

A Taxonomy of Questionable Research Practices in Quantitative Humanities

Luke Plonsky, Northern Arizona University
Tove Larsson, Northern Arizona University
Scott Sterling, Indiana State University
Merja Kytö, Uppsala University
Kate Yaw, University of South Florida
Margaret Wood, Northern Arizona University

Abstract

A growing body of research has begun to address ethical issues in the context of Applied Linguistics (e.g., De Costa, 2016; Isbell et al., 2022). One of the messages inherent in this line of inquiry is that ethical concerns are embedded throughout the research cycle from study conceptualization to realization, dissemination, application, and beyond (see Bernstein et al., this volume). With this concern in mind, the present study sought to catalog and develop a taxonomy of what are often referred to as ‘questionable research practices’ (QRPs; Steneck, 2006) and related decisions that come into play in the conduct of quantitative Applied Linguistics research. These include practices such as selective reporting and obscuring of methodological details to limit criticism. Using existing taxonomies developed in neighboring disciplines as a starting point (e.g., Tauginienė et al., 2019), we employed the *Delphi method* to elicit responses on potential QRPs in an iterative fashion from an expert panel as well as from peer scholars. The analyses of these data resulted in a domain-specific taxonomy that laid the groundwork for a large-scale survey that assessed the prevalence and perceived severity of ethical issues and QRPs found specifically in quantitative Applied Linguistics research (Larsson et al., 2023). The results are also be used to inform materials for methodological training in research ethics in Applied Linguistics and related disciplines (see De Costa et al., 2021; Wood et al., 2024, in press).

Introduction

Attention to meta-research in Applied Linguistics has grown exponentially in the last decade. We view this movement, which includes research ethics, as both an indicator and a consequence of the field's maturity (see Gass et al., 2021, and Marsden & Plonsky, 2018).

Evidence of methodological awareness and the momentum therein can be seen on many fronts of both an official/institutional as well as a more grassroots nature. At the institutional level, for example, the American Association for Applied Linguistics (AAAL) has recently formalized a strand on research methodology for its annual conference, and in 2021, Elsevier launched a journal titled *Research Methods in Applied Linguistics*. Other journal-based initiatives include updated guidelines for reporting (e.g., Norris et al., 2015), formal recognition of authors' efforts at transparency through Open Science badges, and new strands for research methods in *International Journal of Corpus Linguistics*, *Language Learning*, and *Studies in Second Language Acquisition*. (See corresponding editorials, notably all published in the same year, by Crossley et al., 2020, Gass & Plonsky, 2020, Paquot & Callies, 2020).

Efforts to improve the methodological capacity in Applied Linguistics at the grassroots level are too numerous and diverse to be done any justice here. However, a few examples of such 'bottom-up' reforms include (a) formal assessments of methodological literacy (e.g., Loewen et al., 2014), (b) the introduction and fortification of new techniques such as Bayesian data analysis, mixed effects modeling, structural equation modeling, and bibliometrics (e.g., Gass et al., 2022; Gries, 2021; Larsson et al., 2021; Norouzian et al., 2018), (c) reconsideration of familiar techniques such as data visualization, multiple regression, and sampling (e.g., Andringa & Godfroid, 2020; Larson-Hall, 2017; Plonsky, 2023; Plonsky & Ghanbar, 2018), (d) an earnest push for more replication studies (Marsden et al., 2018; Porte & McManus, 2019), and (e) methodological syntheses across a range of substantive and technical domains (e.g., Crowther et al., 2021; Hu & Plonsky, 2021; Sudina, in press). More anecdotally, we have also noticed that it is becoming more common to see scholars list methodological interests along with their substantive areas of expertise in bio statements and university profiles.

Some scholars may not immediately detect the ethical dimension at play in such discussions. It is not common, in fact, for researchers active in this domain frame the methodological advances that they are contributing to as purely methodological in nature. It is our position, however, that one can scarcely consider the many methodological choices made in the course of conducting a study without bringing in an ethical dimension. Moreover, in our view, a study can only be considered to be of high quality to the extent that it is carried out ethically (see Banegas, this volume; Plonsky, in press). And at the same time, a study can only be considered to be carried out ethically to the extent that it was

carried out according to the highest standards of methodological rigor and transparency (see Gass et al., 2021, for a discussion on quality from a largely methodological angle). In other words, ethics and methods are co-dependent.

The project described in this chapter builds on the few recent works that have fallen squarely and explicitly at the intersection of methods and ethics (e.g., De Costa et al., 2021; Sterling & Gass, 2017; Sterling et al., 2023; Yaw et al., in press). In particular, we perceived a need to pull a bit further on the thread introduced by Isbell et al. (2022). The current chapter reports on one stage of a larger, funded project seeking to assess the presence of questionable research practices (QRPs) in quantitative humanities research in the US and Sweden, that is, the “ethical grey areas” (Yaw et al., in press) that one might encounter in the conduct of an empirical study. In particular, this chapter introduces a set of field-specific QRPs and briefly discusses the process through which we arrived at this list.

Literature Review

Our project is not the first of its kind. Indeed, several scholars present compilations of QRPs, representing various disciplines and covering the different phases of the research cycle. It is important to be clear from the outset that, contrary to the very limited treatment of ethics found in most textbooks on research methods (Wood et al., in press), the range of methodological concerns that are ethical in nature is expansive yet difficult to pin down in any exhaustive fashion.

Part of what defines QRPs in such works is what they are not (Hall & Martin, 2019). For example, existing lists tend to clearly—and appropriately—delineate between QRPs and more blatantly or categorically inappropriate behaviors such as falsification, fabrication, and plagiarism (FFP) often grouped together and labeled as ‘research misconduct’. At the other end of the spectrum of researcher behavior, we often find what is labeled as ‘responsible conduct of research’ (RCR), which refers quite generally to research practices that are understood to be entirely acceptable (Steneck, 2006). QRPs lie in the murky and vast space between these two extremes and, therefore, can encompass a wide range of activities, decisions, and practices. They are behaviors that are sometimes justifiable or that may be reasonable under certain circumstances, but that may also be considered problematic or debatable in terms of their appropriateness for a given domain or a given study.

Resnik et al. (2015) examined the policies associated with research misconduct in a survey of 183 U.S.-based institutions. Beyond acts widely agreed to be inappropriate such as FFP, the authors identified thirteen different types of conduct of a questionable nature. Critically and unfortunately, however, the most common behaviors other than FFP were

described by Resnik et al. in fairly general language such as “other serious deviation” or “unethical authorship other than plagiarism”. Though useful as a point of departure, the lack of specificity in this kind of language does little to pinpoint the specific activities we might consider in attempting to list out the breadth of QRPs.

More recently, Tauginienè et al. (2019) sought to consolidate previous attempts such as Resnik et al.’s (2015) in order to define and classify different types of researcher misconduct. Their proposed taxonomy distinguishes between data-related instances of misconduct (e.g., suppression of data; *p*-hacking) and non-data-related behaviors (e.g., plagiarism). Recognizing that many of the concepts related to research(er) integrity differ across fields, settings, and regions (see also Jordan, 2013), the authors also review existing taxonomies and propose a glossary of terms and definitions that fall under the umbrella of research integrity.¹

One of the primary applications of these lists and taxonomies has been to develop surveys designed to gauge the frequency and/or perceived severity of different researcher behaviors and practices. Indeed, such surveys can be found across disciplines ranging from pharmacology to sociology. Looking across this body of work, Fanelli (2009) sought to estimate the prevalence of scientific researcher misconduct as well as a range of QRPs. Overall, this widely cited meta-analysis based on 18 primary studies found that approximately 2% of the researchers surveyed admitted to one or more form of FFP. Admission rates for QRPs, by contrast, were as high as 34%, with a mean rate across the sample of approximately 10%. Two important caveats to these findings are worth noting. First, the true frequency of such practices is unknown. It is potentially much higher due to self-selection, dishonest responses, and memory decay on the part of participants. At the same time, the 34% may be overstated, given that a QRP may be justified in some cases depending on the context. Second, substantial heterogeneity was observed as a function of survey design and item wording as well as discipline. For example, FFP admission rates were higher in medicine and when funding was involved, perhaps due to increased pressure, whether self- or agency-imposed, perceived or real. Considering the different types of forces at play and types of research found in different domains, these findings highlight the need for discipline-specific taxonomies and instruments for assessing these practices, which is one of the goals of the present study.

Intrigued by Fanelli’s findings and curious to know how their own field might compare, Isbell et al. (2022) conducted the first survey of FFP and QRPs in Applied Linguistics. Their questionnaire was based largely (but somewhat loosely) on the aggregate of items found in Fanelli (2009) and included both misconduct (i.e., data falsification or fabrication) and several QRPs. The survey was administered to a sample of 351 applied linguists who identified as users of quantitative and/or mixed (quantitative and qualitative) methods. The rate of self-reported misconduct, 17%, was dramatically higher than the 2% average across other fields reported in Fanelli (2009). The rate of self-reported QRPs was

much higher in many instances, and nearly all of the participants (94%) admitted to at least one QRP. Examples of QRP admission rates included (a) not reporting findings that run contrary to the literature (14%), (b) choosing the type of analysis to be run based on an anticipatedly favorable result (40%), and (c) excluding non-significant results (43%). However, the same team also conducted a follow-up study of the participants' qualitative comments in response survey items, which revealed a great deal of nuance behind their numerical responses (Plonsky et al., 2024).

The current project drew much of its impetus from Isbell et al. (2022). We, too, were very much interested in assessing the prevalence and perceived severity of different QRPs. However, in order to do so, there was a need to establish a humanities-specific taxonomy of QRPs and a corresponding instrument that would more closely align with the types of practices relevant to quantitative humanities research and Applied Linguistics, in particular. The goal of the present study was, therefore, to develop a taxonomy of QRPs for quantitative humanities research. That taxonomy, once developed, could then be employed to design an instrument for assessing the prevalence and severity of QRPs in the quantitative humanities² (see Larsson et al., 2023).

Method

In order to address the goal stated above, we employed the Delphi method (Linstone & Turoff, 2002). The Delphi method is comprised of a set of largely qualitative bottom-up data elicitation techniques. Delphi studies such as ours rely on an iterative process of community-generated responses, ideas, and feedback. One way to think of the Delphi method is to imagine an asynchronous focus group. For a thorough introduction to the Delphi method in the context of Applied Linguistics, readers are directed to Sterling et al. (2023; see also Rodríguez-Lifante & Pereira, 2021, for a review of applications of the Delphi method in language-related research). In this case, we sought to arrive at some consensus around the range of possible QRPs in the quantitative humanities. Readers interested in conducting a Delphi study should seek out Sterling et al. (2023). In that publication, we provide more details on how to conduct a Delphi study, consider strengths and weaknesses of the methodology, and include discussions related to decisions made as we collected data. While the Sterling et al. publication represents an introduction to the Delphi method, the remainder of this chapter will focus on the outcomes of our Delphi project.

The Expert Panel

Given the focus of our study, we invited 10 individuals to serve as expert informants/panelists. Their backgrounds included one or more of the following: (a) humanities research with an explicit interest in ethical issues, (b) researcher training, and

(c) quantitative research in quantitative humanities. Given our particular focus on QRPs in the US and Sweden, we also ensured that scholars from both countries were represented on the panel. As explained below, panel membership involved a substantial commitment of time and expertise, and we are very grateful for the wealth of insight that each panelist contributed.

QRP Item Generation and Revision

The panel was sent a survey that described QRPs and invited them to nominate as many QRPs as might come to mind within six different categories: (a) funding, (b) research design and data collection, (c) data analysis, (d) dissemination, (e) service, and (f) mentorship. The six categories, based loosely on existing taxonomies from other disciplines, were meant to prompt the panelists to consider different facets of the research cycle without unduly leading or restricting them.

Once this initial list was collected, three steps were taken. First, the proposed items that we received were read, combined/consolidated for any redundancies, and revised for consistency/clarity. Second, we set aside items related to two of the categories—service and mentorship—on the grounds that they are not necessarily or entirely questionable *research* practices, but rather scholarly or researcher practices. We are not at all trying to indicate a lack of questionable territory in the realms of service and mentoring (see Sterling, this volume for a discussion of related issues). Rather, we decided that these areas were outside the main scope of our study which was focused on the behaviors of researchers carrying out and reporting research. Finally, with a slightly reduced set of items in hand, we cross-checked the panel-produced list with our experiences as authors, researchers, and funding-seekers as well as with Isbell et al.'s (2022) survey and several taxonomies from other disciplines (Al-Marzouki et al., 2005; Hall & Martin, 2019; Jordan, 2013; Kumar, 2008; Tauginienė et al., 2019). The total list at that point consisted of 62 items, the vast majority of which (90%) were collected from the panel.

Several rounds of iterative revisions then began. The process went as follows. All items, grouped into the four remaining categories, were imported into a Qualtrics survey. The panel was provided with three options for each QRP/item: “accept”, “accept, minor revision”, and “significant revision or reject” (see Figure 1). If a panelist was unhappy with an item, they were instructed to provide comments and suggestions for how to revise any item or to simply suggest that we remove the item from our list. Their responses were then tallied and discussed by our team. A threshold score of 80% (8 out of 10 panelists) was set. Any item that was rated as “accept” by 80% of the panel was moved to the final set of QRPs and not included in subsequent rounds of commentary and revision. Items with an

acceptance rate of less than 80% were revised based on the panel’s feedback. If it was not possible to revise an item to the specifications of the panel or if the panel agreed that an item should be eliminated, it was removed from consideration. Three rounds of revision then ensued, which we will now describe.

Data analysis and interpretation				
	Accept	Accept, minor revision	Significant revision or reject	Please comment on any revisions or rejection here.
Removing whole items/cases knowingly/purposefully to obtain favorable results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Using unjustified methods of handling outliers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Not being transparent with regard to the reporting on what steps were taken for data cleaning (e.g., removing cases/items without a stated criterion or justification for doing so)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Being ambiguous about whether an exploratory vs. confirmatory analysis was employed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
HARKing (i.e., hypothesizing after results are known)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Cherry-picking data to analyze	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>
Choosing a method of analysis that will likely lead to favorable outcome (e.g., in favor of the researcher’s hypothesis)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="text"/>

Figure 1. Screenshot of part of the QRP revision survey administered to the expert panel

As previously stated, round 1 of item revisions began with 62 items collected from the panel as well as the sources mentioned above. 43 of those surpassed the 80% threshold set for our study and were immediately deemed acceptable. Three items were removed entirely based on panel feedback, and one item was added. 15 items remained and were moved on to round two for further revision. In round two, nine items met our threshold score and were deemed acceptable, and one item was removed. Six items received feedback and were then moved on to round three. All six of the items that moved into round three

were found to be acceptable, yielding a final set of 58 QRPs. Since all items either met our threshold score or were rejected, the Delphi was ended at three rounds.

To bring this process to life, consider the item that was removed in round two, which was worded as follows: “Publishing multiple times with a combination of old and new data (i.e., data augmentation)”. One of the panelist’s feedback in round 1 asked “Is this really a QRP?” We then revised the item to read as follows: “Publishing multiple times with a combination of old and new data to increase one’s publication record (i.e., data augmentation)”. Given the round 2 feedback we received which expressed concerns about this item overlapping with ‘salami’ publication item and the fact that a single dataset can quite broad (e.g., an entire corpus of linguistic data/texts; see Brookes & McEnergy, this volume), the item was then removed.

Table 1 highlights the evolution of another item as it made its way from its original to its final, accepted form (see also Sterling et al., 2023). Several issues were raised in the path of this item from its initially proposed wording to the final accepted version. In round 1, for example, we see concern from one panelist regarding the potential lack of clarity around the word “nudging”, which we subsequently removed. The panelist also asks, perhaps rhetorically, whether the pervasiveness of a practice indicates that is no longer questionable. As this item did not initially receive enough votes for accept, we revised it using comments from the panel. In particular, we removed the potentially unknown term “nudging” and provided an example of the kind of (questionable) appeal a researcher might make (e.g., an emotional story) during participant recruitment. In response to the revised item, a panelist suggested referring to shaming potential participants into agreeing to participate or using coercion. Neither of these suggestions were taken up because we felt that both were categorically unethical and no longer in the realm of questionable. For the same reason, we did not incorporate a panelist’s round 3 suggestion to intensify “difficult” with “unduly”. In our third and final wording of them item, however, we did take on the panelist’s wording related to recruiting “in a way that makes refusal difficult or uncomfortable”. (See further discussion on ethical considerations around recruiting participants in Kayi-Aydar, this volume.) We hope to have given a glimpse here into the development of this item—and the larger process to which it belongs—by detailing its development. Many of the other multi-round items went through a similarly involved process.

Not evident in this description are the rich discussions among our group, consisting of active researchers at various stages in our careers and with a wide range of substantive interests within Applied Linguistics. During the several months that the Delphi study took place, we poured over and deliberated on each of the items as well all the responses that we received from our panel. In order to arrive at each round’s list of items, we debated and argued over the wording and the multitudes of contexts, examples, and counterexamples under which each potential QRP may or may not apply. (See Marino et al., this volume, for

an in-depth discussion of ethical considerations specific to the context of computer-assisted language learning, social media, and other online spaces.) It was an unusually time- and labor-intensive exercise, especially for a group more accustomed to the processing, handling, and analysis of numeric data. In the end, though, it was well worth the effort.

Table 1

Example of item commentary and revision based on panel feedback

Item wording	Example panel comment	Decision
Round 1. Nudging participants to join a study using monetary (or other) incentives	“Need to better define nudging. Is it a QRP if we all do it?”	Revise
Round 2. Recruiting participants to join a study in a way that saying “no” feels wrong (e.g., using an emotional story during classroom recruitment)	“Maybe change to 'in a way that makes refusal difficult or uncomfortable' 'Emotional story'....hmmm, I'm all in favor of emotion. How about '...using coercion during classroom recruitment, or by pleading or shaming participants into agreeing)'?”	Revise
Round 3. Recruiting participants to join a study in a way that makes refusal difficult or uncomfortable	“...I might add some kind of intensifier/qualifier before "difficult or uncomfortable" when thinking about this more - "unduly" might work. ...”	Accept

Results

The result of this Delphi study, involving three rounds of item development, feedback, and revision, was a set of 58 QRPs. The full set of items is presented in Table 2 which classifies each QRP as pertaining to the following headings/themes: ‘Funding’ (11 items); ‘Design and Data Collection’ (11 items); ‘Data Analysis and Interpretation’ (14 items); and ‘Write-up and Dissemination’ (22 items).

Table 2

Taxonomy of Questionable Research Practices in Quantitative Humanities (K = 58)

Section I: Funding (k = 11)

1. Cherry-picking samples/data/results that favor the funder
 2. Choosing a topic on the grounds that the funder might expect the study to portray them in a positive light
 3. Not reporting a conflict of interest (financial or otherwise)
 4. Misrepresenting researcher qualification/experience in the proposal
 5. Misrepresenting study importance in the proposal (e.g., exaggeration of impact and value of proposal to society)
 6. Over budgeting (i.e., knowingly asking for more money than is actually needed)
 7. Not producing the promised project outcomes due to project mismanagement (e.g., producing fewer articles than promised)
 8. 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)
 9. Misrepresenting literature in the proposal (e.g., over-emphasizing previous research that supports the proposal and/or ignoring conflicting evidence)
 10. Making changes in context after proposal submitted (e.g., proposing to carry out funded work in one context/with one population but then actually carrying it out elsewhere/with another population)
 11. Not disclosing impacts that funder directly had on research decisions (e.g., using particular datasets, selection of published outcomes)
-

Section II: Design and Data Collection (k = 11)

1. Selecting variables out of convenience and/or familiarity when more theoretically grounded variables are available
 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
 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)
 4. Employing instruments/measures without a strong validity argument
 5. Not being transparent with regard to the decisions made in the data collection phase
 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)
 7. Not reporting the effect of decisions about method, design, or instrumentation on study outcomes (e.g., operationalizing proficiency as grade level instead of using an accepted measure of language proficiency)
 8. Leaving out known/likely moderator variables or covariates from the study design without explanation or acknowledgment
-

-
9. Fishing for results by collecting information on unnecessary variables
 10. Having an unnecessarily long/burdensome data collection for participants
 11. Recruiting participants to join a study in a way that makes refusal difficult or uncomfortable
-

Section III: Data Analysis and Interpretation (k = 14)

1. Removing whole items/cases knowingly/purposefully to obtain favorable results
 2. Using unjustified methods of handling outliers
 3. Not being transparent with regard to the reporting on what steps were taken for data cleaning (e.g., removing cases/items without a stated criterion or justification for doing so)
 4. Being ambiguous about whether an exploratory vs. confirmatory analysis was employed
 5. HARKing (i.e., hypothesizing after results are known)
 6. Cherry-picking data to analyze
 7. Choosing a method of analysis that will likely lead to favorable outcome (e.g., in favor of the researcher's hypothesis)
 8. p-hacking (i.e., running analyses in a manner that produces statistical significance)
 9. Ignoring alternate explanations of 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)
 11. Categorizing continuous variables without sufficient justification
 12. Using too many statistics tests without correction (e.g., Bonferroni)
 13. Using incorrect statistical methods (e.g., tests that are not appropriate for the type of data being analyzed)
 14. Interpreting statistical results inappropriately (e.g., claiming equivalence between groups based on a non-statistically significant difference; undue extrapolation)
-

Section IV: Write-up and Dissemination (k = 22)

1. Failing to refer to relevant work by other authors
 2. Not providing sufficient description of the data analyses or other procedures
 3. Not providing sufficient description of the data and the results (e.g., exact p-values, SD)
 4. Not reporting or publishing results because they are not statistically significant (i.e., the 'file drawer' issue)
 5. Employing selective reporting of results/instruments
 6. Not sharing data when allowable
 7. Not sharing scripts used to analyze the results
 8. Not sharing instruments/coding schemes
 9. Not attempting to publish results in a timely manner
 10. Presenting misleading figures of data (e.g., displaying a truncated entire y-axis)
 11. Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)
-

12. 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
 13. Presenting same presentation at multiple conferences
 14. Employing excessive self-citation
 15. Intentionally omitting relevant work because it does not align with one's theoretical or methodological approach
 16. Inappropriately including or excluding authors
 17. Inappropriately attributing author roles when listed in publications
 18. Inappropriate ordering of authors
 19. Not giving research assistants due credit in publications
 20. Exaggerating the implications and/or importance of findings in order to increase likelihood of publication
 21. Lifting short phrases from others without quoting directly
 22. Irresponsibly co-authoring (e.g., not being involved enough to be able to verify accuracy of analysis)
-

Discussion

This study presents the first attempt to catalogue and classify QRPs in quantitative humanities research. We believe this process to have led to a fairly comprehensive set of researcher behaviors that merit—and are currently receiving—further empirical attention (e.g., Larsson et al., 2023).

The QRPs identified here can and will also go a long way toward informing researcher training in Applied Linguistics and beyond. We hasten to add, though, that the list may not be exhaustive. In fact, active researchers are very likely to encounter decision points and situations that they might classify as questionable and that do not fit neatly under any of the items in this list. We are open to suggestions, and future research will almost certainly expand and revise the taxonomy proposed here. That said, a list of hundreds of QRPs would be overwhelming to read and potentially less impactful.

With respect to the structure of our taxonomy, the proportion of items that belong to each theme is somewhat telling. What might we infer, for example, from the larger number of QRPs that belong to the theme of 'Write-up and Dissemination'? Perhaps this is due to the large degree of latitude (and the lack of agreed-upon conventions) left to authors at this stage of the process. We might also invoke at that phase of research the relatively high stakes associated with academic publishing, especially in the highly competitive 'publish or perish' environment, whether perceived or real, that many of us currently occupy. The larger number of items in that final section of the taxonomy may also reflect the lack IRB-

style norms for writing up research. The smaller number of items in our list that are associated with funding may indicate a small place for QRPs associated with financial interests. However, some of those QRPs may be of greater weight.

To be sure, we make no claims in this paper about the relative severity of these QRPs. Some are certainly more serious than others and some are more common than others. We take up both of these issues in our Phase II study of the pervasiveness and perceived severity of this set of QRPs among quantitative humanities researchers (see Larsson et al., 2023).

We would also note that these categories, though largely functional, are not 100% exclusive of each other. One can easily imagine, for example, crossover between interpretive behaviors (Section III) and what happens during the write-up of a study (Section IV). Likewise, one can also imagine making design-related and analytical choices (Section II) that serve or appeal to a funding agency (Section I). The point is that sections of this taxonomy are not to be viewed as rigid; they are at least somewhat fluid and/or flexible and are simply meant to give some structure to the QRPs they contain.

It may also be useful to consider our taxonomy in relation to others. In comparison with previously-published taxonomies, our quantitative humanities-specific list has some similarities other quantitative research domains, such as medical clinical trials (Al-Marzouki et al., 2005) and biomedical sciences (Kumar, 2008). In all three lists, there were items related to research design, data collection, data analysis, and reporting, while Kumar (2008) and our list also included funding. In the category of data analysis, all the taxonomies included information on HARKing, *p*-hacking, and other questionable quantitative analysis practices. As demonstrated in Table 3, there was some similarity in other items included in the taxonomies, although the wording may have been less descriptive (in the case of Kumar, 2008) and/or targeted to a specific field (in the case of Al-Marzouki et al., 2005).

There are also some key differences between our taxonomy and these other two in relation to item content. For instance, Al-Marzouki et al. (2005) included items specific to clinical trials that would be irrelevant for quantitative humanities research. Some examples are:

- *Selective withdrawals on basis of knowledge of [treatment] allocation*
- *Ignore data on side-effects*
- *Trial stopped for marketing and not scientific reasons*

In fact, numerous items related to allocation of treatment condition, including who is allowed to know which participants are part of the experimental treatments, what schedule of data collection is followed, and how participants are assigned to treatment groups. While quantitative humanities researchers conducting experimental studies will need to consider randomized group assignment, this may not play as central a role as it does in medical clinical trials. On the biomedical sciences side, Kumar (2008) shared two topics that would

be less relevant to quantitative humanities researchers: photo-manipulation (e.g., manipulating images of stem cells in support of research findings), and unethical animal experimentation (i.e., “curiosity-driven animal research”, p. 220). These examples illustrate the benefits of a field-specific taxonomy, namely that items included are relevant to the quantitative humanities and any examples or explanation included in the items match plausible quantitative humanities research scenarios.

Finally, it may be worth comparing our taxonomy with the list of QRP items included in Isbell et al. (2022), which served as a jumping-off point for our current project. The Isbell et al. (2022) survey contained 10 QRP items, all of which fell into the categories of data analysis/interpretation and write-up/dissemination. There was a lot of overlap in the content of the items, though the wording on the current taxonomy tends to offer additional explanation in each item (see Table X). This may be a result of the Delphi process, which integrated feedback from a range of experts into the wording of each item.

Table 3
Wording differences among QRP items across lists

Source	QRP item	Correlate in our taxonomy
Al-Marzouki et al. (2005)	Excluding patients or results to exaggerate effects or remove adverse events	Removing whole items/cases knowingly/purposefully to obtain favorable results
	Inappropriate analysis for example comparison of survival time by <i>t</i> -test	Using incorrect statistical methods (e.g., tests that are not appropriate for the type of data being analyzed)
	Selective reporting of (i) subgroups (ii) outcomes (iii) time points	Employing selective reporting of results/instruments
Kumar (2008)	Divided publication (salami slicing or data fragmentation)	Salami publication (e.g., dividing up the results of a single study into multiple manuscripts in order to publish more)
	Irresponsible co-authorship	Irresponsibly co-authoring (e.g., not being involved enough to be able to verify accuracy of analysis)
Isbell et al. (2022)	Choose analysis based on favorable outcome	Choosing a method of analysis that will likely lead to a favorable outcome (e.g., in favor of the researcher’s hypothesis)
	Delete data (e.g., outliers) based on a gut feeling	Using unjustified methods of handling outliers

Not attempting to publish a study
with nonsignificant results

Not reporting or publishing results
because they are not statistically
significant (i.e., the ‘file drawer’
issue)

Conclusion and Future Directions

This chapter presents a newly developed taxonomy of QRPs. This taxonomy and set of QRPs is unique in at least two respects. First is the process by which it was developed (i.e., the Delphi method). And second, our taxonomy is the first to our knowledge designed specifically for the quantitative humanities.

This study moves provides a greater understanding of the range of potentially problematic practices and behaviors that researchers in quantitative humanities may encounter. However, the goals of the larger project to which this study belongs are broader than simply drawing awareness of QRPs. A logical next step is to use this taxonomy to assess the prevalence and perceived severity of the QRPs it includes overall in the quantitative humanities, across individual disciplines, and across countries (e.g., Sweden and USA). We also intend to develop a set of materials for research methods training that address many of the QRPs in our taxonomy. These materials will be openly accessible and able to be adapted in part or whole.

As surveys have shown and will likely continue to show, QRPs are not likely to disappear any time soon. What we hope to provide through taxonomy, however, is a greater understanding of the range of issues that we all encounter as researchers as a means to improve methodological quality and ultimately to improve the state of our science.

Endnotes

¹ See Jordan (2013) for additional and compelling argumentation on the need for terminological and taxonomical clarity with respect to research ethics, both for the sake of empirical researchers at large as well as for those scholars active in this particular domain, regardless of the discipline they might belong to.

² The development of this taxonomy is the first in the three-phase project. As stated here and described in the subsequent text, the Phase I involved designing a taxonomy of QRPs that are specific to the quantitative humanities. Building naturally on Phase I, in Phase II we designed, piloted, and administered an instrument for assessing the prevalence and perceived severity of the QRPs among quantitative humanities researchers in the US and Sweden. We also examined participant responses in relation to their backgrounds and several demographic and educational variables. Phase III of the project involves developing

a set of materials for researcher training that takes the findings from Phases I and II into account (see Wood et al., 2024).

Acknowledgment

This project was 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).

References

- Al-Marzouki, S., Roberts, I., Marshall, T., & Evans, S. (2005). The effect of scientific misconduct on the results of clinical trials: A Delphi survey. *Contemporary Clinical Trials*, 26(3), 331-337. <https://doi.org/10.1016/j.cct.2005.01.011>
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics* 40, 134-142. <https://doi.org/10.1017/S0267190520000033>
- Crossley, S., Marsden, E., Ellis, N., Kormos, J., Morgan-Short, K., Thierry, G. (2020). Introduction of Methods Showcase articles in *Language Learning*. *Language Learning*, 70, 5-10. <https://doi.org/10.1111/lang.12389>
- De Costa, P. I., Sterling, S., Lee, J., Li, W., & Rawal, H. (2021). Research tasks on ethics in applied linguistics. *Language Teaching*, 54, 58-70. <https://doi.org/10.1017/S0261444820000257>
- Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 54, 245-258. <https://doi.org/10.1017/S0261444819000430>
- Gass, S., & Plonsky, L. (2020). Introducing the *SSLA* Methods Forum. *Studies in Second Language Acquisition*, 42, 667-669. <https://doi.org/10.1017/S0272263120000364>
- Gass, S. M., Plonsky, L., & Huntley, E. (2022). Taking the Long view: A bibliometric analysis. In A. Benati & J. W. Schwieter (Eds.), *Second language acquisition as shaped by the scholarly legacy of Professor Michael Long* (pp. 9-27). John Benjamins. <https://doi.org/10.1075/bpa.14.02gas>
- Gries, S. Th. (2021). (Generalized linear) mixed effects modeling: A learner corpus example. *Language Learning*, 71, 757-798. <https://doi.org/10.1111/lang.12448>
- Hall, J., & Martin, B. R. (2019). Towards a taxonomy of research misconduct: The case of business school research. *Research Policy*, 48, 414-427. <https://doi.org/10.1016/j.respol.2018.03.006>
- Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, 37, 171-184. <https://doi.org/10.1177/0267658319877433>
- Isbell, D., Brown, D., Chen, M., Derrick, D., Ghanem, R., Gutiérrez Arvizu, M. N., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices:

- The ethics of quantitative data handling and reporting in applied linguistics. *Modern Language Journal*, 106, 172-195. <https://doi.org/10.1111/modl.12760>
- Jordan, S. R. (2013). Conceptual clarification and the task of improving research on academic ethics. *Journal of Academic Ethics*, 11, 243–256. <https://doi.org/10.1007/s10805-013-9190-y>
- Kumar, M. N. (2008). A review of the types of scientific misconduct in biomedical research. *Journal of Academic Ethics*, 6, 211-228. <https://doi.org/10.1007/s10805-008-9068-6>
- Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *Modern Language Journal*, 101, 244–270. <https://doi.org/10.1111/modl.12386>
- Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the prevalence and perceived severity of Questionable Research Practices. *Research Methods in Applied Linguistics*, 2, 100064. <https://doi.org/10.1016/j.rmal.2023.100064>
- Loewen, S., Lavolette, E., Spino, S., Papi, M., Schmidtke, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, 48, 360–388. <https://doi.org/10.1002/tesq.128>
- Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, 68, 321–391. doi:10.1111/lang.12286
- Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad, & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219-228). John Benjamins.
- Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, 68, 1032-1075. <https://doi.org/10.1111/lang.12310>
- Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, 65, 470-476. <https://doi.org/10.1111/lang.12104>
- Paquot, M., & Callies, M. (2020). Promoting methodological expertise, transparency, replication, and cumulative learning: Introducing new manuscript types in the *International Journal of Learner Corpus Research*. *International Journal of Learner Corpus Research*, 6, 121–124. <https://doi.org/10.1075/ijlcr.00014.edi>
- Plonsky, L. (2023). Sampling and generalizability in Lx research: A second order synthesis. *Languages*, 8, 75. <https://doi.org/10.3390/languages8010075>
- Plonsky, L. (in press). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*.

Plonsky, L., Larsson, T., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2024). Developing a taxonomy of ethical decisions in applied linguistics research. In P. I., De Costa, A., Rabie-Ahmed, & C., Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship* (pp. 10-27). John Benjamins.

- Plonsky, L., Brown, D., Chen, M., Ghanem, R., Gutiérrez Arvizu, M. N., Isbell, D. R., & Zhang, M. (2024). "Significance sells": Applied linguists' view on questionable research practices. *Research Methods in Applied Linguistics*, 3, 100099. <https://doi.org/10.1016/j.rmal.2024.100099>
- Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.
- Rodríguez-Lifante, A., & Pereira, M. M. B. (2021). The Delphi method in applied linguistics in the light of a theoretical and critical analysis. *Revista Brasileira de Linguística Aplicada*, 21, 271–293. DOI:10.1590/1984-6398202116351
- Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, 12(1), 53–74. <https://doi.org/10.1007/pl00022268>
- Sterling, S., & Gass, S. (2017). Exploring the boundaries of research ethics: Perceptions of ethics and ethical behaviors in classroom-based applied linguistics research. *System*, 70, 50–62. <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, 100040. <https://doi.org/10.1016/j.rmal.2022.100040>
- Sudina, E. (2023). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, 47, 1427-1455. <https://doi.org/10.1017/S0272263122000560>
- Tauginienė, L., Gaižauskaitė, I., Razi, S., Glendinning, I., Sivasubramaniam, S., Marino, F., Cosentino, M., Anohina-Naumeca, A., & Kravjar, J. (2019). Enhancing the taxonomies relating to academic integrity and misconduct. *Journal of Academic Ethics*, 17, 345–361. <https://doi.org/10.1007/s10805-019-09342-4>
- Turoff, M., & Linstone, H. A. (2002). *The Delphi method: Techniques and applications*. Addison-Wesley.
- Wood, M., Larsson, T., Plonsky, L., Sterling, S., Kytö, M., & Yaw, K. (2024). *Addressing questionable research practices in applied linguistics: A practical guide*. Applied Linguistics Press.
- Wood, M., Sterling, S., Larsson, T., Plonsky, L., Kytö, M., & Yaw, K. (in press). *Researchers training researchers: Ethics training in Applied Linguistics*. *TESOL Quarterly*.
- Yaw, K., Plonsky, L., Larsson, T., Sterling, S., & Kytö, M. (2023). Timeline: Research ethics in applied linguistics. *Language Teaching*, 56, 478-494. <https://doi.org/10.1017/S0261444823000010>