Annual Review of Applied Linguistics (2024), 1–15
doi:10.1017/S0267190524000059

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Study quality as an intellectual and ethical imperative: A proposed framework

Luke Plonsky

Northern Arizona University, USA
Email: lukeplonsky@gmail.com

Abstract

What is quality in the context of applied linguistic research? Addressing this question is both an intellectual and ethical imperative for the field. Toward that end, I propose a four-part framework for understanding and evaluating study quality: Quality research is (a) methodologically rigorous, (b) transparent, (c) ethical, and (d) of value to society. The bulk of the paper is devoted to describing each of these four elements, first conceptually and then as observed in the field. I also articulate some of the many connections among the four elements within the framework. In addition, I suggest next steps for further addressing the notion of quality in terms of the framework itself as well as the ways it might be used to advance our field's research, training, and legitimacy as an academic discipline that is still in some ways coming of age.

Keywords: study quality; methodological rigor; ethics

What is quality in the context of applied linguistics research? We all care about quality. We all want to read and produce quality research. And I think we'd all agree that there's little point to any discipline that doesn't insist on quality by means of peer review, scholarly mentoring, and/or other mechanisms, whether structural or more grassroots in nature. However, we, as a field, have yet to arrive at an agreed-upon understanding of this notion. This paper offers a framework for conceptualizing and assessing study quality. I don't expect full or immediate consensus, but I hope the proposed framework will serve as a step toward understanding and operationalizing the notion of quality in a way that will support both the intellectual and ethical imperative of the scholarly discipline we call applied linguistics.

What is quality?

One (perhaps obvious) way to start addressing this question would be to turn to the Mertonian Norms of Science. Proposed in 1942 by sociologist Robert Merton, this set of four principles – communism, universalism, disinterestedness, and organized

skepticism – was meant as a guide for the modern, institutionally based scientific enterprise. These norms are certainly valuable and are worthy of consideration. And striving to embrace them might satisfy – or at least open up conversations with – some of our colleagues in sociology or philosophy of science. But I don't think the Mertonian Norms of Science can be used as a definition of quality for applied linguistics; they're too abstract and would be very difficult – if not impossible – to operationalize.

I should mention that this is not my first attempt to define study quality. I tried to take this on in my dissertation (Plonsky, 2011) and in the articles that were based on it. In Plonsky (2013), for example, I defined quality as "the combination of (a) adherence to standards of contextually appropriate methodological rigor in research practices and (b) transparent and complete reporting of such practices" (p. 657). That definition and its corresponding operationalization served as a useful starting point and led to a number of insights into a wide range of quantitative research practices found in the field. That work also has led to dozens of methodological syntheses seeking to assess the quality within and across different domains (e.g., Burston et al., 2024; Li, 2023; Sudina, 2023b). Looking back, however, that definition was much too narrow both for a concept as broad as quality and for a field as broad as applied linguistics. For example, the quality of a given study (or set of studies) can also be conceived of and assessed in terms of its contribution to society. In addition, my earlier definition, to some extent, and certainly the way I operationalized it were focused very heavily on quantitative research.

Partly recognizing the limitations of Plonsky (2013, 2014), Sue Gass, Shawn Loewen, and I have argued more recently that a definition of quality in the context of quantitative research should include a concern with estimating the magnitude of the effects or relationships of interest as opposed to their mere presence or absence (Gass et al., 2021). The inclusion of this facet of quality can be linked directly to what Cumming (2014) referred to as "estimation thinking" (p. 8), which he contrasts to the "dichotomous thinking" manifest in the (mis)use of *p*-values which is so prevalent in applied linguistics (e.g., Cohen, 1997; Klein, 2013; Lindstromberg, 2022; Norris, 2015). This facet can also be tied to the notion of *synthetic mindedness* argued for in Norris and Ortega (2006) as a perspective on research that prioritizes the cumulative evidence available (rather than any single study) and that prioritizes the extent of an effect or a relationship.

Gass et al.'s (2021) definition of quality also expanded on the construct of transparency, making space for thorough (as opposed to selective) reporting of results and for the sharing of materials and data whether for reanalysis, secondary analysis, replication, training, or other purposes (see Nicklin & Plonsky, 2020). Concern for reproducibility and for open science practices more generally is certainly worthwhile (Marsden et al., 2018); however, even with these additions, the definitions of study quality available to date fail to capture the full construct of interest and have focused almost exclusively on more quantitatively oriented paradigms.

A proposed framework for study quality

The framework I propose here views study quality as a multidimensional construct comprised of the four following elements or subconstructs: (a) methodological rigor,

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

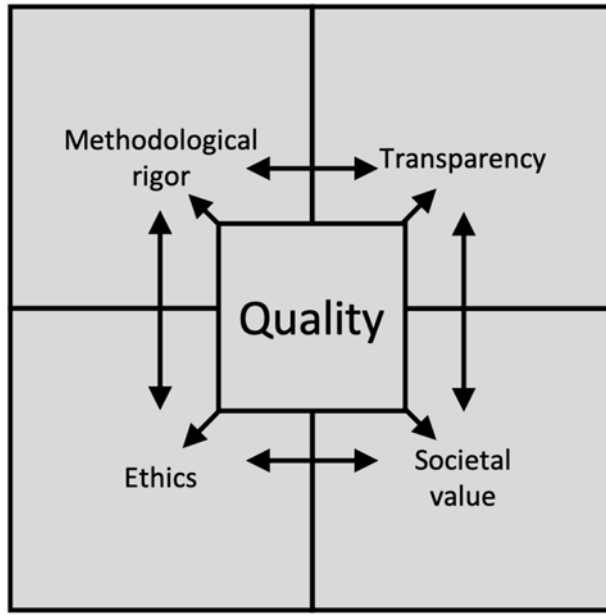

Figure 1. Proposed framework for study quality in applied linguistics.

(b) transparency, (c) societal value, and (d) ethics (see Figure 1). The first two of these overlap with previously proposed definitions. To those, however, I've added societal value and ethics.

At first glance, each of the four might appear distinct from the other three. However, I view them as inextricably intertwined. In some cases, the relationships among the four elements are hierarchical; for example, in order for a study to contribute meaningfully to society, it must have been designed and carried out using rigorous methods. In other cases, two or more elements simply coincide or overlap as I illustrate throughout this paper. As I introduce each of the four elements, I will also refer to some of the relevant evidence to date that has assessed them. Although there is reason to believe that we, as a field, are improving, there is also substantial evidence of a lack of quality in a number of areas.

Transparency

As I alluded to above, transparency is what allows us to evaluate – and is therefore a prerequisite for – every other facet of quality. Indeed, as argued by the Open Science Collaboration (2015), transparency is also critical to the trust that society places in scientific institutions and outputs. And from a synthetic perspective (i.e., one that looks for overarching patterns and trends across studies), transparency in terms of thorough description of procedures, analyses, and data is necessary for secondary research and for replicability and reproducibility (Hiver & Al-Hoorie, 2020; In'nami et al., 2022; Marsden, 2020; Norris & Ortega, 2000; Porte & McManus, 2019).

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

Recognizing the value of transparency, a number of institutional and fieldwide initiatives that encourage transparency have been observed in recent years. Sin Wang Chong and Meng Liu recently launched a Research Network (ReN) with the Association Internationale de Linguistique Appliquée (or "AILA," the International Association of Applied Linguistics) on Open Scholarship, for example, and the theme of the British Association for Applied Linguistics 2023 conference was "Open Applied Linguistics." Journals have certainly led here as well. Over 20 years ago, *TESOL Quarterly* published guidelines tailored specifically to quantitative and qualitative research (Chapelle & Duff, 2003), which they then updated and expanded upon in 2016 (Mahboob et al., 2016). And nearly a decade ago, *Language Learning* commissioned a fairly thorough set of guidelines for reporting quantitative results (Norris et al., 2015). Another prime example of a journal-led initiative is that some titles, such as *Language Testing, Language Learning*, and *Applied Psycholinguistics*, have begun requiring authors to employ certain open science practices (Harding & Winke, 2022). *Applied Psycholinguistics* also recently appointed one of its Associate Editors, Amanda Huensch, as the journal's "open science guru" (my term).

We, as a field, have also seen a number of different researcher-led initiatives toward greater transparency. I'll name just a few that I'm familiar with, but there are surely others worthy of recognition. Kris Kyle's suite of NLP tools (https://www.linguisticanalysistools.org/) comes immediately to mind, along with the many resources for second language (L2) speech research curated and hosted by Kazuya Saito and colleagues (http://sla-speech-tools.com/), the Task Bank (https://tblt.indiana.edu/index.html), hosted by Laura Gurzynski-Weiss, and the IRIS Database (https://www.iris-database.org/), launched in 2011 by Alison Mackey and Emma Marsden, well before anyone was talking about "open science" in applied linguistics (Marsden et al., 2016). Recognizing the importance of transparency-related practices, some authors now flag efforts such as open data and materials on their websites and CVs.

In light of these bottom-up and top-down efforts, it is not surprising that several aspects of our reporting practices have improved in recent years (e.g., Wei et al., 2019). But we have a long way to go in terms of reporting, visualizing, sharing, and making data available (e.g., Larson-Hall, 2017; Vitta et al., 2022). Methodological syntheses that investigate reporting practices have invariably observed deficiencies in, for example, the availability of reliability estimates (Al-Hoorie & Vitta, 2019; Sudina, 2021, 2023b), statistical assumptions (Hu & Plonsky, 2021), and potential conflicts of interest (Isbell & Kim, 2023). Failing to report these types of information obstructs our ability to assess methodological rigor, thus demonstrating the link between these two elements of study quality.

There is also survey-based evidence linking transparency and the other elements of quality. In a recent survey-based study, 94% of the sample in Isbell et al. (2022; $N = 351$) admitted to one or more "questionable research practices" (QRPs), many of which concerned reporting practices. For example, 11% had not reported a finding because it ran counter to the literature and 14% had avoided reporting a finding because it contradicted their own or a colleague's previous research. Even more concerning, 43% excluded nonsignificant results and 44% withheld methodological details in a write-up to avoid criticism (see similar results for the prevalence of these QRPs in Larsson et al., 2023). Critically, these "sins of omission" and other QRPs are not just a

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

matter of transparency; they introduce "research waste" (Macleod et al., 2014; see also Isaacs & Chalmers, in press), they distort the published record, and they pose a serious ethical dilemma for the authors and for the field.

Summing up on this first of four facets for the proposed framework of study quality, there is some momentum behind and evidence for recent increases in transparency. However, both synthetic and survey-based data point to fact that deficiencies in this area are widespread, a problem I attribute at least in part to a lack of fieldwide reporting standards. Compounding my concern here is the fact that thorough and honest reporting is a prerequisite for assessing the three other elements of study quality including methodological quality, which I will now address.

Methodological rigor

The inclusion of methodological rigor in a framework for study quality probably seems like a foregone conclusion. Methodological flaws naturally present threats to the validity of our findings and any corresponding inferences or implications we might draw from those findings, but there are parts of this element that may be less obvious. For example, the methodological choices that we make – many of which may seem equally viable – often have direct effects on study outcomes.

As shown in numerous meta-analyses (e.g., Li, 2015; Plonsky et al., 2020), and as articulated by Vacha-Haase and Thompson (2004), "effect sizes are not magically independent of the designs that created them" (p. 478). It follows naturally, then, that our methodological decisions should be based on the quality of the evidence they provide rather than convenience or convention. There are, of course, also practical considerations that will play a role in our methodological choices. For example, we might identify a particular school as ideal for collecting data, but we cannot conduct a study there if the administration will not grant us access. Larsson et al.'s (2023) survey found that applied linguists' choices regarding designs, samples, instruments, and analyses are regularly based on ease and familiarity. Findings like these, along with compelling arguments put forth by Kubanyiova (2008, 2013) and Ortega (2005), among others, have led me and others to view virtually all methodological choices through the lens of ethics (see Plonsky et al., in press and Yaw et al., 2023).

The methods–ethics link is particularly striking in the context of two types of choices related to sampling. First is size. In the context of quantitative research, larger samples are needed to arrive at more accurate and stable results. As the graduate students in my classes have heard me say many times, smaller samples are too "bouncy" – a metaphor I use to emphasize the instability in quantitative outcomes when considering smaller groups of participants. Despite frequent calls to rectify the situation, small samples are exceedingly common (e.g., Hiver et al., 2022; Loewen & Hui, 2021; Nicklin & Vitta, 2021; Norouzian, 2020; Plonsky, 2013). Of course, a smaller sample can allow for a richer, fuller set of analyses when working with qualitative data. However, quantitative findings based on small samples present a direct threat to internal (and, by extension, external) validity. Publishing such results, unmitigated and unaccounted for (i.e., without sufficient recognition of the corresponding limitations and threats to validity), introduces noise and error into the published record and is, in my view, unethical.

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

The second choice regarding sampling that I want to highlight concerns not *how many* but *who* is included in our samples. Here, too, methodological syntheses, meta-analyses, and three second-order reviews of sampling and participant demographics in applied linguistics have shown that, "Most of what we know […] pertains to formal learning by (highly literate) adolescents and adults in schools and universities" (Ortega, 2009, p. 145; Andringa & Godfroid, 2020; Bylund et al., 2023; Plonsky, 2023b; see also Bigelow & Tarone, 2004, for evidence of longstanding concerns in this area in applied linguistics, and Henrich et al., 2010, for similar concerns elsewhere in the social sciences). Also striking is our lack of empirical attention to L1–L2 pairings that don't involve English. For example, 23 of the 27 studies in Goetze and Driver's (2022) meta-analysis on the relationship between L2 achievement and self-efficacy were concerned with English as the target language. The fact that many meta-analyses and other secondary analyses explicitly limited themselves to papers written in English further exacerbates this problem.

At first glance, the population of interest for a given study might seem somewhat innocuous or arbitrary. "Learners are learners," we might rationalize. But this is not true, unless we only care about language learners from within a tiny sliver of humanity. In other words, sampling is not a *neutral* choice. The decisions we make regarding who to study greatly limit our ability to contribute to theory as well as to practice beyond narrow and often very privileged settings. This methodological choice puts us in a position where we have not adequately been able to serve the scientific or practitioner communities (see again, compelling arguments to this effect by Bigelow & Tarone, 2004, among others), thus failing in two of the adjacent elements of study quality: societal value and ethics.

Before moving on, I need to recognize here that studying populations beyond what we have been doing may require special considerations in terms of educational and research cultures, instrument (re)validation (e.g., measurement invariance; see Sudina, 2023a), and so forth. Recruiting non-convenience samples will be challenging for some, but it is the right thing to do, both for the findings we will obtain and for the societal and scientific contributions we will be able to make.

Instrumentation represents another often-overlooked aspect of methodological quality. One particular concern that I and others have is related to the validity of our tools, which cannot be assumed, given the fact that most of what we measure is both latent and qualitative in nature. I'm not at the first to make this observation. In fact, to reinforce this point, I've complied a short collection of quotes that express concern over the lack of validity evidence in the field.

1. Chapelle (2021): "[scale] validation should be of central importance for the credibility of research results" (p. 11);
2. Cohen and Macaro (2013): "There is perhaps an unwritten agreement that readers will accept measures used in an SLA study at face value …" (p. 133);
3. Ellis (2021): "While researchers have always recognized this issue [validity in SLA measurement], they have largely ignored it …" (p. 197);
4. Norris & Ortega (2012): "Problematic … is the tendency to assume – rather than build an empirical case for – the validity for whatever assessment method is adopted" (pp. 574–575);

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

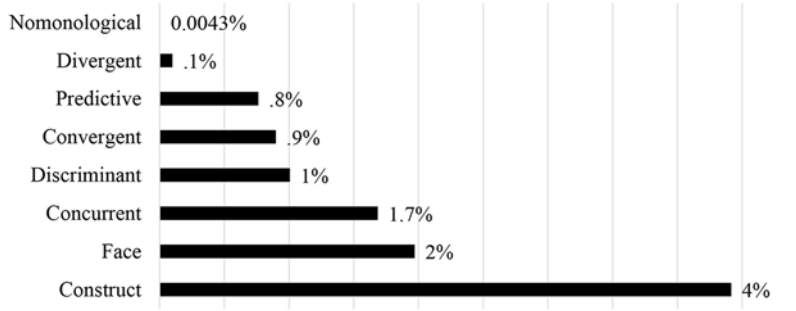

Figure 2. Percentage of applied linguistics articles that mention different facets of validity.

5. Schmitt (2019): "Most vocabulary tests are not validated to any great degree" (p. 268).

Are the concerns of these prominent scholars justified? This question has been addressed in at least four different ways. First, taking a synthetic approach, Sudina (2023b) found very little evidence of convergent, divergent, construct, and criterion-related validity in the context of studies of L2 anxiety and L2 willingness to communicate. Second is collecting (or reanalyzing) primary data to empirically examine the psychometric properties of scales that are in use, as exemplified by Al-Hoorie et al. (in press). Third is addressing researchers directly, as Larsson et al. (2023) did in their survey-based approach asking about their engagement with 58 QRPs. The second most frequently reported QRP was "Choosing a design or instrument type that provides comparatively easy or convenient access to data instead of one that has a strong validity argument behind it."

A fourth approach to understanding the extent to which the field has supplied a sufficient validity argument for its instruments involves a combination of corpus and synthetic methods (see Plonsky et al., 2023). With the help of several graduate students and RAs, I first collected a corpus of research published in 22 mainstream applied linguistics journals ($K = 23{,}142$). I then converted the PDFs to txt files and queried them using AntConc for terms that represent eight different facets of validity (e.g., "nomological, divergent, … + validity"). Figure 2 presents the percentage of articles in which each token appears at least once. The most frequent among these is "construct validity," which appears in just 4% of the sample. In other words, only one in every 25 articles even mentions construct validity. I'm sure there are articles that presented evidence of their instruments' construct validity without explicitly using the term "construct validity," but the converse is also true; that a given article mentioned concurrent validity, for example, does not mean that it necessarily provided evidence thereof.

There are clear connections to be drawn between this example of (a lack of) methodological rigor and the other elements of study quality in the model I'm proposing. Perhaps most immediate is the link between rigor – which appears, here, in the form of addressing instrument validity – and transparency. It is incumbent upon researchers to provide explicit evidence of the validity of their measures (see, for example, Arndt, 2023; Driver, in press). Second, a lack of validity evidence for a study's measures puts

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

into question its potential value to scholarly and/or non-scholarly stakeholders. And third, research that is of low methodological quality is, in my view, unethical in that it wastes time, energy, money, and other resources that could be spent on producing meaningful, higher-validity evidence to inform theory and/or practice. Low quality research is also unethical in that it can mislead future empirical efforts, leading to future inefficiencies.

Societal contribution

High quality research necessarily contributes to society and to the public good. It is not enough for us to produce rigorously executed studies and to report them thoroughly. Our research needs to lead to meaningful if also incremental advances in knowledge.

What do I mean here by "society?" I'm not saying that research can only be considered to be of high quality if it is relevant to the general public; we (academics) are part of society, too. But we are already pretty good at producing research that we, as applied linguists, care about and learn from. We need to expand our audience. Applied linguistics has a long history of borrowing from and relying on neighboring disciplines as a source of both theory and methods. This practice has served us well, but we have done very little, in my view, to give *back* to those same fields. Despite the broad relevance of language and the many language-related phenomena that we study, the field of applied linguistics is virtually unknown to our colleagues across campus. One exception here is the work of Kazuya Saito and colleagues whose cross-disciplinary collaboration and exceptionally rigorous empirical efforts have led to inroads outside of applied linguistics (e.g., Saito et al., 2022). Other noteworthy examples include the works of scholars such as Scott Jarvis, Aneta Pavlenko, and Jesse Egbert, who work with and contribute in very meaningful ways to current legal scholarship and to high-stakes cases being argued in court (e.g., Pavlenko et al., 2020; Tobia et al., 2023).

If applied linguistics remains largely unknown to our university colleagues, it is virtually invisible to the general public – even to some of the nonacademic audiences our research is most relevant for, such as language teachers (Marsden & Kasprowicz, 2017; Sato & Loewen, 2022). To me, our inability to reach these audiences speaks to a lack of quality both on at the level of individual studies as well as our field as a whole (see Coss & Hwang, 2024, for an analysis of the quantity, salience, and quality of pedagogical implications sections in 118 articles published in *TESOL Quarterly*).

I also want to recognize, very briefly, five public-facing projects that applied linguists have launched in recent years.[1]

1. The *OASIS Database* (https://oasis-database.org/) provides freely accessible, one-page summaries of applied linguistics articles written in non-jargony prose (Marsden et al., 2018a). The repository contains over 1,600 summaries which have been downloaded over 65,000 times as of this writing.
2. *TESOLgraphics* (https://tesolgraphics5.wixsite.com/tesolgraphics), currently led by Sin Wang Chong and Masatoshi Sato, provides infographic summaries of secondary research that is relevant for practitioners (see Chong, 2020). The infographics are attractive, professional, and can be read in less than 5 minutes.

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

The project directors have recently started hosting talk-show-styled interviews with authors to discuss timely topics such as the use of chatbots in the L2 classroom.

3. Developed and hosted by graduate students at University of Hawaii, *Multiʻōlelo* is a multimedia, multilingual platform for sharing language-related projects ranging from poems to podcasts (https://multiolelo.com/). One of the founders, Huy Phung, received American Association for Applied Linguistics (AAAL)'s 2022 Distinguished Service and Engaged Research Graduate Student Award for his work on Multiʻōlelo.
4. *Heritage by Design* (https://rcs.msu.edu/2023/05/24/heritage-by-design-podcast/) is a podcast, available on major streaming platforms, that seeks to "build up the community [of heritage speakers] and show the struggles and the beauty of being heritage by design." The hosts – Gabriela DeRobles, Jade Wu, and Meagan Driver – are scholars but the episodes are personal, disarming, and free from the airs of "academese" (i.e., the highly specialized and dense language typically used in academic settings).
5. Háblame Bebé (https://hablamebebe.org/), launched by Melissa Baralt and collaborators, is a mobile app designed to help Hispanic mothers use more Spanish to enhance the amount and type of language input ("language nutrition"), promote bilingualism, and assess linguistic and developmental milestones (see Baralt et al., 2020).

Producing new knowledge that is meaningful and useful (i.e., that contributes to society), whether for theory advancement, for practical matters, for the public good, or to advance justice, equity, diversity, and inclusion, is also an *ethical* issue (see Ortega, 2012). We all use public resources of one form or another, so we owe it to society to give back. In addition, most of us in applied linguistics have undergone extensive graduate studies and specialized training. To not at least attempt to contribute to society, therefore, is a waste of those resources and, hence, a breach of ethics. Another one of the speeches I often give in my graduate classes goes something like this:

> I fully hope and expect you all to publish your final papers from this class. Doing so is not only good for your careers as academics, it's your ethical duty. If your research is well motivated and well conducted, you owe it our community to make your findings known. Not doing so amounts to withholding potentially valuable knowledge and is unethical.

My soapbox speech (approximated here, as I've given the same one many times) aligns with one of Macleod et al.'s (2014) areas of "research waste," namely "publication and dissemination of results," discussed recently in Isaacs and Chalmers (in press). At the same time, we shouldn't be content to spend years of our lives producing papers that live (and die?) on a server somewhere in the Pacific Ocean either. That's wasteful too.

There are also immediate links between this facet of study quality and both rigor and transparency. As Gass et al. (1998) put it, "Respect for the field […] can only come through sound scientific progress" (p. 407). In other words, if a given study is not methodologically sound, it cannot contribute to any corner of society, scholarly or

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

otherwise. Nor can our research contribute if the reporting is opaque or unavailable to its target audiences.

Ethics

Study quality, in my view and according to the framework I'm proposing, involves more than methodological rigor, clear reporting, and an ability to contribute to scholarly and/or lay communities. Quality research must also be ethical.

There are many obvious ways for researchers to fail to meet and/or violate ethical norms. Acts such as plagiarism and data falsification are considered misconduct and are clearly wrong. They are also more common than we might expect. In their survey-based study, Isbell et al. (2022) found that 17% of the sample admitted to one or more of these forms of misconduct.

There are even more ways to find oneself in an ethical gray area, the vast majority of which are not covered by the "macro ethics" addressed by institutional review boards (in the U.S. context or "ethics boards" elsewhere; Kubanyiova, 2008). A growing body of recent work in applied linguistics has sought to catalog the so-called QRPs and to assess their frequency, prevalence, and perceived severity (e.g., Larsson et al., 2023; Plonsky et al., in press; Plonsky et al., 2024; Sterling et al., in preparation). These works, I should note, build on the momentum for greater attention to ethics fostered by Peter De Costa and collaborators (e.g., 2016, 2021), Maggie Kubanyiova (2008), Lourdes Ortega (2005), and others within applied linguistics (e.g., Sterling & Gass, 2017; see timeline in Yaw et al., 2023) as well as many others from outside of it (e.g., Fanelli, 2009). There is also a special issue underway in *Research Methods in Applied Linguistics*, edited by Dan Isbell and Peter De Costa, that seeks to expand our understanding of ethical concerns in applied linguistics far beyond the QRPs I've largely focused on here.

Several of the findings related to QRPs have been mentioned elsewhere in this paper in relation to other areas of study quality. For example, it is ethically questionable to suppress nonstatistically significant findings or to omit methodological complications in order to avoid receiving challenging comments during peer review, both of which are also matters of transparency. It is also questionable, in my view, to rely on public resources – whether through grants or simply by virtue of studying or working at a public institution – to produce research that fails to contribute meaningfully to the public good. I admit that it's hard to assess whether or not we are meeting this standard. At the very least, though, as a field and as individuals, we need to take a hard look in the mirror to ask whether what we're doing is really worthwhile and meaningful. I, myself, wonder about this all the time.

The future of quality

I've been thinking and writing about study quality for about 15 years. But that doesn't mean my framework is right. In fact, as the saying goes, "All models are wrong," including the one I've proposed here, I'm sure. I invite anyone who cares about study quality to work with me to refine the model, the elements that it is comprised of, and the different ways those elements can be assessed. The rest of that saying is, of course, "… but some are useful." And I very much hope that that part applies here as well! Before

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

concluding, I want to lay out very briefly a few different uses that I envision for this model:

1. *Graduate and ongoing professional training*. Training in graduate programs in applied linguistics focuses mainly on just one facet of quality – methodological rigor – and to varying degrees of depth and breadth (Gönülal et al., 2017; Loewen et al., 2020). Most textbooks and courses in research methods do address ethics but they tend to be limited to the practicalities of ethics boards (Wood et al., in press; see notable exceptions in Mackey, 2020 and Mackey & Gass, 2022). There is plenty of room for expanded and more explicit consideration of study quality in graduate curricula and in the professional development offerings of organizations, such as AAAL.
2. *Journal and fieldwide standards*. An agreed upon model of study quality in applied linguistics could also be used to develop a set of field-specific publication guidelines. Those guidelines could then be used by researcher trainers, journal reviewers, editors, and individual researchers. I would like to see such a resource – likely a *living* document that is frequently updated and revised – from an established organization such as AAAL or AILA, which could draw on the expertise of its members to produce it. To date, however, the AAAL leadership has not shown a real interest in developing any such standards despite calls, encouragement, and willingness to do so from its membership. Of course, well thought-out guidelines exist from other disciplines, such as education and psychology (e.g., Appelbaum et al., 2018), but we are not education or psychology. I also feel that devising our own field-specific guidelines will contribute to our legitimacy and establishment in the wider academic community.
3. *Future studies of study quality*. As exemplified throughout this paper, a large-and-growing body of synthetic and survey-based research has assessed different aspects of study quality with a primary focus on methodological rigor and transparency. However, this work has been carried out somewhat inconsistently. I would like to see a more organized agenda and one that addresses the other two elements of quality in this framework: ethics and societal value. Similarly, future meta-analyses, methodological syntheses, and bibliometric analyses (see Plonsky, 2023a) might consider taking on this framework as a way to decide which aspects of quality to code for. The element I think we know the least about is our value, as a field, to society. It would be useful to assess the extent to which we have contributed to other disciplines. What evidence is there that the field of applied linguistics has made meaningful and demonstrable contributions to practical realms and/or other scientific domains? Does anyone even know that we exist?

Conclusion

My main goal in writing this paper was to lay out a conceptual framework for the notion of study quality in applied linguistics. The framework is multidimensional, consisting of four subconstructs: methodological rigor, transparency, societal value, and ethics. I also believe that this model is practical (realistic), operationalizable (comprised of

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

measurable constructs), and actionable (relevant for training and professional development).

I want to be clear, though, that I am entirely open to suggestions for how the framework could be modified, expanded, or reconceived. I'd like to think that the principles of methodological rigor, transparency, societal value, and ethics pertain to all areas of this "big tent" field of ours. But I'm happy to be told that I'm wrong in the name of arriving at a more comprehensive definition and operationalization of quality for all of applied linguistics. That precise task is for this or any field, I believe, both an intellectual and ethical imperative.

Note

1. I have begun drafting and plan to write a book on language and linguistics for kids. If anyone is reading this, please feel free to hold me accountable!

References

Al-Hoorie, A. H., Hiver, P., & In'nami, Y. (in press). The validation crisis in the L2 motivational self system tradition. *Studies in Second Language Acquisition*.

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*(6), 727–744.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 3–25.

Arndt, H. L. (2023). Construction and validation of a questionnaire to study engagement in informal second language learning. *Studies in Second Language Acquisition*, *45*(5), 1456–1480.

Baralt, M., Mahoney, A. D., & Brito, N. (2020). Háblame Bebé: A phone application intervention to support Hispanic children's early language environments and bilingualism. *Child Language Teaching and Therapy*, *36*(1), 33–57.

Bigelow, M., & Tarone, E. (2004). The role of literacy level in second language acquisition: Doesn't who we study determine what we know? *TESOL Quarterly*, *38*(4), 689–700.

Burston, J., Athanasiou, A., & Giannakou, K. (2024). Quantitative experimental L2 acquisition MALL studies: A critical evaluation of research quality. *ReCALL*, *36*(1), 22–39.

Bylund, E., Khafif, Z., & Berghoff, R. (2023). Linguistic and geographic diversity in research on second language acquisition and multilingualism: An analysis of selected journals. *Applied Linguistics*.

Chapelle, C. (2021). Validity in language assessment. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 11–20). Routledge.

Chapelle, C., & Duff, P. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly*, *37*(1), 157–178.

Chong, S. W. (2020). The role of research synthesis in facilitating research–pedagogy dialogue. *ELT Journal*, *74*(4), 484–487.

Cohen, J. (1997). The earth is round ($p < .05$). In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21–35). Lawrence Erlbaum.

Cohen, A. D., & Macaro, E. (2013). Research methods in second language acquisition. In E. Macaro (Ed.), *The Bloomsbury companion to second language acquisition* (pp. 107–133). Bloomsbury.

Coss, M., & Hwang, H.-B. (2024). Issues with pedagogical implications in applied linguistics research: A mixed-methods systematic evaluation. *Research Methods in Applied Linguistics*, *3*(1), 100094.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

De Costa, P. I. (Ed.) (2016). *Ethics in applied linguistics research: Language researcher narratives*. Routledge.

De Costa, P., Sterling, S., Lee, J., Li, W., & Rawal, W. (2021). Research tasks on ethics in applied linguistics. *Language Teaching*, *54*(1), 58–70.

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

Driver, M. Y. (in press). Measuring and understanding linguistic insecurity in heritage and foreign language contexts: Design and validation of a novel scale. *Journal of Multilingual and Multicultural Development*.

Ellis, R. (2021). A short history of SLA: Where have we come from and where are we going? *Language Teaching*, *54*(2), 190–205.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*, *4*(5), e5738.

Gass, S., Fleck, C., Leder, N., & Svetics, I. (1998). Ahistoricity revisited: Does SLA have a history? *Studies in Second Language Acquisition*, *20*(3), 407–421.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*(2), 245–258.

Goetze, J., & Driver, M. (2022). Is learning really just believing? A meta-analysis of self-efficacy and achievement in SLA. *Studies in Second Language Learning and Teaching*, *12*(2), 233–259.

Gönülal, T., Loewen, S., & Plonsky, L. (2017). The development of statistical literacy in applied linguistics graduate students. *International Journal of Applied Linguistics*, *168*(1), 4–32.

Harding, L., & Winke, P. (2022). Innovation and expansion in language testing for changing times. *Language Testing*, *39*(1), 3–6.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Hiver, P., & Al-Hoorie, A. H. (2020). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*(1), 48–102.

Hiver, P., Al-Hoorie, A. H., & Evans, R. (2022). Complex dynamic systems theory in language learning: A scoping review of 25 years of research. *Studies in Second Language Acquisition*, *44*(4), 913–941.

Hu, Y., & Plonsky, L. (2021). Statistical assumptions in L2 research: A systematic review. *Second Language Research*, *37*(1), 171–184.

In'nami, Y., Mizumoto, A., Plonsky, L., & Koizumi, R. (2022). Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics*, *1*(3), 100030.

Isaacs, T., & Chalmers, H. (in press). Reducing 'avoidable research waste' in applied linguistics research: Insights from healthcare research. *Language Teaching*.

Isbell, D., Brown, D., Chen, M., Derrick, D., Ghanem, R., Gutiérrez Arvizu, M. N., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *Modern Language Journal*, *106*(1), 172–195.

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics*, *2*(3), 100060.

Klein, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). American Psychological Association.

Kubanyiova, M. (2008). Rethinking research ethics in contemporary applied linguistics: The tension between macroethical and microethical perspectives in situated research. *Modern Language Journal*, *92*(4), 503–518.

Kubanyiova, M. (2013). Ethical debates in research on language and interaction. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 2001–2008). Blackwell.

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *Modern Language Journal*, *101*(1), 244–270.

Larsson, T., Plonsky, L., Sterling, S., Kytö, M., Yaw, K., & Wood, M. (2023). On the frequency, prevalence, and perceived severity of questionable research practices. *Research Methods in Applied Linguistics*, *2*(3), 100064.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*(3), 385–408.

Li, S. (2023). Working memory and second language writing: A systematic review. *Studies in Second Language Acquisition*, *45*(S3), 647–679.

Lindstromberg, S. (2022). P-curving as a safeguard against p-hacking in SLA research: A case study. *Studies in Second Language Acquisition*, *44*(4), 1155–1180.

Loewen, S., Gonulal, T., Isbell, D. R., Ballard, L., Crowther, D., Lim, J., & Tigchelaar, M. (2020). How knowledgeable are applied linguistics and SLA researchers about basic statistics? Data from North America and Europe. *Studies in Second Language Acquisition*, *42*(4), 871–890.

https://doi.org/10.1017/S0267190524000059 Published online by Cambridge University Press

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *Modern Language Journal*, *105*(1), 187–193.
Mackey, A. (2020). *Interaction, feedback and task research in second language learning: Methods and design*. Cambridge University Press.
Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design* (3rd ed.). Routledge.
Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *Lancet*, *383*(9912), 101–104.
Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., & De Costa, P. I. (2016). TESOL quarterly research guidelines. *TESOL Quarterly*, *50*(1), 42–65.
Marsden, E. (2020). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 15–28). Routledge.
Marsden, E., Alferink, I., Andringa, S., Bolibaugh, C., Collins, L., Jackson, C., & Plonsky, L. (2018a). *Open Accessible Summaries in Language Studies* (OASIS) [Database]. https://www.oasis-database.org.
Marsden, E., & Kasprowicz, R. (2017). Foreign language educators' exposure to research: Reported experiences, exposure via citations, and a proposal for action. *Modern Language Journal*, *101*(4), 613–642.
Marsden, E., Mackey, A., & Plonsky, L. (2016). Breadth and depth: The IRIS repository. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). Routledge.
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*(2), 321–391.
Merton, R. K. (1942). The normative structure of science. In R. K. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–278). University of Chicago Press.
Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data reanalysis. *Annual Review of Applied Linguistics*, *40*, 26–55.
Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *Modern Language Journal*, *105*(1), 218–236.
Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *42*(4), 849–870.
Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Supp. 1), 97–126.
Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.
Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3–50). Benjamins.
Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). Routledge.
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.
Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, *349* (6251).
Ortega, L. (Ed.) (2005). Methodology, epistemology, and ethics in instructed SLA research SLA research: An introduction. *Modern Language Journal*, *89*(3), 317–327.
Ortega, L. (2009). *Understanding second language acquisition*. Routledge.
Ortega, L. (2012). Epistemological diversity and moral ends of research in instructed SLA. *Language Teaching Research*, *16*(2), 206–226.
Pavlenko, A., Hepford, E., & Jarvis, S. (2020). An illusion of understanding: How native and non-native speakers of English understand (and misunderstand) their *Miranda* rights. *The International Journal of Speech, Language and the Law*, *26*(2), 181–207.
Plonsky, L. (2011). *Study quality in SLA: A cumulative and developmental assessment of designs, analyses, reporting practices, and outcomes in quantitative L2 research*. [Unpublished doctoral dissertation]. Michigan State University.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687.
Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*(1), 450–470.
Plonsky, L. (2023a). Introducing bibliometrics in applied linguistics. *Studies in Second Language Learning and Teaching*, *13*(4), 721–726.
Plonsky, L. (2023b). Sampling and generalizability in Lx research: A second order synthesis. *Languages*, *8*(1), 75.
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.
Plonsky, L., Hu, Y., Sudina, E., & Oswald, F. L. (2023). Advancing meta-analytic methods in L2 research. In A. Mackey & S. Gass (Eds.), *Current approaches in second language acquisition research* (pp. 304–333). Wiley Blackwell.
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. De Costa, A. Rabie-Ahmed & C. Cinaglia (Eds.), *Ethical issues in applied linguistics scholarship*. Benjamins.
Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (2020). A methodological synthesis of judgment tasks in second language research. *Second Language Research*, *36*(4), 583–621.
Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. Routledge.
Saito, K., Petrova, K., Suzukida, Y., Kachlicka, M., & Tierney, A. (2022). Training auditory processing promotes second language speech acquisition. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(12), 1410–1426.
Sato, M., & Loewen, S. (2022). The research–practice dialogue in second language learning and teaching: Past, present, and future. *Modern Language Journal*, *106*(3), 509–527.
Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274.
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.
Sterling, S., Yaw, K., Larsson, T., Plonsky, L., & Kytö, M. (in preparation). *A qualitative analysis of questionable research practices*. Manuscript in preparation.
Sudina, E. (2021). Study and scale quality in second language survey research, 2009–2019: The case of anxiety and motivation. *Language Learning*, *71*(4), 1149–1193.
Sudina, E. (2023a). A primer on measurement invariance in L2 anxiety research. *Annual Review of Applied Linguistics*, *43*, 140–146.
Sudina, E. (2023b). Scale quality in second-language anxiety and WTC: A methodological synthesis. *Studies in Second Language Acquisition*, *45*(5), 1427–1455.
Tobia, K., Egbert, J., & Lee, T. R. (2023). Triangulating ordinary meaning (May 8, 2023). *Georgetown Law Journal Online*, *112*, 23–54.
Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481.
Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, *44*(5), 1424–1448.
Wei, R., Hu, Y., & Xiong, J. (2019). Effect size reporting practices in applied linguistics research: A study of one major journal. *SAGE Open*, *9*(2).
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*(4), 478–494.

Cite this article: Plonsky, L. (2024). Study quality as an intellectual and ethical imperative: A proposed framework. *Annual Review of Applied Linguistics*, 1–15. https://doi.org/10.1017/S0267190524000059