STATISTICAL REPORT REFORM IN SECOND LANGUAGE RESEARCH: A CASE OF EXPERIMENTAL DESIGNS

Eka Fadilah
Universitas Widya Kartika Surabaya, Indonesia
ekafadilah@widyakartika.ac.id

Abstract: This survey aims to review statistical report procedures in the experimental studies appearing in ten SLA and Applied Linguistic journals from 2011 to 2017. We specify our study on how the authors report and interpret their power analyses, effect sizes, and confidence intervals. Results reveal that of 217 articles, the authors reported effect sizes (70%), Apriori power and Post Hoc power consecutively (1.8% and 6.9%), and confidence intervals (18.4%). Additionally, it reveals that the authors interpret those statistical terms counted 5.5%, 27.2%, and 6%, respectively. The call for statistical report reform recommended and endorsed by scholars, researchers, and editors is inevitably echoed to shed more light on the trustworthiness and practicality of the data presented.

Keywords: statistical reforms, power analyses, effect size, confidence interval, second language research

INTRODUCTION

Studies in quantitative second language acquisition (SLA) research have provided ample evidences to the use of statistical procedures employed such as *t* test, ANOVAs, correlations, Structural Equation Modelling (SEM), etc. Gass (2009) claimed that there were approximately 86% of empirical research articles leading second language acquisition (SLA) in quantitative approach which embraced some utmost importance of best statistical analysis. In many cases, however, some SLA researchers have relied on a very narrow range of such statistical procedures (Gass, 2009; Plonsky, 2013, 2015). Those researchers are often viable to attest their research questions with the heavily reliance on Null hypothesis and *p-value* which are set up prior to the analysis on the one hand, however, they encounter the problems associated with some technical terms regarding statistical assumption on the other hand i.e., failing to meet statistical assumption or their statistical value results in accepting or rejecting Null hypothesis Significance Testing (NHST) with a certain *p*-value (Plonsky, 2015).

Additionally, it is often exacerbated by the insufficient knowledge of statistical tools to report and interpret the vetted data they have already gauged with such tools to provide comprehensive data to the readers which, eventually, becomes a further problem to impede the quality of statistical report in providing the accurate and precise data.

Some scholars have elucidated the limitation of NHST and *p-value* and recommended a better technique than NHST. Notably, reporting and interpreting meaningful statistic procedures such as power analyses, effect sizes, and confidence intervals instead of "Null ritual" as advocated in NHST (Gigerenzer, 2004). Our main aim is to survey how those statistic procedures are reported and interpreted used in leading SLA journals published in the interval Years 2011-2017. Cook (1999, p.267) underscores that the proper use of SLA research should follow some requirements. The two of them are "validity of the research, ethics in obtaining results" Inevitably, the shift heavily reliance from the serious flaws of NHST to the better statistical procedures should be continuously echoed and endorsed to shed more light to the quality of research method in our field.

Previous survey of statistical reports in SLA

The increase of the statistical package use in SLA has been increasing in SLA literatures even though "SLA is not an innovator but an increasingly knowledgeable borrower and adapter of statistical procedures" (Loewen & Gass, 2009, p.181). Based on 1.411 SLA articles in three SLA journals surveyed, Loewen and Gass reported that the use of inferential statistics i.e., both single and multiple. Based on the research timeline reported, the use of single and multiple inferential statistics had a steady increase between 1970 and 1984 but a significant increase occurred between 1984 and 2006. However, the lack of adequate and knowledgeable statistical concepts and procedures arise. Lazaraton, Riggenbach, and Ediger (1987) and recently reported by Loewen, Lavolette, Spino, Papi, Schmidtke, Sterling, and Wolff (2013) provided such an evidence. For instance, of 121 respondents in (a university professor, Ph.D students, and MA students), in Lazaraton et al.'s survey, 26% respondents felt that they have adequate and knowledgably concepts and procedures of statistics.

Likewise, in Loewen et al.'s survey on 163 PhD students and 162 professors reveals that only 13% of PhD students and 29% of professors felt that their statistical knowledge is adequate. No wonder that such a lack knowledge influence the way of researchers in reporting and interpreting their statistical reports such as effect size, confidence interval, and power analysis.

Lindstromberg (2016) review all (quasi) experimental studies published in *Language Teaching research (LTR)* between 1997 and 2015. The finding reveals that the authors reported confidence intervals only counted 2 articles in result sections for the population effect sizes estimation. While, effect sizes report and interpretation counted 49% and 55%, respectively. Interestingly, it was reported that no clear evidence was found in regard to conduct power analyses across the articles to predict the sample sizes. In a similar vein, Plonksy and Gass (2011) surveyed 174 interactionist research articles across 15 journals published from 1980 to 2009. The finding reveals that Mean and standard deviation are reported 64% and 52%, respectively. While effect size and confidence interval were consecutively reported 41% and 3%. A small report percentage is found in power analysis accounted only 2% across articles in the journals.

In his another study, Plonsky (2013) reported that among 606 quantitative studies in SLA, one or more means were reported without standard deviation. Some papers reported effect sizes when *p<.05* but eliminated when *p>.05*. CI and power analysis are reported counted 5% and 1%, respectively. Furthermore, Plonsky's (2014) survey on journals between 1990s and 2000s reveals that the authors report consecutively reported mean 73% and 79%, standard deviation 48% and 69%, mean without SD 41% and 24%, CI 0.4% and 7%, effect sizes 3% and 42%, and power analysis 0% and 2%.

In light of the above survey reports, the present study aims at investigating the papers (experimental designs) published in some journals appearing from 2011 to 2017. The election of such interval years is based on the assumption how the authors adhere the recommendation endorsed by APA manual guide (2010) as well as suggestions echoed by foregoing SLA researchers and editors.

Likewise, the ways of the authors interpret their statistical reports i.e., effect sizes are investigated which are not provided by the foregoing survey reports. Accordingly, I formulate two research questions:

RQ 1 : to what extent do the authors in the experimental studies report their statistical report procedure?

RQ 2 : How do the authors in the experimental studies interpret their statistical report procedure?

LITERATURE REVIEW

Statistical report reform

Prior to the release of the fifth edition of the Publication Manual of the American Psychological Association (APA), a Task Force of Statistical Inference (TFSI) was formed as a response to the controversy of NHST. Some recommendations were made to the fifth edition of APA by reporting statistical reports such as power analysis, effect size, and confidence interval along with descriptive statistics. Fidler (2002) reported that the fifth edition of APA manual covered as recommended by the TFSI teams i.e., power analysis, CI and effect size, however, it did not elucidate how to present and interpret them. Mostly, the fifth edition of APA publication manual still made emphasis on hypothesis testing.

After getting a massive criticism from the advocates of statistical reform, fortunately, the sixth edition of APA publication manual make some radical changes in reporting and presenting statistical reports. Even though, the concept of hypothesis testing is still attached, some recommendation regarding the statistical reform reports have been made such as reporting statistic parameters, power analysis, effect size and confidence interval. It is followed by the examples of how to "present" them in a table by adding up column cells for CI and effect size (see the 6thpublication manual edition of APA, 2010, p. 143).

With the regard to NHST relied on the acrimonious debates among psychology and other disciplines scholars as in *What if there were no significant tests?* (See i.e., Harlow, Mulaik, & Steiger), it is cited that the degree of the decision of individual journal editors to emphasize or de-emphasize NHST is based on the journal policies themselves. However, the sixth APA manual guide underscores that "*complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectations for all APA journals*" (p. 33, italic added). Additionally, regarding power analysis, it is cited "State how this intended sample size was determined (e.g., *analysis of power or precision)*" (p. 30, italic added).

What should begin from here

Following the recommendation of some SLA researchers, The Publication Manual of the American Psychological Association (6th ed.), and editorial policy of some journals such as *Language Learning, Language Learning & Studies in Second Language Acquisition,* and *TESOL Quarterly* (Byrnes, 2013; Chapelle & Duff, 2003; DeKeyser & Schoonen, 2007; Ellis, 2000; Norris, Plonsky, Ross, Schoonen, 2015), it is strongly required to encompass meaningful statistical reports in addition to Null Hypothesis

Significance Testing (NHST) and *p-Value* such as Power analysis, effect size (i.e., Cohen's *d*) and Confidence Interval (CI).

A synthesis study conducted by Plonsky (2014) on 606 primary reports of quantitative SLA research from two journals – *Language Learning and Studies in Second Language Acquisition* – reveals that the problems found such as means based analysis, missing data, null hypothesis significance testing, the power problem, and design preference. A call for reform is then addressed to SLA researchers, journal editors, and SLA/EFL teachers/students to include effect size and confidence interval along with NHST and *p-value* (Cumming, 2014; Larson-Hall, 2010).

Plonsky (2011) reveals that one of the important things which is missed by SLA researchers when conducting quantitative research is missing data including basic descriptive statistics i.e., Mean and Standard Deviation, which are needed to calculate an effect size. Additionally, Plonksy (2013) underscored that missing descriptive statistics potentially weaken the progress in the SLA field in two ways such as unreported data restricts our ability to interpret the findings of primary studies and it impedes the calculation certain effect sizes i.e., Cohen's d. The consequences of such an insufficient statistical report results in the exclusion of some data in meta-analysis study (Norris & Ortega, 2006; Plonsky, 2011; Russel & Spada, 2006). In Plonsky's (2011) meta-analysis study, 157 of 218 data of L2 strategy instructions was excluded due to some missing data reported by the previous research findings.

Many SLA researchers, however, seem to resist such an endorsement by reliance heavily only on *p-value* and NHST. In fact, *p-value* is heavily relied on sample size which is prone to type I error – rejecting Null Hypothesis, while the fact is true. While, if researchers ignore the power and effect size, it would lead to type II error – accepting Null hypothesis, while the fact it is false.

Power Analysis

Power – the probability of detecting a statistical result when there are in fact differences between groups or relationship between variables (Larson-Hall, 2010, p. 104) – will ensure that the real differences are found and lead to a correct conclusion about the null hypothesis. Murphy and Myors (2004) opine that a consensus of the power level should be above .50 (50%) and be adequate at .80 (80%). It means that with a power level of 80% signifies four out of five times a real effect in the population will be found with the rest of 20% will not be found.

Power analysis prior to study, for instance, can be carried out to give a description about the number of samples each groups required. TFSI suggests "Because power computations are most meaningful when done before data are collected ..., it is important to show how effect-size estimates [to be used in power calculations] have been derived from previous research and theory" (Wilkinson et al., 1999, p. 596). Larson-Hall (2012) provides an alternative in addition of increasing the size of samples that is raising the accepted alpha level to 0.10 instead of 0.05 to increase the statistical power. Button et al. (2013) asserts that the adherence to an alpha level 0.05 (Type I error) results in partly the lack consideration of statistical power (Type II error).

Methodologically speaking, the choice of more stringent alpha level at 0.05 rather than 0.1 leads to reductions in power. Cascio and Zedeck (cited in Murphy, 2010) proposed equations the trade-off of Type I error and Type II error called as apparent relative seriousness (ARS) as $(p\ (H_1)\ (1\text{-power}))\ /\ (1\text{-}p\ (H_1))\ \alpha)$ in which $p\ (H_1)$ stands for probability that H_0 is false. So, in an experimental study, for instance, if a researcher believes that the treatments have some effect is 0.7, $\alpha = 0.05$, and power = 0.80, it can be $[(0.7 \times 0.2)\ /\ (0.3 \times 0.5) = 9.33]$ meaning that false rejection of the null hypothesis i.e., Type I error is 9.33 times as serious as false rejection of the null hypothesis when it is wrong i.e., Type II error. By contrast, if using alpha level at 0.1 with the same conditions with aforementioned effect and power, the value is 4.66 indicating that Type I errors are treated as if they are 4.66 times as serious as a Type II error.

From the calculation above, it implies that if a researcher believes (derived from the previous studies) that Type I errors are nine times as serious as type II errors, then alpha level=0.05 is a best choice, by contrast, If s/he believes that Type I errors are only four or five times as serious as Type II errors, the alpha level at 0.1 is preferable.

How to calculate the power including effect size if the study has not been carried out yet. Cumming and Fidler (2010) provide their suggestions by relying on the previous relevant research and likely from the pilot the study. Let's assume that the previous research conducted by Fadilah (2018) in investigating the effect of oral corrective feedback focused prompt group and unfocused prompts group on the acquisition of grammatical accuracy provided that the sample sizes are 20 participants in each group. The post-test results in effect size *d*=0.8 (large effect size), alpha=0.05, and power value=0.6 (medium power). Suppose further studies want to increase the power value, lets' say 0.8 (ideal power value), it requires to increase the sample sizes into 25 for each group (by assuming that the conditions, statistical assumptions are identical).

Effect Size

A massive criticism was addressed to the routine and narrow adherence to traditional statistical reports i.e., NHST and *p*-value, within SLA scholars and researchers (Larson-Hall, 2010; Norris, et al., 2015; Plonsky, 2013; Plonsky & Gass, 2013; Plonsky & Oswald, 2014). If the difference really exists, *how much*? It is what effect size covers with. It seems to make a sense because *NHST and p-value* are often associated with sample size in which any mean difference between groups leads to statistical significance due to the large enough sample. However, when the magnitude of the relationship is not reported, it does not reliably reflect the size of its associated effect. In contrast, Effect size – "the magnitude of the impact of the independent variable on the dependent variable" (Kline, 2004, p. 97) will not be shaken by those traditional statistical reports although one cannot reach statistical significance conclusion of empirical data he/she has (see illustrations above). Effect size magnitudes i.e., *d, f, eta-squared* or *r* are not "swayed toward statistical significance by a particular large sample, nor are they deflated by a small one" (Oswald & Plonsky, 2014, p. 879).

Confidence Interval

Next recommendation on statistical report in SLA research is reporting confidence interval (CI), along with effect size, as a vital consideration (Cumming, 2014; Kline, 2004; Larson-Hall, 2010). CI represents "a range of plausible values for the corresponding parameter" (Kline, 2004, p. 27). It indicates how far from zero the difference lies on which the width of CI indicates the precision with which the difference can be calculated. It also signifies that the wider CI resulted from a lot of sampling error in a study, the worse statistical estimates.

CI denotes "a set of two numbers that represent a range of values where, with 95% confidence, we would expect a difference in mean scores between the groups to appear if we repeated the same study again and again" (Larson-Hall & Plonsky, 2015, p. 136). Cumming (2012) suggests that if the range of values goes through zero, it means that no statistically significant difference, however, if the range of values doesn't go through zero, it yields statistically significant difference. Indeed, such a value is the same as *p-value* of *t or F* test, but CI provides much more information i.e., the location and width.

Larson-Hall and Plonsky (2015) provide some illustrations of how to interpret CI values. For instance, 95% CI [4 , 56.3] and [4.9 , 10.6]. Both CI range values indicate statistical difference. However, the former indicates a poor precision of estimate in which there might be as little as 4 points differences between the groups in the population, or as much as 56.3 point of differences. Such a large width of CI entails further research to, for instance, use larger sample size to decrease the width of CI. While the latter elicits a narrow CI which signifies statistical significant differences between groups of the population with the difference for the population means can be as little as 4.9 points, or as much as 10.6 points with 95% confidence. Such a narrow interval postulates that the difference between groups is real. By contrast, 95% CI [-5.1 , 8.3] entails that there is no statistical difference between groups because the CI ranges values go through zero. However, the width of CI range scores is narrow and precise (no matter that the negative values).

In a nutshell, CI provides more explanation than *p-value* which entail *effect/No effect* or *difference/no difference* options which covers the information provided by *p-value* with or without p-value itself. CI doesn't provide enough information unless effect size and power analysis are reported.

METHOD

Criteria for inclusion of journals

Following the survey studies reported by previous researchers in regard to statistical reports (Gass & Plonksy, 2011; Lindstromberg, 2016; Plonsky, 2013, 2014), I randomly selected some SLA and Applied Linguistics Journals. Table 1 describes the number of articles elected and the year of the articles published. Two journals were dropped (ASIAN EFL journal and ELT) due to beyond our criteria. ASIAN EFL journal was dropped due to the difficulty in accessing the papers with reference to experimental studies. While ELT journal provide less information with reference to experimental research design. Accordingly, ten journals were selected as a main focus of the present study (see Table 1).

Articles' inclusion criteria

I restricted our survey to empirical studies with experimental designs published from 2011 to 2017 volumes with a total article counted 217 experimental designs. Some articles which are beyond those experimental studies were excluded. The publication years elected were aimed at investigating the adherence of the authors with reference to the statistical recommendation endorsed by some SLA scholars and manual APA guide (2010).

Table 1. Sources of experimental studies in journals

Journal	k	%	Year	Impact factor CJR (2016)
Language Learning	36	16,6	2011-2017	2.079
System	34	15,7	2011-2017	NA
Tesol Quarterly	11	5,1	2011-2017	2.704
Asia TEFL	11	5,1	2011-2017	0.21
JSLW	16	7,4	2012-2017	1.591
LTR	30	13,8	2011-2015	1.741
RELC	9	4,1	2012-2017	0.83
reCALL	19	8,8	2011-2017	2.333
IJAL	19	8,8	2011-2017	NA
SSLA	32	14,7	2011 - 2017	2.044
Total	217	100		

Note: k= number of articles, NA = not avalilable, LTR=Language Teaching Research, JSLW=Journal of second language writing, IJALL=Indonesian Journal of Applied Linguistic, SSLA=Studies in Second Language Acquisition

Statistical report and interpretation classification

I searched the terms either “power analysis”, “effect size”, “CI 95% or 90%” in the introduction, method, result, and discussion sections across the articles. Cumming and Fidler (2010) provide a concise description in regard to which sections of articles (manuscript) provides information about such terms.

I coded whether any power analyses, ES, and, CI are reported in the text, table, or figure. Additionally, any interpretation of power, ES, and CI were analyzed when appearing in the text. I classified the articles based on their report and interpretation of the following categories

Power analysis report: Apriori Power and POST Hoc Power were reported to estimate the sample size as well as to the planning studies in the future. Additionally, the frequency of post power report was also calculated. In the regard to power interpretation, I counted that the researcher at least interpreted their power in the either method, result, or discussion sections.

Confidence Interval report: At least the author reported one CI in a table, figure, or text. Likewise, the interpretation of CI was also analyzed when it was explicitly mentioned in the results or/and discussion. Any mention of the width of CI i.e., narrow or wide were vetted such as "the extent of CI suggests ....", "95% CI includes [..,...] suggests ...".

Effect size report: the author at least reported one measure of effect size either Cohen's *d* or *f*, partial eta-squared (η_p^2), or eta-squared (η^2) when appeared in text, table, or figure. Furthermore, we analyzed the effect size interpretation at least one effect size was mentioned in the results or/and discussion. Any mention of the effect size's benchmarks *small, medium, large* were not coded as interpretation. Rather, we elected to the any mention of practical effect elucidated by the authors such as "the (any mention of effect size's benchmarks) effect suggests ....", or "the effect indicates ....", etc.

Coding

All selected articles were coded by using inter-rater agreement to measure the consistency of measurement. A second rater was involved in coding the statistical report and interpretation criteria counted 10% of the whole articles across the journals. This second rater was trained extensively before undertaking the coding. Agreement was at 98.5% for the first rater, and 93.8% for the second rater. When inter-rater agreement ranges between 85% and 90% is considered acceptable (Miles, Huberman, & Saldana, 2014).

Analysis

To answer research question 1, I calculated the statistical reports (mean, standard deviation, effect sizes, Apriori and Post

Hoc power analyses, and confidence intervals) provided by the authors in the results (tables or statements). The straightforward frequencies and percentages were calculated for any mention of such statistical reports in which at least one statistical report was provided by marking 1 as “reported” and 0 as “not reported”. with the regard to the research question 2, we restricted the interpretation of the three statistical reports “effect size”, “confidence interval”, and “power analysis” by marking 1 as “interpreted” and 0 as “not interpreted”. I searched such reports across the methods, results, and discussions. Whenever we found such statistical reports mentioned in the texts, we came to check and analyzed how the author interpreted them (if any). The frequency and percentage were provided for both total journals and each journals providing their statistical reports and interpretations.

FINDINGS

In this section of article, I present the statistical reports presented by the authors in the forms of frequency and percentage. I divide such forms into four sections: a) any reported and not reported statistical reports for all journals, b) any reported and not reported statistical reports for each journal) any interpreted and not interpreted of statistical reports for all journals, and d) any interpreted and not interpreted of statistical reports for each journal.

Chart 1 Data of data reported and not reported of all journals

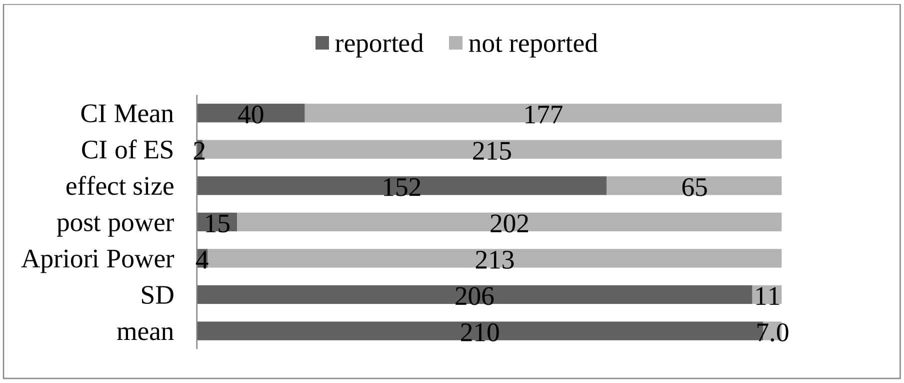

Chart 1 indicates that most articles report their descriptive statistics i.e., mean and standard deviation. Of 217 articles 210 (96.8%) reported mean, while 11 (5.1%) didn't and 206 (94.9%) reported standard deviation, while 7 (3.2%) didn't. With the regard to effect size reports, 152 (70%) were reported, while 65 (30%) were not. Only a few articles report their power analysis. Apriori power analysis was only reported 4 (1.8%), while 15 (6.9%) reported post hoc power analysis. Confidence interval for mean was reported 40 (18.4%), while CI for effect size was only 2 (0.9%) across articles.

Chart 2 statistical interpretation across journals

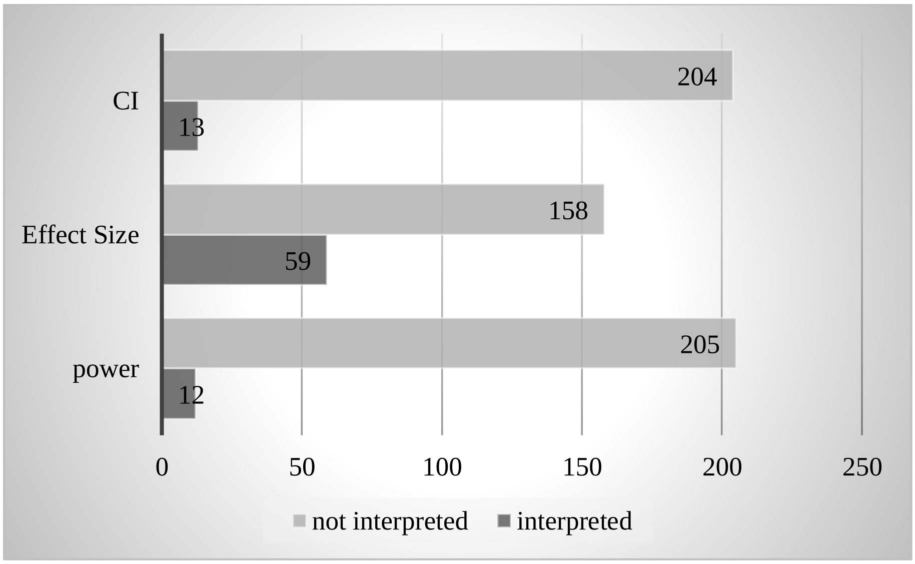

Chart 2 indicates that there is a small number and percentage of authors across journals interpret their power analysis, effect size, and confidence interval. Only 12 (5.5%) interpreted their power analysis and 13 (6%) interpreted CI. While the interpretation for effect size signifies low to medium number accounted 59 (27.2%) across 217 experimental design papers.

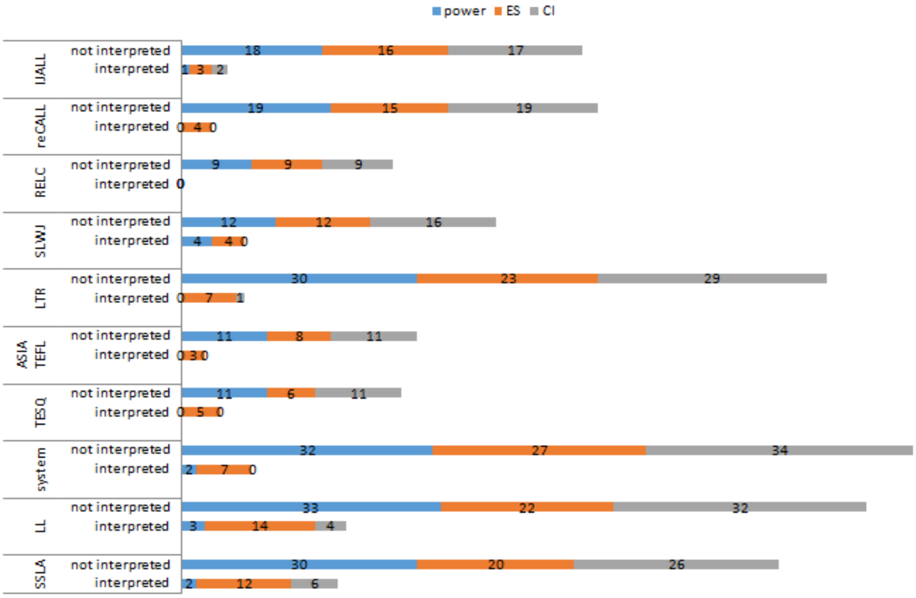

Chart 3 the writers' interpretation across statistical report reform

Chart 3 indicates the writers' interpretation of the statistical terms power, effect size, and CI. Five journals report their interpretation with reference to power analysis consecutively 1 of 19, 4 of 16, 2 of 34, 3 of 36, and 2 of 32 articles for each journal. While such power analysis was not found in the rest four journals. Eight journals provide the interpretation of the effect size which was found in the discussion section signifying the practical effect of the treatment provided during the experiments, while only one journal does not provide such an interpretation at all. Likewise, only four journals interpret their CI accounted small number of articles with ranged between 1 and 6 interpretations.

DISCUSSION

The term "new statistics" yielded by Cumming (2014) was proposed as a response to the several flaws of NHST, therefore, the need to shift from reliance on NHST to the estimation effect sizes, confidence intervals, power analyses is urgently required. However, some researchers still seem to hold a narrow concept of the new statistics in which adopting them widely will be highly beneficial.

In this study, even though there is a double increase in regard to descriptive statistic reports i.e., mean and standard deviation compared to Lindstromberg's (2016), Plonksy and Gass' (2011) study, Plonsky's (2013, 2014) study from 2011 to 2017 counted 96.8% and 94.9%, respectively. While, a somewhat increase is found in effect size report accounted 70% in our study. With the regard to CI report, there is a low increase counted 18.4% in our study. Interestingly, a steady increase is signified in power analyses accounted 1.8% for Apriori power and 6.9% for post hoc power. Plonsky (2014) asserts that when aforementioned statistical reports are left, readers and secondary reviewers have no adequacy to assess the quality of methods. Additionally, future researchers are lack of source of comparison due to the low quality of internal validity in regard to the instruments and samples.

This study, at least, provides ample evidences that there is an increase in statistical reports, specially, in terms of mean and standard deviation, effect sizes and CI. It seems that the authors have adhered the endorsement of statistical report reforms prompted by journal editors and other SLA researchers. In other words, the recommendation has been effective in raising awareness of the importance of such basic statistical reports.

However, with the regard to the power analyses, there is no significant increase in percentage compared to previous surveys. Compared to Plonsky and Gass' (2011), Plonksy's (2013, 2014) surveys, with consecutively 2%, 1%, and 2% of power reported, this survey indicates that the authors are still persistent to report such power analysis with Apriori power and post-hoc power counted 4% and 15%, respectively. It seems that what Crooker's (1991) assertion of power analyses is true that power analyses "are almost never used" (cited in Plonsky & Gass, 2011, p.351).

Additionally, Plonsky and Gass (2011) reported that overall power found in the research articles accounted approximately 0.56 – "only 56% chance to appropriately detect statistical significance" (p.350) - which far cry from Cohen's (1988) recommendation for minimum power is 0.80. Indeed, the heavy reliance on NHST, specially, the achievement of statistically significance dominates the present findings which is deemed as underlying causes of the published research finding are false (Ionnidas, 2005).

Interestingly, with reference to the power analyses, it reveals that that post-hoc power (15%) is more dominating the report than Apriori power (4%). The term "post-hoc" power refers to the observed power default in SPSS. Zumbo and Hubbley (1998) pinpoint that the invariability of power calculation after the results found is not required. They go on recommending that "post hoc power calculation can be value in designing follow up studies, but should not be used in reports (p. 104). Likewise, Greenland (2012) problematized the arbitrariness of power analyses in regard to post hoc power. The problem arises when a post-hoc power is used to judge any study to be underpowered which is irrelevant contextually compared to apriori power.

Furthermore, post hoc power is only a fixed transformation of the *p* value which rises when the such a value is shrinking. In a nutshell, if a *p* value is far from the false-positive (type I error) rate, it results in a large distance for post power from the true-positive rate of the test. It is stated that "because post power is a merely a fixed transformation on the null *p* values, it adds no new statistical information (Greendland, 2012, p.366)." In a similar vein, Kline (2004, p. 43) cites that post hoc power analysis is "more likely an autopsy than a diagnostic procedure" which is better think about the power analysis before the data collection.

Unlike post hoc power, Apriori power is strongly recommended by some researchers (Cohen, 1988; Kline, 2004; Larson-Hall, 2010; Murphy & Myors, 2003; Murphy, 2010). Such scholars provide excellent overviews of the methods, assumptions, and applications of Apriori power analysis. The consensus of the scholars assert that power analysis is applied to design studies i.e., determining sample sizes and evaluate research i.e., understanding the particular studies to reject or fail to reject the null hypothesis (Murphy, 2010). Cohen (1988) suggest that studies should be designed in regard to the apriori power analysis considering hypothesized effect size, exact alpha level, and sample sizes (see also Larson-Hall, 2010 as a main review).

The ideal power level should be 0.80 or greater meaning at least an 80% chance of rejecting a false null hypothesis but if a power is less than 0.50, it will tend to lead to type II errors (Murphy & Myors, 2003). Cohen (1988) provides complete tables available in calculating power whereas Murphy and Myors (2003) provide smaller sets of tables of power calculation. A manual and complete calculation of apriori power analysis are provided by Larson-Hall (2010) by using a sophisticated-open source software *R* software which can be found and downloaded at http://cran.r-project.org/. Likewise, Murphy and Myors (2003) provide an open source software *G*Power* in calculating power analysis developed by Faul, Erdfelder, Lang, and Buchner (2007 which can be reached at: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3.

A somewhat-significant increase is found in CI reported. A survey conducted by Lindstromberg (2016), Plonsky and Gass (2011), Plonsky (2013, 2014) indicate that the authors report CI counted consecutively 2%, 3%, 5%, and 0.4-7% (1990s-2000s) compared to this study accounted 40% and 2% reporting CI for mean and effect size, respectively. It seems that statistical literacy endorsed by Loewen et al. (2014) is true that most researchers i.e., professors and PhD students do not have adequate knowledge about statistical reports.

The most extreme comments are likely dependent on Cohen's (1994) claim that ""I suspect that the main reason they are not reported is that they are so embarrassingly large!"(Cohen, 1994, p.1002) is likely true. A continuous and simultaneous endorsement to use a new statistical report should be echoed. No definitive answer is provided by one study, therefore, further research is required to refine and redefine it to shed more light our statistical report. If we do not provide sufficient statistical information in regard to statistical report, it will impede the research development in our field.

Interpreting effect size, confidence interval, and power

Lindstromberg's (2016) survey among quasi-experimental studies surveyed between 1997 and 2015 in the journal of *Language Teaching research (LTR)* reveals that the effect size interpretation is accounted 55%. There is no an increase compared to our survey accounted only 59%. A small percentage is reported for both CI and Power analyses interpretation accounted 13% and 12%, consecutively.

Although ESs are regularly reported by SLA researchers, they are not often interpreted and even less often interpreted meaningfully (Larson-Hall & Plonsky, 2015; Plonsky & Oswald, 2014). In interpreting effect sizes, Norris and Ortega (2006, p.33) suggested to interpret them instead of "mathematical meaning" as Cohen's *d bencmarks* small ($d \leq 0.20$), medium ($0.20 < d < 0.80$), and large effects ($0.80 \leq d$ (Cohen, 1988). They go on to suggestion what "needed is a frame of reference for interpreting effect sizes that can be understood by readers, users, and researchers alike".

Likewise, Cumming (2014) emphasizes to make a judgment and reasons in interpreting the effect sizes instead of falling back those benchmarks. Similarly, American Educational Research Association (AERA) recommends "a qualitative interpretation of the effect" (AERA, 2006, p.10). The effect size benchmark recommended mostly by researchers is *standardized ES* of Cohen's *d* – typically measured in units of some relevant pooled standard deviation (SD) and means between experimental and control groups. Grissom and Kim (2005) provide excellent sources to calculate and present a variety of ES measures (see also Larson-Hall, 2010).

We code some interpretation of effect sizes which are mostly reported in discussion section. The term "suggesting or suggest" in interpreting effect sizes arise among the researchers such as *"the effect size of task type was greater on the picture description* *test than on the translation test, which suggests that task condition plays a greater role in performance on the picture description test than on the translation test"* (LTR, 2014, p.13). The aforementioned effect size interpretations postulate the authors' stance for not entrapping to the effect size benchmarks – small, medium, and large – it denotes the "practical significance" of the effects of the interventions conducted (Thompson, 2002, p. 65). Practical significance is more meaningful than statistical significance when it should be interpreted to fit the context in which the study is carried out.

Furthermore, Thompson (2001) makes a caveat "if people interpreted effect sizes with the same rigidity that α=.05 has been used in statistical testing, we would merely be being stupid in another metric". Additionally, the recommendation on effect sizes for further research is interpreted as in *"Effect sizes are taken into consideration and should be followed up by the future studies that compare relative effects of SMC and F2FC feedback …*(LL, 2013, p.29). Durlak (2009) asserts that if previous research consistently reported inadequacy of ES's magnitude i.e., .5, then the sample sizes for further studies can be anticipated. For example, in an experimental study, if ES is .5 with Sig. Level 0.05 (two-tailed), and expected power is .80, a researcher requires 64 participants in each group.

Likewise, Another study interprets effect size instead of relying on significance or not significance such as "*effect size of 0.60 suggesting a considerable mean difference between the two groups with the NHL learners being more capable of self-correcting their PA scores than their HL counterparts*" (LTR, 2015, p.14). We exclude any mention of the authors to interpret effect size as Cohen's benchmarks offered. Volker (2006) criticizes the use of such any mention of benchmarks as a general rule of thumb devoid sufficient knowledge of fitting the area discussed. Cohen (1988) urges to interpret ESs in context rather than the trivial magnitude i.e., large effect. Even, when a researcher finds a small effect in his/her treatment, it may have enormous implications in a practical context for the further research. What should be noted is that it is not only the magnitude of the effect that is important, but also practical and clinical value that must be interpreted (Durlak, 2009).

With reference to power interpretation, some authors interpret their power in regard to consider sample sizes (see Larson-Hall, 2010 as a main review), such as *"A priori power analysis was conducted to compute the minimum number of participants required by using G*Power3.1* (LL, 2017, p.10). It seems that the author is aware of the implementation of apriori power in electing the number of sample sizes. The lack adequacy of power analysis, often expressed as 1 – β, results in the likelihood of committing Type II error.

Some studies also report their post hoc power as the recommendation for future research in deciding the number of sample sizes such as *"future studies should increase sample size in order to increase the power of their statistical procedures"* (LL, 2013, p.30). Cumming (2014) problematizes the post power analyses as not telling us about the results. The use of power computation which is default in SPSS i.e., observed power, after the study is done and considered that way as misleading way, specially, when the studies reach non-significant result, is uninformative.

In regard to CI interpretation, Cumming (2012, 2014) and Plonsky and Larson-Hall (2015) provide a full discussion on it. For instance, with 95% CI [17.2 , 100.8], the study is reported to have low precision and little value which signifies 1.4 times the average of lengths of the CI's on two means. Indeed, "it is much better approach than declaring the result statistically significant, *p=0.24"* (Cumming, 2014, p. 19). Maxwell, Kelley, and Rausch (2008) provide a complete guidance of the CI's use which should entail simultaneously the direction, magnitude, and accuracy.

One study reports CI in regard to statistical significance such as *"... 95% confidence interval (CIs p<0.05) indicated no statistically significant difference which may be due to the small sample size"* (JSLW, 2015, p.10). Another study reports the precision of the CI such as *"the finding is also supported by the relatively narrow* *confidence intervals of difference in the mean gains [0.03 , 0.51] which indicates that there is a 95% chance that the difference in the mean gains between the two feedback conditions lay somewhere between 0.03 and 0.51"* (LTR, 2014, p.15). Note, interpreting CI is not merely relied on whether it contains zero value or doesn't, rather it should elucidate the precision and accuracy of a population parameter can be estimated as Thompson (1998) argued, "If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing" (p.800).

As an interval estimate of a population effect size, CI "indicates the precision of the point estimate" (Cumming & Fidler, 2010, p. 79). Commonly, CI is asymmetric with the standard error, sample size, and Margin of Error (MOE) – the length of one arm of a CI. If standard error increases, then MOE will increase which results in the width of CI increases too (less precision). Additionally, if the sample sizes increase, the MOE will decrease which leads to the narrow width of CI (more precision) meaning that the shorter the CI the better. Cumming and Fidler (2010) provides a detail information to interpret CI, while Cumming (2012) provide detail concepts, calculations, and interpretation of CI.

CONCLUSION

A large number of quantitative SLA articles using statistical tools have led to some caveats to use such tools accurately and sufficiently in reporting and presenting the reported data. Small changes of researchers in the way to report and present data can lead to large differences in the way research is comprehended (Larson-Hall & Herrington, 2010). The main purpose in this article is the introduction of the "new statistical report" mostly used in our field. Following the discussion of SLA researchers, scholars, and journal editors (hopefully the editors of this journal), it is likely necessary to "reform" the way we analyze, interpret, and present data in our quantitative articles. Presenting sufficient data and thinking about further replication and meta-analysis will lead to the improvements of our statistical analysis in our field (See the 6th Publication Manual Edition of APA, 2010).

Even though there is a significant increase in the statistical report such as mean, standard deviation, and effect size, CI and power analyses reports seem to be neglected. It seems to favor Plonsky's (2013) claim "much of the field's efforts have been underpowered and therefore unreliable" (p.453). Positively thinking, we do not favor Ionnidas' (2005) claim that "most published research findings are false" or Tressoldi, Giofre, Sella, and Cumming's (2016) claim that "most published research findings have a high probability of being false". Instead, this article suggests "most published research findings should have been a guidance for not being false in the future research". One of the ways to that is likely continuous "statistical literacy" for the researchers and graduate students in our fields (Loewen et al., 2014).

REFERENCES

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

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14 (5)*, 365–376. doi:10.1038/ Nrn3475

Brown, J., D. (2015). Why bother learning advanced quantitative methods in L2 research. In: L Plonsky (ed.) *Advancing quantitative methods in second language research.* New York: Routledge, pp. 9–20.

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal, 97*, 825–827.

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

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

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49, 97–1003.

Cook, V. (1999). Using SLA research in language teaching. *International Journal of Applied Linguistics, 9 (2),* 267-284.

DeKeyser, R. and Schoonen, R. (2007), Editors' announcement. *Language Learning, 57,* IX–X. doi:10.1111/j.1467-9922.2007.00396_2.x

Cumming, G. & Fidler, F. (2010). Effect sizes and confidence intervals. in G.,R. Hancock & R.O.Mueller (Eds.): *The reviewer's guide to quantitative methods in the social sciences* (pp. 79-91). New York: Routledge

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

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

Durlak, J.A., (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34 (9),* 917–928. doi:10.1093/jpepsy/jsp004

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

Fadilah, E. (2018). Oral corrective feedback on students' grammatical accuracy and willingness to communicate in EFL classroom: the effects of focused and unfocused prompts. *ASIAN EFL JOURNAL 20 (4),* 199-220.

Faul, F., Erdfelder, E., Lang, A.,G., and Bushner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39 (2)*, 175-191.

Fidler, F. (2002). The fifth edition of the APA Publication Manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, 62, 749–770.

Gass, S. (2009). ‘A survey of SLA research’ in W. Ritchie and T. Bhatia (eds): *Handbook of Second Language Acquisition*. Emerald, pp. 3–28.

Gigerenzer, G. (2004). Mindless statistic. *The Journal of Socio-Economics, 33,* 587-606.

Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Ann Epidemiol*, *22 (5)*, 364–368.

Harlow, L.L., Mulaik, S.A., 1935- & Steiger, J.,H. (1997). *What if there were no significance tests?* Mahwah, N.J. Lawrence Erlbaum Associates Publishers

Howell, D. C. (2002). *Statistical methods for psychology*. Pacific Grove, CA: Duxbury/Thomson Learning.

Ioannidis J.,P.,A. (2005). Why most published research findings are false. *PLoS Medicine*, *2 (8)*, 1-24

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

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

Larson-Hall, J. (2012). Our statistical intuitions may be misleading us: Why we need robust statistic. *Language Teaching, 45 (4),* 460-474.

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

Larson-Hall, J. & Herrington, R.(2010). ‘Improving data analysis in second language acquisition by utilizing modern developments in applied statistics,’ *Applied Linguistics*,*31*, 368–90.

Lazaraton, A., Riggenbach, H., & Ediger, A. (1987). Forming a discipline: Applied linguists’ literacy in research methodology and statistics. *TESOL Quarterly, 21*, 263–277.

Lindstromberg, S. (2016). Inferential statistics in *Language Teaching Research*: A review and ways forward. *Language Teaching Research, 20 (6),* 741-768.

Loewen, S. & Gass, S. (2009). The use of statistics in SLA. *Language Teaching, 42 (2),* 181-196.

Loewen, S., Lavolette, E., Spino, L. A., Papi, M., Schmidtke, J., Sterling, S. and Wolff, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. TESOL Quarterly, *48*, 360–388. doi:10.1002/tesq.128.

Maxwell, S.E., Kelley, k., & Rausch, J., R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review Psychology, 59,* 537-563.

Miles, M.B., Huberman, A.M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook (3rd ed.).* California: SAGE.

Murphy, K. R., & Myors, B. (2004). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests (2nd ed.)*. Mahwah, NJ: Erlbaum.

Murphy, K., R. (2010). Power analysis. in G.,R. Hancock & R.O.Mueller (Eds.): *The reviewer's guide to quantitative methods in the social sciences* (pp. 329-336). New York: Routledge

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

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

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning, 65(S1)*, 97–126.

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

Plonsky, L. & Gass, S. (2011). 'Quantitative research methods, study quality, and outcomes: The case of interaction research.' *Language Learning,* 61, 325–66.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687

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

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.

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

R Core Team. (2012). *R: A Language and environment for statistical computing. r foundation for statistical computing*, available at http://www. R-project.org/.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 133–164). Amsterdam: John Benjamins.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development,* 70, 434–438.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80–93.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes? *Educational Researcher*, *31 (3)*, 25-32.

Tressoldi, P. E., Giofre, D., Sella, F., & Cumming, G. (2016). High impact = high statistical standards? not necessarily so. *PLoS ONE 8(2), 1-7. doi:10.1371/journal.pone.0056180*

Volker, M. A. (2006). Reporting effect size estimates in school psychology research. *Psychology in the Schools, 43,* 653–672.

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

Zumbo, B.D. & Hubley, A.,M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician, 47(2)*, 385–388.