



On the frequency, prevalence, and perceived severity of questionable research practices

Tove Larsson^{b,c,*}, Luke Plonsky^b, Scott Sterling^a, Merja Kytö^c, Katherine Yaw^d, Margaret Wood^b

^a Indiana State University

^b Northern Arizona University

^c Uppsala University

^d University of South Florida

ARTICLE INFO

Keywords:

Questionable Research Practices
Research ethics
Quantitative humanities research
Delphi method
Applied Linguistics

ABSTRACT

Questionable Research Practices (QRPs) fall in the gray zone between responsible research conduct and absolute misconduct (e.g., falsification and fabrication). Whereas other fields such as medicine have a long tradition of discussing and studying QRPs, there has been very limited focus on this topic in quantitative humanities research, and in applied linguistics specifically. Drawing on a community-generated list of quantitative humanities-specific QRPs, the present study investigates the self-reported frequency, prevalence, and perceived severity of QRPs among researchers in the US and Sweden. We also explored relationships between frequency of QRP engagement and researcher background factors, such as years since Ph.D. and publication rate. With regard to prevalence, the results showed that 96% of the respondents reported having used one or more of the practices listed. The most prevalent item was also the one that occurred with the highest frequency and the one that was reported as the least severe ('Presenting the same presentation at multiple conferences'). Overall, there was a strong negative correlation between frequency and severity ($\rho < -.77$) of QRPs, suggesting that it is uncommon for researchers to engage in an activity considered to be severe. With this exploratory study, we hope to contribute to an open and respectful discussion about QRPs.

1. Introduction

Ethical concerns arise throughout the entire research process (e.g., De Costa, 2016; Isbell et al., 2022; Sterling & Gass, 2017). That is, whether or not we address or discuss them explicitly, ethical decision points pertaining, for example, to data collection, study design, dissemination and application of results are omnipresent in our day-to-day work. While there is a broad consensus that researchers should not fabricate data or manipulate results, there is less agreement about how to approach ethical issues that fall in the gray zone between ideal behavior and absolute misconduct. Researchers may, for example, ask themselves to what extent it is ethical to be selective when deciding which results to report, or if it is fair to include their own students as study participants for course credit. Such practices are examples of potential Questionable Research Practices (QRPs; see Steneck, 2007), which are the focus of this paper.

QRPs are often viewed as the murky waters of research ethics, and as such, they are not particularly well understood. Fields like the medical sciences have a long tradition of discussing ethical concerns and QRPs (see, e.g., Allum et al., 2022; Gopalakrishna et al., 2022; Ljubenkovic et al., 2021; Tjldink et al., 2014). However, there is a dearth of studies in the context of quantitative humanities

* Corresponding author.

E-mail address: tove.larsson@nau.edu (T. Larsson).

research. In applied linguistics specifically, to the best of our knowledge, there is only one empirical study of QRPs, Isbell et al. (2022), although there has been some discussion of the topic of ethics in more general terms (see, e.g., Andringa & Godfroid, 2020; Yaw et al., in press). The lack of field-specific knowledge of QRPs is potentially problematic in that humanities research looks very different from medical research or research in other fields (Ravn & Sørensen, 2021). On the bright side, there is a seemingly growing interest in the field for (research) ethics in general (Yaw et al., in press), which suggests that the time is ripe for deepened discussions of this kind.

The present paper aims to contribute to that discussion by presenting results from an international, grant-funded project¹ designed to map out more of the territory of what constitutes a QRP in quantitative humanities research² in the US and Sweden, two countries that represent somewhat different research and publishing cultures (see Section 2). Specifically, we wish to work toward a better – and field-specific – understanding of the self-reported frequency, prevalence, and perceived severity of QRPs among active researchers with a Ph.D. We seek to answer the following research questions:

1. How widespread are QRPs?
2. What is the perceived severity of QRPs?
3. To what extent are researcher characteristics (e.g., years since Ph.D., country of residence) associated with the self-reported frequency of engagement in, and perceived severity of, QRPs?

2. Background

With a few notable exceptions (Allum et al., 2022; Isbell et al., 2022; Ravn & Sørensen, 2021), most of the research on QRPs has been conducted in fields and disciplines outside the humanities and often with a quantitative framework in mind. Indeed, when it comes to studies of QRPs, surveys have covered a wide range of disciplines including the medical sciences (Gopalakrishna et al., 2022; Ljubenkovic et al., 2021; Tijdink et al., 2014), biomedical sciences (Bouter et al., 2016), psychology (Bottesini et al., 2022; Fiedler & Schwartz, 2016; Fraser et al., 2018; John et al., 2012; Swift et al., 2022), communication (Bakker et al., 2021), economics (Makel et al., 2021), business (Hall & Martin, 2019), social sciences (Allum et al., 2022; Bouter et al., 2016; Ravn & Sørensen, 2021), natural sciences (Allum et al., 2022; Ravn & Sørensen, 2021), ecology (Fraser et al., 2018), and quantitative criminology (Chin et al., 2021).

Although different taxonomies and instruments were used across these studies, meaning that any comparisons of results should be made with caution, certain general findings have emerged. First, in terms of frequency, it generally seems that a majority (>50%) of researchers surveyed have reported engaging in at least one QRP from the list used in the respective study. In terms of the specific percentage, there is a relatively wide range. For example, while Isbell et al. (2022) and Artino et al. (2019) report frequencies over 90% (94% and 90.3%, respectively), Swift et al. (2022) report frequencies of 65% for faculty and 50% for students.

Second, when it comes to which QRPs are most prevalent, three main areas stand out in the literature: 1) QRPs relating to authorship (e.g., who is included as an author of a paper and in what order), 2) the handling of statistical (non-)significance, and, to a lesser degree, 3) HARKing (i.e., hypothesizing after results are known). Several studies found that QRPs related to authorship are particularly prevalent. In fact, of these studies, Allum et al. (2022), Artino et al. (2019), Braun & Roussos (2012), Hofmann et al. (2020), and Ljubenković et al. (2021) reported this as *the* most prevalent type of QRP. QRPs on the topic of data compilation and statistical significance also appear frequently on these lists. The most commonly occurring of these was collecting more data after determining whether or not results were statistically significant (e.g., Agnoli et al., 2017; Bakker et al., 2021; Fraser et al., 2018; John et al., 2012; Ljubenković et al., 2021; Swift et al., 2022). The range of admission to this particular QRP ranged from 11% (Swift et al., 2022) to 55% (Agnoli et al., 2017). Choosing not to report certain results because they were not statistically significant was also commonly listed (e.g., Bakker et al., 2021; Chin et al., 2021; Fraser et al., 2018; Isbell et al., 2022; Makel et al., 2021; Swift et al., 2022). The latter practice was the most prevalent QRP in Fraser et al. (2018) (64%), Makel et al. (2021) (61.69%), Rabelo et al. (2020) (54.7%), Chin et al. (2021) (43%), and Swift et al. (2022) (35% faculty, 25% students). Finally, HARKing was also reported with some regularity (e.g., Agnoli et al., 2017; Artino et al., 2019; Bakker et al., 2021; Chin et al., 2021; Fraser et al., 2018; Holm & Hofmann, 2018; Ljubenković et al., 2021; Makel et al., 2021; Rabelo et al., 2020; Swift et al., 2022), though this was infrequently the most common QRP that respondents reported to have engaged in. Frequency of engagement in this QRP ranged from 3% (Swift et al., 2022; for students) to 51% (Fraser et al., 2018).

While not a focus of the present study, several studies also reported that respondents' estimation of the prevalence of QRPs in general in a field is higher than their own self-reported frequency. For example, Fraser et al. (2018) showed that rates of 'expected prevalence' were substantially higher than reported self-use for half of the QRPs in their questionnaire (underreported covariates, rounding *p*-values, excluding data, undisclosed problems, fabrication). Rabelo et al. (2020) found that for eight of the ten QRPs on their list, the 'prevalence estimate' percentage was double the 'self-admission' percentage. When it comes to specific QRPs, several studies found that respondents perceived the prevalence of inappropriate authorship to be high in their field: Ljubenković et al. (2021) and Saberi-Karimian et al. (2018) both reported this as the QRP with the highest perceived frequency. It appears that for many scholars, QRPs are something that others do.

¹ *Questionable Research Practices: The (un)ethical handling of data in quantitative humanities research* (Larsson, Plonsky, Sterling, Kytö, Yaw, & Wood, 2021-2023). Co-funded by the Swedish Research Council, the Bank of Sweden Tercentenary Foundation, and the Royal Swedish Academy of Letters, History and Antiquities (Project ID: FOE20-0017).

² While we sought to represent researchers active in any subdiscipline of quantitative humanities in these countries, approximately half of our sample is comprised of researchers in (applied) linguistics, meaning that our results primarily represent this population.

For applied linguistics specifically, [Isbell et al. \(2022\)](#) found that the most common QRPs were ‘reporting p-values as inequalities (e.g., $p < .05$)’, ‘reporting effect sizes only for significant results’, and ‘not attempting to publish a study with nonsignificant results’, all of which the researchers reported doing in 1–20% of their studies on average. The least commonly occurring QRPs were ‘excluding findings contrary to own/colleagues’ previous research’ and ‘not reporting findings contrary to literature’.

Third, there seems to be general consensus that those who reported QRP engagement tended to view their behavior as defensible ([Agnoli et al., 2017](#); [Fraser et al., 2018](#); [John et al., 2012](#)). [Bakker et al. \(2021\)](#) also found a strong positive relation between the degree of perceived acceptability of a QRP and its prevalence. Certain QRPs were considered more acceptable than others. For example, in [Chin et al.’s \(2021\)](#) study, the researchers surveyed felt that ‘selectively choosing not to publish null findings’ and ‘looking at p-values before deciding whether to collect more data’ were acceptable in some situations (67% and 65%, respectively). The researchers surveyed in [Bottesini et al. \(2022\)](#) rated QRPs as either ‘acceptable’ or ‘indifferent’ for the following items: ‘p-hacking/cherry picking results’, ‘selective reporting of studies/file-drawer’, and ‘HARKing’.

Fourth, there seems to be at least somewhat of a relationship between researcher characteristics/background, such as their research context and career stage, and the prevalence and frequency of engagement in QRPs. The results for the relationship between QRPs and career stage were mixed. [Fraser et al. \(2018\)](#), [Makel et al. \(2021\)](#), and [Bakker et al. \(2021\)](#) all reported that career stage was not a strong predictor of how often researchers engaged in QRPs. However, some studies found that the rates were higher for junior researchers or early career researchers ([Gopalakrishna et al., 2022](#), [Isbell et al., 2022](#)), while others found the reverse: faculty generally reported QRPs at a higher rate than students ([Swift et al., 2022](#); see also [Xie et al., 2021](#), meta-analysis). When it comes to a possible relationship between number of publications and QRPs, [Swift et al. \(2022\)](#) found that the number of publications was not a predictor of the number of QRPs researchers engaged in, whereas [Isbell et al. \(2022\)](#) showed that for applied linguistics, the number of publications was positively correlated ($\rho \geq .10$) with five of the ten QRPs investigated.

Fifth, and finally, of the studies that made comparisons between researchers from different countries, there seems to be some consensus that US-based researchers report less frequent engagement in QRPs than researchers from other countries. For example, [Allum et al. \(2022\)](#) found higher admission of QRPs in Europe compared to the US; [Braun & Roussos \(2012\)](#) reported a higher rate for Europe and Latin America. The only study looking specifically at the Swedish context ([Holm & Hoffman, 2018](#)) noted significantly higher rates of scientific misconduct (including some QRPs) for Sweden compared to Norway. However, other studies found that geographic location was only a factor for certain QRPs. For example, [Agnoli et al. \(2017\)](#) reported very similar mean admission rates between the US and Italy overall (29.9% and 27.3%, respectively), though with a higher rate for US researchers on half of the QRPs. [Rabelo et al. \(2020\)](#) found that around half of the QRPs were reported at a lower rate in Brazil compared to the US and Italy. Furthermore, [Makel et al.’s results \(2021\)](#) showed that the non-US participants reported to have looked at results to form hypotheses and inform decisions (i.e., data peeking) slightly more frequently than the US participants (40% and 34%, respectively).

Differences among countries may have several explanations, but two that stand out in the context of the present study are procedural differences and differences related to publication requirements. At most US universities, there is an ethics review board (IRB) that requires all studies involving human subjects to be submitted for approval or exemption. Furthermore, the expectations at research-focused institutions (so-called R1 and R2 universities) are generally that their employees are active researchers, and it is not uncommon for promotion guidelines to include information on a required number of publications. In Sweden, by contrast, review boards are only recently beginning to expect submissions from researchers in the humanities, which may influence the frequency with which ethical issues are brought up to discussion in researcher training. In addition, while there are incentives for having an active research agenda in Sweden, external factors (e.g., heavy teaching loads and less focus on quantifiable output) may place less pressure on publishing. As (perceived or actual) pressure of this kind – from funders and/or other employers – could lead to questionable changes in design, methodology, or results ([Saber-Karimian et al., 2018](#)), we may expect differences across the two countries included in the present study. Overall, our literature review points to a multifaceted picture, with factors such as geographic context, career stage, and publication rate being potentially influential. QRPs have been reported to be both prevalent and frequent, with an inverse relationship between frequency and perceived severity. However, at a more general level, our review of the literature also pointed to an important issue, namely that there is no broad consensus on what constitutes a QRP – even within studies focusing specifically on the humanities: [Isbell et al. \(2022\)](#) included ten QRPs, [Ravn and Sørensen \(2021\)](#) looked at 15 QRPs, and [Allum et al. \(2022\)](#) included eight QRPs – from lists with very limited overlap. One of the goals of the present paper therefore is to evaluate a more comprehensive list of QRPs introduced in [Plonsky et al. \(in press\)](#) that was developed specifically for quantitative humanities research, as outlined in the next section.

3. Method

In this study, we took a descriptive and exploratory approach as much less is known about QRPs in the humanities than in other fields and disciplines. Additionally, few pre-existing materials on QRPs have been developed specifically for research in the humanities. As such, the current study used material developed in earlier stages of this project as will be discussed in [Section 3.2](#) (see [Plonsky et al., in press](#); [Sterling et al., 2023](#)).

3.1. Participants

The population of interest was researchers in the quantitative humanities in the US and Sweden. When identifying which disciplines would be considered under the umbrella of “humanities,” we took guidance from the National Endowment for the Humanities (2022) and the National Humanities Center (NHC, 2022). The disciplines we included were: Archaeology (including anthropology in cases

where archeology was not a separate department); Arts; Cultural studies (including area studies, women's and gender studies); Ethics; History; Language, classical; Language, modern; Law; Linguistics (including applied); Literature; Media (including communication, journalism, digital media); Philosophy; and Religion.

Once we identified the disciplines to be included in our sample, we developed a sampling frame for universities in the US and Sweden. For the Swedish context, we included all 15 higher educational institutions designated as *universitet* (i.e., research-intensive institutions), though one was later excluded because it did not have any humanities-related departments. For the US context, we used the 2021 *Carnegie Classifications list* (Indiana University, 2021) to sample from universities designated as *doctoral universities – very high research activity* ($n = 137$) and *doctoral universities – high research activity* ($n = 133$). These designations were chosen to access institutions known for high research output. We then randomly sampled 15 institutions from this list to match the size of the Swedish sample. Due to a low US response rate from the first 15 institutions, a second set of 15 was randomly sampled for a follow-up data collection, as explained below.

For each university, we identified the departments that corresponded to our discipline list and gathered publicly-available email addresses for the faculty in those departments. We had three (self-) inclusion criteria for participation in this project: respondents should 1) hold a doctoral or other terminal degree (e.g., PhD, EdD, JD, MFA, MBA); 2) conduct quantitative research (i.e., research with data collection resulting in numerical data); and 3) be currently affiliated with a university or institution in Sweden or the US; participants who did not meet these criteria were automatically screened out after providing demographic information.

3.2. Instrument

With the goal of developing a comprehensive list of QRPs specific to quantitative humanities research, we used the Delphi method (Hsu & Sandford, 2007; see Sterling et al., 2023, and Plonsky et al., in press, for a more complete account of this process). Somewhat simplified, the Delphi method is a bottom-up, qualitative elicitation technique developed to function as an asynchronous focus group. In our case, it was used to generate a list of QRP items through multiple iterations of ideas and feedback from an expert panel. In brief, a panel of ten experts in quantitative humanities research and/or ethics were asked to provide a list of possible QRP topics. The panelists were researchers who have a stated interest in ethical issues, researcher training, and research in quantitative humanities. We attempted to achieve broad geographical representation, with a particular focus on Sweden and the US, given the focus of the larger project. The research team subsequently read through all the responses, consolidated them, and revised them for consistency and clarity. The next step involved cross-checking the list from the panel with existing surveys (e.g., Isbell et al., 2022) and our own expertise as researchers, editors, and funding-seekers. This procedure resulted in a list of 62 items, where the vast majority of items (approximately 90%) came from the panelists. The subsequent steps involved seeking consensus on which items should remain on the list. Here, for each item, the panelists were asked to decide whether to 'accept', 'accept, with minor revisions', or 'require significant revisions or reject', with room for comments and suggestions for each item. Our cut-off point for consensus was, somewhat arbitrarily, 80%. After three rounds of revisions, the panel reached consensus on including 58 QRPs (many of which were substantially revised since the first round) and excluding the remaining items. This community-generated list of QRPs, each of agreed importance and written for the target audience, comprised the survey items for the current project (available on IRIS: <https://www.iris-database.org/details/s5Uye-pS8WF>, and on our website: <https://sites.google.com/view/qrp-humanities>).

For the current study, we designed a survey in Qualtrics that had three pages (following the informed consent page). The first page asked respondents to provide demographic information and information on their research background (e.g., years since Ph.D., number of quantitative courses taken, number of publications in the past five years). The second page asked respondents to only consider the research projects that they have worked on in the past five years, and indicate what proportion of these involved the activities listed in the 58 items. For each item, they were given six options: (1) never, (2) 1-20%, (3) 21-40%, (4) 41-60%, (5) 61-80%, (6) 81-100%. The third page showed the 58 items again, but this time the respondents were asked to rank how severe they considered these activities among the following answers: (1) completely acceptable, (2) mostly acceptable, (3) neither acceptable nor unacceptable, (4) mostly unacceptable, (5) completely unacceptable, and (6) prefer not to answer. In keeping the frequency page separate from the severity page (with no option of returning to the previous page), we attempted to mitigate the possibility that answers to one of these sets of questions would influence the responses to the other set. The respondents were able to leave optional comments on all items on pages 2 and 3. Only respondents who indicated that they had experience in grant funding (one of the sections) gave responses for that section.

It is worth pointing out that with this design, we (the research team) are not – nor do we consider ourselves to be – judges of what is considered a QRP. As our data will show, most scholars have likely engaged in QRPs, and we (the authors) are in no way different. By contrast, the Delphi model produced a community-generated list of potential QRPs, and this project asked the broader community to weigh in on the extent to which these practices are considered severe, with the assumption that the more severe the item, the more problematic it is.

3.3. Procedures

Survey data were collected in three waves over a period of seven months. Swedish sampling frame data were gathered in June 2022, and US sampling frame data were gathered from June to August, 2022. This yielded responses from 437 individuals affiliated with Swedish universities and 138 affiliated with US universities. Given the relatively low response rate on the US side, we decided to conduct a third round of data collection. For this round, we took two separate approaches. First, from October to November, 2022, we expanded our sample of US institutions by an additional 15, but instead of writing to individual researchers, we identified and

Table 1
Sample characteristics.

	n	M	SD	95% CI	Median	IQR
Country (current institution)						
All	230					
Sweden	139					
US	91					
Country of completed Ph.D.						
Sweden	110					
US	89					
Other	31					
Years since Ph.D.						
All	230	17.38	12.17	[15.81, 18.96]	14	17
Sweden	139	15.45	10.48	[13.70, 17.19]	13	14
US	91	20.34	12.93	[17.48, 23.20]	16	19.5
Pub. per year						
All	230	2.27	1.83	[2.04, 2.51]	2	2
Sweden	139	2.30	1.92	[1.98, 2.62]	2	2
US	91	2.23	1.69	[1.88, 2.58]	2	2
No. of quant courses						
All	230	2.33	2.55	[2.00, 2.66]	2	2
Sweden	139	1.85	2.07	[1.51, 2.19]	1	3
US	91	3.07	3.03	[2.45, 3.68]	2	3
What kind of data/analysis*						
All	230				3	2
Sweden	139				3	2
US	91				2	1

* 1 = all quantitative, 2 = mostly quantitative, 3 = equal parts quantitative and qualitative, 4 = mostly qualitative.

wrote directly to department chairs and program directors of the relevant departments, asking them to forward the information. This strategy was not very successful, and only yielded one additional response. Second, we used social media and language-related listservs to reach additional scholars who may meet our inclusion criteria.

From the total number of responses from all three waves, 296 were screened out for not meeting inclusion criteria (primarily for not doing quantitative research) and a further 150 provided only demographic information and no responses to the survey itself. There were also 107 who were outside of the target geographic range and whose data were not included for the current study. This left a total sample of 230 (139 Swedish responses and 91 US responses). A breakdown of the sample characteristics can be found in Table 1. Some participants did not answer all the frequency and severity questions; for transparency, the *n* size for each item is included in the tables reporting frequency and severity.

Finally, as all our data are self-reported, we have to rely on the participants to provide accurate responses to our questions. In making it clear that all responses are anonymous, we hope to have encouraged honest responses. Nonetheless, whether intentionally or not, it may be the case that the participants have downplayed the frequency with which they have engaged in a certain activity (i.e., social desirability bias).

4. Results and discussion

We present the results for each of the three research questions in turn: the question of how widespread the QRPs are in Section 4.1, their perceived severity in Section 4.2, and the possible association between researcher characteristics and the frequency and severity of QRPs in Section 4.3. Given our exploratory, descriptive design, and following Isbell et al. (2022) and most other studies of QRPs, we use descriptive statistics (e.g., means, standard deviations) to report on frequency and severity, and Spearman's ρ^3 to assess the strength of the correlations among researcher characteristics and frequency.

4.1. How widespread are QRPs?

The question of how widespread these QRPs are will be approached from two slightly different, but complementary perspectives: prevalence and frequency. We define prevalence as the percent of non-zero responses, meaning that it enables us to answer the question of what proportion of our sample has reported that they have carried out these activities in at least 1% of their research in the past five years.

³ As we did not meet the normality assumption for Pearson's correlation coefficient, we used the non-parametric counterpart, Spearman's ρ , to measure the strength and direction of association between two ranked variables. It ranges from -1 to 1, where values closer to -1 indicate a strong negative correlation, values closer to 0 indicate a weak correlation, and values closer to 1 indicate a strong positive correlation.

Table 2
The ten items that were reported with the highest frequency in all the data.

Item category and code	Item	n	M	SD	95% CI	Median	IQR
FRQ_WRIT_13	Presenting the same presentation at multiple conferences	217	2.06	1.18	[1.90, 2.21]	2	1
FRQ_DSGN_3	Defaulting to convention (e.g., choosing a design or instrument type because it is used in previous research, without making sure that it is the most appropriate design or instrument for the target relationships and/or constructs)	220	2.00	1.33	[1.82, 2.18]	2	1
FRQ_DSGN_2	Choosing a design and/or instrument type that provides comparatively easy or convenient access to data instead of one that has a strong validity argument behind it	220	1.92	1.26	[1.76, 2.09]	2	1
FRQ_DATA_5	HARKing (i.e., hypothesizing after results are known)	214	1.73	1.13	[1.58, 1.88]	1	1
FRQ_DSGN_1	Selecting variables out of convenience and/or familiarity when more theoretically grounded variables are available	221	1.65	1.02	[1.51, 1.78]	1	1
FRQ_DSGN_4	Employing instruments/measures without a strong validity argument	220	1.63	1.06	[1.49, 1.77]	1	1
FRQ_WRIT_6	Not sharing data when allowable	217	1.63	1.26	[1.46, 1.80]	1	1
FRQ_WRIT_4	Not reporting or publishing results because they are not statistically significant (i.e., the 'file drawer' issue)	220	1.63	1.10	[1.48, 1.77]	1	1
FRQ_WRIT_11	Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)	218	1.62	0.99	[1.49, 1.76]	1	1
FRQ_WRIT_7	Not sharing scripts used to analyze the results	213	1.62	1.31	[1.44, 1.79]	1	1

The results show that 96% (220/230) of the respondents reported engaging in one or more of the QRPs on the list. This is slightly higher than the 94% that [Isbell et al. \(2022\)](#) reported for applied linguistics specifically. It is perhaps not very surprising that we should note somewhat higher prevalence, given the fact that the length of our list of QRPs is almost six times longer than theirs. In more detail, no single item had a mean of zero, and the overall mean was 22%, meaning that for any given item, on average, a little over a fifth of the researchers in our sample reported to have done the activity listed (22% for the US; 23% for Sweden). For individual items, there was a considerable range. The activity that had the lowest means was 'Not reporting a conflict of interest, financial or otherwise' (overall percentage: 1%; US: 0%; Sweden: 2%).

The activity with the highest means was 'Giving the same presentation at multiple conferences' (overall percentage: 65%; US: 71%; Sweden: 62%).

In terms of the four contexts/stages of the research process (funding, research design/data collection, data analysis, writing/dissemination), the overall percentage for the items grouped per context was lowest for funding (overall percentage: 16%; US: 12%; Sweden: 18%), meaning that of those who reported having experience in this area, a smaller proportion reported having carried out the activities from this category in the past five years.

Moving forward, and due to the wealth of data that come from having 58 items to report on, we had to make some decisions for the present article in the interest of space and minimizing redundancy: First, although not inexistent, the results showed that the differences between the Swedish and US sample were very minor overall. Therefore, we will henceforth merge the results from the two geographical contexts, and only comment on them where relevant. Second, as the rank order of the items with regard to mean prevalence proved to map onto to the rank order of the items for mean frequency very well, we will now move on to the results from frequency analysis and provide more detailed results for the most and least frequent items. For issues of space, we chose to focus on the ten most and least frequent items ([Tables 2](#) and [3](#), respectively); the full list with the results for all the items can be found in the Supplementary materials. The tables list the code/category for the item (to enable cross-reference with the tables in the Supplementary materials), the item itself, the number of respondents who responded to it, the mean, standard deviation, 95% confidence interval, the median, and inter-quartile range.

As explained in [Section 3.2](#), the survey was set up such that a 1 corresponds to "never" and a 2 corresponds to "1-20% of the projects". As shown in [Table 2](#), among the most frequently occurring items, only three met the threshold for 2 for the median, meaning that roughly half the sample engaged in a QRP at least 1-20% of the time. The two items that were reported to occur with the highest frequency were "Presenting the same presentation at multiple conferences" and "Defaulting to convention (e.g., choosing a design or instrument type because it is used in previous research, without making sure that it is the most appropriate design or instrument for the target relationships and/or constructs)".

In terms of the contexts/stages of the research process for the most frequently occurring items, writing/dissemination is the category that is most well represented, with five items (ranks 1, 7, 8, 9, 10), followed by research design and data collection with four items (ranks 2, 3, 5, 6) and data analysis with one (rank 4). Only one category is not represented on this list: grant funding.

[Table 3](#) provides an overview of the ten least frequently occurring items. All these items have a mean frequency that is very close to 1 (= never). The two least frequent items on average are 'Using unjustified methods of handling missing data (e.g., imputing / inserting values for missing data that are not justified and/or that are more likely to yield desired outcomes)' and 'Not reporting

Table 3
The ten items that were reported with the lowest frequency in all the data.

Item category and code	Item	n	M	SD	95% CI	Median	IQR
FRQ_DATA_10	Using unjustified methods of handling missing data (e.g., imputing / inserting values for missing data that are not justified and/or that are more likely to yield desired outcomes)	215	1.04	0.39	[0.99, 1.09]	1	0
FRQ_FUND_3	Not reporting a conflict of interest (financial or otherwise)	159	1.04	0.56	[0.95, 1.12]	1	0
FRQ_FUND_11	Not disclosing impacts that funder directly had on research decisions (e.g., using particular datasets, selection of published outcomes)	161	1.07	0.57	[0.98, 1.16]	1	0
FRQ_WRIT_19	Not giving research assistants due credit in publications	212	1.07	0.45	[1.01, 1.13]	1	0
FRQ_WRIT_17	Inappropriately attributing author roles when listed in publications	214	1.07	0.46	[1.01, 1.14]	1	0
FRQ_DATA_2	Using unjustified methods of handling outliers	217	1.08	0.41	[1.02, 1.13]	1	0
FRQ_WRIT_10	Presenting misleading figures of data (e.g., displaying a truncated entire y-axis)	214	1.08	0.41	[1.03, 1.14]	1	0
FRQ_FUND_4	Misrepresenting researcher qualification/experience in the proposal	160	1.09	0.59	[1.00, 1.19]	1	0
FRQ_DSGN_11	Recruiting participants to join a study in a way that makes refusal difficult or uncomfortable	220	1.10	0.43	[1.04, 1.15]	1	0
FRQ_DATA_1	Removing whole items/cases knowingly / purposefully to obtain favorable results	217	1.11	0.49	[1.05, 1.18]	1	0

a conflict of interest (financial or otherwise)' (see [Gonulal, 2019](#), for evidence of the prevalence of less-than-optimal missing data handling in second language research).

In terms of contexts/stages of the research process, three of the ten items deal with data analysis (ranks 1, 6, 10), three with grant funding (ranks 2, 3, 7), three with write-up/dissemination (ranks 4, 5, 7), and only one with research design and data collection (rank 8). Based on the combined results from these tables, we can conclude that none of the items are particularly frequent, and that items relating to grant funding are comparatively infrequent on average in that they did not occur among the most frequent items and are well represented in the list of the least frequent items, as was also clear from the results from the prevalence analysis above.

When we compare our results to those of previous work, we see that like previous studies (e.g., [Fraser et al., 2018](#); [Holm & Hofmann, 2018](#), [Rabelo et al., 2020](#)), HARKing is a commonly occurring QRP. In addition, the item 'Not reporting or publishing results because they are not statistically significant' was among our most frequent activities, similar to findings from [Fraser et al. \(2018\)](#), [Isbell et al. \(2022\)](#), and [Makel et al. \(2021\)](#). However, we do not see items relating to authorship among our ten most frequent activities, unlike studies such as [Allum et al. \(2022\)](#), [Artino et al. \(2019\)](#), and [Hofmann et al. \(2020\)](#), who all reported this as the most frequent category. One possible reason that we see this less in our data might be that researchers in the quantitative humanities tend to work on smaller research teams than researchers in other fields such as medicine; with fewer people involved, authorship may not be as thorny an issue. We will now turn to our second research question that looks at the perceived severity of the items.

4.2. What is the perceived severity of QRPs?

As outlined in [Section 3.3](#), the researchers were asked to rate the severity of the activities on a scale from 1 to 5 (1 = completely acceptable, 2 = mostly acceptable, 3 = neither acceptable nor unacceptable, 4 = mostly unacceptable, 5 = completely unacceptable). As mentioned above, due to the fact that there were so few and minor differences, we have chosen to merge the results for the US and Sweden, with the promise that we would comment on them where relevant. Although the differences are still not large, we note that the overall mean for perceived severity is slightly higher in the US sample (mean = 4.27; SD = 0.78) than in the Swedish sample (mean = 4.08; SD = 0.89). Furthermore, there was one item in particular that stood out: 'Using funds for a purpose other than what was stated in the proposal (e.g., for a research assistant instead of for participant-related expenses)'. The mean for the US researchers was 4.16 (SD: 1.06) compared to 3.44 (SD: 1.22) for the researchers who are active in Sweden. The remaining items had more similar scores, and we will therefore present the results together in this section, focusing on the ten items that were considered most or least severe, on average. The ten most severe items are shown in [Table 4](#), and the ten least severe in [Table 5](#).

As can be seen, there is a great deal of overlap between these tables and the frequency tables in [Section 4.1](#) above, such that the less frequent items are perceived as less severe and vice-versa. This suggests that researchers seem to show a tendency to avoid practices that they consider severe. In fact, when we computed the mean frequency and severity scores for each QRP and then correlated these scores, we saw a strong negative correlation (Spearman's $\rho = -.77$) between overall frequency and perceived severity, in line with findings from previous studies (e.g., [Bakker et al., 2021](#)).

Returning to the ten practices that are considered the most/least severe, we can see that the ten most severe practices all are rated as >4.5 on average, meaning that their mean is closer to 5 (= completely unacceptable) than to 4 (= somewhat unacceptable). In the quantitative humanities, we thus seem to have strong feelings about the inappropriateness of these. The three practices

Table 4
The ten items reported as most severe.

Item category and code	Item	n	M	SD	95% CI	Median	IQR
SEV_FUND_3	Not reporting a conflict of interest (financial or otherwise)	179	4.83	2.06	[4.53, 5.13]	5	0
SEV_DATA_1	Removing whole items/cases knowingly / purposefully to obtain favorable results	187	4.74	1.97	[4.46, 5.02]	5	0
SEV_DATA_10	Using unjustified methods of handling missing data (e.g., imputing / inserting values for missing data that are not justified and/or that are more likely to yield desired outcomes)	182	4.73	1.99	[4.44, 5.02]	5	1
SEV_FUND_4	Misrepresenting researcher qualification/experience in the proposal	178	4.71	2.05	[4.41, 5.02]	5	1
SEV_DSGN_11	Recruiting participants to join a study in a way that makes refusal difficult or uncomfortable	189	4.68	1.95	[4.41, 4.96]	5	0
SEV_WRIT_17	Inappropriately attributing author roles when listed in publications	188	4.67	1.94	[4.39, 4.95]	5	1
SEV_DATA_2	Using unjustified methods of handling outliers	186	4.65	1.97	[4.37, 4.93]	5	1
SEV_WRIT_10	Presenting misleading figures of data (e.g., displaying a truncated entire y-axis)	186	4.65	1.94	[4.37, 4.93]	5	1
SEV_DSGN_6	Biasing the design/instrument so that outcomes are favorable to researcher beliefs (e.g., choosing a design/instrument that will likely lead to similar outcomes as previous research)	188	4.62	1.94	[4.34, 4.89]	5	1
SEV_DATA_14	Interpreting statistical results inappropriately (e.g., claiming equivalence between groups based on a non-statistically significant difference; undue extrapolation)	185	4.62	1.94	[4.34, 4.90]	5	1

Table 5
The ten items reported as least severe.

Item category and code	Item	n	M	SD	95% CI	Median	IQR
SEV_WRIT_13	Presenting the same presentation at multiple conferences	189	2.82	1.57	[2.60, 3.04]	3	2
SEV_DSGN_3	Defaulting to convention (e.g., choosing a design or instrument type because it is used in previous research, without making sure that it is the most appropriate design or instrument for the target relationships and/or constructs)	188	3.41	1.70	[3.17, 3.65]	4	2
SEV_WRIT_11	Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)	186	3.45	1.70	[3.21, 3.70]	3	1
SEV_WRIT_9	Not attempting to publish results in a timely manner	183	3.47	1.74	[3.22, 3.72]	3	1
SEV_DSGN_2	Choosing a design and/or instrument type that provides comparatively easy or convenient access to data instead of one that has a strong validity argument behind it	190	3.51	1.68	[3.27, 3.75]	4	1
SEV_WRIT_4	Not reporting or publishing results because they are not statistically significant (i.e., the ‘file drawer’ issue)	183	3.60	1.79	[3.34, 3.85]	4	1
SEV_DSGN_1	Selecting variables out of convenience and/or familiarity when more theoretically grounded variables are available	190	3.74	1.73	[3.49, 3.98]	4	1
SEV_DATA_5	HARKing (i.e., hypothesizing after results are known)	182	3.81	1.91	[3.54, 4.09]	4	2
SEV_WRIT_6	Not sharing data when allowable	185	3.82	1.82	[3.56, 4.08]	4	2
SEV_WRIT_7	Not sharing scripts used to analyze the results	185	3.84	1.85	[3.57, 4.10]	4	2

that had the highest means were ‘Not reporting a conflict of interest (financial or otherwise)’, ‘Removing whole items/cases knowingly/purposefully to obtain favorable results’, and ‘Using unjustified methods of handling missing data (e.g., imputing / inserting values for missing data that are not justified and/or that are more likely to yield desired outcomes)’.

For several of the practices found on the list of the ten least severe ones, previous studies have reported similar findings. For example, the corresponding item to our ‘Not reporting or publishing results because they are not statistically significant (i.e., the ‘file drawer’ issue)’ and ‘HARKing’ were both deemed as ‘acceptable’ or ‘indifferent’ in [Bottesini et al.’s \(2022\)](#) study, and the former was listed as acceptable in some situations in [Chin et al.’s \(2021\)](#) study.

It is also relevant to mention that while the list of survey items was created through a rigorous process, we remain open to the possibility that the broader research community did/does not agree that these particular items were considered QRPs or that the

Table 6
Relations (Spearman's ρ) between frequency and researcher background.

Item	Years since Ph.D. [95% CI]	No. of quant. courses [95% CI]	Num. of pubs per year [95% CI]	Data type (quant./qual) [95% CI]
Defaulting to convention (e.g., choosing a design or instrument type because it is used in previous research, without making sure that it is the most appropriate design or instrument for the target relationships and/or constructs)	-0.10 [-.22, .04]	0.02 [-.12, .15]	0.07 [-.06, .20]	-0.17 [-.29, -.04]
Ignoring alternate explanations of data	-0.11 [-.24, .04]	0.05 [-.07, .17]	0.19 [.06, .32]	-0.10 [-.23, .03]
Using too many statistical tests without correction (e.g., Bonferroni)	-0.07 [-.21, .05]	0.14 [.02, .27]	0.11 [-.03, .24]	-0.33 [-.43, -.21]
Using incorrect statistical methods (e.g., tests that are not appropriate for the type of data being analyzed)	-0.09 [-.25, .03]	0.06 [-.08, .19]	0.07 [-.08, .21]	-0.24 [-.35, -.11]
Not providing sufficient description of the data and the results (e.g., exact p-values, SD)	-0.05 [-.18, .08]	-0.04 [-.18, .10]	0.08 [-.05, .21]	-0.16 [-.28, -.03]
Not sharing data when allowable	-0.15 [-.27, -.02]	0.13 [.01, .25]	0.09 [-.04, .22]	-0.16 [-.28, -.02]
Not sharing instruments/coding schemes	-0.16 [-.27, -.02]	0.10 [-.03, .24]	0.09 [-.04, .21]	-0.10 [-.23, .04]
Employing excessive self-citation	0.07 [-.07, .20]	-0.11 [-.24, .03]	0.20 [.08, .33]	0.05 [-.09, .18]
Inappropriately attributing author roles when listed in publications	0.07 [-.07, .20]	-0.03 [-.21, .14]	0.29 [.17, .40]	-0.13 [-.26, .00]
Not giving research assistants due credit in publications	0.03 [-.08, .12]	0.10 [-.04, .23]	0.21 [.09, .31]	-0.15 [-.25, -.04]
Irresponsibly co-authoring (e.g., not being involved enough to be able to verify accuracy of analysis)	0.06 [-.10, .19]	0.14 [.00, .27]	0.23 [.10, .36]	-0.16 [-.28, -.04]
Misrepresenting researcher qualification/experience in the proposal	-0.11 [-.24, .02]	0.05 [-.12, .21]	0.18 [.03, .32]	0.02 [-.13, .16]
Misrepresenting study importance in the proposal (e.g., exaggeration of impact and value of proposal to society)	-0.10 [-.25, .07]	-0.04 [-.21, .12]	0.15 [-.02, .30]	-0.11 [-.26, -.04]
Misrepresenting literature in the proposal (e.g., over-emphasizing previous research that supports the proposal and/or ignoring conflicting evidence)	-0.16 [-.30, -.01]	-0.01 [-.17, .16]	0.15 [-.01, .29]	-0.09 [-.23, .06]

severity of the QRP would change depending on context. An overall average in the range between 1 and 2 (completely acceptable to mostly acceptable) may have been an indication of such an item. However, no item had an average in this range, and we thus draw the tentative conclusion that all items fall into the realm of potentially questionable in that none of them was considered (close to) completely acceptable. Nonetheless, as shown in Table 5, we did have four items that had a mean lower than 3.5 (i.e., closer to 'neither acceptable nor unacceptable' than to 'mostly unacceptable'), which may indicate that these items are not considered particularly questionable by the research community. These items are 'Presenting the same presentation at multiple conferences', 'Defaulting to convention (e.g., choosing a design or instrument type because it is used in previous research, without making sure that it is the most appropriate design or instrument for the target relationships and/or constructs)', 'Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)', and 'Not attempting to publish results in a timely manner'. However, that said, there are certainly arguments in favor of treating these as potentially questionable. For example, it could be argued that giving the same presentation at multiple conferences or salami publication may lead to unfair advantages on, say, the job market, if some candidates do so to boost their CVs whereas other candidates with comparable merits do not.

4.3. Are researcher characteristics associated with the frequency of engagement in, and perceived severity of, QRPs?

To answer our third research question, we look at the potential role of four researcher characteristics in relation to (a) the frequency and (b) perceived severity of the QRPs: years since Ph.D., number of quantitative courses taken, number of publications per year, and the extent to which they do quantitative vs. qualitative studies. Following Isbell et al. (2022) to be able to compare results wherever possible, we used Spearman's ρ to assess the strength of these relations. The full tables – Tables S3 and S4 – can be found in the Supplementary materials. In this section, we will focus on providing an overview of the strongest relations found.

Like Isbell et al. (2022), we used $\rho \geq .10$ as guidance for exploring correlations of potential interest. For the sake of conciseness given our long list of QRPs, Table 6 highlights QRPs that correlated with at least one researcher characteristic at $\rho \geq .15$, but we have presented all correlations $\geq .10$ in bold. The bootstrapped 95% confidence intervals (CIs) were calculated in R using the RVAide-

Memoire (Hervé, 2021) package. As can be seen, many of our weaker correlations have CIs that cross zero, meaning, as usual, that these findings should be interpreted with caution.

The first thing to note is that the correlations are relatively weak overall, as none of them reach Plonsky and Oswald's (2014) proposed benchmark of $|\rho| \geq .30$ for a moderate, or medium, correlation. Nonetheless, there are some clear trends in the data. For years since Ph.D., only three items had stronger correlations than $\rho \geq .15$: 'Not sharing data when allowable', 'Not sharing instruments/coding schemes', and 'Misrepresenting literature in the proposal (e.g., over-emphasizing previous research that supports the proposal and/or ignoring conflicting evidence)'. As the correlations are negative, it means that the longer it has been since the researchers defended their dissertations, the less likely they are to do these at a higher frequency. In fact, all the $\rho \geq .10$ correlations in this category are negative, again suggesting that recent Ph.D. holders report to having carried out these activities at higher frequency than researchers who have been in the field longer, in line with findings by Gopalakrishna et al. (2022) and Isbell et al. (2022).

There are several possible explanations to why that may be. One might be that researchers who have been in the field longer may be less likely to take the lead on projects (when working with graduate students, etc.) and may therefore be less likely to be faced with decisions that may lead to QRPs and/or be less aware of such decisions being made by the lead researchers on the team. Furthermore, recent Ph.D. holders are perhaps less likely to have data of their own, and so when they compile some, they may want to have a chance to publish on it before they share it with others, meaning that they may become more likely to share data and instruments the longer they have been in the field.

Another reason could be related to increasing pressure in the field to publish, which may lead to somewhat lower ethical standards, especially among more junior researchers who may need publications to secure a job. The fact that publication rate ('number of publications per year' in the table) is consistently positively correlated with frequency of these items may offer some support for this claim. However, one may also argue that researchers who are more prolific are more likely to encounter situations where QRPs may arise than researchers who publish less frequently. The items that have the strongest correlations (all positive) with this researcher characteristic all fall in the writing/dissemination category, and mostly pertain to authorship issues: 'Employing excessive self-citation', 'Inappropriately attributing author roles when listed in publications', 'Not giving research assistants due credit in publications', and 'Irresponsibly co-authoring (e.g., not being involved enough to be able to verify accuracy of analysis)'. Our results are, thus, similar to those of Isbell et al. (2022) who found that for applied linguistics, the number of publications was positively correlated ($\rho \geq .10$) with many of the QRPs investigated.

Out of the four researcher characteristics, the one that had the weakest correlations overall, was the number of quantitative courses taken. Surprisingly, all $\rho \geq .10$ correlations except one was positive, which somewhat counter-intuitively suggests that the more courses you take, the more likely you are to report higher frequencies for the items. The one exception was 'Employing excessive self-citation', which was negatively correlated with frequency. While, again, the correlations are weak, the fact that many of them are positive may suggest that there is more work to be done in methods courses to discuss research ethics and QRPs. In fact, a recent study (Wood et al., submitted) of textbooks and syllabi for methods courses in applied linguistics showed that the focus on research ethics is relatively rare and largely limited to IRB-issues that fall under the purview of IRBs.

The final characteristic, quantitative vs. qualitative focus, exhibits negative correlations for all $\rho \geq .10$ correlations, suggesting that the more of a qualitative focus you have in your research, the less likely you are to carry out one of the items at a higher frequency. This relation is particularly strong for 'Using incorrect statistical methods (e.g., tests that are not appropriate for the type of data being analyzed)'. This finding may be counter to what we might expect. While it may well be the case that more qualitatively oriented researchers are less likely to carry out QRPs, our research design does not preclude the possibility that the reason these researchers report lower frequencies is that they may not encounter these situations as frequently as those who exclusively do quantitative research; our instrument was designed for quantitative research, after all. It would be helpful in future studies if actual opportunity and frequency could be teased apart.

Turning to the relations between the researcher characteristics and perceived severity, we see a somewhat similar picture. The relations for 23 items are shown in Table 7; correlations $\rho \geq .10$ are bolded.

We see almost exclusively positive correlations between years since Ph.D. and the items included, meaning that the longer you have been in the field, the more serious you find these actions. This can be interpreted in two ways: Either you become more ethically conscious as your seniority increases, or the field itself has changed such that activities of this kind are increasingly considered more acceptable (that is, researchers who were 'brought up' academically many years ago have different views and expectations from more junior researchers). More qualitative and quantitative work would be needed to fully understand the situation, but the fact remains that the field looks very different now than it did 20, or even ten, years ago (Gass et al., 2021; Plonsky, 2014).

When it comes to research focus, with one exception – 'Cherry-picking data to analyze' – all the correlations were positive, meaning that the more qualitatively oriented researchers viewed the items as more severe on average than their more quantitatively oriented counterparts. If the assumption holds that you are less likely to carry out an activity you find to be severe (see the overall correlation from Section 4.2), then this finding coupled with the correlations for frequency above suggests that qualitatively oriented researchers actually differ somewhat from quantitatively oriented researchers, not just in terms of the frequency with which the groups are likely to encounter these situations.

For the $\rho \geq .10$ correlations for number of publications per year, 'HARKing' is the only positive one. The remaining three are all negative: 'Ignoring alternate explanations of data', 'Using too many statistical tests without correction (e.g., Bonferroni)', and 'Lifting short phrases from others without quoting directly'. While the correlations are weak, there is some tendency for researchers who publish more to consider these QRPs less severe than those who publish less.

The results for the number of quantitative courses taken is a mixed bag, with an almost perfect split between positive and negative $\rho \geq .10$ correlations. Focusing on the positive correlations, we can see that the results indicate that the higher number of courses

Table 7
Relations (Spearman's ρ) between perceived severity and researcher background.

Item	Years since Ph.D. [95% CI]	No. of quant. courses [95% CI]	Num. of pubs per year [95% CI]	Data type (quant./ qual.) [95% CI]
Selecting variables out of convenience and/or familiarity when more theoretically grounded variables are available	0.07 [-.07, .21]	0.10 [-.05, .24]	0.04 [-.10, .18]	0.18 [.04, .32]
Leaving out known/likely moderator variables or covariates from the study design without explanation or acknowledgment	0.10 [-.04, .24]	-0.13 [-.25, .01]	-0.08 [-.23, .06]	0.16 [.01, .31]
HARKing (i.e., hypothesizing after results are known)	0.00 [-.13, .13]	0.16 [.00, .30]	0.10 [-.06, .25]	0.00 [-.15, .16]
Cherry-picking data to analyze	0.06 [-.08, .21]	0.21 [.06, .34]	-0.02 [-.17, .13]	-0.11 [-.27, .04]
Ignoring alternate explanations of data	0.16 [.01, .30]	-0.10 [-.25, .28]	-0.12 [-.28, .02]	0.21 [.06, .36]
Using unjustified methods of handling missing data (e.g., imputing / inserting values for missing data that are not justified and/or that are more likely to yield desired outcomes)	0.16 [.01, .29]	-0.12 [-.25, .02]	-0.08 [-.24, .07]	0.16 [.03, .29]
Using too many statistical tests without correction (e.g., Bonferroni)	0.10 [-.05, .24]	-0.08 [-.23, .05]	-0.13 [-.28, .01]	0.23 [.09, .38]
Using incorrect statistical methods (e.g., tests that are not appropriate for the type of data being analyzed)	0.21 [.08, .35]	-0.04 [-.17, .10]	0.02 [-.13, .17]	0.18 [.04, .32]
Not providing sufficient description of the data analyses or other procedures	0.06 [-.09, .19]	0.15 [.01, .29]	-0.02 [-.17, .11]	0.07 [-.07, .21]
Not sharing data when allowable	0.17 [.01, .31]	-0.09 [-.23, .04]	-0.05 [-.20, .09]	0.20 [.06, .35]
Not sharing scripts used to analyze the results	0.18 [.04, .30]	-0.05 [-.19, .08]	-0.09 [-.22, .07]	0.14 [.00, .29]
Not sharing instruments/coding schemes	0.15 [.02, .28]	0.02 [-.10, .15]	-0.05 [-.20, .10]	0.09 [-.06, .23]
Not attempting to publish results in a timely manner	0.24 [.10, .36]	-0.11 [-.24, .02]	0.00 [-.15, .15]	0.24 [.09, .38]
Presenting misleading figures of data (e.g., displaying a truncated entire y-axis)	0.16 [.02, .29]	-0.01 [-.15, .12]	0.07 [-.08, .21]	0.09 [-.05, .23]
Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)	0.18 [.04, .30]	-0.01 [-.15, .13]	-0.01 [-.14, .13]	0.10 [-.05, .25]
Not managing time well for one's own conference presentations, resulting in less time for other presenters, limiting discussion, and impacting others at the conference	0.16 [.02, .30]	-0.01 [-.15, .13]	0.02 [-.12, .16]	0.12 [-.02, .27]
Intentionally omitting relevant work because it does not align with one's theoretical or methodological approach	0.18 [.05, .31]	0.05 [-.10, .19]	-0.01 [-.16, .14]	0.02 [-.12, .16]
Inappropriately including or excluding authors	0.18 [.04, .31]	-0.08 [-.22, .07]	0.01 [-.14, .16]	0.19 [.05, .33]
Not giving research assistants due credit in publications	0.19 [.05, .32]	-0.14 [-.27, -.01]	-0.08 [-.24, .08]	0.14 [-.01, .28]
Exaggerating the implications and/or importance of findings in order to increase likelihood of publication	0.09 [-.06, .22]	-0.08 [-.22, .05]	-0.05 [-.19, .10]	0.15 [.00, .29]
Lifting short phrases from others without quoting directly	0.06 [-.08, .20]	-0.07 [-.21, .06]	-0.10 [-.25, .06]	0.17 [.04, .32]
Irresponsibly co-authoring (e.g., not being involved enough to be able to verify accuracy of analysis)	0.15 [.02, .29]	-0.09 [-.23, .06]	-0.09 [-.24, .06]	0.22 [.07, .36]
Misrepresenting researcher qualification/experience in the proposal	0.17 [.03, .30]	-0.08 [-.21, .07]	-0.09 [-.25, .07]	0.07 [-.08, .22]

you have taken, the more severe you consider the following four items: 'Selecting variables out of convenience and/or familiarity when more theoretically grounded variables are available', 'HARKing', 'Cherry-picking data to analyze', and 'Not providing sufficient description of the data analyses or other procedures'.

5. Concluding discussion

The present study has sought to investigate the prevalence, frequency, and perceived severity of a list of 58 QRPs among researchers conducting quantitative humanities research in the US and Sweden. We also looked at the possible role of researcher characteristics such as years from Ph.D. With some minor exceptions, the differences noted between the two countries were very small, leading us to present the combined results.

With regard to how widespread the practices are, the results showed that 96% of the respondents reported to having used one or more of the practices listed in the past five years. The most prevalent item was also the one that occurred with the highest frequency *and* the one that was reported as the least severe ('Presenting the same presentation at multiple conferences'). More broadly, we noted a general trend that the more frequent practices were also the ones that were considered less severe and vice-versa. The results also showed that some researcher characteristics appeared to be associated with the frequency of engagement in, and perceived severity of, the practices listed. For example, years since Ph.D. tended to be negatively correlated with frequency and positively correlated with perceived severity of several items.

Overall, we note that QRPs are pervasive in the sense (a) that all of our items were found to occur within the sample, though not every participant has carried out the activity listed in each item, and (b) that they occur in all stages of research. Their frequency, by contrast, was relatively low on average (mostly reported to be occurring in fewer than 1-20% of research projects). On the one hand, we can take the fact that the practices were not more frequent as good news: yes, we all do them sometimes, but at least most researchers are not doing these activities very frequently. On the other hand, they definitely seem to occur, meaning that we would want to ask ourselves what we consider (un)acceptable as a field. This is where researcher training comes in: Which of these practices should we teach students to avoid and in what contexts? There may not be any clear-cut answers to this question. However, as part of this project, we have developed a set of materials that provide researchers, researcher trainers, and researchers in training with materials and guidance for developing an informed view of many of the issues that come into play in such situations (Wood et al., 2023; available open access at <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-502481>). We view these materials as a starting point, while recognizing that much more work is needed.

As mentioned in the introduction, QRPs are often considered the 'murky waters' of ethics. Part of the reason is that these are not always well or easily defined as a category. However, this is also due to the fact that there are no absolute answers to what constitutes a QRP, as the question of 'in what context?' is always lurking. A frequent theme of comments we received on the survey and when presenting the study at conferences is versions of 'it depends'. For many of these items, there may be external circumstances that would make an action fall into the realm of 'fully acceptable' rather than 'questionable'.

Against this background, an avenue for future research that will likely prove fruitful would be to investigate the association between context and frequency and perceived severity in a more systematic manner. Future studies may also want to address the question of to what extent intention matters: Is it a QRP if you did not know better? Also, and on the topic of limitations, it should be noted that our survey design did not allow us to differentiate between a hypothetical and an actual event; that is, we do not know how large a proportion of our respondents actually had, for example, 'a conflict of interest' in the past five years and chose to report or not to report it – just that some proportion of our sample reported *not* to have failed to report it, or at least did not remember not reporting it. In addition, it would be helpful for future studies to complement self-reported data with data obtained from other sources, as self-reported data are always going to be dependent on the participants giving an accurate picture of the question at hand. Some of the activities asked about in the present study have been found to be relatively frequent in observational studies. For example, by looking at published articles, recent studies and systematic reviews in applied linguistics have shown that conflict of interest statements are exceedingly rare among language test developers (Isbell & Kim, 2023) and that the handling of outliers frequently goes against best practice (Nicklin & Plonsky, 2020). Finally, there are decisions that we made in this study that may, by some, be considered questionable. For example, we did not pre-register the present study for the simple reason being that it is not yet a sufficiently common practice in the field that we thought to do so.

All in all, successfully carrying out ethical research means navigating a myriad of decision trees, views of colleagues, and external pressure (promotion, etc.). With this article, and with the larger project that it stems from, we hope to contribute to an open and respectful discussion about QRPs and what steps should be taken when it comes to researcher training and publishing policies to ensure honest reporting and replicable results.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is co-funded by the Swedish Research Council, the Bank of Sweden Tercentenary Foundation, and the Royal Swedish Academy of Letters, History and Antiquities (Project ID: FOE20-0017). The authors are also very grateful to Amanda Lindqvist at Linköping University, Sweden, whose rigorous work provided a starting point for the literature review for this project.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.rmal.2023.100064](https://doi.org/10.1016/j.rmal.2023.100064).

References

- Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS one*, 12(3), Article e0172792. [10.1371/journal.pone.0172792](https://doi.org/10.1371/journal.pone.0172792).
- Allum, N., Reid, A., Bidoglia, M., Gaskell, G., Bonn, N. A., Buljan, I., Fuglsang, S., Horbach, S. P. J. M., Kavouras, P., Marušić, A., Mejlgaard, N., Pizzolato, D., Roje, R., Tjldink, J., & Veltri, G. A. (2022). Researchers on research integrity: A survey of European and American researchers. *F1000 Research*, 12, 187–187. [10.12688/f1000research.128733.1](https://doi.org/10.12688/f1000research.128733.1).
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, 40, 134–142. [10.1017/S0267190520000033](https://doi.org/10.1017/S0267190520000033).
- Artino, A. R., Driessen, E. W., & Maggio, L. A. (2019). Ethical shades of gray: International frequency of scientific misconduct and questionable practices in health professions education. *Academic Medicine*, 94, 76–84. [10.1097/ACM.0000000000002412](https://doi.org/10.1097/ACM.0000000000002412).
- Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, 71(5), 715–738. [10.1093/joc/jqab031](https://doi.org/10.1093/joc/jqab031).
- Bottesini, J. G., Rhemtulla, M., & Vazire, S. (2022). What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society Open Science*, 9(4), Article 200048. [10.1098/rsos.200048](https://doi.org/10.1098/rsos.200048).
- Bouter, L. M., Tjldink, J., Axelsen, N., Martinson, B. C., & Ter Riet, G. (2016). Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*, 1(1), 1–8. [10.1186/s41073-016-0024-5](https://doi.org/10.1186/s41073-016-0024-5).
- Braun, M., & Roussos, A. J. (2012). Psychotherapy researchers: Reported misbehaviors and opinions. *Journal of Empirical Research on Human Research Ethics*, 7, 25–29. [10.1525/jer.2012.7.5.25](https://doi.org/10.1525/jer.2012.7.5.25).
- Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2021). Questionable research practices and open science in quantitative criminology. *Journal of Quantitative Criminology*, 1–31. [10.1007/s10940-021-09525-6](https://doi.org/10.1007/s10940-021-09525-6).
- De Costa, P. I. (Ed.). (2016). *Ethics in applied linguistics research: Language researcher narratives*. Routledge.
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1), 45–52. [10.1177/1948550615612150](https://doi.org/10.1177/1948550615612150).
- Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS one*, 13(7), Article e0200303. [10.1371/journal.pone.0200303](https://doi.org/10.1371/journal.pone.0200303).
- Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 54, 245–258. [10.1017/S0261444819000430](https://doi.org/10.1017/S0261444819000430).
- Gonulal, T. (2019). Missing data management practices in L2 research: The good, the bad and the ugly. *Erzincan University Journal of Education Faculty*, 21, 56–73. [10.17556/erziefd.448559](https://doi.org/10.17556/erziefd.448559).
- Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS one*, 17(2), Article e0263023. [10.1371/journal.pone.0263023](https://doi.org/10.1371/journal.pone.0263023).
- Hall, J., & Martin, B. R. (2019). Towards a taxonomy of research misconduct: The case of business school research. *Research Policy*, 48, 414–427. [10.1016/j.respol.2018.03.006](https://doi.org/10.1016/j.respol.2018.03.006).
- Hervé, M. (2021). RVAideMemoire: Testing and Plotting Procedures for Biostatistics. R package version 0.9-80. <https://CRAN.R-project.org/package=RVAideMemoire>
- Hofmann, B., Bredahl Jensen, L., Brandt Eriksen, M., Helgesson, G., Juth, N., & Holm, S. (2020). Research integrity among PhD students at the Faculty of Medicine: A comparison of three Scandinavian universities. *Journal of Empirical Research on Human Research Ethics*, 15(4), 320–329. [10.1177/1556264620929230](https://doi.org/10.1177/1556264620929230).
- Holm, S., & Hofmann, B. (2018). Associations between attitudes towards scientific misconduct and self-reported behavior. *Accountability in Research*, 25(5), 290–300. [10.1080/08989621.2018.1485493](https://doi.org/10.1080/08989621.2018.1485493).
- Hsu, C., & Sandford, B. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research & Evaluation*, 12, 1–8. <https://scholarworks.umass.edu/pare/vol12/iss1/10/>.
- Indiana University Center for Postsecondary Research. (2021). *The Carnegie Classifications of Institutions of Higher Education, 2021 Edition* Bloomington, IN. Available at <https://carnegieclassifications.acenet.edu>.
- Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, 106(1), 172–195. [10.1111/modl.12760](https://doi.org/10.1111/modl.12760).
- Isbell, D., & Kim, J. (2023). Development and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics*. [10.1016/j.rmal.2023.100060](https://doi.org/10.1016/j.rmal.2023.100060).
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. [10.1177/0956797611430953](https://doi.org/10.1177/0956797611430953).
- Ljubenković, A. M., Borovečki, A., Čurković, M., Hofmann, B., & Holm, S. (2021). Survey on the Research Misconduct and Questionable Research Practices of Medical Students, PhD Students, and Supervisors at the Zagreb School of Medicine in Croatia. *Journal of Empirical Research on Human Research Ethics*, 16(4), 435–449. [10.1177/15562646211033727](https://doi.org/10.1177/15562646211033727).
- Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. A. (2021). Both questionable and open research practices are prevalent in education research. *Educational Researcher*, 50(8), 493–504. [10.3102/0013189x211001356](https://doi.org/10.3102/0013189x211001356).
- National Endowment for the Humanities. (2022). *What are the humanities?* January 19 <https://www.neh.gov/about>.
- National Humanities Center. (2022). *What are the humanities?* January 19 Humanities in action <https://action.nationalhumanitiescenter.org/what-are-humanities/>.
- Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, 40, 26–55. [10.1017/S0267190520000057](https://doi.org/10.1017/S0267190520000057).
- Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, 98, 450–470. [10.1111/j.1540-4781.2014.12058.x](https://doi.org/10.1111/j.1540-4781.2014.12058.x).
- Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. [10.1111/lang.12079](https://doi.org/10.1111/lang.12079).
- Plonsky, L., Larsson, T., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (in press). A taxonomy of questionable research practices in quantitative humanities. In P. I. DeCosta, A. Rabie-Ahmed, & C. Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship*. John Benjamins.
- Rabelo, A. L., Farias, J. E., Sarmet, M. M., Joaquim, T. C., Hoersting, R. C., Victorino, L., Modesto, J. G. N., & Pilati, R. (2020). Questionable research practices among Brazilian psychological researchers: Results from a replication study and an international comparison. *International Journal of Psychology*, 55(4), 674–683. [10.1002/ijop.12632](https://doi.org/10.1002/ijop.12632).
- Ravn, T., & Sørensen, M. P. (2021). Exploring the gray area: Similarities and differences in questionable research practices (QRPs) across main areas of research. *Science and engineering ethics*, 27(4), 40. [10.1007/s11948-021-00310-z](https://doi.org/10.1007/s11948-021-00310-z).
- Saberi-Karimian, M., Afshari, R., Movahhed, S., Amiri, F., Kaykhaee, F., Mohajer, F., Noormandipour, M., Lamsehchi, A., Nasiri, M., Barkhidarian, B., & Norouzy, A. (2018). Different aspects of scientific misconduct among Iranian academic members. *European Science Editing*, 44(2), 28–31. [10.20316/ESE.2018.44.17020](https://doi.org/10.20316/ESE.2018.44.17020).
- Steneck, N. H. (2007). *Introduction to the responsible conduct of research*. Department of Health and Human Services, Office of Research Integrity <https://ori.hhs.gov/sites/default/files/rcrintro.pdf>.
- Sterling, S., & Gass, S. (2017). Exploring the boundaries of research ethics: Perceptions of ethics and ethical behaviors in applied linguistics research. *System*, 70, 50–62. [10.1016/j.system.2017.08.010](https://doi.org/10.1016/j.system.2017.08.010).

- Sterling, S., Plonsky, L., Larsson, T., Kytö, M., & Yaw, K. (2023). Introducing and illustrating the Delphi method for applied linguistics research. *Research Methods in Applied Linguistics*, 2, Article 100040. [10.1016/j.rmal.2022.100040](https://doi.org/10.1016/j.rmal.2022.100040).
- Swift, J. K., Christopherson, C. D., Bird, M. O., Zöld, A., & Goode, J. (2022). Questionable research practices among faculty and students in APA-accredited clinical and counseling psychology doctoral programs. *Training and Education in Professional Psychology*, 16(3), 299. [10.1037/tep0000322](https://doi.org/10.1037/tep0000322).
- Tijdink, J. K., Verbeke, R., & Smulders, Y. M. (2014). Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics*, 9(5), 64–71. [10.1177/1556264614552421](https://doi.org/10.1177/1556264614552421).
- Wood, M., Sterling, S., Larsson, T., Plonsky, L., Kytö, M., & Yaw, K. (submitted). *Researchers training researchers: Ethics training in Applied Linguistics* Manuscript submitted for publication.
- Wood, M., Larsson, T., Plonsky, L., Sterling, S., Kytö, M., & Yaw, K. (2023). Research ethics training materials: Questionable research practices in the quantitative humanities. Available at <http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-502481>
- Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: a systematic review and meta-analysis. *Science and engineering ethics*, 27(4), 1–28. [10.1007/s11948-021-00314-9](https://doi.org/10.1007/s11948-021-00314-9).
- Yaw, K., Plonsky, L., Larsson, T., Sterling, S., & Kytö, M. (in press). Timeline: Research ethics in applied linguistics. *Language Teaching*, 1-17. [10.1017/S0261444823000010](https://doi.org/10.1017/S0261444823000010).