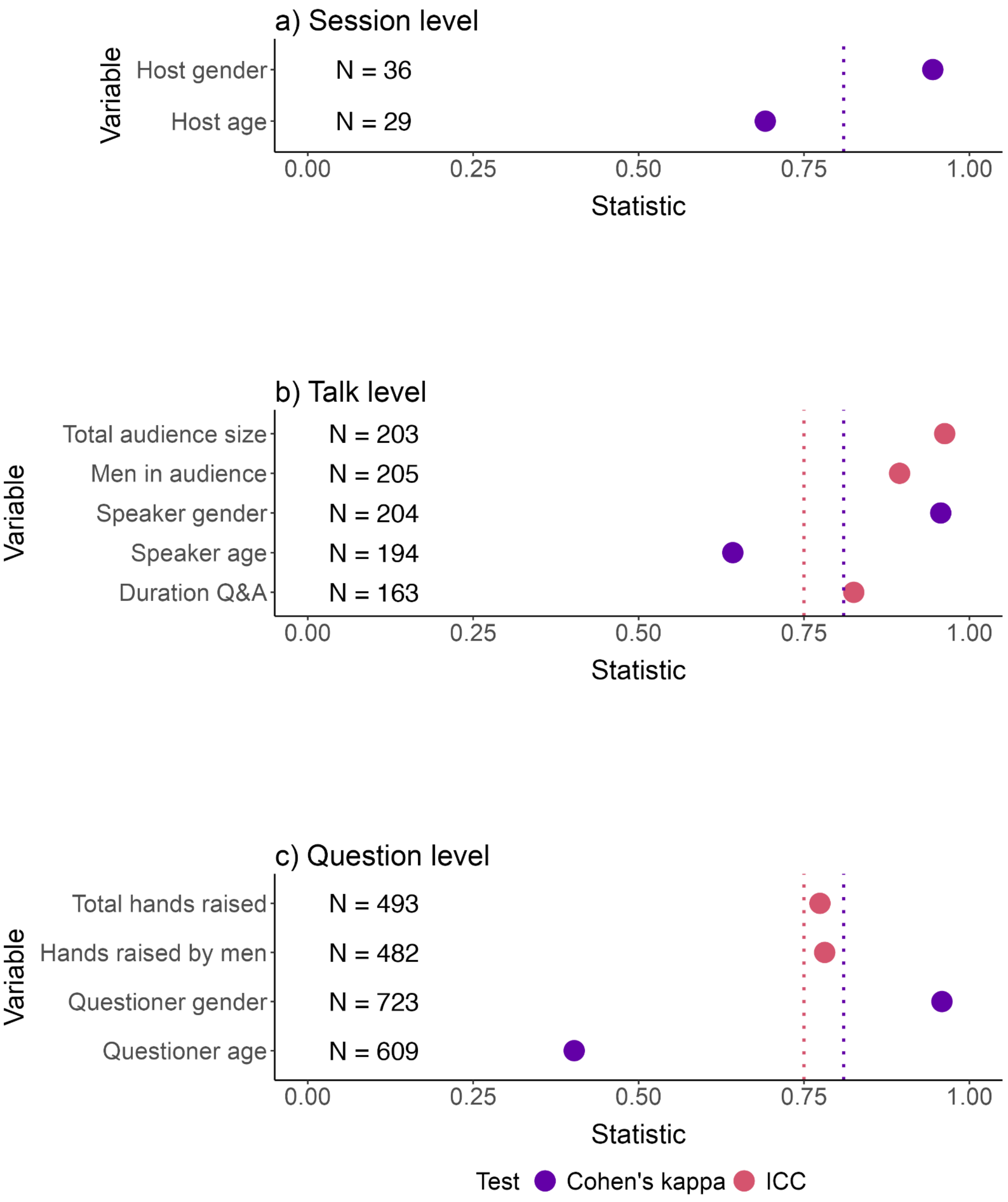
Model	**Question**	**Formula**
i) Gender effects on question asking motivation and hesitation		
Motivations (“mot_or_hes”): Relevance own research, Making voice heard, Interest in topic, Deeper understanding, Appreciate work **Hesitations (“mot_or_hes”)**: Too introverted, Rather in private, Phrasing, Not clever, No time, No confidence, Misunderstanding, Irrelevance/unimportant, Intimidation setting, Intimidation audience		
PCS.1	What motivations and hesitations are affected by gender?	mot_or_hes hesitation ~ gender + career
PCS.2	Which motivations and hesitations are predictors of whether a person asked a question at the congress or not?	ask_question ~ mot_or_hes + gender + career
ii) How do different social identities experience the conference?		
Social identities/controlling variables (“identity”): LGBTQ+, Nationality, Affiliation, Expat, English comfort, Expertise		
PCS.3	Which social identities/controlling variables were associated with the statement “felt heard during the conference”?	felt_heard ~ identity
PCS.4	Which social identities/controlling variables were associated with the statement “felt comfortable being myself during the conference”?	be_yourself ~ identity
PCS.5	Which social identities/controlling variables were associated with the statement “felt like I belong in my research field by attending the conference”?	social_belonging ~ identity
iii) Perception of equity, diversity and inclusivity among congress attendees		
Social identities/controlling variables (“identity”): Gender, LGBTQ+, Nationality, Affiliation, Expat, English comfort, Age		
PCS.6	Which social identities/controlling variables were associated with the statement “the conference attendees represented the diversity of researchers in our field”?	diversity ~ identity

PCS.7	Which social identities/controlling variables were associated with the statement “our research field experiences equity, diversity and inclusion related issues”?	edi_issue ~ identity
PCS.8	Which social identities/controlling variables were associated with the statement “the questions asked after the talks were equally divided across genders”?	no_disparity_qa ~ identity

S2 Table 16.

Variable	**How data were combined**
Gender questioner/speaker/host	If observers disagree, set to N/A
Age category questioner/speaker/host	If observers disagree, set to N/A
Audience size	Mean of the audience sizes, if the disagreement was high (SD > 20), put both audience sizes to N/A
Duration Q&A session	Mean of the durations
Number of hands raised	The maximum number of hands raised
Was a compliment given?	If one of the observers said yes, then yes
Did the speaker talk for longer than the allocated time slot?	If one of the observers said yes, then yes
Was the question type ‘challenging’?	If one of the observers said yes, then yes

S2 Figure 1.


a) Session level
Variable
Host gender
Host age
N = 36
N = 29
0.00
0.25
0.50
0.75
1.00
Statistic
b) Talk level
Variable
Total audience size
Men in audience
Speaker gender
Speaker age
Duration Q&A
N = 203
N = 205
N = 204
N = 194
N = 163
0.00
0.25
0.50
0.75
1.00
Statistic
c) Question level
Variable
Total hands raised
Hands raised by men
Questioner gender
Questioner age
N = 493
N = 482
N = 723
N = 609
0.00
0.25
0.50
0.75
1.00
Statistic
Test
Cohen's kappa
ICC

S3 File. Supplementary analysis on age

As previously noted (51,52,54), the gender disparity in question-asking could be explained by age-related effects. More specifically, if senior scientists ask more questions compared to junior scientists, and if there are more senior men present than senior women due to demographic inertia, we might observe that women ask less questions than men because of these age-related effects. In this analysis, we explore the potential for age-related effects to bias our interpretation of gender disparity in question asking.

Do senior scientists ask more questions compared to junior scientists?

First, we investigated whether senior scientists ask more questions than junior scientists. We built a binomial GLMM similar in structure to QA.1. In model QA.1, we use the gender of the questioner as the response variable, and correct for the gender proportion in the audience. In the current model which investigates age rather than gender, we use the seniority category (0 = junior, 1 = senior) as the response variable, and correct for the proportion of juniors in the audience, as well as the non-independence of talks within a session. Seniority category was based on the age category which was noted for each questioner (< 35, 35-50, >50) where a junior was defined as having a perceived age category of < 35 or 35-50, and a senior defined as having a perceived age category > 50. Moreover, we did not record the age category of audience members, and therefore base this proportion of juniors in the audience on the registration data. More specifically, we used the *offset* function to correct for the logit of the proportion of registrants who were junior (anything but "Professor" or "Associate Professor"). Note that this analysis therefore: 1) assumes that the distribution of juniors and seniors across all attendees was similar across all talks and 2) assigns a certain seniority category using two independent sources of information, of which one is collected through self-reports (career stage) and the other is perceived by observers (age category) which may not always correlate perfectly. We found that there was a trend for senior scientists being less likely to ask a question corrected for the number of senior scientists at the conference (estimate = -0.88, SE = 0.47, z-value = -1.87, *p* = 0.06).

Are there more female senior scientists than male senior scientists at the congress?

Second, we calculated the proportion of senior women who attended the congress based on collected data on career stage and pronouns during registration. We defined a female scientist as someone who uses she/her pronouns, a male scientist as someone who uses he/him pronouns, and we defined senior as someone with a "Professor" or "Associate Professor" title. Across the entire congress, 7.7% of attendees were female senior scientists, whereas 5.0% of attendees were male senior scientists. Across senior scientists only, 61% were female and 39% were male.

Is the gender disparity in question-asking dependent on seniority?

Third, we investigated the gender disparity in question asking separately for junior and senior researchers. We built a binomial GLM identical to QA.1 which uses the gender of the questioner as the response variable (1 = woman, 0 = man) and corrects for the proportion of women in the audience. However, in this analysis, we split up the dataset between juniors and seniors. To execute this separation, we used the data collected during registration to calculate the proportion of women who were junior (0.88) and senior (0.12), and the proportion of men who were junior (0.85) and senior (0.15). We used these proportions based on the registration data to adjust the observed proportion of women in the audience to what we assume was the observed proportion of junior women out of all juniors. To estimate the number of female juniors, we would multiply the proportion of junior women based on the registration data by the number of observed perceived women in the total audience. To estimate the number of all juniors (female and male), we would multiply the proportion of juniors based on the registration data by the number of observed people in the audience. Based on the estimated number of female juniors and number of juniors in total, we corrected for this proportion of female juniors in the model. We found that women ask less questions when subsetting the data only to include junior attendees (intercept = -0.67, *p* < 0.001) and observe an even higher gender disparity in the subsetted data that includes only senior attendees, although with marginal significance (intercept = -0.78, *p* = 0.06).

So, senior scientists are not more likely to ask questions compared to junior scientists. Further, there are more female senior scientists than male senior scientists present at the congress. Lastly, the gender disparity is still present when subsetting the data to only include junior scientists. These three lines of evidence therefore suggest that age effects and demographic inertia are unlikely to explain the gender disparity in question asking. However, these analyses were based on a number of assumptions: (i) age correlates with seniority, (ii) the distribution of seniority classes and genders was homogenous across oral sessions, (iii) observers can reliably estimate a questioner's age category. As we are not confident that any three of these assumptions are valid, we do not describe these models and their outputs in the main text.