Effect Size Reporting Practices in Applied Linguistics Research: A Study of One Major Journal

Rining Wei¹, Yuhang Hu², and Jianhui Xiong³

Abstract
Many surveys of effect size (ES) reporting practices have been conducted in social science fields such as psychology and education, but few such studies are available in applied linguistics. To bridge this gap and to echo the recent calls for more robust statistics from scholars in applied linguistics and beyond, this study represents the first attempt, in the field of applied linguistics, to focus upon ES reporting practices. With an innovative “two-standards” approach for coding, which overcomes the limitations with similar studies in other social science fields (e.g., communication), this study assesses the ES reporting practices over a span of 6 years in a major journal. Findings include the following: (a) the ES reporting rate is about 50% and (b) some improvement of ES reporting over time is in evidence. Future research directions (e.g., examining whether and how ES is interpreted after being reported) are suggested.

Keywords
effect size, null hypothesis significance testing, quantitative method, p value, effect size reporting, effect size interpretation

Introduction
The importance of effect size vis-à-vis the inherent limitations of Null Hypothesis Significance Testing (NHST; including the significance level, viz. the p value) has been underlined since five decades ago (e.g., Hays, 1963) not only in psychology (American Psychological Association [APA], 1994, 2010; Wilkinson & Task Force on Statistical Inference, APA Board of Scientific Affairs, 1999) and education (American Educational Research Association, 2006) but also, more recently, in applied linguistics (Larson-Hall, 2010; Norris & Ortega, 2007; Oswald & Plonsky, 2010). The limitations of NHST are not delineated here due to space constraints (see Kline, 2004, for an excellent and detailed account, and Henson, 2006; Oswald & Plonsky, 2010; Sun & Fan, 2010, for summaries). One major limitation is that the p value depends upon the sample size; in other words, an increased sample size will eventually yield a small enough p value (viz. p being .05 or smaller; Biskin, 1998; Wasserstein & Lazar, 2016).

On the contrary, effect size, simply put, is “an objective and (usually) standardized measure of the magnitude of observed effect” (Field, 2009, p. 56). Compared with the p value, effect size is much less influenced by sample size (cf., Fan & Konold, 2010). In this sense, effect size is as important as, if not more important than, the significance level. Fan (2001) uses the two-sides-of-one-coin analogy to argue that the p value and effect size complement but do not substitute for each other, suggesting that researchers report both in their quantitative studies. Larson-Hall (2012) goes a step further by stating that “effect size is much more important than a null significance hypothesis test” (the p value included) (p. 472).

The importance of effect size notwithstanding, only a handful of journals in applied linguistics make such reporting practices mandatory in their editorial policies. While Larson-Hall (2010, p. 114) claims that “currently, the only journal in the second language research field which requires effect sizes is Language Learning,” to our knowledge, TESOL Quarterly has required the reporting of effect size measures in quantitative studies since the early 2000s; since about the same time, The Modern Language Journal has released similar editorial policies with respect to effect size reporting (Plonsky & Gass, 2011). To date, another five
In view of the undue neglect of effect size reporting in applied linguistics, this article aims to contribute to our understanding by surveying such practices in System, subtitled An International Journal of Educational Technology and Applied Linguistics. This journal is selected because of two considerations. First, it is a major journal in the field, as reflected in the fact that it is indexed in Social Sciences Citation Index (SSCI) and has been regarded as “major” by previous studies (e.g., Benson, Chik, Gao, Huang, & Wang, 2009; Jung, 2004; Wang & Gao, 2008). Second, it does not mandate effect size reporting in its editorial policy, as is the case with most journals in the field, which means that it may better reflect the general situation of effect size reporting in applied linguistics, than the few above-mentioned journals that mandate effect size reporting.

This exploratory study focuses upon the effect size reporting practices concerning five statistical procedures: t test, analysis of variance (ANOVA),2 correlation, regression, and chi-square (χ²) test. They are selected primarily because they are “the top five” most frequently used methods in four major second language acquisition (SLA) academic journals (Gass, 2009), and thus presumably most frequently used in the wider field of applied linguistics (Larson-Hall, 2012). Furthermore, these five tests are also the focal methods in other fields of social sciences such as communication (Sun & Fan, 2010) and education (Alhija & Levy, 2009); findings from our study may have wider implications beyond the field of applied linguistics per se.

Three research questions are pursued:

Research Question 1: To what extent are measures of effect size reported?

Research Question 2: Do the effect size reporting practices vary across the years?

Research Question 3: For each of the five focal statistical methods, what is the effect size reporting rate and what effect size measures are typically reported?

In the remainder of this article, after providing a more detailed introduction to the definition and use of effect size with an illustrative example in a published study, we review relevant studies from such fields as education and psychology as well as from applied linguistics. We then report upon the data collection and analysis methods of our study. After presenting and discussing major findings, we conclude the article by offering suggestions for effect size reporting practices and for further studies that help contribute to the ongoing methodological reform in applied linguistics.

Effect Size

Definitions of Effect Size

Effect size is defined as an objective and standardized measure of the magnitude of an observed effect, with the wording “(usually)” removed from the above-cited concise definition by Field (2009, p. 56). Although some other definitions (e.g., Meline & Wang, 2004; Sun & Fan, 2010) are so broad that they include nonstandardized forms (e.g., raw mean difference), it is strongly recommended that effect size measures be confined to standardized forms only, so as to maximize the benefits of these forms such as letting “the reader compare effects across groups” and “meta-analysts compare studies even if they use different original measures” (Larson-Hall & Plonsky, 2015, p. 135). Furthermore, the danger of relying on raw mean difference, vis-à-vis the benefits of drawing upon standardized forms of effect size, will be illustrated with one authentic example of effect size reporting below.

Dozens of effect size measures are available, each with relative strengths and weaknesses for particular purposes (Ellis, 2010; Henson, 2006; Kirk, 1996). Two types3 of effect sizes highly relevant to applied linguistic research are the d family and the r family. The d family is based on standardized measures of mean differences (e.g., Cohen’s d), whereas the r family includes standardized measures of strength of relations based on the proportion of variance accounted for (e.g., the r squared in regression) or correlation between two variables. Table 1 provides some frequently used effect size measures for the focal statistical methods, where only Cohen’s d and Hedges’s g belong to the d family and the others to the r family.

Table 1 also lists some benchmarks for interpreting effect sizes recommended by Cohen (1988) and by researchers in applied linguistics. Cohen’s (1988) benchmark system had better be reserved “as a last resort” (Ellis, 2010, p. 42), although it has been used by too many researchers as ironclad criteria without reference to the measurements taken, the study design, or the practical importance of the findings. Whenever possible, researchers should try to interpret effect sizes by grounding them in a meaningful context (e.g., comparisons with previous studies vis-à-vis the measurements and study design) or by assessing their contribution to knowledge (e.g., in terms of practical or clinical value). The two benchmark systems from researchers in applied linguistics

---

1. effect size
2. test
3. types

---
(see Table 1) provide more nuanced guidance in interpreting the effect size in question than Cohen’s system: Plonsky and Oswald’s (2014) benchmarks are highly relevant to experiment-based studies in what they called “L2 research,” and Wei and Hu’s (2018) to survey-based studies examining the effects of sociobiographical variables (e.g., gender and multilingualism) on (socio-)psychological variables (e.g., L2 joy and tolerance of ambiguity).

Consequences of Not Reporting Effect Size

As Zientek, Capraro, and Capraro (2008) point out, “not reporting effect size can be detrimental” (p. 212). Presenting one authentic example helps drive home the consequences of failing to report effect sizes. Table 2 is adapted from Wei and Su’s (2015) analysis of the respondents’ self-reported data concerning their English spoken proficiency and other variables from the largest language survey in China. The major modification made to Wei and Su’s (2015) original table was that we added a column containing Cohen’s d values. We suggest that an effect size from either the r or the d family can be used, and in fact, one can be easily converted into the other (see Larson-Hall, 2010, pp. 117-119, for conversion formulas). Take t tests as an example. As indicated in Table 1, both r and Cohen’s d can be used as effect size measures. Although many textbooks on statistical procedures rigidly recommend Cohen’s d for t tests, Field’s (2009) textbook is one interesting exception, in which he writes that “I’m going to stick with the effect size r because it’s widely understood, frequently used, and yes, I’ll admit it, I actually like it!” (p. 332).

The corresponding research question for Table 2 asks, with regard to English spoken proficiency, was there a significant difference between the national average and the city...
average for each of the seven selected cities? These authors answer the question with results (see Table 2) from a series of one-sample *t* tests.

Two important observations can be made regarding Table 2. First, if one relies on the raw mean differences (viz. the city mean minus the national mean) for Beijing (0.269) and Shenzhen (0.256), one might reach a conclusion that Beijing performed better than Shenzhen with the national average being a baseline. But an entirely opposite conclusion that Shenzhen performed better than Beijing is true because the effect size for the former (0.326) was higher than that for the latter (0.295). In this example, effect size, rather than raw mean difference, is the appropriate measure reflecting the magnitude of the real difference. In other words, failure to use effect size and reliance upon unstandardized measures (e.g., raw mean difference) can lead to a completely opposite conclusion. Second, many researchers with traditional training tend to erroneously believe that “the smaller the *p* value, the larger the effect” (Zhang, 2009, p. 68), and consequently many might conclude that Shanghai, Dalian, and Tianjin performed equally well because of the same level (0.000) of their *p* values. However, the true scenario revealed by the effect sizes is that Shanghai (0.238) scored higher than the national average, Dalian (0.475) was better than Shanghai, and Tianjin (0.572) was even better than Dalian; put differently, the actual mean difference (reflected by effect size rather than raw mean difference) between Tianjin and the nation was largest, whereas that between Shanghai and the nation was smallest. All in all, failure to report effect size along with the *p* value masks a lot of (more) important information in the results of an inferential statistical procedure.

It is noteworthy that Wei and Su (2015) explain why they do not attempt to interpret the *r* values after reporting these effect sizes. Their main justification is that simply drawing upon the frequently used Cohen’s (1988) guidelines is not instructive and there have been no similar studies reporting relevant effect sizes for their analysis to compare and contrast. Considering the random sampling approach and the large sample size used in their study, Wei and Su (2015) suggest future studies of similar topics use their effect size values as a baseline for cross-study comparisons. These practices resonate with the earlier suggestion that referring to effect size values from relevant previous studies is a better practice for interpreting effect size, compared with using Cohen’s (1988) benchmarks.

**Literature Review**

Many surveys of effect size reporting (and to a less extent, interpreting) practices have been conducted in such fields as psychology, education, and communication. For example, in gifted education research, with all the 723 papers from six full volumes of three selected journals, Paul and Plucker (2004, p. 69) report that “28.9% of the quantitative research blocks contained effect size estimates”; the so-called “quantitative research blocks” include three subgroups: descriptives, univariate blocks, and multivariate blocks; the effect size reporting rates for the latter two blocks were 17.9% and 52.2%, respectively. To these authors, there is no need for papers utilizing only descriptive statistics to report effects sizes.

More recently, in the fields of education and psychology, Sun, Pan, and Wang’s (2010) survey of 1,243 articles published in 14 journals from three full volumes (2005-2007) reveals an effect size reporting rate of 49%. In the field of communication, after examining four full volumes (2003-2006) of four influential journals, Sun and Fan (2010) find a relatively high effect size reporting rate (about 75%) in their 224 sampled papers. One major limitation with Sun and Fan’s (2010) study is that their coding method tends to overestimate the effect size reporting rate. If, in one particular article, two or more focal statistical procedures (say, *t* test and ANOVA) are used but only one procedure (say, *t* test) has an effect size reported, Sun and Fan (2010, p. 333) give “benefit of the doubt” to that article by coding this article as one that reports effect size. Another coding standard, which is more stringent than Sun and Fan’s (2010), is that one article using two or more focal statistical procedures has to report effect sizes for all the procedures to qualify as one paper that reports effect size in the coding process. This more stringent standard is likely to yield a lower effect size reporting rate, compared with the situation when Sun and Fan’s (2010) standard is adopted. However, in the field of applied linguistics, no studies focus upon effect size reporting practices. Although Plonsky (2013) finds an effect size reporting rate of 25% by examining 606 articles from *Language Learning and Studies in Second Language Acquisition*, which mandate effect size reporting in their submission guidelines, his study does not focus on effect size reporting, but rather on a wider range of features reflecting the study quality (e.g., designs, statistical analyses, reporting practices, and outcomes). Understandably, no sufficient coding details are provided regarding papers with two or more statistical procedures, although in Plonsky’s (2013, p. 669) sample, up to 60% of the articles use multiple statistical techniques. Another study that sheds light upon effect size reporting is Lindstromberg (2016), who finds 49% of the 96 (quasi) experimental studies reported in 90 articles in 19 volumes (1997-2015) of *Language Teaching Research* report effect sizes. Again, it is unclear how a particular paper/study with two or more statistical procedures is dealt with. Based on the single effect size reporting rate from Plonsky (2013) and Lindstromberg (2016), it seems that only one standard for coding papers with multiple statistical procedures is used, although neither study clarifies whether a relatively loose standard such as Sun and Fan’s (2010) or a more stringent one (e.g., the above-proposed “two-standards” approach) is adopted.
To date, no studies concerning effect size reporting in applied linguistic research endeavor to make explicit the coding standard regarding papers with multiple statistical procedures, let alone adopt two standards to arrive at a more comprehensive picture of the reporting practices. Furthermore, no studies have surveyed papers from journals that do not mandate effect size reporting, as all the journals covered in Plonsky (2013) and Lindstromberg (2016) have such a mandate.

The Study

Sampling

To contribute to the current understanding of effect size reporting in the field, six full volumes (2011-2016) of System were examined. A span of 6 years was decided upon for three reasons. First, publications over a span of 6 years should be sufficient for studies of an exploratory nature such as the present one to show whether any systematic change in the effect size reporting practices took place across the time (see Research Question 2). Second, similar (and often shorter) time spans were adopted in studies of effect size reporting from other social science areas such as communication (e.g., 4 years, see Sun & Fan, 2010), education (e.g., 2 years, see Alhija & Levy, 2009), and psychology (e.g., 2 years, see Dunleavy, Barr, Glenn, & Miller, 2006). Third, the number of articles used for coding (see below) in this study was comparable with similar research in other fields (e.g., Sun & Fan, 2010), which should be manageable in studies of an exploratory nature.

The sampling frame for this study was based on all of the 414 full-length research articles from the six selected volumes of System. Our first two rounds of initial review, as per the first six questions on a checklist developed for the present study (see the appendix), led to the identification of a total of 217 articles that are supposed to report effect sizes, which formed the core dataset of this study. Specifically speaking, our first initial review identified 17 nonempirical research articles, which were irrelevant to the development of the core dataset. Empirical research articles refer to those that are data-based, characterized by systematic collection and analysis of data (cf., Gao, Li, & Lü, 2001). Our second review identified 96 empirical articles utilizing only qualitative data and another 83 that use purely descriptive statistics (e.g., frequencies and percentages; cf., Dunleavy et al., 2006; Meline & Wang, 2004). After these articles were excluded, the remaining articles totaled 218 in which there is one meta-analytic research articles, which were irrelevant to the development of the core dataset. Following Keaton and Bodie’s (2013, p. 117) rationale that “reporting conventions for meta-analytic reviews are remarkably different from those for individual (primary) studies,” we removed this meta-analysis paper in the development of the core dataset.

Finally, the remaining 217 articles formed the core dataset, each of which should have effect size(s) reported. This total number was used as the denominator to generate the overall effect size reporting rates for Research Question 1.

Coding

The unit of analysis was each article. Each article in the core dataset was coded in terms of the research topic, publication year, nature of paper (empirical or not), types of statistical procedures, practices of effect size reporting, types of effect size measures, and authors’ awareness of effect size (see the appendix). Two coding standards are used for the situations where two or more of the focal statistical procedures are used in one single paper, so as to achieve a more comprehensive picture of effect size reporting in applied linguistics and facilitate comparisons with findings from other fields: one is Sun and Fan’s (2010) standard, which tends to give “benefit of the doubt” and hence is relatively loose, and the other is the more stringent one proposed in “Effect Size” section.

The use of these two standards might introduce an element of subjectivity, although most of the coded variables are dichotomous and involve little subjective judgment (e.g., reported vs. not reported). To ensure consistent application of the checklist, the first and second authors independently coded a common set of articles (20.3% of the core dataset, totaling 44). The intercoder agreement rate was 93.1% (with an acceptable agreement rate ranging between 85% and 90%; cf., Miles, Huberman, & Saldana, 2014), and the points of disagreement were resolved through collegial discussion. Once consistency was established, the second author continued to record information pertaining to the other articles.

Data Analysis

After data were coded, both descriptive and inferential statistics were generated with the statistical package SPSS 21.0. For Research Question 1 regarding the extent of effect size reporting practices, only descriptive statistics in the forms of percentage and frequency were generated. To answer Research Question 2 concerning whether the effect size reporting practices change over time, a series of chi-square tests were performed, using Cramer’s $V$ as an effect size. For Research Question 3 concerning the types of effect size typically reported, frequencies and percentages were used, supplemented with qualitatively enumerated examples.

Findings and Discussion

Research Question 1: To What Extent Are Measures of Effect Size Reported?

As Table 3 shows, overall, 73.27% of the sampled papers that should have effect size reported do report effect size(s), when Sun and Fan’s (2010) standard is adopted for situations involving papers with two or more statistical procedures.
This effect size reporting rate diminishes to 52.07% when a standard more stringent than Sun and Fan’s (2010) is adopted. It is unfortunate that effect sizes, the importance of which is no less than that of the \( p \) value, get reported in only half of the papers that are supposed to report effect sizes. These findings are indicative of an alarming tip of the iceberg, namely, the troubling situation of a lack of effect size reporting in applied linguistics research.

These remarks may seem overly critical toward the field of applied linguistics. To be fair, we need to situate the discussion in a broader context by reiterating that the underreporting of effect sizes has also been observed in other fields. One comparable study is a survey of 256 papers from the Journal of Counseling & Development over 11 years, where Bangert and Baumberger (2005) find an effect size reporting rate of less than 50% among the papers that conduct statistical significance tests and, hence, need to report effect sizes.

Research Question 2: Do the Effect Size Reporting Practices Vary Across the Years?

A chi-square test, \( \chi^2(5) = 3.533, p = .618 \), based on the effect size reporting frequencies according to Sun and Fan’s (2010) standard, revealed a small-to-medium level of association (Cramer’s \( V = 0.128 \)) between the reporting rates and the publication year. Another chi-square test, \( \chi^2(5) = 10.730, p = .057 \), based on the effect size reporting frequencies according to the more stringent standard, also revealed a small-to-medium level of association (Cramer’s \( V = 0.222 \)) between these two variables. Although the corresponding \( p \) values (e.g., .057) were higher than the conventional statistical significance level (\( p = .05 \)), this does not diminish the importance of the results reflected by effect size. In the words of authorities on statistics, “surely, God loves the 0.06 nearly as much as the 0.05” (Rosnow & Rosenthal, 1989, as cited in Ellis, 2010, p. 49). The \( p \) value will become small enough in a future replication study based on a large enough sample.

The effect size values reported above are higher than those reported in previous studies. The counterpart Cramer’s \( V \) in Sun et al. (2010) was only 0.07, although their corresponding \( p \) value was small, \( p = .06, \chi^2(2) = 5.66 \), probably because of their large enough sample size. This suggests that the variable of interest (effect size reported or not) was associated with the publication year at a negligible-to-small level according to Cohen’s benchmarks, depending on the discipline.

All in all, the answer to Research Question 2 is that effect size reporting practices do vary across time. The strength of association lies between Cohen’s (1988) small and medium benchmarks.

Research Question 3: For Each of the Five Focal Statistical Methods, What Is the Effect Size Reporting Rate and What Effect Size Measures Are Typically Reported?

For papers that used correlation analysis, 94.29% of them (see Table 4) reported an effect size measure. This extremely high reporting rate can be attributed to the fact that the test statistic (i.e., correlation coefficient) in itself is effect size (Sun et al., 2010). Similarly high effect size reporting rates for correlation analysis can be found in other fields. For instance, in the field of communication, Sun and Fan (2010, p. 334) note that “nearly 100% of studies” that used Pearson correlation reported effect size measures, whereas the corresponding rate in Alhija and Levy (2009) reached 100% in the field of education. In this study, the effect size measures typically used were correlation coefficients such as Pearson’s \( r \) and Spearman’s rho.

For the papers that used regression analysis, about 84.00% (refer to Table 4) reported effect sizes. High effect size reporting rates for regression analysis can be found in other fields such as communication (nearly 100%, see Sun & Fan, 2010) and education (100%, see Alhija & Levy, 2009). The effect size measure mostly used was adjusted \( R^2 \), which is consistent with observations from other fields (e.g., Alhija & Levy, 2009; Sun & Fan, 2010). The underlying reason might be that, as noted by Kirk (1996), adjusted \( R^2 \) is readily available in the statistical output for regression analyses generated by popular statistics packages such as SPSS.

Table 3. No. of Articles Reporting Effect Size in Selected Years.

| Year | Total no. of articles that need to report effect size | No. of articles reporting effect size (%) |
|------|-----------------------------------------------|-----------------------------------------------|
|      |                                               | By Sun and Fan’s (2010) standard | By a more stringent standard than Sun and Fan’s (2010) |
| 2011 | 19                                            | 13 (68.42)                           | 11 (57.89)                           |
| 2012 | 27                                            | 21 (77.78)                           | 16 (59.26)                           |
| 2013 | 42                                            | 28 (66.67)                           | 17 (40.48)                           |
| 2014 | 45                                            | 32 (71.11)                           | 19 (42.22)                           |
| 2015 | 41                                            | 34 (82.92)                           | 29 (70.73)                           |
| 2016 | 43                                            | 31 (72.09)                           | 21 (48.83)                           |
| Total| 217                                           | 159 (73.27)                          | 113 (52.07)                          |
More than 60% (64.10%, see Table 4) of the papers using ANOVA reported effect size measures. This was highly similar to its counterparts, namely, 56.5% and 57%, respectively, from Sun and Fan (2010) and Alhija and Levy (2009). The reporting rate for ANOVA was lower than that for regression, partly because effect sizes for ANOVA are not as readily available as those for regression in statistics packages. Take SPSS as an example. In SPSS, ANOVA can be realized through three ways. The most common way is to initiate the test by clicking “Compare Means → One-way ANOVA,” but an effect size measure for ANOVA, eta-squared, cannot be generated in the output this way, misleading many researchers into believing that SPSS does not provide eta-squared for ANOVA (Zhang, 2009). However, this effect size can be generated in the two less commonly used ways in SPSS (cf., Plonsky & Oswald, 2017). Therefore, when effect sizes were reported, (partial) eta-squared was unsurprisingly most reported, which is consistent with the earlier findings (e.g., Sun & Fan, 2010). In light of the observation that “researchers arbitrarily selected one of these two” (i.e., eta-squared and partial eta-squared) from the field of communication (Sun & Fan, 2010, p. 338) and a most recent discussion of the misuses of (partial) eta-squared in the field of L2 research (Norouzian & Plonsky, 2018), future research needs to investigate whether these effect sizes have been correctly used when being reported.

Twenty-three (34.32%) of a total of 67 articles that used $t$ tests reported effect sizes. Similarly, moderate reporting rates are found in other fields. In Alhija and Levy’s (2009) sampled papers from five educational journals that do not mandate effect size reporting, the corresponding rate was 31%. In Sun and Fan’s (2010) sampled papers from two communication journals that do not effect size reporting requirements, none of the five papers that used chi-square tests reported effect size; to account for this, the authors speculate that “it is likely that neither Cramer’s $V$ nor $\varphi$ is well known to communication researchers” (Sun & Fan, 2010, p. 338). In our sample, most papers correctly used Cramer’s $V$, with only two using odds ratios.

### Table 4. No. of Articles Using the Focal Statistical Procedures and Reporting Effect Sizes.

| Procedure            | No. of articles in which the procedure is used | No. of articles in which effect size is reported |
|----------------------|-----------------------------------------------|-----------------------------------------------|
| $t$ test             | 67                                            | 23 (34.32%)                                    |
| Analysis of variance | 78                                            | 50 (64.10%)                                    |
| Chi-square           | 23                                            | 7 (30.43%)                                     |
| Correlation          | 70                                            | 66 (94.29%)                                    |
| Regression           | 25                                            | 21 (84.00%)                                    |

Conclusion

This study has examined the effect size reporting practices in one major applied linguistics journal. The effect size reporting practices seem to have improved in the past few years, while the identified reporting rate of about 50% is inadequate. Although such improvement is encouraging, evidence from other disciplines suggests that such advances of effect size reporting can be lost without continued vigilance (Loewen et al., 2014). Therefore, journal editors, researchers, and researcher trainers need to (continue to) encourage and/or implement good reporting practices (e.g., reporting effect sizes along with the exact $p$ values).

Although this exploratory study is innovative in terms of the “two-standards” approach for coding and the target journal selection, it has three major limitations. First, it would have benefited from a larger sample size. The above findings and conclusions are tentative, which require verification and/or falsification in future research. In terms of generalizability, the results may not be representative of the use of effect sizes in applied linguistics in general, as this study has only focused on one journal in the field. Second, the findings here provide limited information about effect size reporting practices for statistical procedures (such as factor analysis and
structural equation modeling) other than the five focal ones. Third, the present study has provided evidence of frequency of application of effect sizes in the focal journal, but does not indicate whether these effect sizes have been correctly applied or not (see Norouzian & Plonsky, 2018, for a review of the misuses of [partial] eta-squared in L2 research).

To contribute to the on-going methodological reform in applied linguistics (Larson-Hall & Plonsky, 2015), more studies on effect size reporting are needed. Future studies will stand to gain by expanding the sample size and/or comparing the reporting practices across journal types (journals with vs. without a requirement for effect size reporting). It is also useful to examine whether effect sizes are reported more frequently for statistically significant results than their non-significant counterparts, as Plonsky (2013) notices that some authors tend to report effect sizes solely for statistically significant results, although such information was “not coded for throughout the entire sample” in his study. Furthermore, future studies of effect size reporting need to incorporate an element of effect size interpretation in a more systematic way, as the reporting of effect sizes should not be treated “as an end in itself” (Larson-Hall & Plonsky, 2015, p. 135). It is useful to know how effect sizes are interpreted after being reported.

Ellis (2010) predicts that “If history is anything to go by, statistical reforms adopted in psychology will eventually spread to other social science disciplines” (p. xiv). Recently, the editors of Basic and Applied Social Psychology (Trafimow & Marks, 2015) have issued a journal-wide ban on NHST. This ban represents a natural progression of the long-standing critiques of NHST and a strong call for the use of more robust statistics (e.g., effect size) in our reporting practices. It is our firm belief that applied linguistics will soon be one of these disciplines in Ellis’s prediction.

Appendix

A Checklist for Analyzing the Sampled Articles.

| No. | Item | Note/check |
|-----|------|------------|
| 1   | Title |            |
| 2   | Year |            |
| 3   | Issue number |            |
| 4   | Is this paper empirical? |            |
| 5   | Which type of empirical research was adopted (quantitative, qualitative, or mixed methods)? |            |
| 6   | Are the statistics purely descriptive? |            |
| 7.1 | t test | □Yes□No |
| 7.2 | Analysis of variance | □Yes□No |
| 7.3 | Chi-square | □Yes□No |
| 7.4 | Correlation | □Yes□No |
| 7.5 | Regression | □Yes□No |
| 7.6 | The other procedures | □Yes□No |
| 8.1 | Is effect size reported? | □Yes□No |
| 8.2 | Is effect size reported with clear awareness on the part of the author(s)? | □Yes□No |
| 8.3 | What effect size measure(s) is(are) used? | □Yes□No |
| 8.4 | Is effect size interpreted? | □Yes□No |
| 9   | Overall, does this paper report effect sizes according to Sun and Fan’s (2010) standard? | □Yes□No |
| 10  | Overall, does this paper report effect sizes according to a more stringent standard? | □Yes□No |

Note. This checklist is adapted from Alhija and Levy (2009) and Sun, Pan, and Wang (2010). A major statistical procedure is defined as a method that is used to directly address at least one research question in the article concerned.
Acknowledgments
The authors would like to extend their sincere thanks to the anonymous reviewers and the article editor for their constructive comments on an earlier version of this article. All the remaining inadequacies are the authors’ responsibility.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The writing of this article was supported by the Educational Science Research Development Fund of Xi’an Jiaotong-Liverpool University (RDF-16-01-61). The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The writing of this article was supported by the Educational Science Research Development Fund of Xi’an Jiaotong-Liverpool University (RDF-16-01-61).

Notes
1. This general statement is about both sample effect size and population effect size. While some authors (e.g., Melone & Wang, 2004, p. 204) state that effect size is “unaffected by sample size,” it is instructive to point out that such a statement is applicable to population effect size only. As “sample effect size itself is a random variable” (Fan & Konold, 2010, p. 448), sample effect size is therefore affected by sampling variability, which is inversely related to sample size. Put differently, when we wish to estimate effect size as a population parameter based on different values of sample effect size, sample size matters.

2. In this exploratory study, analysis of variance (ANOVA) was confined to one-way ANOVA; the other variants (e.g., analysis of covariance [ANCOVA] and multivariate analysis of variance [MANOVA]) were not covered.

3. Although, for the most part, the dozens of available effect size measures can be categorized into these two types, some indices such as odds ratio and the $I^2$ index for hit rate (cf., Huberty & Lowman, 2000) do not fall neatly into these two categories.

4. In SPSS, the two ways to generate an effect size for ANOVA are (a) Analyze→General Linear Model→Univariate→Choose “Dependent Variable” and “Fix Factor(s)”→Click “Option”→Check “Estimates of effect size,” and (b) Analyze→Compare Means→Means→Choose variables into the “Dependent List” and “Independent List”→Click “Option”→Check “ANOVA table and etc.” Please note that SPSS mislabeled partial eta-squared as eta-squared until version 11.0 (Larson-Hall, 2016).

5. We tend not to distinguish between these two indices here as some papers may have misreported one for another, partly due to the mislabeling problem in SPSS (cf., Levine & Hullett, 2002; Norouzian & Plonsky, 2018).

6. As a result of the long-standing critiques of Null Hypothesis Significance Testing (NHST), most recently in the field of applied linguistics, the Bayesian hypothesis testing approach has been recommended as an alternative to NHST. See Norouzian, de Miranda, and Plonsky (2018) for the Bayesian estimation of effect sizes, and Norouzian, de Miranda, and Plonsky (2019) for Bayesian equivalents of $p$ values.

ORCID iD
Yuhang Hu https://orcid.org/0000-0003-3867-1179

References
Alhija, F., & Levy, A. (2009). Effect size reporting practices in published articles. Educational and Psychological Measurement, 69, 245-265.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. Educational Researcher, 35(6), 33-40.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the Journal of Counseling & Development: 1990-2001. Journal of Counseling & Development, 83, 480-487.

Benson, P., Chik, A., Gao, X., Huang, J., & Wang, W. (2009). Qualitative research in language teaching and learning journals, 1997–2006. The Modern Language Journal, 93, 79-90.

Biskin, B. H. (1998). Comment on significance testing. Measurement and Evaluation in Counseling and Development, 31, 58-62.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. M. (2006). Effect size reporting in applied psychology: How are we doing. The Industrial-Organizational Psychologist, 43(4), 29-37.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge, UK: Cambridge University Press.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. The Journal of Educational Research, 94, 275-282.

Fan, X., & Konold, T. R. (2010). Statistical significance versus effect size. In P. Peterson, E. Baker, & B. McGaw (Eds.), International encyclopedia of education (Vol. 7, pp. 444-450). Oxford: Elsevier.

Field, A. P. (2009). Discovering statistics using SPSS: (and sex and drugs and rock ‘n’ roll) (3rd ed.). Los Angeles, CA: Sage.

Gao, Y., Li, L., & Lü, J. (2001). Trends in research methods in applied linguistics: China and the west. English for Specific Purposes, 20, 1-14.

Gass, S. (2009). A historical survey of SLA research. In C. R. William & K. B. Tej (Eds.), The new handbook of second language acquisition (pp. 3-28). Bingley, UK: Emerald.

Hays, W. L. (1963). Statistics for psychologists. New York, NY: Holt, Rinehart and Winston.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. The Counseling Psychologist, 34, 601-629.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. Educational and Psychological Measurement, 60, 543-563.
Jung, U. O. (2004). Paris in London revisited or the foreign language teacher’s top-most journals. *System*, 32, 357-361.

Keaton, S. A., & Bodie, G. D. (2013). The statistical and methodological acuity of scholarship appearing in the “international journal of listening” (1987-2011). *International Journal of Listening*, 27, 115-135.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavior research* (1st ed.). Washington, DC: American Psychological Association.

Koga, T. (2010). Dynamicity of motivation, anxiety and cooperativeness in a semester course. *System*, 38, 172-184.

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

Larson-Hall, J. (2012). How to run statistical analyses. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 245-274). Chichester, UK: Wiley-Blackwell.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. New York, NY: Routledge.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1), 127-159.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, 28, 612-625.

Lindstromberg, S. (2016). Inferential statistics in language teaching research: A review and ways forward. *Language Teaching Research*, 20, 741-768.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtko, J., Sterling, S., & Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, 48, 360-388.

Meline, T., & Wang, B. (2004). Effect-size reporting practices in ASJSLP and other ASHA journals, 1999–2003. *American Journal of Speech-Language Pathology*, 13, 202-207.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Los Angeles, CA: Sage.

Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, 68, 1032-1075.

Norouzian, R., de Miranda, M. A., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *The Modern Language Journal*, 103, 248-261.

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

Norris, J. M., & Ortega, L. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly*, 41, 805-815.

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, 30, 85-110.

Paul, K. M., & Plucker, J. A. (2004). Two steps forward, one step back: Effect size reporting in gifted education research from 1995–2000. *Roeper Review*, 26, 68-72.

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, 35, 655-687.

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

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878-912.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 39, 579-592.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.

Sun, S., & Fan, X. (2010). Effect size reporting practices in communication research. *Communication Methods and Measures*, 4, 331-340.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102, 989-1004.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1-2.

Wang, W., & Gao, X. (2008). English language education in China: A review of selected research. *Journal of Multilingual and Multicultural Development*, 29, 280-299.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA’s statement on p-values: Context, process, and purpose. *The American Statistician*, 70, 129-133.

Wei, R., & Hu, Y. (2018). Exploring the relationship between multilingualism and tolerance of ambiguity: A survey study from an EFL context. *Bilingualism: Language and Cognition*. Advance online publication. doi:10.1017/ S1366728918000998

Wei, R., & Su, J. (2015). Surveying the English language across China. *World Englishes*, 34, 175-189.

Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.

Zhang, S. L. (2009). Xiaoying fudu: Waiyu dingshang yanjiu buneng hushi de cedu zhi [Effect size: Measures that cannot be ignored in L2 quantitative research]. *Waiyu Jiaoxue Lilun Yu Shijiandun / Foreign Language Learning: Theory and Practice*, 3, 67-70, 96.

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at the evidence cited in the AERA panel. *Educational Researcher*, 37, 208-216.

**Author Biographies**

**Rining Wei** (Tony), PhD, teaches courses related to bilingualism and research methods at undergraduate and postgraduate levels, at the Department of English, Xi’an Jiaotong-Liverpool University. He has supervised master’s and doctoral dissertation projects concerning bilingual education, TESOL, and language policy. He has published in journals including English Today, and World Englishes. He serves on the editorial board of the TESOL International Journal.
Yuhang Hu (Sophie) is a master student at the Department of Linguistics with a concentration in Applied Linguistics, Georgetown University. Her areas of research include (socio-)psychological variables in bilingualism and quantitative methodology. She will commence her PhD study in Applied Linguistics at Northern Arizona University this Fall.

Jianhui Xiong, PhD, conducts research concerning educational policy and comparative education at the National Center for Education Development Research, Ministry of Education of the People’s Republic of China. His recent research interests include internationalization of education and the use of big data in education.