The Path in Behind and the Challenges in Front: A Methodological Synthesis of Iranian L2 Papers

Akbar A. Jahanbakhsh (a.jahanbakhsh@tabrizu.ac.ir)
English Department, Faculty of Persian Literature and Foreign Languages, University of Tabriz, Iran https://orcid.org/0000-0001-5735-1328
Parviz Ajideh
University of Tabriz

Original article

Keywords: Experimental studies, Iranian Journals, Methodological synthesis, Study quality

Posted Date: April 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-141389/v2

Abstract

The present study is a methodological synthesis aiming to evaluate the adherence of Iranian L2 papers to the study quality standards. Ten Iranian journals were selected based on the latest ranking of Iran's Ministry of Science, Research, and Technology (MSRT), and all experimental papers (N = 367) published from their beginnings were explored for study quality with regards to sampling, design, statistical tests, reporting practices and data sharing, and visual presentation. In the evaluation of the papers, the protocols proposed by Gass and Plonsky (2011) and Pagout and Plonsky (2017) were moderated and some recent issues proposed by APA's (2018) Journal Article Reporting Standards and some scholars (e.g., Hu & Plonsky, 2019; Khany & Tazik, 2019; Larson-Hall, 2017) were added. The results showed that while there were issues, like acceptable sample size, use of pre-testing, reporting descriptive and inferential statistics, and ensuring the reliability of instruments, that were acceptably adhered to quality standards, problematic areas existed in all five facets of quality, and the majority of them stayed constant or changed slightly over time. The shortcomings caused by such lack of adherence are discussed to identify the challenges in the way of improving the papers' quality. Although the results are obtained from a specific context, the implications may be generalizable to other contexts where English is taught and researched as a foreign language.

1. Introduction

Our knowledge of the world is a collective phenomenon gathered by pieces of information provided by exploring the existing patterns or purposeful manipulation of the conditions to reach a conclusive understanding. Every piece of information works like a brick necessary to build the ivory tower of science.

The task of gathering together the pieces of information and reaching a comprehensive results is what meta-analytic research has been focused on. Meta-analysis is "a statistical method for calculating the mean and the variance of a collection of effect sizes across studies, usually correlations (r) or standardized mean differences (d)" or broadly as "not only these narrower statistical computations, but also the conceptual integration of the literature and the findings that gives the meta-analysis its substantive meaning" (Plonsky & Oswald, 2015 , p.106). Recent examples of such studies in the field of language education are Derakhshan and Shakki (2021) and Yousefi and Nassaji (2019).

The classical meta-analyses, as introduced above, aim to provided aggregated knowledge on different concepts in hand. However, they miss the essential question of how these pieces of knowledge are gathered. In other words, the question of quality in practicing the research within the scientific framework and ahereing to its standards are not focused upon. It is, thus, essential, to make sure of the quality of every piece of knowledge we are relying upon. This gap is addressed by what is known and methodological synthesis.

Methodological meta-analysis, or methodological synthesis, to use Plonsky and Gonulal's (2015, p. 10) definition, is a kind of study that uses "synthetic methods to describe and evaluate the presence of research and reporting practices in a given domain, whether broadly or narrowly conceived". The methodological meta-analyses, thus, "treat primary studies as participants that are surveyed to collect methodologically oriented data". Study quality, as the target of methodological synthesis, according to Plonsky (2011, p. 5), is defined as "adherence to standards empirical rigor, appropriateness, and transparency in study design, analysis, and reporting practices", which in turn provides the necessary means to evaluate and rely upon the previous studies as a building block to step on and venture into the exploration of the next unknowns.

According to Plonsky (2013), the quality in methodological meta-synthesis is the combination of (1) respect to standards of contextually appropriate, methodological rigor in research practices and (2) transparent and full reporting of these practices. Study quality studies are usually in the form of meta-syntheses aiming to a) describe the practices and identify the methodological culture; b) describe and evaluate the results for the purpose of improving future research; c) examining the relationship among the facets of the research; and d) inspection of changes over time (Plonsky & Gonulal, 2015). The present study both describes and evaluate the adherence to the facets of quality and the changes in these facets over time. By doing so, the study provides a comprehensive description of the strong and weak areas of study quality in the research practices of Iranian authors. It also draws attention of journals and authors to the concerns that have not been changed/solved over the last decade.

Moreover, as the previous research syntheses (e.g., Hu & Plonsky, 2019; Khany & Tazik, 2019; Larson-Hall & Plonsky, 2015; Larson-Hall, 2017; Norris, et al., 2015; Plonsky, 2011, 2013, 2014a, 2014b; Plonsky & Gonulal, 2015), except for Plonsky's (2013, 2014a, 2014b) works, have usually addressed separate statistical and methodological issues in the field, including examination of quality in (a) study design, (b) instrumentation, (c) statistical analyses, and (d) reporting practices, a study which focus on all of these aspects together seems essential. Although Plonsky (2013, 2014a, 2014b) did a similar job, his study was conducted on papers published from 1990 to 2010. After almost a decade, replication of such study would give a good perspective on what have been changed.

Moreover, the previous studies, except for Khany and Tazik (2019), used a limited number of journals as the sources for the evaluated papers. Even Plonsky's (2013) work only included two journals in its analysese. None of these studies also examined quality in local journals. In this study, we have examined experimental papers published in 10 Iranian journals from their beginnings (most of them started from around 2010). We have also integrated the categories of assumption checking used by Hu and Plonsky (2019), types of tests explored by Khany and Tazik (2019), type and purpose of visual presentation in Larson-Hall (2017), and data sharing as emphasized by APA's Journal Article Reporting Standards (2018) with the quality protocols used by Gass and Plonsky (2011) and Pagout and Plonsky (2017) to reach a comprehensive framework.

Finally, papers published in Iranian journals were examined as a token of research practices in an EFL context. The previous works on study quality were done almost exclusively on the papers published in high-ranked international journals (e.g., The Modern Language Journal, Language Learning, and Studies in Second Language Acquisition in Larson-Hall, 2017; Language Learning and Second Language Research in Hu and Plonsky, 2019; Language Learning and

Studies in Second Language Acquisition in Plonsky, 2013). This study aims to evaluate study quality in locally-published journals to depict how they are catching up in adhering to the standards of quality that have been emphasized.

Research Questions

1. How is study quality adhered to in Iranian L2 papers? What are the most-adhered and most-challenging areas?
2. What quality aspects have changed over time in L2 papers published in Iranian journals?

2. Literature Review

The terms research synthesis, research review, systematic review, and meta-synthesis, according to Cooper and Hedges (2009), have been used interchangeably in the literature. Such studies, although seemingly involved with confusing analyses and intimidating appearance, evaluate papers and the difference of groups (Rosenthal & DiMatteo, 2001). Methodological synthesis, on the other hand, "seeks not only to describe but to evaluate and comment on the field's practices with the intention to improve future research as well" (Plonsky & Gonulal, 2015, p. 12).

As a focus of meta-synthesis, study quality refers to the adherence to standards to practice a "contextually appropriate, methodological rigor in research" combined with "a transparent and complete reporting of such practices" (Plonsky, 2013. P.657). As Plonsky (2011) asserts, there are numerous factors, depending on the context and focus of any primary study that might be influencing each individual study. However, assigning weight to each of these factors seems an impossible task. That is why the methodological meta-synthesis seems an appropriate mean to evaluate these influencing factors.

The emphasis on study quality issues has been accelerated by the works of Plonsky and his colleagues in the last decade (e.g., Gass, Loewen, & Plonsky, 2020; Hu & Plonsky, 2019; Plonsky, 2013, 2014a, 2014b; Plonsky, Egbert, & Laflair, 2015; Plonsky & Gass, 2011; Plonsky & Gonulal, 2015; Norouzian & Plonsky, 2018). Other scholars (e.g., Hudson & Lisoa, 2015; Kany & Tazik, 2019; Larson-Hall, 2012, 2017; Norris, 2015) added essential information and guidelines with regards to the adherence to quality facets. In what follows, we will summarize the findings with respect to the five features, i.e., sampling, design, statistical tests, reporting practices and data sharing, and visual presentation of data, which are the focus of this study.

With regards to the sampling, the findings of previous studies (Plonsky, 2013, 2014b; Plonsky & Gass, 2011) showed the possibility of lack of required power for yielding significant results (Type I error) in a large proportion of L2 studies. Their results indicate that it is possible that large amounts of L2 research frequently lack the required power to yield statistically significant results. Moreover, the meta-synthesis reports of Plonsky (2013) and Plonsky (2014b) showed the rarity of power analysis (about 1%) in L2 papers. Similar results were also reported by other studies (2% in Plonsky & Gass, 2011; 7% in Ziegler, 2013). This is followed by the commonality of convenience sampling in these papers, which in turn results in limited generalizability of these studies. Reported results (Norris & Ortega, 2000; Plonsky, 2014b) indicated that the majority of the participants in L2 research are young adult university students who live in the USA, west Europe, or East Asia whose first or second language is English. Therefore, no matter how sufficient the sample is selected or how large the effect size is, there is no guarantee that the results may be generalizable for a large number of other contexts (Ortega, 2005, 2009).

With respect to design issues, several issues have been pointed out as shortcomings of L2 research. Chaudron (2001), for example, note the low reliability, poor design, and regularity of using intact groups in research. Other studies (e.g., Plonsky, 2013; Plonsky & Gass, 2011) reported that a small portion of classroom-oriented experimental researches was conducted in a classroom environment. While studies (Gass, 2009; Plonsky, 2014a) have shown an increase in relying on quantitative data, some features, such as random selection/assignment are relatively concerning. Plonsky (2013) reports that, in his sample of twenty years, only 47% of the studies used random assignments (37% individual assignment and 10% group assignment) and 38% of them used delayed posttests. However, the use of control group, pretesting, and delayed posttesting has been increased over time (Plonsky, 2014a).

Concerning the statistical tests, the first issue is related to the power analysis addressed above, which can directly affect the results of statistical tests. Next is the use of multiple statistical tests on the same data which causes the change in alpha level, which is regularly ignored in studies of social science (Wilkinson, 1999). Plonsky (2013) reports that 60% of the papers in his study used multiple tests. Khany and Tazik (2019) also reported that 78.77% of applied linguistics papers use basic statistical tests (e.g., descriptive, chi-square, t-tests, and one-way ANOVA). Moreover, he reported that the assumptions of running these tests were only checked in 17% of the cases. Similarly, Hu and Plonsky (2019) reported that 17% followed stringent standards (reporting all required assumptions) and 24% of the quantitative studies in their sample followed lenient standards (meeting one or more of the assumptions). The final issue related to statistical analysis is the over-reliance on null hypothesis significant testing (NHST) and the dichotomous interpretation of the p-value. The use of robust statistics (e.g., Larson-Hall, 2012) and new statistics, i.e., effect size and confidence intervals, (e.g., Cumming, 2012, Norris, 2015) were recommended, as a result. However, Plonsky's (2013) reports show that studies of SLA research included 35 p-values on average, while in 26% of the cases effect size was reported and, shockingly, only 5% of the studies reported confidence intervals. APA's (2018) Journal Article Reporting Standards recommends reporting effect size and confidence intervals alongside the statistical difference. Larson-Hall (2012) explains how relying solely on p-value might be misleading. According to her, confidence intervals "functions like the p-value but also gives more information about how large or small the difference between groups might be" (p. 470). She also explains that the role of effect size stays the same no matter what arbitrary cut-off value is set as the alpha. Therefore, putting the analysis based on effect size and confidence intervals would both prevent the type II error and give us more detailed data about the existing relationships/differences among our variables.

The next facet concerns the reporting practices and data sharing. Larson-Hall and Plonsky (2015, p. 131) refer to the issue of not reporting the descriptive statistics as "a practice that harms our field as a whole" since it prevents secondary level analysis (meta-analysis). Plonsky's (2013) results show that the most frequently reported descriptive statistics was the sample size (reported in 99% of the articles), followed by means (77%). However, in 17% of the cases, the mean was reported without standard deviation, leaving only 60% of papers with the basic and necessary information for running meta-analyses. Besides, the descriptive statistics gives the readers a primary picture about the data and how they look or change during the study. Therefore, ignoring these basic

information from the readers may not be promoted. The next issue is the pre-determination of alpha, which needs prior power analysis. As reported above, the rarity of power analysis exists in L2 papers. The prior level of alpha, which, according to Plonsky (2013), was done only in 22% of the cases. Similar results (16-26%) were obtained by Plonsky (2014a). The next concern in missing reports addresses the omission of non-statistical results. Plonsky's (2013) meta-analysis showed that p-value was not reported in 13% of his studies. Besides, the exact value of p was only reported in 49% of the sample. Ajideh, Zohrabi, and Jahanbakhsh (in press) also reported that Iranian authors considered issues like reporting reliability, validity, and inferential statistics as the ones highly associated with study quality. The final issue is data sharing. APA Ethical Standard 8.14 "stipulates that psychologists do not withhold their data from other competent professionals who seek to verify substantive claims" (Breckler, 2009, para. 7). Plonsky (2011) reports that only about one-third of his request, from the study authors, of the descriptive statistics, resulted in the provision of them. Other studies also reported a small proportion of successful retrieval of raw data. For example, only 14% replied to the data request of Plonsky et al. (2015).

The final concern of the quality in the practice of L2 research is the use of visual presentations. The use of graphics is argued as a necessary means to understand and convey the findings of the research (Larson-Hall & Plonsky, 2015). Despite the importance of graphics, the use of them in L2 papers is concerning. Norris and Ortega (2000), for example, reported that graphic presentation did not appear in 46% of the papers they studied. Plonsky (2013) reported that about two-thirds (66%) of the studies he surveyed did not use visual displays Similarly, Larson-Hall (2017) found a fairly low percentage of graphical presentation in three well-known L2 journals, i.e., 24% in The Modern Language Journal, 34% in Language Learning, and 48% in Studies in Second Language Acquisition. She also reported, among the papers which used graphic presentation of data, a large proportion (70 to 79%) used either line graphs or bar plots. She calls these types of graphs as "data poor" and encourages authors to use either "data accountable" grpahs, like scatterplots or pirate plots which "plot all of the relevant details of the dataset[…] as well as the individual data points", or at least "data rich" graphics, like boxplots, which "show the distribution of the data and necessarily present a large amount of information about the data set to the reader, although they do not show individual points" (Larson-Hall, 2017, p. 244).

Having reviewed the existing challenges in the literature, the present study aims to both describe and evaluate L2 papers published in Iran against the above-mentioned concerns and also identify the existing changes over time.

3. Methodology

3.1. The Corpus

The study examines study quality in papers published in 10 Iranian Journals from their beginnings. Eight of these journals, including Applied Research in English Language (AREL) published by University of Isfahan, Iranian Journal of Applied Linguistics (IJAL) by University of Kharazmi, Iranian Journal of Applied Language Studies (IJALS) by University of Sistan and Balouchestan, Issues in Language Teaching (ILT) by University of Allameh Tabataba'i, Journal of English Language Teaching and Learning (JELTL) by University of Tabriz, Journal of Research in Applied Linguistics (RALS) by Shahid Chamran University, Journal of Teaching Language Skills (JTLS) by University of Shiraz, and Journal of Teaching English Language (TEL) by Teaching English Language and Literature Society, were ranked A in the latest report of Ministry of Science, Research, and Technology (MSRT) by the time (ranked in 2017) data were collected. The two others, i.e., Iranian Journal of Language Teaching Research (IJLTR) published by the University of Urmia and Journal of Language and Translation (JLT) by Islamic Azad University, which were not yet evaluated by MSRT, were included for their quality and history of publication. The number of studies included from each journal as well as their years of publications are presented in Table 1, below.

Table 1

Description of the Corpus

		Year											Total
		Before	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Journal Name	AREL	NP*	NP	NP	2	2	3	3	5	7	7	5	34
	IJAL	7	4	3	4	1	7	5	4	4	1	NP	40
	IJLAS	2	2	2	2	3	3	3	1	3	2	5	28
	IJLTR	NP	NP	NP	NP	1	6	2	2	2	3	1	17
	ILT	NP	NP	NP	5	4	5	3	3	5	5	6	36
	JELTL	NP	2	4	1	5	3	5	5	5	2	4	37
	JLT	NP	6	3	2	7	5	5	5	5	7	2	47
	JTLS	NP	3	3	8	6	11	7	8	7	3	8	64
	RALS	NP	0	1	2	1	5	3	3	5	4	3	27
	TEL	NP	4	5	3	5	3	3	3	2	6	3	37
Total		10	21	21	29	35	51	39	39	45	40	37	367
* Not Published													

3.2. Selection Criteria

In the selection of the papers, three criteria were taken into account. First, only the journals, not the books, were selected as they are known as "the medium of choice for publishing primary L2 research [and] they are also generally accessible through hard copy and electronic library resources" (Plonsky, 2013, p. 664). Second, as the focus of the research is on the quality of quantitative L2 research, the qualitative papers and the ones which did not directly focus on L2 learning and/or teaching were excluded. Finally, the corpus included all experimental designs, discarding non-experimental ones, "due to the wide range of observational or nonexperimental design types" (Plonsky, 2013, p. 666), as well as the case studies to adhere to the coding protocol proposed by Plonsky & Gass (2011). Accordingly, 367 papers met the criteria for selection.

3.3. Data Coding

An integrated-moderated coding scheme was developed based on the recommendation made by APA's (2018) Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report to evaluate the papers. First, the coding schemes used by Gass and Plonsky (2011) and Paquot and Plonsky (2017), which were developed based on APA's sixth edition and recommendations provided by other scholars, were integrated. Then, the new recommendations provided by the Journal Article Reporting Standards (2018) were added.

Table 2

The Coding Scheme

Category	Feature	Coding options
Identification		Journal; Year; Volume; Issue; Article No
Sampling	Type	Random Sampling; Convenience with random assignment; Convenience with intact class; convenience without assignment; Not reported
	Power Analysis	Yes; No
	Sample Size	
	N of Participants per Groups	
Design and Data Collection	Comparison Group(s)	Yes; No
	Control Group(s)	Yes; No
	Checking Homogeneity	Yes; No
	Pretesting	Yes; No
	Delayed Posttest	Yes; No
	Setting	Classroom; Lab; Others
	Triangulation by Qualitative Inquiry	Yes; No
Statistical Analyses	Type of analyses	Basic; Intermediate; Advanced
		Correlation; Chi-Square; Paired t-test; Independent t-test; One-Way ANOVA; Factorial ANOVA; ANCOVA; MAN(C)OVA; RM ANOVA; others
	Number of Statistical Tests per Paper	
	Normality Assumption Checking	Yes; Minimal information; No; Not Applicable
	Test-Specific Assumption Checking	Stringent; Lenient; Minimal Information; None
	Robust Statistics	Bootstrapping; M-Estimator; Trimmed Mean
Reporting and Sharing	Pre-determined Alpha	Yes; No
	Validity	Yes; No
	Reliability	Yes; No
	Descriptive/Frequency Statistics	(Gain) Mean; (gain) SD; Percentage
	p-value for significant results	Exact p-value; Exact $p < \alpha$; $p<\alpha$; Not reported
	p-value for non-significant results	Exact p-value; Exact p-value > α; $p>\alpha$; Not reported
	Referential Statistics	Yes; No
	New Statistics	Confidence Intervals; Effect size
	Sharing Data	Yes; No
Visual Presentation	Number of visualizations	
	Type	Pie-chart; Histograms; bar plot; line graphics; boxplot; scatter plot; pirate plot; beeswarm plot; others
	Data	Poor; Rich; Accountable

Moreover, for the assumption checking, the listed assumptions provided by Hu and Plonsky (2019) were added to the coding scheme. Then, the moderations were done to include some other scholars' concerns, such as sampling and design issues (Plonsky, 2015), reliance on NHST (Larson-Hall, 2012; Norris, 2015), types of statistical analyses (Khany & Tazik, 2019), reporting practices (Norris et al., 2015, Gass, et al., 2020) and visual presentation of data (Larson-Hall & Plonsky, 2015, Larson-Hall, 2017). The developed coding scheme, was, then, emailed to two leading experts in this area and they provided approval after some minor changes. The final approved coding scheme is provided in Table 2, above.

The data collection was done on the papers that met the criterion of selection based on the above coding scheme. The two authors, first, separately coded 50 papers (around 14% of the sample) and put the results in a Kappa formula to reach the inter-coder agreement. The results of the test, $\kappa = 0.686$, SE = 0.074, 95% CI [0.54, 0.83], p = 000 < .05, showed significant and strong agreement between the coders.

The analysis of the results were done with SPSS version 24. It should not be left unmentioned that the datasets generated and analyzed during the current study are available in the Mendeley Data repository, https://data.mendeley.com/datasets/2c3w6h95n4/1.

4. Results

As mentioned before, the study aims to both depict the overall picture of adherence to study quality facets and the changes over time in the adherence. For the purpose of manageability, the result for each category will be presented separately. Moreover, to keep a balance in the number of papers examined over time, three time spans which included a close number of papers were set: a) from the beginning to 2013 (including 116 papers); b) from 2014 to 2016 (including 129 papers); and c) from 2017 to 2019 (including 122 papers).

4.1. Issues Related to Sampling

Table 3 summarizes the results related to the facet of sampling. As reported, a small proportion of studies used random sampling. Convenience sampling by random assignment of individuals into groups showed a small fall in the mid-way. A considerable amount (18.3%) of studies either did not use any randomization or not reported it. Random assignment by individuals was frequent in studies before 2014, but slightly decreased in the second period, 2014-2016, giving room to random assignment by groups. In the last period, however, the trend changed back and random assignment by individuals was more favored. There was almost no power analysis throughout the path while the proportion of participants and groups throughout the path was kept close with a slight fall down in the mid-way.

Table 3

Description of the Sampling Features Over Time

		Year			Total
		Before 2014	2014 to 2016	2017 to 2019	
Random Sampling		5 (4.3%)	6 (4.7%)	4 (3.3%)	15 (4.1%)
Convenience Sampling	Random by Individuals	63 (54.3%)	55 (42.6%)	56 (45.9%)	174 (47.4%)
	Random by Group	32 (27.6%)	42 (32.6%)	37 (30.3%)	111 (30.3%)
	No Random Assignment	3 (2.6%)	7 (5.4%)	9 (7.4%)	19 (5.2%)
Not Mentioned		13 (11.2%)	19 (14.7%)	16 (13.1%)	48 (13.1%)
Power Analysis		0 (0%)	1 (0.8%)	0 (0%)	1 (0.3%)
N of Participants	Range	10-350	8-200	7-300	7-350
	Mean	85.66	69.64	73.43	75.97
	SD	51.01	34.45	47.63	45.03
Participants per Groups (P/G)	Average N of Groups	2.82	2.62	2.75	2.73
	Average P/G	31.78	28.02	28.36	29.32

4.2. Issues Related to Design

Five factors were examined with regard to the design issues. First of all, the setting in which the experiments were done were either in classrooms (N = 360, F = 98.1%) or in laboratories (N = 7, F = 1.9%). The number of lab-based experiments has decreased from 4 in the first period to 2 in the second period to 1 in the last one. Therefore, the changes over time, based on this small sample, may not be reliable. Overall, six cases of pretesting (86%) happened in lab-based experiments while the use of comparison groups, delayed posttest, and qualitative triangulation (2 cases for each) were low. With regards to the classroom experiments (Figure 1, below), an almost constant pattern exists. That is to say, the issues of design have shown slight if no changes over time. Pretesting is the issue of high-preservation followed by ensuring the pre-treatment homogeneity of participants while delayed posttesting and triangulation of the results have the lowest rate. The use of control or comparison groups is also moderate.

4.3. Issues Related to Statistical Analyses

With regards to the statistical tests, five features were examined. First, the overall frequency and number of tests per papers were investigated. The types of tests were categorized, based on the classification used by Khany and Tazik (2019), into three classes of basic (including descriptive, correlation, chi-square, t-tests, and one-way ANOVA), intermediate (including univariate ANOVAs, regressions, and non-parametric tests), and advanced (including multivariate tests) analyses.

Table 4

Percentage of the Type and Number of Statistical Tests Over Time

Class			Year			
			Before 2014	2014 to 2016	2017 to 2019	Total
Basic	Test	Frequency	0 (0%)	0 (0%)	4 (2.01%)	4 (0.59%)
		Correlation	3 (1.28%)	3 (1.25%)	2 (1.01%)	8 (1.19%)
		Chi-Square	2 (0.85%)	29 (12.08%)	5 (2.51%)	36 (5.35%)
		Paired t-test	45 (19.23%)	53 (22.08)	44 (22.11%)	142 (21.10%)
		independent t-test	104 (44.44%)	82 (34.17%)	70 (35.18%)	256 (38.04%)
		One-way ANOVA	80 (34.18%)	73 (30.42%)	74 (37.18%)	227 (33.72%)
	Total		234 (74.29%)	240 (70.38%)	199 (61.04%)	673 (68.53%)
Intermediate	Test	Factorial ANOVA	27 (48.21%)	5 (9.80%)	12 (13.4%)	44 (22.11%)
		ANCOVA	13 (23.21%)	23 (45.10%)	51 (55.44%)	87 (43.72%)
		Multiple Regression	0 (0%)	0 (0%)	1 (1.09%)	1 (0.50%)
		Non-parametric	16 (28.57%)	23 (45.10%)	28 (30.44%)	67 (33.67%)
	Total		56 (17.78%)	51 (14.96%)	92 (28.22%)	199 (33.66%)
Advanced	Test	MAN(C)OVA	4 (16%)	21 (42%)	13 (37.14%)	38 (34.55%)
		RM ANOVA	21 (84%)	29 (58%)	22 (62.86%)	72 (65.45%)
	Total		25 (7.94%)	50 (14.66%)	35 (10.74%)	110 (11.20%)
Total			315 (32.07%)	341 (34.73%)	326 (33.20%)	982 (100%)
Robust Statistics per Paper		Bootstrapping	0 (0%)	0 (0%)	0 (0%)	0 (0%)
		M-estimator	1 (0.86%)	0 (0%)	0 (0%)	1 (0.27%)
		Trimmed Mean	0 (0%)	0 (0%)	1 (0.82%)	1 (0.27%)
		Total	1 (0.86%)	0 (0%)	1(0.82%)	2 (0.55%)
Number of Tests		Range	1-11	1-17	1-16	1-17
		Mean	2.72	2.65	2.69	2.68
		SD	1.71	1.95	1.82	1.83

As reported in Table 4, above, the basic analyses had the highest percentage (68.53%), among the three classifications, followed by intermediate (33.66%) and advanced (11.2%) ones. Within the basic analyses, independent samples t-test (38.04%) and one-way ANOVA (33.72%) were the two most commonly used analyses. Meanwhile, ANCOVA (43.72%) and repeated measures ANOVA (65.45%) were the most frequently-used analyses within intermediate and advanced analyses, respectively. Moreover, the use of robust statistics was rare (use in only two papers of the sample) while the average number of statistical tests used in papers was 2.68.

Looking into the numbers over time, it is evident that the use of intermediate tests has increased from 17.78% in papers published before 2014 to 28.22% in papers published from 2017 to 2019 with a slight fall-down in the mid-way. This coincided with the decrease in the use of basic analyses from 74.29% to 61.04%. This indicates a shift from using basic analyses to intermediate ones. The changes in the number of advanced tests and the total number of tests per paper were very slight.

The next issue related to statistical analyses was checking the normality and test specific assumptions. To do so, the profile used by Hu and Plonsky (2019) was used. Papers with strict adherence to the assumptions were considered *stringent*; those with partial adherence as *lenient*; those which provided information without reporting the results of assumption checking as *minimal information*. Overall, normality was checked in 36.88% of the papers, increasing from 21.55% in papers before 2014 to 33.33% in papers from 2014 to 2016, and then, to 55.37% in papers from 2017 to 2019. However, checking test-specific assumptions was not reported in 60.65% of the papers. Although the number decreased from 71.55% in papers before 2014 to 62.79% in papers from 2014 to 2017, and then, to 47.9% in papers from 2017 to 2019. Moreover, only 4.37% of the papers had stringent assumption checking while lenient checking was 28.14% and minimal information 6.83%. Figure 2 depicts the percentage of assumption checking in the papers.

4.4. Issues Related to Reporting Practices

The next issue is the reporting practices and data sharing in the papers. Table 5 summarizes the results. As reported, there was a rarity of reporting pre-determined alpha in all three periods. Reliability was a highly-adherend issue with some slight increases from 2014 on. The validity, however, showed slight decreases over time, and the total percentage of papers reporting it was 39.62%. The descriptive statistics were reported in a large proportion of the papers with slight positive changes from 2014 onwards while inferential statistics were fully reported in about half of the sample and cases of partial reporting remained constant over time. P values were reported in 96.72% of the papers. However, exact p values were only reported in 71.58% of them and no significant change in reporting exact p values was observed over time.

Table 5

Percentage of the Adherence to Reporting Issues and Data Sharing Over Time

		Year			Total
		Before 2014	2014 to 2016	2017 to 2019	
Pre-determination of α		6 (5.17%)	4 (3.10%)	3 (2.48%)	13 (3.55%)
Reliability		85 (73.28%)	100 (77.52%)	96 (78.69%)	281 (76.57%)
Validity		52 (44.83%)	45 (34.88%)	48 (39.67%)	145 (39.62%)
Descriptive statistics	M and SD/ Frequency	103 (88.79%)	119 (92.25%)	112 (91.80%)	334 (91.01%)
	Only M	1 (0.86%)	1 (0.78%)	3 (2.46%)	5 (1.36%)
	Not reported	12 (10.34%)	9 (6.98%)	7 (5.74%)	28 (7.63%)
Inferential statistics	Fully reported	61 (52.59%)	67 (51.94%)	63 (52.07%)	191 (51.86%)
	Partially reported	45 (38.79%)	51 (39.54%)	48 (39.67%)	144 (39.34%)
	Not reported	10 (8.62%)	11 (8.52%)	10 (8.27%)	31 (8.47%)
P-value	Exact *p*	45 (38.79%)	55 (42.64%)	49 (40.50%)	149 (40.71%)
	p <alpha	27 (23.28%)	32 (24.81%)	33 (27.27%)	92 (25.14%)
	Exact *p* < alpha	37 (31.89%)	38 (29.46%)	38 (31.41%)	113 (30.87%)
	Not reported	7 (6.04%)	4 (3.10%)	1 (0.83%)	12 (3.28%)
Non-significant results	Reported	3 (18.75%)	5 (19.23%)	3 (12%)	11 (16.41%)
	Ratio per paper	13 (11.21%)	23 (17.83%)	16 (13.22%)	52 (14.21%)
Effect size		39 (33.62%)	55 (42.64%)	56 (46.28%)	150 (40.98%)
Confidence Intervals		40 (34.48%)	45 (34.88%)	65 (53.71%)	150 (40.98%)
Data Sharing		0 (0%)	0 (0%)	0 (0%)	0 (0%)

With regards to reporting non-significant results, from 67 cases identified in 52 papers, only 16.41% of the cases were reported. The numbers showed decreases over time. Effect size and confidence intervals were reported similarly in 40.98% of the papers while in some cases only either of them was reported. The increase in the reporting percentage of both of them was observed from 2014 on. Finally, no single instance of data sharing was found in the papers.

4.5. Issues Related to Visual Presentation

The final issue to be addressed was the use of visual graphics in the papers. Table 6 summarizes the obtained results.

Table 5

Percentage of Visual Graphics Used in Papers Over Time

		Year			Total
		Before 2014	2014 to 2016	2017 to 2019	
Data Poor	Bar chart	25 (7.88%)	35 (51.47%)	36 (46.75%)	96 (36.09%)
	Line plot	38 (57.58%)	30 (44.12%)	40 (52.95%)	108 (40.6%)
	Pie	3 (4.54%)	3 (4.41%)	1 (1.3%)	7 (2.63%)
	Number of papers	32 (27.59%)	39 (30.23%)	40 (32.79%)	111 (30.25%)
Data Rich	Boxplots	6 (60%)	2 (28.57%)	0 (0%)	8 (3.01%)
	Histograms	4 (40%)	5 (71.43%)	5 (100%)	14 (5.26%)
	Number of papers	8 (6.9%)	7 (5.43%)	3 (2.46%)	18 (6.77%)
Data Accountable	Scatter plot	5 (100%)	1 (100%)	7 (100%)	13 (4.89%)
	Beesworm	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Pirate	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Number of papers	5 (4.31%)	1 (0.78%)	5 (4.1%)	11 (3%)

As reported, 30.25% of the papers used data poor graphics and the use of data rich (6.77%) and data accountable (3%) graphics was rare. The rest of papers did not use any visual presentations. The most frequent graphics used were line plots (40.6%) and line plots (36.09%) while the only data accountable graphic was scatterplot (4.89%). The only two types of data rich graphics, i.e., boxplots (3.1%) and histograms (5.26%) were also rarely used. No significant change was observed in the use of graphs over time.

5. Discussion

The obtained results showed almost low adherence to the majority of study quality facets in Iranian papers. The pattern stayed constant or had slight changes in a large proportion of the cases. In what follows we will discuss each facet in detail.

First, with regards to the sampling, although the average number of participants per group seemed adequate, the lack of power analysis to determine the sample size was a serious problem. These results were in line with previous findings in L2 research (e.g., Plonsky, 2013, 2014b; Plonsky & Gass, 2011; Ziegler, 2013). As reported, in the present study, range of participants in studies was 7 to 350. The problem is that, with a sample large enough, any statistical test can be found significant (Hudson & Llosa, 2015; Plonsky, 2013; Norris, 2015). In other words, the lack of appropriate power can easily lead to Type I error – therefore, threatening the internal validity of the research – in the common practice of NHST (Norris, 2015). Interestingly, the same problem happens with a small sample. As effect size and sample size conjointly and inversely impact power (Plonsky & Gonulal, 2015, p. 16), having small samples often lead to type II error – threatening the internal validity, again. The next sampling issue was the dominance of convenience sampling. Although as DeKeyser, Alfi-Shabtay, and Ravid (2010) maintain, "almost every sample has been one of convenience" (p. 416), the samples usually used in L2 research are specific to limited numbers of participants in certain regions with certain cultures or background. This highlights the need for comprehensive reporting of the results, so that the secondary studies may reach a conclusive point about the issue at hand.

The second facet focused on the design-related issues. The first concern, that is closely related to the sampling issues, is the random assignment. The results showed that, unlike the findings of Plonsky (2013), a large proportion (77.7%) of Iranian L2 experiments used random assignments (47.3% by individual and 30.4% by group). The use of classroom studies was also highly-frequent as opposed to the results reported by Plonsky (2013) and Plonsky and Gass (2011). Pretesting was an upheld issue from the beginning and the use of control/comparison groups was collectively large while delayed-posttesting and qualitative triangulation were low. The proportions remained almost constant over time and unlike the findings reported by Plonsky (2014b) no improvement was seen in delayed-posttesting. Using delayed posttesting was also reported by Ajideh et al. (in press), among the items that Iranian authors perceived as a feature with a low association to the study quality.

Third, regarding the statistical tests, the results supported the findings of Khany and Tazik (2019) as the most common analyses in the papers were the basic ones (68.53%). Moreover, the number of tests per paper ranged from 1 to 17 with an average value of 2.68. This indicates multiple statistical tests in the majority of papers. The problematic part in such studies is the ignorance of the fact that within each new, or repeated, inferential analysis from the same sets of data, the errors from the other analysis are compounded; therefore, the alpha level should be adjusted by some omnibus analysis before making the conclusion (Larson-Hall, 2012; Norris, 2015; Norris, et al., 2015). Considering that majority of the tests in the sample were basic ones, which do not lend to adjustment of alpha level, this issue can be considered a challenging one in Iranian L2 papers.

Moreover, with regards to the assumption checking, the results showed a low frequency of checking normality of distribution (36.88) and reporting results for test-specific assumptions (39.45). Although improvements over time were found with respect to both, the results obtained from the most recently-published papers were still low. These results are in line with the findings of Plonsky (2013) and Hu and Plonsky (2019). If the assumptions are not met, the test might be prevented from identifying statistical significance, i.e., type I error (Plonsky, 2013; Plonsky et al., 2015). While not reporting the results of checking assumptions is not necessarily an indicator of not doing them, Plonsky's (2013) cross-tabulation on studies that did or did not report checking assumptions showed that

studies in which the assumptions were checked and reported were five times more likely to employ a nonparametric test than those in which the assumption-checking was not reported. Wilcox (2005, p. 1) addresses the tendency of users to use standard parametric tests as "a false sense of security". The researchers seem to consider these parametric tests robust to violations of assumptions and, thus, not reporting them checked, if not checking at all. An alternative would be using robust statistics as they are robust to violation of assumptions. However, the obtained results showed a rarity in using such statistics.

The fourth facet of quality dealt with reporting practices and data sharing. The first feature is the pre-determination of alpha which can be done with the help of prior power analysis. As there was a rarity of power analysis in the sample, it is not surprising that pre-determination of the alpha level was only done in a handful of cases. On the other hand, reliability, as one of the features considered highly-associated with study quality in the findings of Ajideh et al. (in press), was reported. Moreover, the proportion of the studies that reported descriptive statistics, including both mean and standard deviation (the fundamental values for meta-analyses), was large (91.01%), paving the way for conducting secondary analyses. The inferential statistics, on the other hand, were only fully reported in about half of the papers. Given that the new statistics, i.e., effect size and confidence intervals, were not also reported in about 60% of the sample, the precision of the results may not be evaluated in future meta-analyses. According to Cumming (2012), meta-analysis "gives an overall interval estimate that signals how precise an estimate the weighted average is likely to be" (p. 5). Therefore, reporting intervals, apart from its usefulness in determining the precision of the results obtained from the study, would help secondary research in estimating the overall average of precision; the lack of which is the widespread problem in L2 research (Larson-Hall & Plonsky, 2015).

Furthermore, the fact that p-values were reported in 96.82% of the cases while new statistics were absent in more than half of them indicates the over-reliance on NHST. As quantitative L2 researches are mostly depending on mean-based analyses (Gass 2009; Plonsky 2013, 2014a), a thorough report of inferential statistics is necessary. As stated by Plonsky (2015), "the absence of evidence is not evidence for absence" (p. 235). The other cause for such reluctance in reporting non-significant results. The results showed that non-significant results existed in only 14.21% of the published papers. Moreover, only 16.41% of these results were thoroughly reported. It may be caused by the common-belief among the researcher that if they don't reach the significant results, their study has failed. Therefore, they seem to be reluctant to report these results. However, putting non-significant results aside in reporting the results would generate literature contaminated with *publication-bias* in secondary researches (Rosenthal, 1979), Which, in turn, would result in failure in forming appropriate future theories and practices (Plonsky & Gonulal, 2015). Finally, no single study claimed data sharing. According to Vines et al. (2014), the raw data are often subject to missing over time if they are not stored using external methods of storage. That is why sharing the data must be taken seriously.

The final issue is the use of visual graphics. The results showed that only about 32% of the papers used graphics. These results were in line with the findings of Plonsky (2013) and Larson-Hall (2017). Larson-Hall and Plonsky (2015) argued that graphs should not be considered as a nice accompaniment to the research, but a necessary means to understand and convey the findings of the research. Norris et al. (2015) also recommend considering if graphic techniques are helpful to present both individual variability and main patterns of the results, such as mean and dispersion of data around it, regression lines, confidence intervals, etc.

6. Conclusion

In conclusion, it seems that Iranian journals need to improve their adherence to study quality standards. Although they showed high adherence to standards of study quality in some of the fundamental issues in designing and reporting the results, there are areas that need more consideration to be taken into account by both authors and journal editors. The changes over time also showed that few of these issues are taken more seriously. However, it seems that there is still a long way to go.

Gass et al. (2020) describe the present of study quality in the field of applied linguistics as promising. They point out that discipline-based research books that are focusing on quantitative methods are flowing while journals are publishing special issues to highlight the essentiality of the matter. L2 journals are also refining their submission guidelines, asking the authors to follow the quality standards. Iranian journals are also encouraged to use such updated submission guides requiring the authors to adhere to the standards. For example, by putting emphasis on the precision, rather than merely effectiveness of the results, journals can encourage the authors to publish both non-significant and significant results and make their interpretations based on the new statistics. They can also promote open science by asking for transparent and thorough reporting and sharing the data.

Authors are also encouraged to develop their knowledge of advanced quantitative methods and statistics by taking training and developing methodological specialties. As the results of this study were in line with the findings of Ajiedeh et al. (in press), it is expected that such training would contribute to authors' understandings of the factors associated with high-quality research and, in turn, would result in boosting the quality in their research practices.

Finally, this study showed that the existing concerns in L2 studies published in high-ranked international journals are also the challenges locally-published journals are dealing with. While at present, as maintained by Gass et al. (2020), the issues are highlighted in books, special issues published by journals, updated submission guidelines, published recommendations for conducting and reporting with high-quality, classroom training and workshops, the local journals and authors need to catch up. The slight changes shown in this research might be existing in papers published locally in other EFL contexts. It is recommended that journals and authors take the same path (e.g., publish special issues, update submission guidelines, take part in methodological training courses, etc.) if they wish not fall behind in the movement.

It should not be left unmentioned that this research was limited in its scope as only experimental studies were included. Future researches may contribute to a more comprehensive understanding of quality issues in other types of studies. Moreover, the focus of the study was papers published in Iranian Journals. The issues of quality would be more captured by inspection of papers in other EFL or ESL contexts. Finally, this study examined the papers based on the existing guidelines and recommendation of study quality. Future researches may introduce new aspects and try to incorporate them into the framework use in this study.

Declarations

Availability of Data and Material

The datasets generated and analyzed during the current study are available in the Mandeley Data repository, https://data.mendeley.com/datasets/2c3w6h95n4/1.

Funding

The authors received no funding for this research.

Competing Interest

We hereby declare that the authors have approved the paper for release and are in agreement with its content. There is also no (non-)financial competing interest in relation to this manuscript.

Authors' Contribution

This study is a part of bigger study conducted as a doctoral dissertation of the corresponding author. The second author was responsible for gathering, coding, and analyzing the results. The first author supervised process and provided guidelines. He also contributed as the second coder of the data.

Bionote

First Author holds PhD in Teaching English as a Foreign Language (TEFL). He has published in both local and international journals and his areas of interest are research methodology, quantitative research, and statistical analyses.

Second Author is a full professor in TEFL and has done and supervised researches in different areas, including research methodology, English for specific purposes, cultural studies, etc. He has published both locally and internationally.

Acknowledgments

We are obliged to two leading experts in the field, Dr. Jenifer Larson-Hall and Dr. Luke Plonsky who provided their comments and helped us in obtaining the final coding scheme used in this study.

References

Ajideh, P., Zohrabi, M., & Jahanbakhsh, A. A. (in press). Study quality in quantitative L2 studies: A path Analysis on the perceptions of Iranian published authors. *Journal of Modern Research in English Language Studies*. http://dx.doi.org/10.30479/jmrels.2020.14296.1760.

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. http://dx.doi.org/10.1037/amp0000191.

Breckler, S. (2009). Dealing with data. *American Psychological Association: Science Directions*. http://www.apa.org/monitor/2009/02/sd.aspx

Chaudron, C. (2001). Progress in language classroom research: Evidence from the *Modern Language Journal*, 1916-2000. *Modern Language Journal*, *85*, 57–76. https://doi.org/10.1111/0026-7902.00097.

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.

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. https://doi.org/10.1017/S0142716410000056.

Derakhshan, A., Shakki, F. (2021). A meta-analytic study of instructed second language pragmatics: A case of the speech act of request. *Journal of Research in Applied Linguistics*, *12*(1), 15-32. https://doi.org/10.22055/rals.2021.16722.

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. https://doi.org/10.1017/S0261444819000430.

Hu, Y., & Plonsky, L. (2019). Statistical assumptions in L2 research: A systematic review. *Second Language Research*. https://doi.org/10.1177/0267658319877433.

Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research. *Language Learning, 65*(Suppl. 1), 76–96. https://doi.org/10.1111/lang.12113.

Khany, R., & Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: From 1986-2015. *Journal of Quantitative Linguistics, 26*(1), 48–65. https://doi.org/10.1080/09296174.2017.1421498.

Larson-Hall, J. (2012). Our statistical intuitions may be misleading us: Why we need robust statistics. Language Teaching, 45(4), 460-474. https://doi.org/10.1017/S0261444811000127.

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. https://doi.org/10.1111/modl.12386.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(Suppl. 1), 127–159. https://doi.org/10.1111/lang.12115.

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. https://doi.org/10.1177/0267658316684904.

Norris, J. M. (2015). Statistical Significance Testing in Second Language Research: Basic Problems and Suggestions for Reform. *Language Learning*, *65*(Suppl. 1), 97-126. https://doi.org/10.1111/lang.12114.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136.

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. https://doi.org/10.1111/lang.12104.

Ortega, L. (2005). Methodology, epistemology, and ethics in instructed SLA research: An introduction. *Modern Language Journal, 89*(3), 317 - 327. https://doi.org/10.1111/j.1540-4781.2005.00307.x.

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

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research, 3*(1), 61–94. https://doi.org/10.1075/ijlcr3103paq.

Plonsky, L. (2011). *Study quality in SLA: A commulative and developmental assessment of designs, analyses, reporting practices and outcomes in quantitative L2 research*. [Doctoral dissertation, Michigan State University]. MSU Libraries. https://d.lib.msu.edu/etd/1417/datastream/OBJ/download/Study_quality_in_SLA___a_cumulative_and_developmental_assessment_of_designs__analyses__r

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. https://doi.org/10.1017/S0272263113000399.

Plonsky, L. (2014a). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*(1), 450–470. https://doi.org/10.1111/j.1540-4781.2014.12058.x.

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. https://doi.org/10.1111/j.1467-9922.2011.00640.x.

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. https://doi.org/10.1111/lang.12111.

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 106-128). Routledge.

Plonsky, L., Egbert, J., & Laflair, G. T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics, 36*(5), 591-610. https://doi.org/10.1093/applin/amu001.

Rosenthal, R (1979). File drawer problem and tolerance for null results. *Psychol Bull. 86*, 638-41. https://doi.org/10.137/0033-2909.86.3.638.

Rosenthal, R. and DiMatteo, M.R. (2001) Meta-Analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82. https://doi.org/10.1146/annurev.psych.52.1.59

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. https://doi.org/10.1037/0003-066X.54.8.594.

Yousefi, M., & Nassaji, H. (2019). A meta-analysis of the effect of instructions and corrective feedback on L2 pragmatics and the role of moderator variables: Face-to face vs. computer-mediated instruction. *ITL - International Journal of Applied Linguistics, 170*(2), 277-308. https://doi.org/10.1075/itl.19012.you.

Ziegler, N. (2013). *Synchronous computer-mediated communication and interaction: A research synthesis and meta-analysis*. [Unpublished doctoral dissertation, Georgetown University]. DigitalGeorgetown, https://repository.library.georgetown.edu/bitstream/handle/10822/559497/Ziegler_georgetown_0076D_12341.pdf;sequence=1.

Figures

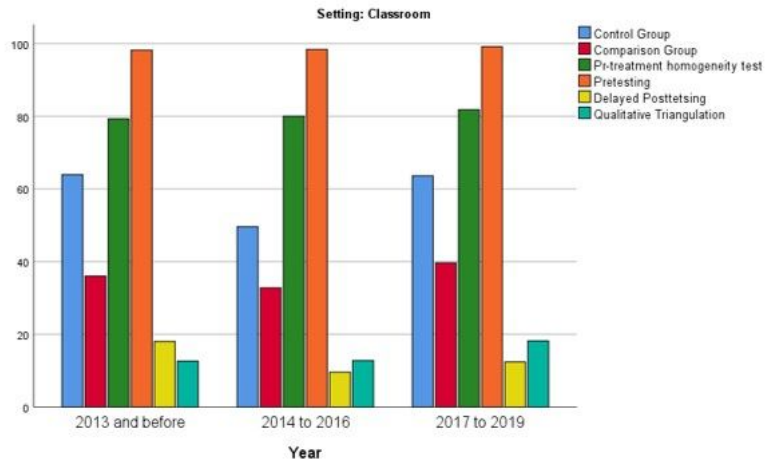

Figure 1

Overtime Changes in the Percentage of Design Features in Classroom-Based Studies

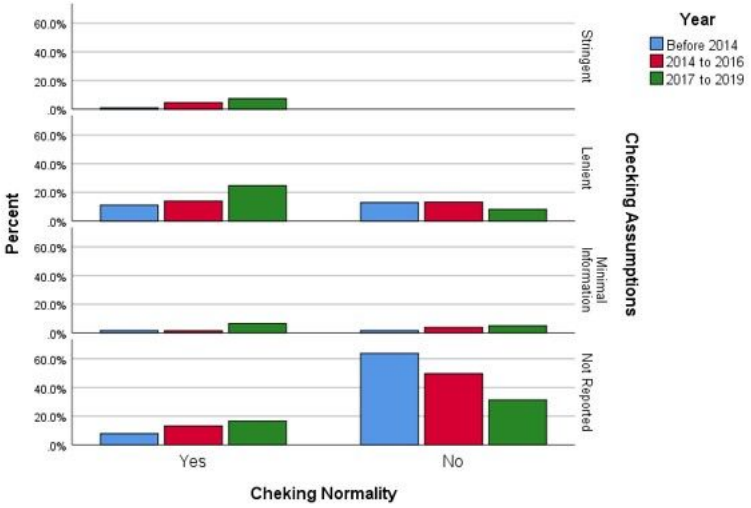

Figure 2

Overtime Changes in the Percentage of Assumption-Checking