
Takumi AOYAMA
University of Warwick
Mami HOSHIKA
Chiba University
Katsuhisa HONDA
Chiba University

Abstract

In recent years, discussion has appeared in several international journals in the field of English language teaching and L2 research about reporting practices and study quality. However, the journal Annual Review of English Language Education in Japan (ARELE) has no reporting standards or guidelines. This paper explores the reporting practices of articles published in ARELE by adopting a method of systematic review. The article pool was generated from a collection of articles reporting on quantitative research published in ARELE between 2002 (Volume 13) and 2017 (Volume 28). The review aims to investigate to the extent to which the reporting practices of ARELE follow the American Psychological Association’s guidelines, which are commonly adopted by ARELE contributors. Finally, drawing upon the results of the review, we propose a basic checklist for reporting quantitative research that future issues of ARELE can adopt to improve the quality of papers.

1. Introduction

In recent years, high-impact international journals have published reporting guidelines either as articles or official statements by the editorial team. In 2015, for example, Language Learning published an article providing guidelines for reporting quantitative research (Norris, Plonsky, Ross, & Schoonen, 2015). The Modern Language Journal also clearly states the requirements for reporting outcomes in its author guidelines, and in 2018, American Psychologist published reporting guidelines for both quantitative and qualitative studies in the field of psychology (Appelbaum et al., 2018; Kazak, 2018; Levitt et al., 2018).

Annual Review of English Language Education in Japan (ARELE), one of the most prestigious journals in Japan, has been publishing articles related to English language teaching and second language (L2) study since 1990. Although it follows the formatting and typesetting standards
published by the American Psychological Association (APA) (American Psychological Association, 2009), the submission guidelines for ARELE volumes 13–28 make no reference to the reporting standards expected (i.e., what should be reported). As is common practice in the literature of English language teaching, most researchers have followed the APA guidelines. However, although the reporting practices of articles published in ARELE seem to follow these guidelines, the absence of specific editorial guidelines means that the reporting practices of certain items lack consistency.

This study aims to systematically review the reporting practices of quantitative research papers published in ARELE to see if these are consistent and follow clear standards by comparing them with the guidelines published by international applied linguistics and L2 research journals that utilize the APA guidelines of 2001 or 2009. We also aim to propose a basic checklist for reporting quantitative studies for future volumes of ARELE. To review reporting practices, the following research questions are examined.

(1) Do papers published in ARELE 13–28 properly report quantitative outcomes by following the standards of the APA?
(2) If not, which items are and are not properly reported in ARELE 13–28?

2. Background

As testimony to the importance of L2 research and corresponding research practices, two major reports were recently published: a synthesis of quantitative research entitled Study Quality in Quantitative L2 Research (1990–2010): A Methodological Synthesis and Call for Reform (Plonsky, 2014), and a set of reporting standards entitled Guidelines for Reporting Quantitative Methods and Results in Primary Research (Norris et al., 2015). Both reports recommend that editors, reviewers, and researchers improve the overall quality of L2 research. The reports emphasize the established expectations for certain reporting practices, which were reiterated in the Publication Manual of the American Psychological Association (American Psychological Association, 2009) and Standards for Reporting on Empirical Social Science Research in AERA Publications (American Educational Research Association [AERA], 2006). These recommendations include reporting inferential statistics, means, standard deviations, \( p \) values, effect sizes, and confidence intervals. Plonsky (2013) assessed reporting practices in quantitative L2 research by reviewing a sample of 606 primary studies published from 1990 to 2010 in Language Learning and Studies in Second Language Acquisition. These displayed incomplete and inconsistent reporting practices, lacking (a) effect sizes, (b) confidence intervals, and (c) reliability and validity coefficients. Larson-Hall and Plonsky’s (2015) findings show that “there are some deficits within L2 quantitative research in reporting in all of the areas we have surveyed” (p. 151), and recent researchers have found inadequate reporting of important data. In descriptive statistics, for example, this includes reporting means without reporting standard deviations and giving insufficient attention to effect sizes,
confidence intervals, reliability measurements, graphics, and data sharing.

In the early 1990s, there was lengthy discussion about the importance of effect sizes and debate on the use of null hypothesis statistical significance testing (NHST). One result of this was the 1996 formation of the APA Task Force to address issues related to statistical significance testing and effect size reporting. Following this, in 1999, Leland Wilkinson and the Task Force on Statistical Inference (for the APA Board of Scientific Affairs) published recommendations for reporting practices (Wilkinson and TFSI, 1999). In 2001, the fifth edition of the APA Publication Manual listed “failure to report effect sizes” as one of the “defects” that editors and reviewers watch for in evaluating a research paper (p. 5).

In 2010, the sixth edition adopted the publication standards recommended by the Task Force, which require the reporting of effect sizes and encourage the use of confidence intervals. The APA manual also prescribes reporting of standard deviations to accompany means, exact \( p \) values, exact test statistics, and degrees of freedom as well as effect sizes and confidence intervals referring to point and interval estimates rather than the binary outcomes of statistical tests (see Result section of 2.07 in the APA manual). Reporting effect sizes contextualizes the impact of the treatment (Thompson, 2008).

Reviews of L2 research (e.g., Larson-Hall, 2010a, 2010b; Larson-Hall & Plonsky, 2015; Plonsky, 2013, 2014) have found that, except for regressions and correlations, researchers infrequently reported effect sizes. Effect sizes should be reported for both statistically significant and statistically nonsignificant results (see Method section of 2.06 in the APA manual). However, it seems that most of the studies reviewed were conducted by univariate or non-parametric tests without including effect sizes. An infinitesimal \( p \) value does not necessarily make the effect magnitude important, and a requisite power analysis should be conducted before beginning a study. According to Thompson (2008), even diminutive effect sizes will be statistically significant, and the null hypothesis will be rejected if the sample size is large enough. Larson-Hall and Plonsky (2015) explained this clearly and succinctly:

The reader says, “Oh! The \( p \) value is less than .05, so the result is important” or “I see … the \( p \) value is greater than .05, so I can ignore the results.” However, \( p \) values depend on sample size and the amount of variability in the data and thus a small (statistically significant) \( p \) value may actually describe an unimportant result while a larger (statistically nonsignificant) one may hide an important result. (pp. 134–135)

Even if research outcomes without sufficient sample size are given an anticipated effect, readers cannot confirm that the statistically nonsignificant result is not just an artefact of sample size (AERA, 2006; Cohen, 1988).

Researchers conducting null-hypothesis testing should be also report confidence intervals, because statistical significance testing is dependent on sample size and does not evaluate result importance or replicability. Although 26% of the studies in Plonsky’s (2013) survey sample reported on effect size, only 5% of the cited articles reported confidence intervals. Confidence
intervals across studies tell us how accurately and consistently effects operate over time. Wilkinson and TFSI (1999) suggested that confidence intervals should always be reported, and the use of confidence intervals is strongly recommended in the APA Publication Manual (APA, 2009). Confidence intervals are “the best reporting strategy” (APA, 2009, p. 34), and they are therefore important tools for describing results and measuring whether we can be sure of our results. The APA manual also indicates that “it is best to use a single confidence level, specified on an a priori basis (e.g., a 95% or 99% confidence interval), throughout the manuscript” (p. 34). As noted in Cumming (2014), if researchers present a range of values with a confidence level of 95%, the same research can be replicated and the same point estimate (the accuracy of correlation coefficients) can be expected.

Reporting descriptive statistics is advocated by the APA manual. These include dispersion statistics, which characterize how well central tendency statistics represent the data. In addition to reporting effect sizes and confidence intervals, both means and standard deviations should always be reported. This allows an understanding of the quality of the mean by characterizing the score spread of the data. Pictorial representations of results are also useful to provide sufficient information of the central tendency of the scores. The APA manual emphasizes using visual displays of data as another way of accurately representing all the scores when dispersion is smaller. Visual representations can aid the detection of trends in data and help identify outliers and score clusters. Many of the quantitative research studies reviewed contained figures (e.g., AERA, 2006; Wilkinson & TFSI, 1999), but these were mainly limited to bar graphs, histograms, or circle graphs. Larson-Hall and Plonsky (2015) provided an example of a box plot for two experimental groups, recommending L2 researchers to make broad use of tables and figures (e.g., box-and-whisker plots) and noting that “a lack of visuals indicates a missed opportunity to convey findings succinctly to readers” (p. 145). Researchers need to provide visual displays so readers can (a) formulate their own well-informed interpretations of results, (b) easily compare groups, (c) view the spread of data, and (d) identify the impact of outliers.

Recent surveys of L2 reporting practices have found that the author guidelines of many journals require the reporting of effect sizes (Larson-Hall & Plonsky, 2015; Norris et al., 2015; Plonsky, 2013, 2014). Larson-Hall and Plonsky (2015) proposed a set of recommendations for five main areas of quantitative reporting:

1. descriptive statistics and other measures of study quality such as power,
2. effect sizes and confidence intervals,
3. instrument reliability,
4. visual displays of data, and
5. raw data.

They recommended nine reporting items for each of these five areas (pp. 152–154).

Over the past decade, consensus on the reporting practices of international journals has emerged. In the present article, we examine the analytic and reporting features of research studies
that used quantitative reporting practices in ARELE journals 13–28. Our purpose is to identify quantitative reporting practices and standards that can be improved in the interests of creating the best possible evidence base for L2 research. Mizumoto, Urano, and Maeda (2013) reviewed three representative aspects (themes, methods, and outcomes) of the published 450 articles in ARELE journals 1–24, and Stapleton and Collett (2010) reported the 297 articles published in JALT journal (the journal of the Japan Association for Language Teaching) over a quarter of a century. These results serve as a benchmark for determining the impact of changing trends on English education in Japan. However, there is little research reporting English education practices in Japan and seeking to identify deficiencies that need to be addressed. Adopting standard reporting practices would guide and facilitate the production of high-quality L2 studies and provide a basis for future research in Japan.

3. Method

3.1 Article Pool

All articles published in ARELE between 2002 and 2017 (K = 399) were downloaded from the J-STAGE online database, and metadata (Author, Year, Title) were recorded in a spreadsheet. Articles not published online were manually scanned to PDF files from the printed editions, and information was recorded. Articles recorded in the spreadsheet were then coded (0, exclude: 1, include: 9, not sure) by two of the authors to eliminate articles that did not report quantitative research. Evaluations by the two raters were gathered (Cohen’s kappa = .708), and disagreements were moderated. As a result, 345 articles were included in the final article pool.

3.2 Coding

For the evaluation of reporting practices, a coding framework (Table 1) was developed drawing upon Plonsky and Gass’s (2011) evaluation criteria and Larson-Hall and Plonsky’s (2015) reporting guidelines.

Table 1
Framework used for the Coding

<table>
<thead>
<tr>
<th>Coding category</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification</td>
<td>Author, Year, Title</td>
</tr>
<tr>
<td>Reported values a</td>
<td>Reliability, Power, p value (exact, &gt;, or &lt;), Mean, Standard deviation</td>
</tr>
<tr>
<td></td>
<td>Confidence intervals, Effect size b</td>
</tr>
<tr>
<td>Visual displays a</td>
<td>Yes/No, Bar chart, Line chart, Box plot, Scatter plot, Others</td>
</tr>
</tbody>
</table>

Note. a Coded as 0, No: 1, Yes. b Coded as 0, No: 1, Reporting effect size(s) other than r: 9, Reporting r only.
The framework consisted of three main components:

1. metadata (i.e., Author, Year, Title) of the article,
2. statistical values reported, and
3. use of visual data presentation.

All the papers were coded independently by two of the authors, and consistency of the coding was checked.

4. Summary of Findings

4.1 Reliability, Power, Descriptive values, CI, p value, Effect size

Table 2 shows the frequencies and percentages of data reported in the articles published in 2002–2009 and 2010–2017. In the whole article pool, means and standard deviations were the most frequently reported items ($k = 311$, 90.14%, and $k = 291$, 84.35%, respectively). A total of 144 articles (85.21%) in the 2002–2009 group reported means, and 131 (77.51%) reported standard deviations, while the reporting rates in 2010–2017 samples were 94.89% ($k = 167$) and 90.91% ($k = 160$), respectively. Many articles included both means and standard deviations in the 2002–2009 and 2010–2017 groups ($k = 130$, 76.92%, and $k = 160$, 90.91%, respectively). This suggests that the practice of reporting means and standard deviations was improving. However, some articles did not include either means or standard deviations ($k = 22$, 6.38% of the whole sample). Although the number of articles in 2010–2017 that reported means without their associated standard deviations was half that of 2002–2009, there were still seven articles. The least frequently reported item was power ($k = 2$, 0.58%). The frequency of articles reporting the results of power analysis did not change over the two periods, suggesting that the practice of reporting the results of power analysis remained unfamiliar.

Approximately half of all articles reported exact $p$ values ($k = 184$, 53.33%), and 35.94% ($k = 124$) reported values as greater or less than a particular $p$ value (e.g., $p < 0.05$; $p < 0.01$) or not significant (n.s.). During 2002–2009, the proportion of articles reporting exact $p$ values was not high ($k = 67$, 39.64%). In this period, almost half the samples reported values as being greater or less than a particular $p$ value ($k = 83$, 49.11%). After the release of the sixth edition of the APA Publication Manual (APA, 2009), the number of articles which reported exact $p$ values increased ($k = 117$, 66.48%); however, there were still 41 articles that presented statistical significance as greater or less than a particular $p$ value ($k = 41$, 23.30%). Compared with the results obtained from the articles published in 2002–2009, the 2010–2017 results showed an increase in the percentage of reported exact $p$ values; however, this percentage was still very low.

Confidence interval data is also required to be reported (American Psychological Association, 2009). Few articles in the whole data pool reported confidence interval data ($k = 31$, 8.99%), and almost all of these were published in 2010–2017 ($k = 30$, 17.05%). This suggests that although the reporting of confidence interval remains infrequent, there is a growing awareness of its importance.
In the entire sample, 117 studies reported effect sizes ($k = 117, 33.91\%)$. As with exact $p$-values and confidence intervals, most of the articles reporting effect sizes were published in 2010–2017 ($k = 97, 55.11\%$). This may be because the *APA Publication Manual* was published in 2009. Before 2009, 20 articles reported effect sizes (11.83\%) for the outcomes of non-correlational analysis (e.g., $t$-test, ANOVA). There were studies that used both correlational analysis and non-correlational methods such as $t$-test, only reporting effect sizes for correlational analysis ($k = 51, 30.18\% \text{ [2002–2009]}, k = 24, 13.64\% \text{ [2010–2017]}$).

Overall, 40.58\% of articles reported either inter-rater reliability for studies evaluating writing or speaking, or Cronbach’s alpha coefficients for questionnaires ($k = 140$). In the period 2002–2009, 32.54\% of articles reported reliability ($k = 55$). This increased to 48.30\% ($k = 85$) in the period 2010–2017. However, some articles failed to report reliability, even though they described questionnaire surveys.

### Table 2

*Summary of Frequencies and Percentages of the Reporting Rates*

<table>
<thead>
<tr>
<th></th>
<th>Overall ($K = 345$)</th>
<th>2002–2009 ($k = 169$)</th>
<th>2010–2017 ($k = 176$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>%</td>
<td>$k$</td>
</tr>
<tr>
<td>Reliability</td>
<td>140</td>
<td>40.58</td>
<td>55</td>
</tr>
<tr>
<td>Power</td>
<td>2</td>
<td>0.58</td>
<td>1</td>
</tr>
<tr>
<td>Exact $p$-values</td>
<td>184</td>
<td>53.33</td>
<td>67</td>
</tr>
<tr>
<td>$p$-values (&lt; or &gt;)</td>
<td>124</td>
<td>35.39</td>
<td>83</td>
</tr>
<tr>
<td>$M$</td>
<td>311</td>
<td>90.14</td>
<td>144</td>
</tr>
<tr>
<td>$SD$</td>
<td>291</td>
<td>84.35</td>
<td>131</td>
</tr>
<tr>
<td>Confidence intervals</td>
<td>31</td>
<td>8.99</td>
<td>1</td>
</tr>
<tr>
<td>Effect sizes</td>
<td>117</td>
<td>33.91</td>
<td>20</td>
</tr>
<tr>
<td>Visual presentation</td>
<td>208</td>
<td>60.29</td>
<td>92</td>
</tr>
</tbody>
</table>

### 4.2 Visual Data

As shown in Table 3, 60.29\% of all articles included visual presentations. The number of articles using visual presentation increased from 92 (54.44\%) in 2002–2009 to 116 (65.91\%) in 2010–2017.

The frequencies and percentages of each of the visual data types are summarized in Table 3. Line plots were the most common type of visual presentation method ($k = 11, 32.46\%$). In 2002–2009, 33.14\% ($k = 56$) of studies used line plots. This decreased slightly in 2010–2017 to 31.82\% ($k = 56$); however, line plots were still frequently adopted. The other common visual presentation style was the bar plot ($k = 93, 26.96\%$ of the total sample). In 2002–2009, bar plots ($k = 34, 20.12\%$) were second to line plots ($k = 56, 33.14\%$) as the preferred visual presentation. In 2010–2017, the
number of bar plots increased to 59 (33.52%), replacing line plots ($k = 56$, 31.82%) as the most popular form of graphic representation.

Box plots and scatter plots were rarely seen in the article pool ($k = 4$, 1.16% and $k = 10$, 2.90%, respectively). None of the articles published in 2002–2009 used box plots, and they were only used in four (2.27%) of the 2010–2017 studies. The number of studies using scatter plots decreased from six in 2002–2009 (3.55%) to four in 2010–2017 (2.27%). Visual presentations categorized as “Other” included histograms, density plots, path models, dendrograms, pie charts, and scatter plot matrices.

Table 3

<table>
<thead>
<tr>
<th></th>
<th>Overall ($K = 345$)</th>
<th>2002–2009 ($k = 169$)</th>
<th>2010–2017 ($k = 176$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k$</td>
<td>$%$</td>
<td>$k$</td>
</tr>
<tr>
<td>Bar plot</td>
<td>93</td>
<td>26.96</td>
<td>34</td>
</tr>
<tr>
<td>Line plot</td>
<td>112</td>
<td>32.46</td>
<td>56</td>
</tr>
<tr>
<td>Boxplot</td>
<td>4</td>
<td>1.16</td>
<td>0</td>
</tr>
<tr>
<td>Scatter plot</td>
<td>10</td>
<td>2.90</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histogram</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density plot</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dendrogram</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pie chart</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scatter plot matrix</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Numbers of items in category “Other” are not recorded in the coding framework.

5. Discussion

5.1 Reported Values

Of the total of 345 articles, more than 90.14% reported means and almost 84.25% reported standard deviations. This implies that the practice of reporting means and standard deviations is common among researchers in the field. However, as reported, 6.38% of the whole sample reported means without reporting standard deviations. Without any information other than means, it is hard to obtain an overall picture of the data presented in the paper, such as its distribution or variability. Papers that do not report standard deviations may also be excluded from research syntheses and meta-analyses, since effect sizes are usually calculated from means and standard deviations, and this would reduce the precision of meta-analyses. Together with descriptive statistics, confidence intervals of the representative values are required to be reported. However, only 8.99% of papers (0.59% in 2002–2009, 17.05% in 2010–2017) reported confidence intervals. Confidence intervals reflect the uncertainty of the point estimate, and this provides the range of values with 95% (or 90%, or 99%) confidence in the study. As mentioned earlier, recent reporting guidelines for L2 studies and other disciplines all include confidence intervals, and the APA Publication Manual (APA, 2009, p. 34) strongly recommends the reporting of confidence intervals. Therefore, we propose that articles which will be published must include confidence intervals of values reported in the studies.
Information about the reliability of instruments or scoring was reported in 40% of the articles in the article pool. In many cases, Cronbach’s alpha coefficients were reported as the measure of internal consistency, and indices such as KR-20/21 and correlation coefficients were used to present the scoring consistency of the marking. The remaining articles (approximately 60%) either did not report on reliability, even though they were using questionnaires or tests, or the nature of the studies did not require reporting on reliability. However, in order to provide validity evidence, information about reliability or validity must be included in study reports.

In the article pool, p values were one of the most common items reported in the papers. Needless to say, p values are the most important index in null hypothesis statistical significance testing (NHST). The APA Publication Manual requires that:

- When reporting p values, report exact p values (e.g., \( p = .031 \)) to two or three decimal places.
- However, report p values less than .001 as \( p < .001 \). The tradition of reporting p values in the form \( p < .10, p < .05, p < .01 \), and so forth, was appropriate in a time when only limited tables of critical values were available. (p. 114)

However, it was observed that the reporting styles of p values of almost half (49.11%) of the papers in 2002–2009 volumes and 23.3% in 2010–2017 volumes did not comply with the reporting standards of the APA Publication Manual. Although a number of papers used both exact p values and the \( p < .10 \) style in a same report, as Norris et al. (2015, p. 475) state, exact p values must be provided “regardless of significance”. In addition, researchers should be aware that p values can be controlled by adjusting sample sizes (known as p-hacking [e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015]), and the reporting of effect size, confidence intervals, and statistical powers are crucial (for more, see Mizumoto & Takeuchi, 2010).

The reporting rate of effect sizes was approximately 11.38% among articles published in 2002–2009. Although the fifth edition of the APA Publication Manual included effect sizes among the items that are required to be reported, the importance of effect sizes was not recognized by the researcher community until around 2010. The reporting rate improved dramatically (11.83% to 55.11%) in the papers published in 2010–2017. This may be because of the release of comprehensive introductory papers on effect sizes (Mizumoto & Takeuchi, 2008, 2010) which described why effect sizes must be reported in papers. Nonetheless, the importance of reporting effect sizes needs to be recognized by more researchers in the field. Although more than 55% of the studies reported effect sizes in 2010–2017 article pool, some of them failed to report effect sizes appropriately. For example, some papers reporting the results of ANOVA and post-hoc tests provided effect sizes for ANOVA only (e.g., eta-squared) but did not report effect sizes for the post-hoc tests (e.g., Cohen’s \( d, r \)). A number of papers still used the evaluation criteria for effect sizes proposed by Cohen (1988), even though Plonsky and Oswald (2014) have proposed evaluation criteria specific to L2 research. Furthermore, Norris et al. (2015) require authors to provide confidence intervals (or the standard error) of effect sizes. In this review, the coding framework does not cover confidence intervals of effect sizes. However, to the best of our knowledge, almost
all papers reported their point estimates of effect sizes. Thus, confidence intervals of effect sizes must also be calculated, together with point estimates (e.g., \( d = .50, 95\% \text{ CI} [.46, .54] \)).

Unlike \( p \) values and effect sizes, there was no improvement in the number of papers reporting statistical powers. In our article pool, two papers reported the statistical power of analyses. Takada (2004) reported powers of two ANOVAs to confirm statistical powers obtained in her study. Tamura and Kusanagi (2015) calculated the sample size from the effect size (\( f = .40 \)), error probability alpha (alpha = .05), and target statistical power (.80), all of which were calculated from the previous studies focusing on the same research topic. This is an effective method for determining the optimal sample size for data collection in advance, and “is advisable when a particular effect size is known or anticipated in advance of the study” (Norris et al., 2015, p. 472). Statistical power can be calculated by recent versions of SPSS, and free computer programs (e.g., G*Power [Faul, Erdfelder, Buchner, & Lang, 2007]) can also calculate statistical powers or sample sizes (when the anticipated statistical power is known). We believe that all researchers should use the readily available software to calculate the statistical power of their analyses.

5.2 Visual Data Presentation

The review revealed that in ARELE 2002–2017, approximately 60% of the papers utilized graphics to present their data. A breakdown of these visual data presentations shows that approximately 30% used bar plots and line plots, 1% used box plots, and 3% used scatter plots. These results were similar to the findings of Hudson (2015) and Larson-Hall (2017), who reported that bar plots and line plots were used more frequently in international journals (e.g., Modern Language Journal, TESOL Quarterly) than box plots and other types of graphics. Table 4 compares the frequencies and percentages of bar plot and line plot usage in the ARELE papers reviewed in the present study and those in the three international journals (Modern Language Journal, Language learning, and Studies in Second Language Learning) reviewed in Larson-Hall (2017).

Table 4

Comparisons Between the Frequencies and Percentages

<table>
<thead>
<tr>
<th>Items</th>
<th>ARELE 2002–2017</th>
<th>MLJ</th>
<th>LL</th>
<th>SSLA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( k )</td>
<td>%</td>
<td>( k )</td>
<td>%</td>
</tr>
<tr>
<td>Bar plot</td>
<td>93</td>
<td>44.71</td>
<td>83</td>
<td>39</td>
</tr>
<tr>
<td>Line plot</td>
<td>112</td>
<td>53.85</td>
<td>79</td>
<td>37</td>
</tr>
</tbody>
</table>


However, as Larson-Hall and Plonsky (2015) argue, many graphics used in our discipline, such as bar charts and line charts, are “data-poor and not very informative” (p. 145). For instance, a
simple bar chart or a line chart can only display limited data points, such as representative values, but cannot display other information essential for statistical analyses, such as the distribution of the data. On this point, Larson-Hall & Herrington (2010) proposed that box plots can be a flexible alternative to bar charts, since box plots display the median, the first and third quartiles (i.e., the 25% and 75% points), and the minimum/maximum values in the same plot area. Violin plots can also be utilized to allow more effective visualization of the data. Figure 1 shows examples of a bar chart, boxplot, and violin plot which visually present the same dataset.

![Figure 1. Examples of bar chart, boxplot, and violin plot with a sample data (Group A; n = 45, M = 4.29, SD = 0.73; Group B; n = 53, M = 3.84, SD = 0.37)](image)

As shown in Figure 1, advanced graphics such as box plots and violin plots can provide much more information about the characteristics of the data than the traditional line and bar plots that have been frequently used in ARELE. These advanced graphics allow readers to interpret the data reported in an article more easily. Thus, we recommend the advanced graphics to be utilized more effectively not only in ARELE submissions, but also other reports of quantitative research.

Authors contributing to ARELE may benefit from using advanced graphics in their papers, since those advanced graphics can provide more information about their data in the same plot area. Currently, ARELE restricts the maximum length of a research article to 16 pages, and this includes all texts, tables, and figures. Consequently, using a figure which presents only a limited amount of information (e.g., the mean only) wastes space that could be used for other information. We strongly recommend that authors should include as much data as possible using advanced graphics, thus allowing more space in the text for narrative discussion of the data.

Combining different types of graphics in a single figure may also be an efficient way to write up quantitative reports in a limited space. For example, Kusanagi et al. (2015) in ARELE used a scatter plot matrix to present correlation coefficients, histograms, and scatter plots with linear
regressions in a single figure. This type of visualization method (i.e., scatter plot matrix or correlation matrix [Figure 2]) can easily be produced using the default function on R (R Core Team, 2017), a free statistical package, or on the web-based platform (langtest.jp) developed by Dr. Atsushi Mizumoto (For more about R and langtest.jp, see Mizumoto & Plonsky, 2015).

Figure 2. An example of a correlation matrix. Data source: Iris dataset (Anderson, 1935).

In summary, although visual data presentations were used in more than 60% of the articles in ARELE published between 2002 and 2017, there still remains the question of whether those figures were appropriately and effectively used. Thus, it is crucial for future contributors to ARELE to adopt advanced, information-rich graphics that can present data more efficiently to improve the reporting practices of quantitative studies, not only in future volumes of ARELE but also in other publications reporting quantitative studies in the area of L2 research.

6. Concluding Remarks

In this study, we comprehensively reviewed the reporting practices of the quantitative research papers in ARELE volumes 13 to 28. This review revealed that the reporting practices of papers for most items in the 2002–2009 volumes improved in the 2010–2017 volumes. However, the reporting rates of statistical powers, confidence intervals, and effect sizes still remain low, and effective visual data representations, namely box plots, were neglected in all volumes.
Based on the outcomes of this study, as well as the reporting guidelines published in *Language Learning* (Larson-Hall & Plonsky, 2015), we conclude this paper by proposing a basic checklist which aims to help researchers and reviewers to report and assess research outcomes appropriately in future ARELE volumes (Table 5). The checklist comprised of seven items that are found to be inconsistent in our review, and all of these items are included both in *APA Publication Manual* (APA, 2009) and Larson-Hall and Plonsky (2015). Due to limited space, we do not discuss the details (e.g., why and how to) of each item; however, for accessible descriptions, please refer to Larson-Hall and Plonsky (2015) and Norris et al (2015).

<table>
<thead>
<tr>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Reliabilities of the instruments are calculated and presented (e.g., Cronbach’s α coefficient, inter-rater reliability).</td>
</tr>
<tr>
<td>2. Descriptive statistics, especially means and standard deviations, are appropriately presented.</td>
</tr>
<tr>
<td>3. Confidence intervals of descriptive statistics are provided.</td>
</tr>
<tr>
<td>4. Exact p values (e.g., p = .012 instead of p &gt; .01) are presented, together with a clear indication of predetermined Type I error rate.</td>
</tr>
<tr>
<td>5. Appropriate effect sizes (e.g., Cohen’s d for t-test) are calculated and presented, together with interpretations. An easily accessible Excel spreadsheet for calculation is made public by Dr. Atsushi Mizumoto³.</td>
</tr>
<tr>
<td>6. Confidence intervals of the effect sizes are calculated and presented.</td>
</tr>
<tr>
<td>7. Visual representations of the data are effectively used where necessary. Advanced visualization methods are particularly recommended for better understandings of data.</td>
</tr>
</tbody>
</table>

Finally, we should note the limitations of our review. The time and resources we could dedicate to researching and writing this paper were necessarily limited; hence, we focused on the more general aspects of the reporting practices of ARELE contributors and their use of visual data displays. Studies focusing on more specific aspects (e.g., the use of effect sizes) should be done in the future, using similar approaches so the outcomes will be comparable with the present study. The checklist we propose does not cover all aspects of the reporting items mentioned in the editorial guidelines of journals such as *Language Learning*. However, we hope that the points discussed in this paper will stimulate discussion among the readership community of ARELE. Also, one of the reviewers pointed out that ARELE publishes practical reports and the checklist we propose may not be applicable to practitioners who submit papers to practical reports section; therefore, careful considerations on the reporting guidelines are required for non-research articles.
Last of all, we would like to conclude by suggesting the editorial team of ARELE that they need to tackle the issues raised in this paper by improving their submission guidelines for future ARELE volumes.

Notes

1. Articles published in the latest Volume 28 (2017) were not available online at the time of data collection. These were therefore scanned from the printed edition and added to the database manually.
2. The maximum length was 10 pages until Volume 21 (2010).
3. A spreadsheet for the calculation of effect sizes is available at Dr. Atsushi Mizumoto’s website (www.mizumot.com/stats/effectsize.xls).

Acknowledgements

We would like to thank the anonymous reviewers for the insightful comments and suggestions that contributed to improving the final version of this paper. We are also grateful to our research assistant, Natsuki Higa, for her efforts in preparing the database for this research.

References


255


Takada, T. (2004). The listening proficiency elementary school (ES) and junior high school (JHS) intend to develop: What teachers should know to bridge the gap between ES and JHS programs. *Annual Review of English Language Education in Japan, 15*, 109–118. https://doi.org/10.20581/arele.15.0_109

