Journal of Modern Research in English Language Studies
Print ISSN: 2676-5357
Online ISSN: 2676-5985

Study Quality in Quantitative L2 Research: A Path Analysis on the Perceptions of Iranian Published Authors

Parviz Ajideh[1], Mohammad Zohrabi[2], Akbar A. Jahanbakhsh[3*]

[1]English department, Faculty of Persian literature and foreign languages, University of Tabriz, Tabriz, Iran, *parvizaj@gmail.com*

[2]English department, Faculty of Persian literature and foreign languages, University of Tabriz, Tabriz, Iran, *mohammadzohrabi@gmail.com*

[3*]English department, Faculty of Persian literature and foreign languages, University of Tabriz, Tabriz, Iran, *a.jahanbakhsh@tabrizu.ac.ir*

Article Info

Article Type:

Research Article

Received:

06/10/2020

Accepted:

14/12/2020

ABSTRACT

Inspired by the works of scholars who have done great efforts to improve study quality in second language (L2) papers in the recent decade (e.g., Hu & Plonsky, 2019; Larson-Hall, 2012, 2017; Plonsky, 2013; Norris, 2015), this paper aims to capture the perceptions of Iranian authors around issues of quality that have been emphasized by publication manuals (e.g., APA, 2010; Wilkinson, 1999) and recommendation (e.g., Norris et al., 2015). The triggering idea behind the study was that in order to adhere to standards of quality, authors, first, need to perceive the issues highly-associated with the concept. Accordingly, a questionnaire of quality developed by Larson-Hall and Jahanbakhsh (inreview) was used to capture respondents' perceptions. Out of the 1029 authors who had published quantitative papers in 10 Iranian journals from 2015 to 2019, 885 authors could be contacted and 128 answered the questionnaire. The results showed that respondents saw fundamental issues like random sampling, reliability, validity, checking normality, and reporting inferential statistics as the most highly-associated features with quality while the concerned issues by scholars, like generalizability in convenience samplings, use of delayed posttest, reporting non-significant results, and importance of visual presentations are less acknowledged. Moreover, using Structural Equation Modeling (SEM), it was revealed that a path of perception exists which starts from sampling issues and going through design and statistical to reporting practices. It was recommended that authors take the issues of study quality more seriously in both their works and the context of education.

Keywords: Iranian Authors' Perception, Quantitative research, Study Quality, Path Analysis

Cite this article: Ajideh, P., Zohrabi, M., Jahanbakhsh, A. A. (2022). Study quality in quantitative L2 research: A path analysis on the perceptions of Iranian published authors. *Journal of Modern Research in English Language Studies, 9*(1), 97-123. DOI: 10.30479/jmrels.2020.14296.1760

Publisher: Imam Khomeini International University

1. Introduction

Today, with growing numbers of researches in different fields, piles of scientific reports, whose aims are to collectively improve the understanding of the world, are pouring in. To use Cooper and Hedge's (2009) analogy, "[l]ike the artisans who construct a building from blueprints, bricks, and mortar, scientists contribute to a common edifice, called knowledge" (p. 4). However, following these huge amounts of studies as well as choosing which one to rely on as a well-developed and rigorously-practiced research is a challenging job.

Research syntheses provide essential information helping to deal with the above-mentioned concern. It is defined as "a continuum of techniques and research procedures that have been developed by social scientists with the aim of reviewing past literature systematically" (Ortega, 2015. p. 219). Although there are different focuses in research syntheses (e.g., meta-analysis, methodological research synthesis, second-order research synthesis, scoping review, etc.), the one which synthesizes and analyzes the methodological aspects of the research papers, i.e., *methodological research synthesis*, has provided invaluable information about the rigorousness or, to use Plonsky's (2013) terminology, *study quality* in quantitative studies.

The motivation behind such studies, according to Plonsky and Gonulal (2015), can be categorized into four categories: a) description of what is practiced in different journals to define the methodological culture of L2; b) description of L2 practices and evaluation of the results to improve future researches; c) examination of the relationship between practical features of the research and obtained outcomes; and d) surveying the changes in the practice of research which happens over time.

In the last decade, several studies (e.g., Gass, 2009; Larson-Hall, 2012, 2017; Norris & Ortega, 2000; Plonsky, 2013; Plonsky & Gass, 2011; Plonsky & Gonulal, 2015) intended to describe or evaluate the methodological issues in L2 research from different aspects. In Iran, Amini Farsani and his colleagues (e.g., Amini Farsani, 2017; Babaii et al., 2017; Amini Farsani & Babaii, 2018, 2020) also explored the conception and practices of TEFL research studies. In this study, relying on the logic that the essentiality of the issues of quality, first, should be realized by authors in order to be applied in the research practice, we will focus on describing the perception of Iranian authors about the viable features associated with high-quality research and examine if these perceptions are related with regards to different aspects of the methodological concerns of quality.

The concerns of quality have also been addressed by publication manuals and research and reporting sources (e.g., APA, 2010; Journal Article Reporting Standards Working Group, 2008, 2018; Wilkinson, 1999). American Psychological Association (APA) has been extensively worked on upholding the features of quality in published papers by constantly updating its standards. APA's (2010) emphasis on reporting "all relevant results" (p. 32), encouragement of the use of *new statistics,* i.e., effect size and confidence intervals as "the best strategy" (p. 34), and recommendation of reporting "direction, magnitude, degrees of freedom, and exact p level, even if no significant effect is reported" (Journal Article Reporting Standards Working Group, 2008, p. 843) are among the emphasized issues that were used by Plonsky (2011, 2013). However, some issues like transparency in open science by sharing the data (Journal Article Reporting Standards Working Group, 2018) use of visual presentation (Larson-Hall, 2017), use of robust statistics (Larson-Hall, 2012) were not included in Plonsky's meta-syntheses.

Moreover, different research synthesis work with different lines of research. For example, Plonsky (2013) examined study quality based on design, analyzing, and reporting practices. Others (e.g., Plonsky, 2009; Wilcox, 2005; Norris & Ortega, 2007) focused on statistical tests while reporting practices were highlighted in the works of Norris & Ortega (2000, 2006). Despite the growing emphasis on the facets of quality, the findings of researchers (e.g., Amini Farsani, 2017; Amini Farsani & Babaii, 2020; Al-Hoorie & Vitta, 2019; Plonsky, 2013, 2014a, 2014b; Plonsky & Ghanbar, 2018; Larson-Hall, 2017; Larson-Hall & Plonsky, 2015; Norouzian & Plonsky, 2018) have shown that lots of concerns still existing in L2 papers, warning about the potential dangers of not adhering to these standards. While these studies depict the current challenges in L2 studies, most of the previous studies have suggested 'what should be done' not what should be conceptualized in practice?

The aspects that have been considered as quality features in researches are the ones that are provided by publication associations. For example, the protocol developed by Plonsky and Gass (2011) which was later modified and used in Plonsky (2013) was developed based on three sources: a) previous instruments which aimed to address quality in research; b) recommendations by publishing associations such as *American Psychological Association* (APA), i.e. Journal Article Reporting Standards Working Group (2008, 2018) and Wilkinson & the Task Force on Statistical Inference (1999); c) the results of quality assessment from other fields; and d) journal guidelines and editorials (e.g., Chapelle & Duff, 2003; Ellis, 2000). However, no attempts, to the best of researchers' knowledge, has been done to examine

how these quality features are perceived by the authors of L2 journals. Given that the authors are the primary producers of research papers that may or may not fully adhere to standards of study quality, and considering the fact that some of these authors also act as reviewers or editors in different journals, their conception of quality may play a determining and influential role in the quality of the final outcome.

Moreover, as Larson-Hall and Jahanbakhsh(inreview) assert, although the different facets of study quality can be examined separately, quality should be considered as a whole phenomenon. That is why in this study, alongside describing the authors' inclinations, the interaction among the constructing facets of study quality, as it is perceived by authors, are analyzed using path analysis. Therefore, the research questions of the study are:

1. How is the features of study quality perceived by Iranian L2 authors?
2. Is there a path among the study quality facets as they are perceived by Iranian L2 authors?

2. Literature Review

Quality is a slippery concept, the measurement of whose aspects is a challenging job. In this research, the methodological aspects are focused as the existing literature provides reliable sources for the identification of the features to be evaluated. In what follows, these aspects are briefly introduced and the concerns around them are addressed

2.1. Sampling and Power Issues

The practical phase of any quantitative study often starts with selecting the sample. The sampling scheme, including the size and the type of sampling, is an essential factor in determining the quality of inference. Small sample sizes would result in debilitating statistical power, especially with many advanced statistical procedures (Brown, 2015). The review of the literature indicates that's scholars have been warning about the typical use of small samples in researches (e.g., Larson-Hall, 2010; Plonsky & Gass, 2011; Plonsky & Oswald, 2012; Plonsky, 2013, 2014b). Although Amini Farsani and Babaii (2020) report that the sample sizes used in Iranian MA theses were sufficiently large, they reported the rarity of systematic procedures like statistical power analysis in determining the sample size. According to Plonsky (2013), statistical power refers to “the probability of observing a statistically significant relationship given that the null hypothesis is false” (p. 29). The problem of sample size is two-folded, both of which threaten the internal validity of the results: a) if the sample is large enough, it is claimed that almost any statistical test may reach significant results or what is known

as Type I error (Hudson & Llosa, 2015; Plonsky, 2013; Norris, 2015); b) with small samples, it is likely that the statistical tests fail to capture the existing significant results (Type II error) since effect size and sample size together have an inverse impact on power (Plonsky & Gonulal, 2015, p. 16), having small samples often lead to type II error. The methodological syntheses of Plonsky (2013) and Plonsky (2014b) showed that only about 1 percent of the examined research papers used power analysis.

The type of sampling is also crucial to the external validity of the research. The convenience sampling falls short in including participants of various demographics. According to Tarone (2013), the consequence of such limited sampling is a severe problem with the generalizability of the results. The reports provided by the research synthesists (e.g., Norris & Ortega, 2000, Plonsky, 2014b) indicates that a large proportion of L2 papers are conducted on young adult university students living in the USA, west Europe, or east Asia. Therefore, the generalizability of the results can be considered limited mainly to those specific contexts (Ortega, 2009). Amini Farsani and Babaii (2020) also reported that in 91% of Iranian theses, the sampling did not follow a probabilistic technique.

The last issue of sampling relates to participants' selection procedure. While probability, or random, samplings are the favorite one among the researchers as they result in both higher levels of generalizability, as they include participants with various demographic background, and better control over extraneous variables by allowing initial variability to be distributed within the groups (Hudson & Llosa, 2015). In many cases, however, random sampling is not possible. The researchers may, in such situations, try to control the effect of extraneous variables by random assignment of the participants into groups or some design controls like the use of pretesting and/or comparison groups to take into account the initial differences and deal with the threats to internal validity. However, the results of Plonsky (2013) show that only 50% of the 609 studies he reviewed used random assignment either by group (intact classes) or individuals.

2.2. Design-Related Issues

Creswell (2014) defined research design as the general plan for linking the conceptual research problems to the relevant and achievable empirical research. It is also an examination that gives specific directions for procedures in research. The design of experimental studies is concerned with setting the experiment conditions in a way that the results are reliable. This includes using controlling the effects of extraneous variables, competent operationalization of the treatment, using multiple measurements to increase the accuracy of results, setting the experiment context compatible with the

nature of the experiment, and, overall, doing whatever can be done to preserve high validity and reliability of the results. L2 research, however, is shown to have concerning problems like low reliability of the measurements and regular use of intact group assignments (Chaudron, 2001). Plonsky (2013) reports that only 47& of the experimental studies use random assignments. Plonsky (2014b) pointed out that the use of pretesting and delayed posttesting, and control groups in L2 researches has increased. However, he warns that only a small proportion of experimental studies that are classroom-oriented are being conducted in the required settings, i.e., actual classes.

2.3. Issues Related to Statistical Tests

The two major types of quantitative L2 research are correlational and comparative studies (Plonsky, 2013). The routine procedure in these studies is to formulate a null hypothesis, which states the lack of relationships between the variables, is formulated. The answer to the research question will be, then, sought by testing this null hypothesis. A null hypothesis testing tries to provide the degree of certainty/uncertainty that the possible changes in the population are due to the effect of the independent variable, or the possible relationship between the variables is not by chance. There are several issues regarding the use of statistical tests, though. Use of statistical tests is almost inevitable in quantitative research.

The first issue is closely related to the sampling and design issues introduced above. Plonsky (2014a) reported that the median sample size per SLA papers increased from 56 to 62 within two decades (i.e., from the 1990s to the 2000s), however, it coincided with the increase in the average number of groups in studies, which moderated the increase in the power of samples. As mentioned before, both too small and too large sample sizes may lead to Type II and Type I errors, respectively. Moreover, the incompetent operationalization of the experiment, like the problem with the length of the treatment can directly affect the results (Hudson & Liosa, 2015).

The second issue is concerned with the problems of multiple statistical tests. The review of L2 papers by Plonsky (2014b) has shown a significant increase in the numbers of statistical tests per paper. While researchers may be willing to include wider scopes in their papers, they may be neglecting the fact that using multiple statistical tests on the same data adds new errors to the tests, which are often not considered by the researchers. Given that a large proportion (78.77%) of statistical tests in applied linguistics studies, as reported by Khany and Tazik (2019), are basic ones (which, unlike advanced analyses, do not compensate for the change in alpha level), this issue must be taken very seriously. Larson-Hall (2012)

explains how these errors are compounded, if they are not adjusted for the errors, and may lead to Type I error.

Next is the issue of assumptions. For the results of a statistical test, especially a parametric one, to be reliably accurate, some pre-requisites have to be met. The common assumption among all parametric tests is the normality of distribution. However, the test-specific assumption gets more and more complicated for tests that take into account more variables and measures. An independent samples t-test, for example, requires checking the error variance equality besides the normality assumption while ANCOVA requires several other assumptions, like the linearity of the relationships, homogeneity of regression slopes as well. Moreover, the results of statistical tests are directly influenced by the violation of general or test-specific assumptions. Without having these assumptions in place, the possibility of type I errors increases (Larson-Hall & Herrington, 2010; Plonsky et al., 2015). This is while the analyses of the papers published in L2 journals have low rates of full assumptions checking (17% in Plonsky, 2013 and 24% in Hu and Plonsky, in press). That is why the use of robust statistics has been encouraged (e.g., Wilcox, 2005, Larson-Hall & Herington, 2010). The measures, like trimmed mean, M-estimator, and Bootstrapping, are robust to the violations of normality and exitance of outliers.

The final issue that has been increasingly warned about by scholars (e.g., Larson-Hall, 2012; Norris, 2015; Plonsky, 2015) is the overreliance on null hypothesis statistical testing (NHST). This refers to the dichotomous understanding of the *p*-value resulted from statistical tests as representing either a significant relation/difference between the variables or no considerable relation at all. As a result, a researcher sees the non-significant results as a failure which is not worth reporting. As Larson-Hall (2012) elaborates, there is a fine difference between *not significant* and *insignificant* results. She explains that the cut-off alpha value of 0.05 is just an arbitrary value and, in some contexts is not a good judge of the existence of an effect or relationship. That is why the researchers are encouraged to rely on *new statistics*, i.e., confidence intervals and effect size, to base their interpretations on the power obtained from the test rather than merely focusing on the value results (Cumming, 2012, Norris, 2015). While the use of new statistics is encouraged in reporting manuals and recommendations (e.g., APA, 2010; Norris et al., 2015; Wilkinson, 1999), Plonsky's (2013) results showed a low percentage for reporting effect size (26%) and confidence intervals (5%) in L2 papers.

2.4. Transparency in Reporting Practices and Data Sharing

Journal guides and standard manuals regularly publish what they see as necessary information to be included in a scientific report. For instance, in its 2018 guideline, APA's Journal Article Reporting Standards (JARS) provided a comprehensive list of what is needed to be included in each part of a quantitative study. The essentiality of these reports has been emphasized by scholars (e.g., Norris et al., 2015). The very first results needed to be reported are the comprehensive presentation of the data through descriptive statistics. These statistics are both necessary to give a general picture of the results in the primary studies and contain the data required for conducting secondary studies (meta-analyses). Plonsky's (2013) results show that more than a third of the examined L2 papers lack the necessary data description required for running meta-analysis.

Next are the inferential statistics and power analyses. As mentioned before, a priori power analysis of the sample is what is almost ignored in L2 research (e.g., Plonsky, 2013, Ziegler, 2013). The predetermination of the power would help researchers to set the alpha level compatible with the study. The logical consequence of not setting the sample power, thus, is the low rate of pre-determination of alpha (22% in Plonsky (2013) and 16-26% in Plonsky (2014a)). The inferential statistics are also missing in large numbers of the L2 studies. According to Plonsky (2013), the p-values were reported in about 35% of the L2 papers examined in his study. This was while 26% of studies reported effect size and only 5% of them reported confidence intervals for their results.

The third concern is the omission of non-significant results. As addressed above in the statistical concerns, authors may see the non-significant results as a failure and tend not to report them. The results in Plonsky's (2013) showed that 13% of studies did not report the *p*-value for non-significant results and the exact *p*-value was not reported in 51% of the papers Amini Farsani and Babaii (2020) also reported that only 31% of the studies they examined reported the exact *p*-value. Aside from not reporting the values for non-significant results, the reluctancy of authors to publishing papers with non-significant results would affect the secondary studies through what is known as *publication bias* (Rosenthal, 1979), since the number of data against the formulated hypotheses would be limited and the Type I error resulted, in turn, would lead us to do ineffective practices or follow falsely supported theories (Plonsky & Gonulal, 2015).

The final concern, although not directly related to the reporting practices, is data sharing. Although previous research syntheses have not included data sharing as a criterion of study quality, acknowledging the importance of *open science*, which is, according to Gass et al. (2020), is

closely related to transparency in reporting of data, we have added data sharing to our coding protocol. APA in its latest Journal Article Reporting Standards (2018), recognizing data sharing as one of the developing domains in the reporting practices, declared that they will venture into its standards in the future. The inspections provided by Plonsky (2015) and Plonsky et al. (2015) showed that data are more likely to get missed over time if not stored externally. The external storage such as IRIS, COCA, SPLLOC2, and MALELC keeps the data for future references and builds up the required knowledge for future researches (Vines et al., 2014).

2.5. Visual Presentation

The final concern, the visual presentation of data, is a reporting practice that we have separated since our analyses showed it is perceived as a separate component of quality by the authors. Graphs are often considered, to use Larson-Hall's (2017, p.264) words, "as a frill in a research article". This is while the graphics are very useful means to describe the frequency and dispersion of data, checking normality and test-specific assumptions, and provision of the main pattern for relationships and differences. Larson-Hall (2017) provides good examples of how graphics can be used for multiple tasks, encouraging the researchers to use graphics like boxplots. Anzures-Cabrera & Higgins (2010) explains how the numerical and visual analyses of the data, collectively, contribute to reaching more reliable and comprehensive results. The use of graphics in L2 research papers is, however, very limited. The findings of Amini Farsani and Babaii (2020), however, indicated that the visual presentation of data is theses is around 60%. It is reported by Plonsky (2013) that only about a third of the L2 papers examined in his study had used graphical presentations. Inspecting two journals of The Modern Language Journal and Language Learning, Larson-Hall (2017) also reported that a large proportion of graphics used in these journals were single-task graphs like line graphs or bar plots.

3. Method

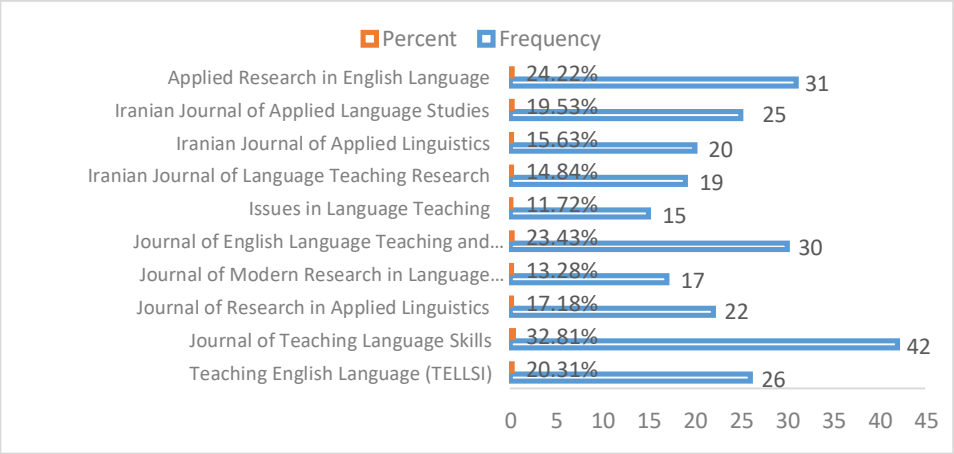
3.1. Participants

The participants of this study were authors who had published in ten high-ranked L2 journals of Iran (based on the 2017 raking provided by Iran's Ministry of Science, Research, and Technology (MSRT), which was the latest available ranking by the time the data collection was being done). The total number of published quantitative authors in Iranian journals was 1029 (. From these 1029, the email addresses for 144 authors were either not found by searching the web or not valid, leaving 885 authors who were contacted by email and asked for participation. Consequently 128 authors submitted

their answers (return rate = 14.46%). They were all Iranians, both males (N = 73) and females (N = 55) with the age range of 24 to 80 (M age = 41.16). Regarding the academic level, 11 of them were MA students/ graduates, 18 Ph.D. students/ holders, 1 postdoc researcher, 19 lecturers, 49 assistant professors, 14 associate professors, 12 professors, 2 professor emeritus, and 2 others. Moreover, 35 (27.3%) of them had published 1-5 papers, 32 (25.0%) of them 6-10, 24 (18.8%) had 11-20 papers, and the published papers of 37 (28.9%) of them were more than 20 to the date. The number of their recent publications in the ten Iranian Journals are presented in Figure 1. Note that the overall count exceeds the number of participants since some of them had published in more than one journal.

Figure 1

Frequency of Participants' Publications in Iranian Journals



It should not be left unmentioned that after initial screening, 14 respondents whose answers had standard deviations below 1 were discarded as unengaged ones, leaving the final sample with 114 participants.

3.2. Instrument

A survey questionnaire recently-developed by Larson-Hall and Jahanbakhsh (inreview) was used to capture the opinions of authors of journals about the practical features of study quality. In their study, Larson-Hall and Jahanbakhsh (inreview) used APA guidelines provided in the Journal Article Reporting Standard (2018) as well as the concerning issues of quality addressed by scholars (e.g., Larson-Hall, 2012, 2017), the summary of which is provided in the Literature Review of this article. In order to make sure of the content validity, the developed questionnaire was first sent to two experts, based on whose opinions the required changes were made. They also

sent the questionnaire to a pilot group and added a few items which were commented on by the pilot group. The final version was sent to authors published in 10 high-ranked international journals and the analyses of the 242 collected answers showed the perceptions of authors with regards to the study quality can be addressed by six factors: sampling and power, related to competence in data collection, design issues related to controlling extraneous variables, statistical testing, reporting practices and data sharing, and visual presentation. The same questionnaire was used in this study. The questionnaire was put on a Google Form page, the access link to which was sent to 50 authors randomly selected from the target group. Thirteen authors who responded to the invitation were asked to answer the items and add their comments in the space provided at the end of the questions. Based on the comments, which mostly focused on issues like provision of more detailed instruction, context-dependability of the items, or lack of sufficient explanation in some items (especially the ones related to visual presentation issues), required information was added. The finalized questionnaire had 41 items, to answer to each of which the respondents were required to select a number from 1 to 10, 1 representing the lowest association of the given item with the high-quality research and 10 presenting the highest association.

The questionnaire aimed to capture quality based on Plonsky's (2013) definition, i.e., "adherence to standards of contextually appropriate, methodological rigor in research practices and transparent and complete reporting of such practices". The results of Exploratory Factor Analysis (see Results, below) showed that the factors found in Larsen-Hall and Jahanbakhsh (inreview) maintained as the constructing ones in the context of Iran, except that the two types of design-related issues were merged into one. The perception of study quality by Iranian published authors, thus, was examined with regards to five factors: a) sampling and power issues, b) statistical tests, c) transparent reporting practices and data sharing, d) design issues, and e) visual presentation of data. The questionnaire is available in the appendix.

3.3. Procedure

The data collection started with the development of the study quality questionnaire. Then the questionnaire was sent to two experts and a pilot group and their opinions were taken into account to develop the final instrument. Then the list of authors who had published a paper from 2015 to 2019 in 10 high-ranked Iranian L2 journals was obtained and they were contacted through email to answer the questionnaire, already uploaded to a Google Form page. The email addresses were obtained through the information provided in the journals or (if not included) by searching the web. Of course, some scholars' contact information could not be obtained.

However, the researchers did their best to contact as many authors as possible.

3.4. Data Analysis

Two statistical software, namely IBM SPSS and IBM SPSS AMOS, were used to analyze the data. First, the data was pre-processed to identify and discard the unengaged participants. Then, the descriptive statistics of the data for each item and the calculated average mean for each section of the questionnaire were presented to capture the perceptions of the authors. Next, running an Exploratory Factor Analysis (EFA) with Maximum Likelihood extraction and the Promax rotation helped the researcher to start the path analysis. The aim of this analysis was to find the existing path among the facets of study quality as they were perceived by the authors; that is to say, the analysis aimed to examine if authors' perception about the association of one facet with study quality affects how the other facets are perceived. Using the pattern matrix obtained in this phase, the Confirmatory Factor Analysis (CFA) was run in AMOS, discarding the items that had low loadings to the factors. The measures of validity, reliability, and goodness of fit were calculated in this phase. Then, a structural model was formulated to examine the causal effect in the proposed path.

4. Results and Discussion

4.1. Results

The data obtained from the authors went through initial screening to identify unengaged (careless) responses. First, the demographic information was eye-screened and no suspicious answers were found. Then, the answers of participants were checked for constant or increasing/decreasing patterns of answering where no concerns were found. Then the variances of answers were examined by calculation of standard deviation for each participant's answers. Fourteen cases (9.72% of the sample) who had standard deviations below 1 were excluded in this phase. Removing unengaged responses is essential in running SEM, as they contaminate the results. In removing the unengaged answers, guidelines provided by Gaskin (2013) were followed. Table 1, below, shows the descriptive statistics of the data.

The items with the highest and lowest mean within each category are highlighted in Table 1. As reported, concerning sampling issues, random sampling, and the use of sufficiently empowered samples received the highest attention while generalizability concerns of convenience sampling had the lowest value. In the design section, ensuring the validity and reliability were the two most attended issues and the use of delayed posttest the least attended one. With regards to statistical test concerns, checking assumptions for running parametric tests was considered as the most highly

associated factor with quality and use of robust statistics the least. With regards to the reporting practices, reporting the sample size and inferential statistics had the highest mean values while reporting non-significant results the lowest. Finally, concerning visual presentation issues, visual presentation for checking normality as well as test-specific assumptions was perceived as the most-highly associated factors with study quality while the least value of association was given to the specification of abnormal cases like outliers by graphs (Table 1, below). Overall, the average total scores for factors showed the respondents' high attention to the issues of design and reporting practices while visual presentation received the lowest attention.

In the next step, in order to do the path analysis, first, an Exploratory Factor Analysis (EFA) with the Maximum Likelihood (ML) extraction method and Promax rotation was run. The KMO value of .774 and significant results of Bartlett's test ($\chi 2$ (820) = 2659.557, p = .000) legitimized proceeding with the test. The initial analysis showed the existence of 10 factors with eigenvalues above 1, which explained 58.295 percent of the cumulative variance. However, looking into the pattern matrix, it was revealed five of the factors had no or only one item with loadings above 0.4. Consequently, the test was re-run with 5 fixed factors. The results explained 46.297 percent of the cumulative variance. The pattern matrix is presented in Table 2.

Figure 2

Standardized Estimates and GOF Values for the Modified Questionnaire

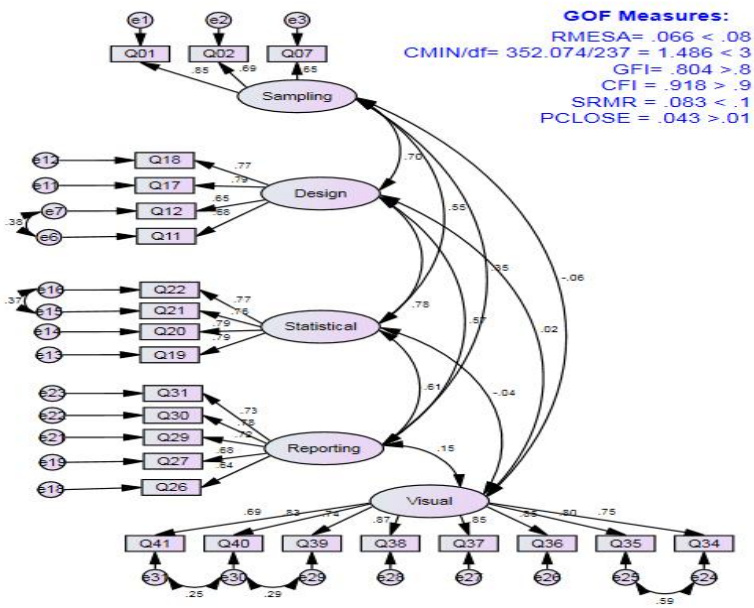


Table 1

Descriptive Statistics of the Data

		N	Min.	Max.	Mean	SD
Sampling	Q01	114	2.00	10.00	7.6140	2.07603
	Q02	114	1.00	10.00	7.5439	2.04870
	Q03	114	1.00	10.00	7.9474	1.84267
	Q04	114	1.00	10.00	6.6579	2.32276
	Q05	114	1.00	10.00	7.5526	2.02678
	Q06	114	1.00	10.00	7.0614	2.59777
	Q07	114	1.00	10.00	6.7719	2.36132
	Total	114	4.571	10.00	7.30696	1.293376
Design	Q08	114	4.00	10.00	9.0263	1.19335
	Q09	114	4.00	10.00	9.2895	.97512
	Q10	114	2.00	10.00	7.9474	2.08170
	Q11	114	2.00	10.00	8.1930	1.78428
	Q12	114	1.00	10.00	8.1754	1.86844
	Q13	114	2.00	10.00	7.7632	1.80119
	Q14	114	1.00	10.00	8.0175	1.81431
	Q15	114	1.00	10.00	8.1228	1.80039
	Q16	114	3.00	10.00	8.3684	1.57561
	Q17	114	4.00	10.00	8.5614	1.45757
	Q18	114	2.00	10.00	8.5175	1.60878
	Total	114	4.455	10.00	8.36204	1.035641
Statistical	Q19	114	1.00	10.00	7.7368	1.82438
	Q20	114	1.00	10.00	8.3596	1.73014
	Q21	114	3.00	10.00	7.8070	1.93193
	Q22	114	3.00	10.00	7.9298	1.72293
	Q23	114	2.00	10.00	7.0702	2.00318
	Q24	114	4.00	10.00	7.9649	1.61262
	Q25	114	2.00	10.00	7.0000	2.00000
	Total	114	3.429	10.00	7.69550	1.261185
Reporting	Q26	114	3.00	10.00	8.3158	1.54757
	Q27	114	3.00	10.00	8.7544	1.43633
	Q28	114	4.00	10.00	8.0351	1.61262
	Q29	114	5.00	10.00	8.8158	1.30055
	Q30	114	4.00	10.00	8.4649	1.56951
	Q31	114	3.00	10.00	8.3070	1.50590
	Q32	114	1.00	10.00	7.3246	2.28692
	Q33	114	1.00	10.00	7.6754	2.29079
	Total	114	5.125	10.00	8.21162	1.044651
Visual	Q34	114	1.00	10.00	6.6930	2.31626
	Q35	114	1.00	10.00	6.7544	2.18831
	Q36	114	1.00	10.00	6.2105	2.41483
	Q37	114	1.00	10.00	6.1316	2.49429
	Q38	114	1.00	10.00	6.3421	2.21351
	Q39	114	1.00	10.00	5.6754	2.53993
	Q40	114	1.00	10.00	6.3947	2.33715
	Q41	114	1.00	10.00	6.0965	2.31184
	Total	114	1.00	10.00	6.28728	1.959274
Valid N (listwise)		114				

Table 2

Exploratory Factor Analysis: Pattern Matrix

	Factor				
	1	2	3	4	5
Q01					.543
Q02					.519
Q03					
Q04					.582
Q05					.444
Q06					
Q07					.480
Q08					
Q09					
Q10					
Q11		.534			
Q12		.619			
Q13		.483			
Q14		.705			
Q15					
Q16		.551			
Q17		.540			
Q18		.802			
Q19				.415	
Q20				.401	
Q21				.514	
Q22				.435	
Q23					
Q24					
Q25				.541	
Q26			.473		
Q27			.675		
Q28			.449		
Q29			.612		
Q30			.751		
Q31			.607		
Q32					
Q33					
Q34	.755				
Q35	.799				
Q36	.840				
Q37	.823				
Q38	.827				
Q39	.784				
Q40	.878				
Q41	.713				

Extraction Method: Maximum Likelihood.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

Comparing the results with the content of the questions, the factors can be named as follows: Factor 1: Visual Presentation; Factor 2: Design Issues, Factor 3: Reporting Practices; Factor 4: Statistical Tests; and Factor 5: Sampling and Power Issues. Note that 10 items (items 3, 6, 8, 9, 10, 15, 23, 24, 32, and 33) did not have loadings above 0.4 to any of the factors, thus, discarded from the analysis.

After extracting the pattern, a Confirmatory Factor Analysis (CFA) was run using IBM AMOS software. This resulted in discarding six more items, three of which had R2 values below 0.5 and the other three endangering the validity and reliability indices. After taking into account the modification indices proposed by the software, the following standardized estimates and measures of Goodness of Fit (GOF) were obtained (Figure 2). As presented in Figure 2, the obtained GOF, based on cut-off values recommended by Hu and Bentler (1999), had acceptable values.

Table 3

Reliability and Validity Measures

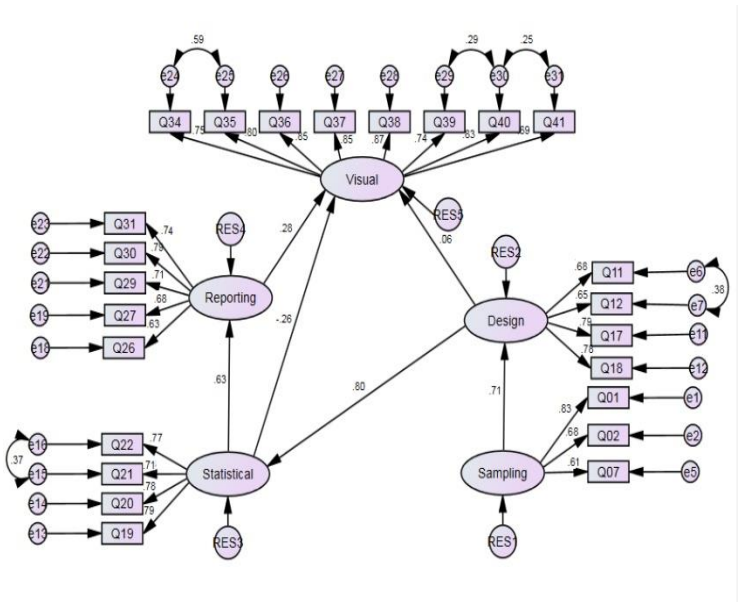
				Fornell Larcker Criterion				
	CR	**AVE**	**MSV**	Sampling	Design	Statistical	Reporting	Visual
Sampling	0.715	0.533	0.492	**0.730**				
Design	0.745	0.523	0.515	0.701	**0.723**			
Statistical	0.802	0.607	0.515	0.547	0.714	**0.779**		
Reporting	0.789	0.532	0.366	0.347	0.567	0.605	**0.729**	
Visual	0.934	0.639	0.024	-0.058	0.025	-0.036	0.153	**0.799**

Subsequently, the reliability and validity indices were checked (Table 3, above). As reported in Table 3, for all five factors, the composite reliability (CR) indices were above the cut-value of 0.7, which are considered acceptable (Hu & Bentler, 1999).

Regarding the convergent validity, the Average Variance Explained (AVE) for each factor was above 0.5 (Henseler et al., 2015). Concerning the discriminant validity, all AVE values were safely above their MSV values and none of the factors had correlations above the squared root of AVE (Bold value under Fornell Larckers' criterion) with another factor. Therefore, the criteria for the discriminant validity (Henseler et al., 2015) were also met. After making sure of the reliability and validity of the questionnaire, the path analysis was run (Figure 3).

Figure 3

Standardized Estimates of the Structural Model



The results of the Structural Equation Modeling are summarized in Table 4.

Table 4

The Path among the Facets of Study Quality Perceived by Authors

			Regression Weight	S.E	Critical Ratio	*p*	*β*	R^2
Design	<---	Sampling	.423	.101	4.199	.000	.708	.501
Statistical	<---	Design	1.109	.223	4.975	.000	.798	.637
Reporting	<---	Statistical	.359	.085	4.204	.000	.626	.391
Visual	<---	Reporting	.584	.342	1.706	.088	.277	.049
	<---	Design	.108	.407	.266	.791	.064	
	<---	Statistical	-.311	.332	-.936	.349	-.256	

Based on the results in Table 4, it can be concluded that Iranian authors' perceptions of study quality issues related to sampling issues significantly affect the design-related ones. The design issues, in turn, are significantly affecting statistical tests, which also, in turn, affect reporting practices. Moreover, neither reporting, statistical tests, nor design issues are significantly affecting concerns with visual presentation. Finally, the inspection of R2 values showed that authors' perceptions of sampling issues predict 50.1 of their perceptions with regards to the design issues while design issues, in turn, can predict 63.7 of the variances in authors' perceptions

of statistical tests. Finally, statistical issues predict 39.1 of the variances in the authors' perceptions of reporting practices issues.

4.2. Discussion

The results of the study showed that Iranian authors perceive issues like random sampling, ensuring the reliability and validity of measurements, checking assumptions for statistical tests, reporting inferential statistics, and checking normality and test specific assumptions through the visual presentation as the issues highly-associated with quality in quantitative researches. These are the primary issues that are emphasized in every research courses or books. It is not, thus, surprising that the authors see them as the primary measures of quality. These features are often taken very seriously by journals and reviewers. However, as Larson-Hall and Plonsky (2015) warn, sometimes authors just put them in their papers "to tick off a list of submission guidelines rather than a meaningful source of information and interpretive value" (p.141).

Next are the items that received the lowest values of association with quality. Among these items are the attention to the low generalizability power in the convenience samplings, the use of delayed posttests to examine the duration of the effect, the use of robust statistics, reporting non-statistical results, and the use of visual graphics to identify abnormal cases. Although according to DeKeyser et al. (2010), "almost every sample has been one of convenience" (p. 416), it is crucial to acknowledge the lower external validity of such studies, although the role of many context-dependent factors in deciding the design of the study should not be ignored. The limited demographic variations in L2 studies generalize the results to that specific context (Ortega, 2009) and highlighted the need for research synthesis for obtaining more generalizable results (Plonsky, 2014b). Moreover, the use of delayed posttest, as a measure of retention of effect, should receive more attention, especially in L2 research context, where the ultimate goal of treatments is to make a long-standing change in the learners. Unfortunately, the results of Plonsky's (2013) also show that the attention to this area is relatively low as only 38% of the experimental studies put it in their designs.

The two other issues which received low amounts of concern from the authors' part are the ones that have been recently encouraged to be included in papers. Reporting of non-significant results is very essential to conduct research synthesis and build the required knowledge for future investigations (Plonsky & Gonulal, 2015). The reluctance of authors shows in reporting these results, as mentioned before, is the consequence of the common false belief that having obtained no significant result is a failure of the research. They might not consider the difference between non-significant and insignificant results, as explained by Larson-Hall (2012) or they may fail to

fully capture the importance of such results in conducting meta-analyses and dealing with publication bias (Rosenthal, 1979). With regards to the robust statistics, the answers of authors may be affected by the fact that these methods of analysis are being used and encouraged recently in L2 studies and their merits are not highlighted enough. These measures are very helpful with L2 data which frequently have problems like lack of normality, outliers, or low sample powers as they are robust to all those problems (Larson-Hall, 2010; Larson-Hall & Herrington, 2010). The final issue with low quality-association perceived will be dealt with in the following discussion of visual presentation issues as these features all had low values in authors' perceptions of features associated with quality.

Looking into the average total scores in each category, it was revealed that design and reporting issues received the highest scores while visual presentation had the lowest. These results also indicate that the quality issue is perceived as adherence to fundamental issues in designing research and rigor reporting of the results. However, the visual presentation seems an issue that needs more attention from authors. All items in this category, even the ones which had the highest mean (i.e., use of visual presentation for checking normality and test specific assumptions) had mean values lower than almost all other items in the questionnaire. Plonsky and Larson-Hall (2015) warn about neglecting the essentiality of visual presentation as a means to comprehend and convey the results by seeing them as a nice accompaniment to the research. Larson-Hall (2017) also recommends the use of multi-task graphics like boxplots, beeswarms, and pirate plots. Aside from checking normality, outlier, or test-specific assumptions like homogeneity of regression slopes, graphs show both individual variations and patterns of the results (Norris, 2015). Together with numerical results, visual presentations present the accuracy of results (Wilkinson, 1999). According to Norris (2015, p. 121), "the actual differences observed between groups and relationships between variables are most directly indicated through these techniques".

Finally, the inspection of the structural model showed that authors' perceptions of sampling issues affect their perception of design issues. The design issues, in turn, affect the statistical issues, which also, in turn, affect the reporting practices. The path, thus, starts with sampling and goes through design and statistical tests to reporting. For authors to take the concerns of quality more seriously, each part should be focused and emphasized. The visual presentation, which could not be predicted by any other factors needs more fundamental attention. Although journals and manuals (e.g., APA, 2010; Journal Article Reporting Standards Working Group, 2008, 2018; Valentine & Cooper, 2008; Wilkinson, 1999), it seems that if those concerns are included in research courses and classes, like the issues of reliability and

validity that are currently taken very serious, the other concerns may be more likely to be attended, too. In other words, as our results show, the way quality is perceived by authors is closely related to how quality is preserved in L2 papers. Therefore, the improvement in the quality should start in the minds of authors before being applied in the practice of the research (Holliday, 2015).

5. Conclusion and Implications

In conclusion, it seems that Iranian authors' higher perceptions of quality in adherence to the fundamental issues in conducting and reporting research need to be improved by acknowledging the importance of other areas of quality. It is hoped that L2 authors soon recognize the significant improvements using these measures can bring in.

The areas which received lower attention could be emphasized in academic research courses. As a study conducted by Loewen et al. (2019) has already shown, the researchers' statistical literacy is affected by the number of statistical courses taken by them and their frequency of reading statistical books. This can also be the case for other issues of quality if implemented.

Another influential factor is Faulty members' conceptions of research practices that are also reflected in the students' practices (Amini Farasani, 2017). The two are, in fact, the primary authors in journals. Therefore, the amount of attention paid by professors to the feature of study quality both affects the students' practices and builds up the intuition in the works of novice authors that look up to their professors' works or the already-published papers that may lack adherence to standards of quality in one aspect or another. The faculty members, thus, are encouraged to take the study quality concerns seriously in their works.

Iranian journals may also contribute to fostering the process of knowing and implementing quality features by updating their guidelines for publications and applying stricter adherence to the standards to both keep the quality and make a good primary source for novice authors who use these journals as one of their primary sources to get the intuition of required quality. They may also contribute by giving more space to the papers focusing on research issues. Having special issues, for example, on these matters may also be very helpful. It is hoped that by adherence to standards of quality, our knowledge of the world proceeds to build up with a higher pace and leveled-up degrees of accuracy.

It should not be left unmentioned that this research was conducted in a specific context and the results may not be generalizable to the other contexts in one aspect or another. Adherence to the concept of quality is still growing (Gass, et al., 2020). The issues investigated in this research were the ones that had been recognized by standard manuals and recommendations. As these standards are regularly updating, other issues may also be the concern of future studies.

Acknowledgements

We are obliged to Dr. Jenifer Larson-Hall and Dr. Luke Plonsky, based on whose professional and insightful comments the validity of the developed questionnaire was ensured. Our sincere gratitude also goes to the participants of this study for the time and efforts they spent on answering the questionnaire. Finally, we would like to thank Dr. Mohsen Moradi at Iran's Analysis Academy whose guidelines were taken into account in the structural equation modeling of the results.

References

Amini Farsani, M. (2017). *Exploring three decades of TEFL research in Iran: Conceptions and practices*. (Unpublished doctoral dissertation). Kharazmi University, Tehran, Iran.

Amini Farsani, M., & Babaii, E. (2018). Mapping past, current, and future TEFL research trends in Iran: A synthetic review of topics and paradigms in three decades. *Journal of Modern Research in English Language Studies, 5*(4), 81-108.

Amini Farsani, M., & Babaii, E. (2020). Applied linguistics research in three decade: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity: International Journal of Methodology, 54*(4), 1257-1283.

Al-Hoorie, A., & Vitta, J. (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.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Author.

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, *1*, 66–80.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 839-851.

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.

Babaii, E., Hashemi, M.R., & Amini Farsani, M. (2017). Exploring EFL applied linguists' conceptions of research: A mixed methods study. *Foreign Language Research Journal, 6*(2), 501–533.

Brown, J. D. (2015a). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 9-20). Routledge.

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

Chaudron, C. (2001). Progress in language classroom research: Evidence from the *Modern Language Journal*, 1916-2000. *Modern Language Journal*, *85*, 57–76.

Cooper, H. & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 147–158). Russell Sage Foundation.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th ed.). Sage.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*(3), 413 - 438.

Ellis, N. C. (2000). Editorial statement. *Language Learning, 50*, xi–xiii.

Gaskin, J. (2013a). SEM Series Part 2: Data Screening. Retrieved on 2 October, 2020, from *http://www.youtube.com/watch?v=1KuM5e0aFgU*.

Gass, S. (2009). A survey of SLA research. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 3–28). Emerald.

Gass, S., Loewen, S., & Plonsky, L. (2020). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, 1-14. doi:10.1017/S0261444819000430.

Henseler, J., C. M. Ringle, and M. Sarstedt (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling, *Journal of the Academy of Marketing Science, 43*(1), 115-135.

Holliday, A. (2015). Analyzing qualitative data. In B. Partridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical gu*ide (pp. 60-73). Bloomsbury Publishing

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Hu, Y., & Plonsky, L. (In press). Statistical assumptions in L2 research: A systematic review. *Second Language Research*. doi:10.1177/0267658319877433.

Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research. Language Learning, *65*(S1), 76-96.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Larson-Hall, J. (2012). Our statistical intuitions may be misleading us: Why we need robust statistics. *Language Teaching 45*, 460–74.

Larson–Hall, J. (2017), Moving Beyond the Bar Plot and the Line Graph to Create Informative and Attractive Graphics. *The Modern Language Journal, 101*, 244-270. doi:10.1111/modl.12386.

Larson-Hall, J., & Herrington, R. (2010). Improving Data Analysis in Second Language Acquisition by Utilizing Modern Developments in Applied Statistics. *Applied Linguistics, 31*, 368-390. doi:10.1093/applin/amp038

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(Supp. 1), 127–159.

Loewen, S., Gönülal, T., Isbell, D., Ballard, L., Crowther, D., Lim, J., . . . Tigchelaar, M. (2019). How knowledgeable are applied linguistics and SLA researchers about basic statistics. *Studies in Second Language Acquisition,* 1-20. doi:10.1017/S0272263119000548.

Long, M. H. (1983). Does second language instruction make a difference? A review of research. *TESOL Quarterly*, *17*, 359–382.

Jahanbakhsh, A. A., & Larson-Hall, J. (inreview). Building a research paper: Researcher perceptions of study quality.

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning, 53*, 123–163.

Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research, 34*(2), 257–271.

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Suppl. 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. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly*, *41*, 805–815.

Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning, 65*(2), 470-476.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Ortega, L. (2015). Research synthesis. In B. Paltridge and A. Phakiti (Eds.), *Research methods in applied linguistics: a practical guide* (pp. 219-237). Bloomsbury Publications.

Plonsky, L., & Oswald, F. L. (2012). How to do a meta-analysis. In A. Mackey& S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 275 – 295). Wiley-Blackwell.

Plonsky, L. (2009, October). *"Nix the null": Why statistical significance is overrated*. Paper presented at the Second Language Research Forum, East Lansing, MI.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*, 993–1038.

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*, 655–687.

Plonsky, L. (2014a). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470.

Plonsky, L. (2014b, February). *Sampling, power, and generalizability in L2 research (Or, why we might as well be flipping coins)*. Keynote presentation at the Second Language Studies Symposium, East Lansing, MI.

Plonsky, L. (2015). Statistical power, *p* values, descriptive statistics, and effect sizes: A "back-to-basics" approach to advancing quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 23-45). Routledge.

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61*(2), 325– 366.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *Modern Language Journal*, *102*(4), 713–731.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(Suppl. 1), 9-36.

Rosenthal R (1979). File drawer problem and tolerance for null results. *Psychol Bull. 86*, 638-41.

Steenbergen-Hu, S., & Moon, S. (2011). The effects of acceleration on high-ability learners: A meta-analysis. *Gifted Child Quarterly, 55*, 39-53.

Tarone, E. (2013). Applied linguistics without borders. *Language Teaching, 46*, 355–364.

Taylor, A., Stevens, J. R., & Asher, J. W. (2006). The effects of Explicit Reading Strategy Training on L2 reading comprehension: A meta-analysis. In J. M. Norris, & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching* (pp. 213-244). John Benjamin.

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods*, *13*, 130–149.

Vines, T., Albert, A., Andrew, R., Debarre, F., Bock, D., Franklin, M., et al. (2014). The availability of research data declines rapidly with article age. *Current Biology, 24*(1), 94–97.

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing*. Elsevier Academic.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594 – 604.

Ziegler, N. (2013). *Synchronous computer-mediated communication and interaction: A research synthesis and meta-analysis*. Unpublished doctoral dissertation, Georgetown University.

Appendix 1: The Questionnaire of Authors' Perception of Study Quality Issues

The present questionnaire aims to capture your perception of the importance of different features in the determination of study quality in the context of quantitative L2 research. Four main categories, namely sampling and design, statistical analyses, reporting and sharing data, and visual presentation of results, are investigated. Your contribution is most-appreciated and your answers will be kept anonymous.

Instruction:

The question you need to ask yourself to answer each item is "***to what extent is each feature associated with high-quality research?***". Please code your answers on a scale from 1 to 10 where 1 represents the lowest association and 10 the highest association of the feature. *Note that some questions may not apply to some designs. Your answers should be based on the presupposition that the type and design of the study are compatible with the features asked.*

1 = The lowest association 10= The highest association

1. Random sampling.
2. Statistical power analysis to decide sample size.
3. Use of sufficiently empowered samples.
4. Avoiding convenience sampling or pointing out the limitation in generalizability.
5. Inclusion of participants with various demographic backgrounds or pointing out the limitation in generalizability.
6. Participants not being aware of their experimental condition/group.
7. Random assignment.
8. Ensuring measurement reliability/dependability.
9. Ensuring the validity of instruments.
10. Pre-treatment comparability of groups in experimental designs.
11. Use of control/comparison group to minimize the effects of intervening variables
12. Compensation for initial differences (pretesting).
13. Use of delayed posttest to show the duration of effects.
14. Taking demographic/proficiency differences into account
15. Use of qualitative inquiry to triangulate the quantitative results
16. Appropriate length of the treatment
17. Competent operationalization of the treatment
18. Data collection setting (e.g., lab, classroom, etc.) compatible with the focus of study (e.g., participants, treatment type, measurement of variables)
19. P-value adjustment in cases of multiple inferential analyses from the same data
20. Checking assumptions for running parametric tests
21. Use of regression or multifactorial/multivariate tests (e.g., two-way ANOVA or MANOVA) to avoid multiple tests (e.g., two one-way ANOVAs or multiple t-tests)
22. Checking all of the test-specific assumptions (e.g., linearity, regression slopes, and equality of error variances for ANCOVA)