DOI:10.6129/CJP.202009_62(4).0006
中華心理學刊 民 109，62 卷，4 期，553-592
Chinese Journal of Psychology 2020, Vol.62, No.4, 553-592

Effect Size Reporting Practices in Taiwanese Psychology and Education Journals: Review and Beyond

Li-Ting Chen[1], Qi-Wen Ding[2], Cheng-Yu Hsieh[2], Yi-Kai Chen[2], Yu-Shan Chiang[2], Ssu-Ching Huang[3, 4], Tong-Rong Yang[2], Che Cheng[2], Pey-Yan Liou[3], and Chao-Ying Joanne Peng[2]

Counseling and Educational Psychology, University of Nevada, Reno, USA[1]

Department of Psychology, National Taiwan University[2]

Graduate Institute of Learning and Instruction, National Central University, Taiwan[3]

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology[4]

The importance of reporting effect sizes (ESs) in quantitative empirical studies has been emphasized in the literature. However, no published study to date has shed light on current ES reporting practices in Taiwanese psychology and education journals. To fill this gap, the present study systematically reviewed 268 articles published in eight Taiwanese psychology journals and nine education journals during 2017 and 2018. All of these 17 journals were highly ranked in their respective fields. Four aspects of ES reporting practices were investigated: (A) the ES reporting rate, (B) the ES type, (C) the ES interpretation, and (D) the resolution of discrepancies between the ES magnitude and statistical significance. The results revealed that 72% of articles reported at least one ES, and more than 65% of ESs reported were the *r*-type, such as Pearson's *r* and $\hat{\eta}^2$. Of the studies that reported ESs, 55% also interpreted the ESs. More than 80% of these interpretations were the mere labeling of an ES as small, medium, or large, according to established benchmarks. Approximately 50% of the articles showed a discrepancy between the magnitude of an ES and its corresponding statistical significance, but only 35% of these articles attempted to explain or resolve the discrepancy. When the data for psychology and education articles were analyzed separately, the psychology articles exhibited a lower rate of both ES reporting and ES interpretation by labeling. In sum, the majority of articles reported at least one ES, but few interpreted ES fully or meaningfully. To assist authors with a full and meaningful ES reporting, we offer five suggestions and one exemplary ES reporting in the Extended Abstract. It is hoped that this paper contributes to an increased practice of meaningfully reporting ES(s) in empirical quantitative studies in Taiwan.

Keywords: *clinical significance, effect size, practical significance, reporting practice, statistical inference*

Extended Abstract

Since 1999, the American Psychological Association (APA) has strongly encouraged researchers to report effect sizes (ESs) to supplement their statistical analysis results and interpretations (Wilkinson, L., & the Task Force on Statistical Inference, 1999). The sixth and seventh editions of the *Publication Manual of the APA* (APA, 2010, 2020) went a step further, providing guidelines for why, how, and where ESs ought to be presented in a quantitative empirical study. Similarly, the American Educational Research Association (AERA) formulated guidelines on ES reporting for its affiliated journals in 2006 (AERA, 2006).

Indeed, the combined impact of the APA/AERA guidelines, editorial policies, and computing software defaults to automatically generate ESs has contributed to increased reporting of ESs in various disciplines (e.g., Peng et al., 2013; Sun et al., 2010), on specific topics (e.g., Zientek et al., 2008), and even in non-APA and non-

AERA journals (e.g., Alhija & Levy, 2009). There has also been an increase in the reporting of ES confidence intervals (CIs), in ES interpretations in terms of practical and clinical significance, and in novel classifications of ESs after 1999 (Peng et al., 2013). However, these findings were based exclusively on reviews of American journals in the subfields of psychology and education. No published study to date has shed light on ES reporting practices in Taiwanese psychology and education journals.

To fill this gap in the literature and promote meaningful ES reporting, the present study investigated four aspects of ES reporting practices in and between Taiwanese psychology and education journals. These four aspects are: (A) the ES reporting rate, (B) the ES type, (C) the ES interpretation, and (D) the resolution of discrepancies between the ES magnitude and statistical significance.

Method

Journals and Articles Reviewed

A total of 268 articles published in 17 Taiwanese psychology and education journals during 2017 and 2018 were reviewed. The eight psychology journals were *Chinese Journal of Psychology* (*中華心理學刊*), *Formosa Journal of Mental Health* (*中華心理衛生學刊*), *Chinese Journal of Guidance and Counseling* (*中華輔導與諮商學報*), *Indigenous Psychological Research in Chinese Societies* (*本土心理學研究*), *Bulletin of Educational Psychology* (*教育心理學報*), *Journal of Education & Psychology* (*教育與心理研究*), *Taiwanese Journal of Psychiatry* (*臺灣精神醫學*), and *Research in Applied Psychology* (*應用心理研究*). The nine education journals were *Bulletin of Special Education* (*特殊教育研究學刊*), *Bulletin of Educational Research* (*教育研究集刊*), *Journal of Educational Research and Development* (*教育研究與發展期刊*), *Educational Policy Forum* (*教育政策論壇*), *Journal of Research in Education Sciences* (*教育科學研究期刊*), *Journal of Educational Media & Library Sciences* (*教育資料與圖書館學*), *Contemporary Educational Research Quarterly* (*當代教育研究季刊*), *Curriculum and Instruction Quarterly* (*課程與教學季刊*), and *Taiwan Journal of Sociology of Education* (*臺灣教育社會學研究*). These journals were rated as reputable by Weng, Huang, and Cheng (2012), Hwang (2009), and the 2017 Taiwanese Social Science Citation Index (TSSCI). Details of the 17 journals and 268 articles are presented in supplemental materials available at https://osf.io/n69xs/.

The articles included in the present study had to be empirical and quantitative in nature and applied at least one statistical analysis to answer their research questions. Simulation studies were excluded because the ESs in these studies are defined theoretically. Meta-analytical review articles or articles with test construction/development as their main focus were also excluded because in these types of study, ESs serve a different purpose than in quantitative empirical studies.

Coding of ES

Articles that met the inclusion criteria from each journal were reviewed by one member of the research team, who extracted information from each article on the four aspects of ES reporting practices according to the coding scheme (see the Appendix). For each of the 17 journals, another member of the research team independently recoded 30% of randomly selected articles. Any differences between the two coders were resolved by discussion until 100% agreement was reached. During the coding process, regular meetings were held to ensure that the coding scheme was consistently and correctly applied.

Results

Each article served as the unit of analysis. All of the analyses were conducted using PROC FREQ in SAS 9.4. An α level of .05 was preselected as the level of statistical significance. Regarding (A) the ES reporting rate, results revealed that 192 articles (72%) reported at least one ES. Psychology articles yielded a lower ES reporting rate (65%) than education articles (79%), and the difference was statistically significant ($\chi^2(1, N = 268) = 6.98$, $p = .01$). The odds of reporting at least one ES in the education articles were 2.09 times higher than those for the psychology articles, with a 95% CI = [1.20, 3.63].

Regarding (B) the ES type, we classified all ESs into three types: *d*-type, *r*-type, and others (Kelley & Preacher, 2012; Kirk, 2005; Rosenthal, 1994). The most frequently reported ESs were the *r*-type (67.0%), such as Pearson's *r* or $\hat{\eta}^2$, while the least reported were the *d*-type (9.1%). The fit indices of structural equation modeling (SEM) were the most frequently reported ESs in the others category. The difference between psychology and education articles in terms of ES type reported was not statistically significant ($\chi^2(2, N = 318) = 2.61$, $p = .27$). The odds of reporting an *r*-type ES in education articles were 0.95 times lower than those in psychology articles, with a 95% CI = [0.59, 1.52].

For (C) the ES interpretation, we defined three types. The first type labeled an ES according to an established benchmark. For example, Cohen's *d* can be labeled small, medium, or large according to Cohen's (1988) criteria. The second type compared the ES with ESs of other published studies. For example, Huang and Chen (2018) cited previous research in interpreting the relative risk (RR), which is as follows:

> Our data also revealed that the estimated RR of alcohol-related injuries in northern Taiwan is 2.54 (95% confidence interval = 1.84-3.51). ⋯ According to the published data of Borges et al, we found that RR was also higher in Taiwan (2.54) than those of western countries with similar proportion of alcohol-related injuries. (Huang & Chen, 2018, pp. 203, 206)

The third type interpreted the ES with reference to its clinical and practical significance (Kendall, 1999; Kirk, 1996). The second and third types of interpretation are informative and align with the APA and AERA guidelines. Among the 192 articles that reported at least one ES, 106 (55%) offered an interpretation. A higher proportion of education articles than psychology articles (58% vs. 52%) interpreted the ESs, although the difference was not statistically significant ($\chi^2(1, N = 192) = 0.66$, $p = .42$). The odds of interpreting an ES in education articles were 1.27 times higher than those in psychology articles, with a 95% CI = [0.72, 2.24].

Approximately 89% of interpretations were a mere labeling of the ES as small, medium, or large. About 9% of interpretations compared the ESs with those of previous published studies, while only 2% discussed the clinical or practical significance of the ESs. After simplifying interpretations into labeling versus non-labeling, the difference between psychology and education articles in labeling ES was statistically significant ($\chi^2(1, N = 106) = 4.82$, $p = .03$). The odds of interpreting ESs by labeling in education articles were 4.23 times higher than those in psychology articles, with a 95% CI = [1.08, 16.64].

Regarding (D) the resolution of discrepancies between the ES magnitude and statistical significance, we first examined each of the 192 articles that reported at least one ES to determine if it contained a discrepancy. A discrepancy was determined if an ES was at least medium yet its corresponding statistical test was insignificant, or vice versa. The judgement of an ES as small, medium, or large was based on published benchmarks, such as those for Cohen's *d* (Cohen, 1988) or for goodness of fit in SEM (Hu & Bentler, 1999). The judgement of statistical significance was based on the author(s)' specification of the α value or *p* level. Seventeen articles reported ESs without a corresponding statistical test, such as the area under the receiver operating characteristic (ROC) curve. Eighty-six (49%) of the remaining 175 (192-17) articles exhibited a discrepancy. Specifically, 43% of the psychology articles and 55% of the education articles exhibited a discrepancy. The difference between these two percentages was not statistically significant ($\chi^2(1, N = 175) = 2.58$, $p = .11$). The odds of exhibiting a discrepancy in education articles were 1.63 times higher than those in psychology articles, with a 95% CI = [0.90, 2.97].

For the 86 articles that exhibited a discrepancy between the magnitude of an ES and its statistical significance, we further investigated whether these discrepancies were explained or resolved. The results showed that only 35% of the articles attempted to explain or resolve such discrepancies. Specifically, 31% of the psychology articles explained or resolved them, compared with 37% of the education articles, and this difference was not statistically significant ($\chi^2(1, N = 86) = 0.31$, $p =$

.58). The odds of explaining or resolving a discrepancy in education articles were 1.30 times higher than those in psychology articles, with a 95% CI = [0.52, 3.22].

Comparisons of Taiwanese and American ES Reporting Practices

The ES reporting rate of Taiwanese education journals was comparable to that of AERA journals (Peng et al., 2013; Sun et al., 2010), and both American and Taiwanese ES reporting rates were higher for education than psychology journals. However, there was great variation in the ES reporting rate among journals in the same field. In terms of ES types, Taiwanese journals reported R^2 and $\hat{\eta}^2$ at a frequency equal to that of APA or AERA journals. Yet, compared with their American counterparts, Taiwanese journals reported Cohen's *d* far less frequently, and far more frequently reported Pearson's *r*, regression coefficients in mediation analysis, and fit indices in SEM.

More than 50% of the Taiwanese and American articles that reported ESs also interpreted them (Alhija & Levy, 2009; Peng et al., 2013; Sun et al., 2010), although the interpretations mostly were a mere labeling of the ESs. In terms of discrepancies between the ES magnitude and statistical significance, Sun et al. (2010) reported lower percentages (10% to 16%) than those found in both Taiwanese psychology (43%) and education (55%) articles. These discrepancies were resolved in 31% of Taiwanese psychology articles and 37% of Taiwanese education articles compared with 22% of APA articles, 12% of AERA articles, and 48% of non-APA/non-AERA articles reported in Sun et al. (2010).

Recommendations and Discussion

In light of the findings of the present study and those reported in Alhija and Levy (2009), Peng et al. (2013), and Sun et al. (2010), we formulated five recommendations to improve current ES reporting practices. First, each ES should be clearly defined along with its supporting reference(s). Second, ESs with sound properties should be preferred over variants or alternatives, whether the ESs are standardized or unstandardized. A sound ES index should be easy to be comprehended and should convey the practical significance of the result or its clinical/theoretical importance. If an ES estimates a population parameter, it should be unbiased (refer to Tables 1a to 1c in the supplemental materials). Third, each ES should be reported along with its CI. The width of a CI directly reflects the precision of a sample ES estimate. Fourth, an ES should be interpreted based on similar past research findings, and/or the clinical or practical importance of the result. Such an interpretation should take into account specific facets of a study, such as the population of interest, the treatment or intervention introduced, and the measurement method(s). For intervention studies, the magnitude of an ES has been shown to be influenced by the study design/procedure (Bakker et al., 2019; Kraft, 2020; Schäfer & Schwarz, 2019; Simpson, 2020), instruments (Cheung & Slavin, 2016; Li & Ma, 2010), the definition of the experimental and control groups (Simpson, 2018; Steenbergen-Hu & Cooper, 2014), and the characteristics of the sample (Simpson, 2018, 2019). To accurately interpret an intervention effect or a treatment manipulation, the context of a study needs to be considered, along with the magnitude of the ES. Merely labeling an ES according to publicized benchmarks, such as Cohen's (1988) criteria, is inadequate and insufficient. Fifth, when there is a discrepancy between the magnitude of an ES and its corresponding statistical significance, authors need to explain or resolve this discrepancy (see Table 2 in Fan, 2001).

Any empirical study that applies quantitative methods to answer research questions can be facilitated by the practical guidelines offered by Sun et al. (2010). The computation of ES can be accomplished by general-purpose software, such as SPSS and SAS, or specialized free software, such as the Practical Meta-Analysis Effect Size Calculator at https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD-main.php, or the Effect Size Calculators at https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html.

This study has some limitations. First, the findings may not be generalizable to other Taiwanese psychology or education journals, because ES reporting practices

were found to vary greatly even within the same journal. Second, the ES reporting practices revealed in this study may be associated with the statistical analysis performed. We did not investigate this potential association. Third, this study did not explore the reasons for certain reporting practices, such as the mere labeling of ES as small, medium, or large or the under-reporting of a few ES indices. Further studies are needed to fully understand the reasons behind the current Taiwanese ES reporting practices.

It is encouraging to note that the majority of empirical research findings published in Taiwanese psychology and education journals during 2017 and 2018 followed the APA/AERA guidelines on ES reporting. The present investigation has documented areas in which current ES reporting practices can be improved. It is hoped that this paper contributes to an increase in meaningful ES reporting in empirical quantitative studies in Taiwan.

Appendix

Effect size coding scheme

Items	Responses
1. What was the reported ES?	_____
2. On what page is the reported ES?	_____
3. Was the ES calculation specified (e.g., equation, statistical software, references)?	☐ 0. No. ☐ 1. Yes_____
4. What was the interpretation of the reported ES?	☐ 0. No interpretation. ☐ 1. The ES was labeled as small, medium, or large. ☐ 2. The ES was compared with those of similar studies. ☐ 3. The ES was interpreted in terms of its practical implications and clinical significance.
5. How did the author(s) assess the ES magnitude?	☐ 1. The author(s) assessed the ES magnitude subjectively or cited published benchmarks to assess ES magnitude (in SEM, fit indices were assessed as ES for the acceptability of the model). ☐ 2. The author(s) did not assess ES magnitude. The coders located references to assess the magnitude of the ES (in SEM, fit indices were assessed as ES for the acceptability of the model). ☐ 3. The literature has no established benchmark for assessing the reported ES. All coders discussed and agreed cutoffs for small, medium, and large ESs.
6. Was there a discrepancy between the ES magnitude and its corresponding statistical significance?	☐ 0. The reported ES did not have a corresponding significance test. ☐ 1. Yes. For SEM, a discrepancy existed when the chi-square test was significant (indicating that the model did not fit the data), but the fit indices indicated an adequate model fit. For other statistical methods, a discrepancy existed when the statistical test was significant but its corresponding ES was small, or when the test was not significant but its corresponding ES was medium or large. ☐ 2. No discrepancy.
7. Was the discrepancy explained or resolved?	☐ 0. Not applicable (the response to Item 6 was 0 or 2). ☐ 1. Neither explained nor resolved. ☐ 2. Explained or resolved.