

## 效果量在臺灣心理與教育期刊的應用：回顧與再思

Li-Ting Chen<sup>1</sup>、丁麒文<sup>2</sup>、謝承佑<sup>2</sup>、陳奕凱<sup>2</sup>、江宇珊<sup>2</sup>、黃思婧<sup>3,4</sup>、楊同榮<sup>2</sup>、鄭澈<sup>2</sup>、  
劉佩艷<sup>3</sup>、彭昭英<sup>2</sup>

Counseling and Educational Psychology, University of Nevada, Reno, USA<sup>1</sup>

國立臺灣大學心理學系暨研究所<sup>2</sup>

國立中央大學學習與教學研究所<sup>3</sup>

國立臺灣科技大學數位學習與教育研究所<sup>4</sup>

雖然效果量在量化實徵研究裡十分重要，但過去尚未有研究探討效果量在臺灣心理學與教育學的應用情況。本研究系統性地回顧了 2017 年與 2018 年發表在臺灣具高評價的八本心理學與九本教育學期刊，一共 268 篇文章，旨在探討四個報告效果量的面向：(1) 效果量報告的比例、(2) 效果量報告的類型、(3) 效果量的解釋，以及 (4) 作者如何處理統計顯著性與效果量強度的落差。結果顯示：72% 的文章報告至少一個效果量，超過 65% 的效果量是  $r$  類型（如：Pearson 相關係數或  $\eta^2$ ）。在報告效果量的文章中，55% 解釋了效果量，80% 以上對效果量的解釋僅以小、中或大的標籤標註。在同時報告統計顯著性與效果量的文章裡，50% 呈現兩者間有落差的問題，其中僅有 35% 對落差提供解釋。就兩學門的比較而言，儘管心理學期刊文章報告效果量的比例顯著地低於教育學，然而教育學文章使用標籤化的方式解釋效果量之比例卻顯著地高於心理學。整體而言，大多數的作者報告了效果量，卻未必對效果量的意涵提供完整的解釋。本文提出五點效果量報告之建議與四個效果量報告的範例，希冀能幫助讀者在研究報告中正確地闡述效果量的意涵，並促進效果量在量化研究之廣泛應用。

**關鍵詞：**研究報告慣例、效果量、統計推論、實務顯著性、臨床顯著性

近二十年來，在心理學界及教育學界的量化研究中，同時報告統計檢定顯著性（ $p$  值）與效果量（effect size, ES）的文章，有漸增的趨勢。《美國心理學會出版手冊》第六版與第七版（American Psychological Association, 2010, 2020）皆強調，學者需在量化的實徵研究裡報告效果量；如果可能的話，也報告效果量的信賴區間（confidence interval）。

究竟效果量提供讀者什麼重要的訊息？為何報告效果量如此重要？又該如何報告效果量？本文回顧過

去探討效果量相關議題之論文，並根據美國心理學會（American Psychology Association, APA）與美國教育學會（American Educational Research Association, AERA）所擬定的效果量報告準則，回答上述幾個問題。此外，本文也依據 APA/AERA 報告準則，回顧 2017 年與 2018 年發表在臺灣心理學與教育學期刊的論文，以便深入瞭解效果量在臺灣這兩學門裡應用的現況。最後，根據回顧的結果，本文提出效果量報告之建議與典範，俾使效果量在臺灣的量化實徵研究中能被正

初稿收件：2019/12/26；一修：2020/06/27；正式接受：2020/07/21

通訊作者：Li-Ting Chen（litingc@unr.edu）William Raggio Building Rm 3038, University of Nevada, Reno/0281, Reno, Nevada 89557-0281, USA

致謝：感謝主編與兩位匿名審查委員提供諸多寶貴修改意見。

確地應用與解釋。

以下的段落將依序討論統計檢定之侷限、效果量之意涵與用途、效果量之類型、APA/AERA 效果量報告準則，以及美國心理與教育期刊報告效果量之回顧研究，進而提出本研究目的。

### （一）統計檢定之侷限

在社會與行為科學的實徵研究裡，學者常會對其研究問題所對應的量化指標進行虛無假設的統計檢定（null hypothesis significance testing），以下簡稱統計檢定。統計檢定為進行統計推論（statistical inference）之常見派典（paradigm），其混合了 Fisher（1925a）的顯著檢定（significance testing）以及 Neyman 與 Pearson（1928）的假設檢定（hypothesis testing）兩大統計推論派典，當今統計教科書對於統計檢定的介紹乃結合了上述兩派典之精神。有關 Fisher 與 Neyman-Pearson 兩學派之比較可參考 Carlson（1976）、Huberty（1987, 1993）及 Lenhard（2006）。

儘管多數研究者在量化研究裡採用統計檢定進行推論，但統計檢定的邏輯和實用性早在 1938 年就被一篇由 Berkson 發表在 *Journal of the American Statistical Association* 名為「Some Difficulties of Interpretation Encountered in the Application of the Chi-square Test」的文章所質疑。最近，Amrhein、Greenland 及 McShane（2019）在 *Nature* 期刊發表一篇名為「Retire Statistical Significance」的文章，獲得全球八百多位來自不同領域的專家連署同意。學者對統計檢定的主要批判為：只要測量工具夠精準，或者樣本數夠大，統計檢定便具有足夠的統計檢定力（statistical power）推翻如「組間的平均數無差異」或是「兩變項間的相關係數為零」之虛無零假設（nil null hypothesis）（其他對統計檢定的批判請見本文的補充材料 <https://osf.io/n69xs/>）。然而，即便學者透過統計檢定推翻了虛無零假設，其結果仍無法提供更多關於母群參數的資訊（Cohen, 1994; Kirk, 1996; Meehl, 1967）。學者在詮釋結果時，更應關注觀察效果的強度，例如：「組間平均數的差異有多大」、「兩變項間的相關係數有多高」，而非僅止於「組間平均數有無差異」，抑或「兩變項間的相關係數是否為零」的結論。因此，許多文獻建議學者除了報告統計檢定的結果之外，也應報告該檢定所對應之效果量（Citrome & Ketter, 2013; Grissom & Kim, 2012; Kirk, 1996; Maxwell, Camp, & Arvey, 1981）。

### （二）效果量之意涵與用途

早在 1905 年，Karl Pearson 即呼籲學者在研究結果中報告效果量。其後 Fisher（1925a）更明確鼓勵學者報告相關比（correlation ratio）或其平方（ $\eta^2$ ）作為描述獨變項與依變項間關聯強度的指標，來補充說明變異數分析（analysis of variance, ANOVA）之統計檢定結果。1960 年代，Cohen 正式提出效果量的指標，如標準化的兩組平均數差異（standardized mean difference），供學者執行事前統計檢定力分析（a priori power analysis）來預估所需的有效樣本數（Cohen, 1962, 1969）。爾後，Cohen（1965）與 Hays（1963）鼓勵心理與教育領域的學者在實徵研究的結果中報告效果量，以補充說明統計檢定之結果。Kirk（1996）則認為效果量數值可以做為實務顯著性（practical significance）的指標，根據 Kirk（1996）的觀點，統計顯著性（statistical significance）表徵研究結果是否由機率或抽樣變異所造成；反之，實務顯著性則著重於研究結果是否具有實用價值。針對臨床的研究，Kendall（1999）建議學者依據診斷的標準，評估病患接受治療後與正常人的差異來說明臨床顯著性（clinical significance），例如：病患接受治療後，是否可以恢復到正常的功能？這些呼籲及觀點促使後續許多學者提出各式各樣的效果量指標，以彌補統計檢定之不足。

針對單一研究的結果，效果量可表徵觀察效果的強度，其數值應與樣本數及統計顯著性無關（Kelley & Preacher, 2012; Rosenthal, 1994）。一個好的效果量應易於理解（Pek & Flora, 2018）、能表達研究結果的實用性（Kirk, 1996），或是可彰顯研究結果在理論上或臨床上的重要性（APA, 2010; Thompson, 2002; Wilkinson & the Task Force on Statistical Inference, 1999）。

效果量不僅能增進學者對單一研究結果的理解，也能協助學者進行後續的整合分析（meta-analysis）來整合個別研究之結果，以便估計母群效果量參數，或是辨識可能調節（moderate）效果量強度的因素。《美國心理學會出版手冊》第七版把整合分析歸為十大論文類別之一，並詳細說明這類論文應如何報告，AERA 亦成立特殊興趣小組（Systematic Reviews and Meta-Analysis Special Interest Group）探討整合分析之相關議題。然而，部分學者指出效果量在整合分析中的誤用與錯誤解釋，並質疑整合分析之合理性（Simpson, 2018, 2019, 2020），本文在討論與建議的（三）效果量的誤用及錯

誤解釋中，進一步探討此議題。

總而言之，效果量所提供的資訊可幫助學者瞭解單一研究效果的強度、執行事前統計檢定力分析、判斷實務或臨床的顯著性，以及進行整合分析。上述藉效果量能完成的分析或判斷，除整合分析外，皆無法由統計檢定達成。

### (三) 效果量之類型

根據 Rosenthal (1994)、Kirk (2005) 及 Kelley 與 Preacher (2012) 的理念，本研究將效果量分為三類： $d$  類型、 $r$  類型及其他類型。 $d$  類型效果量藉分數差異來描述效果的大小，如：平均數差異、Cohen's  $d$ 、Cohen's  $q$ 、百分比差異等。 $r$  類型效果量則藉變項間的關聯強度來描述效果的大小，如： $r$ 、 $R^2$ 、 $\eta^2$  及 Cramér's  $\hat{p}$  等。其他類型的效果量包括勝算比 (odds ratio, OR)、邏輯斯迴歸 (logistic regression) 中的分類正確率 (correct classification rate) 及結構方程模型 (structural equation model, SEM) 裡的適配度指標<sup>1</sup> (fit index) 等。

效果量會隨著研究設計與統計方法的不同，而衍伸出各式各樣的計算方法與指標，涵蓋的範圍相當廣泛。例如：在受試者間設計 (between-subjects design) 中，Cohen's  $d$  可用來描述標準化的兩組平均數差異；在變異數分析中， $\eta^2$  或  $\omega^2$  可用來說明某一因子能解釋依變項之變異量的比例 (proportion of explained variance)。Pearson 相關係數則用來描述兩連續變項間的線性關係 (Funder & Ozer, 2019; Rosenthal, 1994)。此外，以線性迴歸 (linear regression) 所估計的線性模型之參數也有其對應的效果量指標，如：迴歸係數 (regression coefficients,  $\beta$ )、決定係數 (coefficient of determination,  $R^2$ ) 或調整後決定係數 (adjusted  $R^2$ )。邏輯斯迴歸中的類決定係數 (pseudo- $R^2$ ) 也可用來作為效果量指標 (Cox & Snell, 1989; Cragg & Uhler, 1970; Maddala, 1983; McFadden, 1973; Nagelkerke, 1991)。在中介分析 (mediation analysis) 中，當獨變項與中介變項沒有交互作用時，兩迴歸係數的相乘積  $\hat{a} \times \hat{b}$  可視為效果量，在此  $\hat{a}$  為獨變項對中介變項的迴歸係數， $\hat{b}$  為控制獨變項下中介變項對依變項的迴歸係數 (Imai, Keele, & Yamamoto, 2010; Preacher & Kelley, 2011; VanderWeele & Vansteelandt, 2009)。若上述三變項皆為連續變項時，則此相乘積可解釋為：當獨變項增加一單位時，透過中介變項，平均而言依變項的改變

量<sup>2</sup> (Hayes & Preacher, 2014; Preacher, 2015; Preacher & Kelley, 2011)。

在無母數統計 (nonparametric statistics) 的範疇中，優勢機率 (probability of superiority; Grissom & Kim, 2001)、優勢統計量 (dominance statistic; Cliff, 1993)、隨機優劣勢排序的  $A$  測量 ( $A$  measure of stochastic superiority; Vargha & Delaney, 2000) 等指標可用來估計兩母群分配重疊之程度，因此可視為效果量指標。這些指標都和 Wilcoxon rank sum (或 Mann-Whitney) 檢定統計量有關，也與 ROC 曲線下面積 (area under the receiver operating characteristic curve) 有密切的關係 (Peng, Chen, Chiang, & Chiang, 2013)。本文將文獻中較常出現，或有較佳統計特性的效果量 (如不偏性，即 unbiasedness) 呈現在表 1a 至表 1c，供讀者參考。

### (四) APA/AERA 效果量報告準則

APA 於 1999 年發表統計推論的專案報告裡，建議學者在進行統計分析時報告並解釋效果量 (Wilkinson & the Task Force on Statistical Inference, 1999)，此項建議隨後被納入《美國心理學會出版手冊》第六版 (APA, 2010) 及第七版 (APA, 2020) 中。AERA 於 2006 年也在 *Educational Researcher* 一篇名為「Standards for Reporting on Empirical Social Science Research in AERA Publications」的文章中，提出作者發表實徵研究在 AERA 期刊上時，所須遵循的原則。究竟作者該如何報告效果量？本文整理了 APA 與 AERA 所擬定的報告準則，詮釋如下：

- 作者應在研究結果裡報告效果量，以便讀者體認研究效果的強度或結果的重要性 (APA, 2010, p. 34; APA, 2020, p. 89)。
- 報告效果量的基本原則在於提供讀者足夠的資訊，以便讀者評估結果的實用性，或是結果在理論上或臨床上的重要性 (APA, 2010, p. 34; APA, 2020, p. 89)。
- 若測量單位在實務上有其意義 (如：年收入、正確的作答題數)，則作者可根據原始測量單位報告效果量。此外，作者也可以額外報告標準化或無測量單位的效果量 (如：Cohen's  $d$ 、標準化迴歸係數) 以增加效果量的實用性 (APA, 2010, p. 34; APA, 2020, p. 89; Wilkinson & the Task Force on Statistical Inference, 1999, p. 599)。
- 當「單一自由度的效果量指標」所提供的訊息，比「多自由度的效果量指標」更有助於討論結果時，

作者應報告前者<sup>3</sup>（APA, 2010, p. 34; APA, 2020, p. 89）。例如：在趨勢分析（trend analysis）中，若某一等距時間獨變項（equally-spaced time independent variable）有兩個自由度，則可被拆解為線性與二次多項式的正交對比（orthogonal contrast）分別進行統計檢定，並計算其效果量（各為單一自由度）。此分析方式可能比執行兩個自由度的  $F$  檢定並報告其相應的效果量，更有利於結果的解釋（Schad, Vasishth, Hohenstein, & Kliegl, 2020）。

- 作者應該同時報告效果量、效果量的信賴區間或標準誤、統計檢定的結果及其顯著性，這些都是幫助解釋研究結果的重要指標（AERA, 2006, p. 37; APA, 2010, p. 34; APA, 2020, p. 89）。效果量的信賴區間或標準誤是估計效果量精準性的指標，可表示效果量的不確定性。
- 作者應針對研究問題詮釋效果量的意涵，並說明研究結果的不確定性（例如：所估計的效果大到足以宣稱具有教育上的重要性，然而此結果並不排除真實效果

表 1a  $d$  類型的效果量：母群效果量與其樣本估計

母群效果量	樣本估計
<p>兩母群標準化平均數差 (<math>\delta</math>)</p> $\delta = \frac{\mu_1 - \mu_2}{\sigma} \text{ (單尾檢定, 且對立假設為 } \mu_1 > \mu_2 \text{)}$ $= \frac{ \mu_1 - \mu_2 }{\sigma} \text{ (雙尾檢定)}$ <ul style="list-style-type: none"><li>• <math>\mu_1</math> = 第一組的母群平均數</li><li>• <math>\mu_2</math> = 第二組的母群平均數</li><li>• <math>\sigma</math> = 共同母群標準差（假設兩母群的標準差相等）（Cohen, 1969, p.18）</li></ul>	<p>Cohen's <math>d</math> (Cohen, 1969, pp. 64-65)</p> $d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{pooled}^2}} \text{ (單尾檢定)}$ $= \frac{ \bar{X}_1 - \bar{X}_2 }{\sqrt{s_{pooled}^2}} \text{ (雙尾檢定)}$ <ul style="list-style-type: none"><li>• <math>\bar{X}_1</math> = 第一組的樣本平均數</li><li>• <math>\bar{X}_2</math> = 第二組的樣本平均數</li><li>• <math>s_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}</math> = 兩組合併樣本之變異數（pooled sample variance），在此，<math>n_1</math> 為第一組樣本的樣本數，<math>n_2</math> 為第二組樣本的樣本數，<math>s_1</math> 為第一組樣本的標準差，<math>s_2</math> 為第二組樣本的標準差（Cohen, 1969, pp. 17-18，其使用必須符合母群常態性與變異數相等之假設）</li></ul> <p>注意事項：</p> <ol style="list-style-type: none"><li>1. Cohen 之解釋準則（Cohen, 1988, pp. 24-27）：小效果量 = 0.2、中效果量 = 0.5、大效果量 = 0.8。</li><li>2. 信賴區間計算方式可參考 Goulet-Pelletier 與 Cousineau（2018）以及 Odgaard 與 Fowler（2010）。</li><li>3. Cohen's <math>d</math> 會高估 <math>\delta</math>，其估計偏差為 <math>\delta \times \left[1 - \frac{1}{c(m)}\right]</math>，其中 <math>c(m)</math> 之定義請見下欄（Hedges, 1981）。</li><li>4. Cohen's <math>d</math> 在傳達臨床顯著性上有所侷限，讀者可參考 Kraemer 與 Kupfer（2006）來瞭解其他較能傳達臨床顯著性的效果量指標。</li></ol>
	<p>Hedges's <math>g_u</math> (Hedges, 1981, pp. 111-114)</p> $g_u = d \times c(m)$ <ul style="list-style-type: none"><li>• <math>d</math> = 上列 Cohen's <math>d</math></li><li>• <math>c(m) = \frac{\Gamma(\frac{m}{2})}{\sqrt{\frac{m}{2}} \Gamma(\frac{m-1}{2})}</math>，可用 <math>1 - \frac{3}{(4 \times m - 1)}</math> 逼近</li><li>• <math>\Gamma(\cdot)</math> = Gamma 函數</li><li>• <math>m = n_1 + n_2 - 2</math></li></ul> <p>注意事項：</p> <ol style="list-style-type: none"><li>1. Hedges's <math>g_u</math> 是 <math>\delta</math> 的不偏估計。當 <math>m \geq 10</math>，採用 <math>c(m)</math> 的逼近式來計算 <math>g_u</math> 時，逼近誤差會小於 .00033（Hedges, 1981, p. 114）。</li><li>2. 信賴區間計算方式可參考 Goulet-Pelletier 與 Cousineau（2018）。</li><li>3. 文獻中仍有其他估計兩母群標準化平均數差之效果量，讀者可參考 Peng 與 Chen（2014）。</li></ol>

表 1b  $r$  類型的效果量：母群效果量與其樣本估計

母群效果量	樣本估計
<p>兩變項間相關係數 (<math>\rho</math>)</p> $\rho = \frac{\sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y)}{\sigma_X \sigma_Y N}$ <ul style="list-style-type: none"> <li>• <math>X_i</math> = 第 <math>i</math> 個母群個體在變項 <math>X</math> 所測量的值</li> <li>• <math>Y_i</math> = 第 <math>i</math> 個母群個體在變項 <math>Y</math> 所測量的值</li> <li>• <math>\mu_X</math> = 變項 <math>X</math> 的母群平均數</li> <li>• <math>\mu_Y</math> = 變項 <math>Y</math> 的母群平均數</li> <li>• <math>\sigma_X</math> = 變項 <math>X</math> 的母群標準差</li> <li>• <math>\sigma_Y</math> = 變項 <math>Y</math> 的母群標準差</li> <li>• <math>N</math> = 母群的觀察值個數</li> </ul>	<p>Pearson's <math>r</math></p> $r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y n}$ <ul style="list-style-type: none"> <li>• <math>X_i</math> = 第 <math>i</math> 個樣本個體在變項 <math>X</math> 所測量的值</li> <li>• <math>Y_i</math> = 第 <math>i</math> 個樣本個體在變項 <math>Y</math> 所測量的值</li> <li>• <math>\bar{X}</math> = 變項 <math>X</math> 的樣本平均數</li> <li>• <math>\bar{Y}</math> = 變項 <math>Y</math> 的樣本平均數</li> <li>• <math>s_X</math> = 變項 <math>X</math> 的樣本標準差</li> <li>• <math>s_Y</math> = 變項 <math>Y</math> 的樣本標準差</li> <li>• <math>n</math> = 樣本數</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. Cohen 解釋 <math> r </math> 的準則 (Cohen, 1988, pp. 78-81)：小效果量 = .10、中效果量 = .30、大效果量 = .50。</li> <li>2. 讀者可在 SAS 統計軟體的 PROC CORR 程序中設定 FISHER 選項，以便計算 <math>r</math> 的信賴區間，該區間在資料輕度違反常態分配的情況下仍具有穩健性 (robustness)。</li> </ol>
<p>線性迴歸之決定係數</p> $\rho^2 = 1 - \frac{\sigma_M^2}{\sigma_0^2}$ <ul style="list-style-type: none"> <li>• <math>\sigma_M^2</math> 為在預測模型中，所無法解釋的誤差變異數</li> <li>• <math>\sigma_0^2</math> 為在虛無模型 (僅含有截距項 <math>b_0</math>) 下，所無法解釋的誤差變異數</li> </ul>	<p>決定係數 (coefficient of determination, <math>R^2</math>) 或稱多元相關係數 (squared multiple correlation coefficient)</p> $R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$ <ul style="list-style-type: none"> <li>• <math>\hat{Y}_i</math> = 第 <math>i</math> 個樣本個體在效標變項上的模型預測值</li> <li>• <math>Y_i</math> = 第 <math>i</math> 個樣本個體在效標變項上的觀察值</li> <li>• <math>\bar{Y}</math> = 效標變項 <math>Y</math> 的樣本平均數</li> <li>• <math>n</math> = 樣本數</li> <li>• <math>\sum_{i=1}^n (Y_i - \hat{Y}_i)^2</math> 為殘差平方和 (residual sum of squares)</li> <li>• <math>\sum_{i=1}^n (Y_i - \bar{Y})^2</math> 為總離均差平方和 (total sum of squares)</li> </ul> <p>注意事項：</p> <p>Cohen 解釋 <math>R^2</math> 的準則 (Cohen, 1988, pp. 412-414)：小效果量 = .0196、中效果量 = .13、大效果量 = .26。</p>
	<p>調整後決定係數 (adjusted <math>R^2</math> 或 <math>R_{adj}^2</math>)</p> <p>(Fisher, 1925b, Eq. II; Wherry, 1931)</p> $R_{adj}^2 = 1 - \frac{n-1}{n-p} (1 - R^2) = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}$ <ul style="list-style-type: none"> <li>• <math>p</math> = 模型中預測變項的個數 (含截距)</li> <li>• <math>\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-p}</math>，是預測模型下 <math>\sigma^2</math> 的限制最大概似 (restricted maximum likelihood, REML) 估計值 (Liao &amp; McGee, 2003)</li> <li>• <math>\hat{\sigma}_0^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}</math>，是虛無模型下 <math>\sigma^2</math> 的限制最大概似估計值 (Liao &amp; McGee, 2003)</li> </ul> <p>注意事項：</p> <p>當母群 <math>\rho^2 = 0</math> 時，<math>R_{adj}^2</math> 為不偏估計 (Mittlböck &amp; Schemper, 1996)。<math>R_{adj}^2</math> 能矯正 <math>R^2</math> 高估 <math>\rho^2</math> 的問題 (Maxwell et al., 1981; Shieh, 2008)，並考量模型的複雜度 (Liao &amp; McGee, 2003)，比 <math>R^2</math> 更適合用來選擇模型 (Faraway, 2014)。</p>

(續下頁)

表 1b  $r$  類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計
固定效果模型（fixed-effects model）在受試者間（between-subjects）的變異數分析裡，某獨變項可以解釋依變項的總變異之百分比（ $\rho_{fixed}^2$ ） $\rho_{fixed}^2 = \frac{(\sigma_Y^2 - \sigma_e^2)}{\sigma_Y^2}$ <ul style="list-style-type: none"> <li>• <math>\sigma_Y^2</math> = 母群變異數</li> <li>• <math>\sigma_e^2</math> = 母群組內變異數（假設各組內變異數相同）</li> </ul>	$\hat{\eta}^2$ （Pearson, 1905） 單因子變異數分析（one-way ANOVA） $\hat{\eta}^2 = \frac{SS_{Between}}{SS_{Total}}$ <ul style="list-style-type: none"> <li>• <math>SS_{Between}</math> = 樣本組間離均差平方和</li> <li>• <math>SS_{Total}</math> = 樣本總離均差平方和</li> </ul> 單因子變異數分析下的對比（contrast） $\hat{\eta}^2 = \frac{SS_{Contrast}}{SS_{Total}}$ <ul style="list-style-type: none"> <li>• <math>SS_{Contrast}</math> = 樣本對比離均差平方和</li> </ul> 多因子變異數分析（factorial ANOVA） $\hat{\eta}^2 = \frac{SS_{Effect}}{SS_{Total}}$ <ul style="list-style-type: none"> <li>• <math>SS_{Effect}</math> = 樣本單一效果離均差平方和</li> </ul> 注意事項： 1. Cohen 解釋 $\eta^2$ 的準則（Cohen, 1988, pp. 284-288）：小效果量 = .0099、中效果量 = .0588、大效果量 = .1379。 2. 報告與解釋單因子變異數分析的結果之範例請參見 Pek 與 Flora（2018）。 3. 讀者可在 SAS 統計軟體的 PROC GLM 程序中的 MODEL 指令界定 EFFECTSIZE 選項，以便計算信賴區間。
	$\hat{\omega}^2$ （Hays, 1963） 單因子變異數分析（one-way ANOVA） $\hat{\omega}^2 = \frac{SS_{Between} - (p - 1)MS_{Within}}{SS_{Total} + MS_{Within}}$ <ul style="list-style-type: none"> <li>• <math>p</math> = 組數</li> <li>• <math>MS_{Within}</math> = 樣本組內均方</li> </ul> 單因子變異數分析下的對比（contrast） $\hat{\omega}^2 = \frac{SS_{Contrast} - MS_{Within}}{SS_{Total} + MS_{Within}}$ 二因子變異數分析下的正交對比（orthogonal contrast） 因子 A 其單一自由度的對比 $\psi_A$ ：淨 $\hat{\omega}^2 = \frac{F_{\psi_A} - 1}{F_{\psi_A} - 1 + npq}$ 因子 B 其單一自由度的對比 $\psi_B$ ：淨 $\hat{\omega}^2 = \frac{F_{\psi_B} - 1}{F_{\psi_B} - 1 + npq}$ A 與 B 的交互作用之單一自由度的對比 $\psi_{A*B}$ ：淨 $\hat{\omega}^2 = \frac{F_{\psi_{A*B}} - 1}{F_{\psi_{A*B}} - 1 + npq}$ <ul style="list-style-type: none"> <li>• <math>p</math> = 因子 A 的組數</li> <li>• <math>q</math> = 因子 B 的組數</li> <li>• <math>n</math> = 細格內樣本數</li> </ul> 多因子變異數分析（factorial ANOVA） $\hat{\omega}^2 = \frac{SS_{Effect} - (df_{Effect} \times MS_{Within})}{SS_{Total} + MS_{Within}}$ <ul style="list-style-type: none"> <li>• <math>df_{Effect}</math> = 單一效果的自由度</li> </ul> 注意事項： 1. Cohen 之解釋準則（Kirk, 2013, p. 135，引自 Cohen, 1988, pp. 284-288）：小效果量 = .010、中效果量 = .059、大效果量 = .138。 2. Maxwell 等人（1981）建議 $\hat{\omega}^2$ 比 $\hat{\eta}^2$ 更適合用來推論母群的 $\rho^2$ 。 3. 讀者可在 SAS 統計軟體的 PROC GLM 程序中的 MODEL 指令界定 EFFECTSIZE 選項，以便計算信賴區間。

（續下頁）

表 1b  $r$  類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計
隨機效果模型（ <b>random-effects model</b> ）在受試者間（ <b>between-subjects</b> ）的變異數分析裡，獨變項與依變項的關聯強度——組內相關係數（ <b>intraclass correlation coefficient, ICC</b> ）	
單因子隨機效果變異數分析（one-way random effects ANOVA）	（Olejnik & Algina, 2000） 單因子隨機效果變異數分析（one-way random effects ANOVA）
$\rho_{\text{random}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_c^2}$	$\hat{\rho}_{\text{random}} = \frac{p(MS_A - MS_{\text{error}})}{SS_{\text{Total}} + MS_A}$
<ul style="list-style-type: none"> <li>• <math>\sigma_a^2</math> = 母群組間變異數</li> <li>• <math>\sigma_c^2</math> = 母群組內變異數</li> </ul>	<ul style="list-style-type: none"> <li>• <math>p</math> = 組數</li> <li>• <math>MS_A</math> = 樣本效果均方</li> <li>• <math>SS_{\text{Total}}</math> = 樣本總離均差平方和</li> <li>• <math>MS_{\text{error}}</math> = 樣本誤差均方</li> </ul>
二因子隨機效果變異數分析（two-way random effects ANOVA）	二因子隨機效果變異數分析（two-way random effects ANOVA）
因子 A： $\rho_{\text{random}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_{ab}^2 + \sigma_c^2}$	因子 A： $\hat{\rho}_{\text{random}} = \frac{p(MS_A - MS_{AB})}{SS_{\text{Total}} + MS_A + MS_B - MS_{AB}}$
因子 B： $\rho_{\text{random}} = \frac{\sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_{ab}^2 + \sigma_c^2}$	因子 B： $\hat{\rho}_{\text{random}} = \frac{q(MS_B - MS_{AB})}{SS_{\text{Total}} + MS_A + MS_B - MS_{AB}}$
因子 A 與因子 B 的交互作用： $\rho_{\text{random}} = \frac{\sigma_{ab}^2}{\sigma_a^2 + \sigma_b^2 + \sigma_{ab}^2 + \sigma_c^2}$	因子 A 與因子 B 的交互作用： $\hat{\rho}_{\text{random}} = \frac{pq(MS_{AB} - MS_{\text{error}})}{SS_{\text{Total}} + MS_A + MS_B - MS_{AB}}$
<ul style="list-style-type: none"> <li>• <math>\sigma_a^2</math> = 母群因子 A 之組間變異數</li> <li>• <math>\sigma_b^2</math> = 母群因子 B 之組間變異數</li> <li>• <math>\sigma_{ab}^2</math> = 母群因子 A 與 B 交互作用之變異數</li> <li>• <math>\sigma_c^2</math> = 母群誤差變異數</li> </ul>	<ul style="list-style-type: none"> <li>• <math>p</math> = 因子 A 的組數</li> <li>• <math>q</math> = 因子 B 的組數</li> <li>• <math>MS_A</math> = 樣本因子 A 均方</li> <li>• <math>MS_B</math> = 樣本因子 B 均方</li> <li>• <math>MS_{AB}</math> = 樣本因子 A 與 B 交互作用均方</li> <li>• <math>MS_{\text{error}}</math> = 樣本誤差均方</li> <li>• <math>SS_{\text{Total}}</math> = 樣本總離均差平方和</li> </ul>
線性迴歸之未標準化迴歸係數（ <b>unstandardized linear regression coefficients</b> ）	樣本未標準化迴歸係數
一般線性迴歸模型在母群中的效標變項 $y$ 與預測變項 $X$ 若以矩陣表示，則應符合以下關係：	線性迴歸的迴歸係數之樣本估計式可由最小平方方法求得：
$y = X\beta + \epsilon, \epsilon \sim \text{Normal}(0, \sigma^2 I)$	$\hat{\beta} = (X^T X)^{-1} X^T y$
<ul style="list-style-type: none"> <li>• <math>\beta</math> = 母群未標準化迴歸係數組成之向量（<math>p \times 1</math>）</li> <li>• <math>\epsilon</math> = 誤差（error）向量（<math>N \times 1</math>）</li> <li>• <math>\sigma^2</math> = 誤差變異數</li> <li>• <math>I</math> = <math>N \times N</math> 單位矩陣（identity matrix）</li> <li>• <math>N</math> = 母群觀察值個數</li> </ul>	<ul style="list-style-type: none"> <li>• <math>X = n \times p</math> 之設計矩陣（design matrix），其第一行之元素皆為 1，每一列為個體在所有預測變項（predictors, regressors, or covariates）上的數值。<math>n</math> 為樣本數，<math>p</math> 為預測變項個數（含截距）。</li> <li>• <math>y = n \times 1</math> 之向量，為 <math>n</math> 筆個體的效標（criteria, response, or outcome）變項</li> <li>• <math>\hat{\beta} = [\beta_0 \cdots \beta_j \cdots \beta_{p-1}]^T</math>，為 <math>p \times 1</math> 之向量，內含各預測變項之未標準化迴歸係數估計值，第一列為截距項之估計值</li> </ul>
線性迴歸之標準化迴歸係數（ <b>standardized linear regression coefficients</b> ）	樣本標準化線性迴歸係數
$\beta_{\text{std},j} = \beta_j \frac{\sigma_{X_j}}{\sigma_Y}$	$\hat{\beta}_{\text{std},j} = \hat{\beta}_j \frac{s_{X_j}}{s_Y}$
<ul style="list-style-type: none"> <li>• <math>\beta_j</math> = 第 <math>j</math> 個預測變項（<math>X_j</math>）之母群未標準化迴歸係數</li> <li>• <math>\sigma_{X_j}</math> = 第 <math>j</math> 個預測變項之母群標準差</li> <li>• <math>\sigma_Y</math> = 效標變項之母群標準差</li> </ul>	<ul style="list-style-type: none"> <li>• <math>\hat{\beta}_j</math> = 第 <math>j</math> 個預測變項（<math>X_j</math>）之樣本未標準化迴歸係數</li> <li>• <math>s_{X_j}</math> = 第 <math>j</math> 個預測變項之樣本標準差</li> <li>• <math>s_Y</math> = 效標變項之樣本標準差</li> </ul>
	<p>注意事項：</p> <p>因為標準化會改變參數之意義，文獻不建議將虛擬變項（dummy variables）標準化（Pek &amp; Flora, 2018; Schielzeth, 2010）。</p>

（續下頁）

表 1b  $r$  類型的效果量：母群效果量與其樣本估計（續）


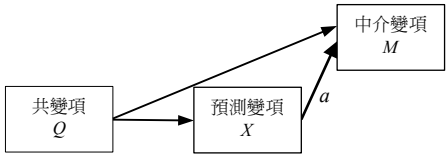
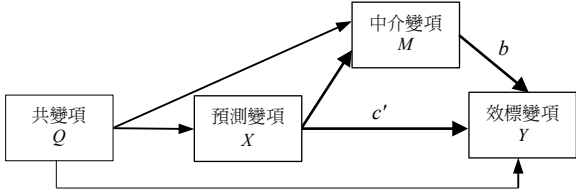
母群效果量	樣本估計
<p>兩迴歸係數之乘積做為間接效果在中介分析的效果量</p> $IE = c - c' = a \times b$ <p>是下列公式 3 的 <math>c'</math> 減去公式 1 的 <math>c</math>，或是公式 2 的 <math>a</math> 乘以公式 3 的 <math>b</math>。此模型應包含以下三個線性迴歸式，並應考量可能的共變項以避免估計偏差：</p> <p>公式 1. <math>Y_i = d_1 + cX_i + q_1Q_i + \varepsilon_{1i}</math></p>  <p>公式 2. <math>M_i = d_2 + aX_i + q_2Q_i + \varepsilon_{2i}</math></p>  <p>公式 3. <math>Y_i = d_3 + bM_i + c'X_i + q_3Q_i + \varepsilon_{3i}</math></p>  <ul style="list-style-type: none"> <li>• <math>i = 1, 2, \dots, N</math></li> <li>• <math>d_1</math>、<math>d_2</math> 及 <math>d_3</math> = 截距</li> <li>• <math>X_i</math> = 母群的預測變項，可為類別變項或連續變項</li> <li>• <math>M_i</math> = 母群的中介變項，須為連續變項</li> <li>• <math>Y_i</math> = 母群的效標變項，須為連續變項</li> <li>• <math>Q_i</math> = 母群的共變項（可有多個），可為類別變項或連續變項</li> <li>• <math>a</math>、<math>b</math>、<math>c</math>、<math>c'</math>、<math>q_1</math>、<math>q_2</math> 及 <math>q_3</math> 皆為迴歸係數</li> <li>• <math>\varepsilon_{1i}</math>、<math>\varepsilon_{2i}</math> 及 <math>\varepsilon_{3i}</math> = 個體的誤差</li> <li>• <math>N</math> = 母群觀察值個數</li> </ul>	$\hat{IE} = \hat{c} - \hat{c}' = \hat{a} \times \hat{b}$ <p>是下列公式 3 的 <math>\hat{c}'</math> 減去公式 1 的 <math>\hat{c}</math>，或是公式 2 的乘以公式 3 的 <math>\hat{b}</math>。同左欄，單一中介模型應包含三個線性迴歸式（<math>i = 1, 2, \dots, n</math>）：</p> <p>公式 1. <math>Y_i = \hat{d}_1 + \hat{c}X_i + \hat{q}_1Q_i + e_{1i}</math></p> <p>公式 2. <math>M_i = \hat{d}_2 + \hat{a}X_i + \hat{q}_2Q_i + e_{2i}</math></p> <p>公式 3. <math>Y_i = \hat{d}_3 + \hat{b}M_i + \hat{c}'X_i + \hat{q}_3Q_i + e_{3i}</math></p> <ul style="list-style-type: none"> <li>• <math>\hat{d}_1</math>、<math>\hat{d}_2</math> 及 <math>\hat{d}_3</math> 為各模型之截距估計值，<math>\hat{a}</math>、<math>\hat{b}</math>、<math>\hat{c}</math> 及 <math>\hat{c}'</math> 為各模型之迴歸係數估計值</li> <li>• <math>X_i</math> = 樣本的預測變項，可為類別變項或連續變項；<math>M_i</math> 為樣本的中介變項，須為連續變項；<math>Y_i</math> 為樣本的效標變項，須為連續變項；<math>Q_i</math> 為樣本的共變項（可有多個），可為類別變項或連續變項。</li> <li>• <math>e_{1i}</math>、<math>e_{2i}</math> 及 <math>e_{3i}</math> = 個體在各模型下之殘差（residuals）</li> <li>• <math>n</math> = 樣本數</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. 其他中介模型之效果量指標，可參考 MacKinnon（2008）與 Preacher 與 Kelley（2011）。</li> <li>2. 報告與解釋單一中介模型之範例請參見 Yzerbyt、Muller、Bataillier 及 Judd（2018）與 Pek 與 Flora（2018）。</li> <li>3. 關於結合傳統中介分析與因果推論（causal inference）的因果中介模型（causal mediation model）之介紹，讀者可參考 Imai、Keele 及 Tingley（2010）以及 Nguyen、Schmid 及 Stuart（2020）。</li> <li>4. Baron 與 Kenny（1986）認為獨變項與依變項的關聯性需達統計顯著（即公式 1 的係數 <math>c</math>）才可進一步執行中介分析。近期學者則認為獨變項與依變項的關聯性未達統計顯著亦可進行中介分析（Imai, Keele, &amp; Yamamoto, 2010; Loeys, Moerkerke, &amp; Vansteelandt, 2015）。</li> <li>5. 早期 Sobel（1982, 1986）提出的間接效果統計檢定，在近期的模擬研究中顯示，Sobel 的方法統計檢定力低，且其信賴區間亦有參數覆蓋率偏低的問題，因此不推薦使用（Biesanz, Falk, &amp; Savalei, 2010; Hayes &amp; Scharkow, 2013; MacKinnon, Lockwood, Hoffman, West, &amp; Sheets, 2002; MacKinnon, Lockwood, &amp; Williams, 2004）。</li> <li>6. 因為偏差校正拔靴法（bias-corrected bootstrap，以下簡稱 BC 法）及偏差加速校正拔靴法（bias-corrected-accelerated bootstrap，以下簡稱 BCa 法）的統計檢定力高且不需要常態假設，在間接效果的統計檢定與計算信賴區間時，以往不少學者推薦採用此兩方法（Hayes &amp; Scharkow, 2013; MacKinnon et al., 2004; Mallinckrodt, Abraham, Wei, &amp; Russell, 2006; Preacher &amp; Hayes, 2008; Shrout &amp; Bolger, 2002）。然而，近期研究發現此兩方法其第一型錯誤率過高（Biesanz et al., 2010; Fritz, Taylor, &amp; MacKinnon, 2012; Hayes &amp; Scharkow, 2013; Yzerbyt et al., 2018）。此外，兩方法所估計的信賴區間在小樣本時的參數覆蓋率偏低（Biesanz et al., 2010; Hayes &amp; Scharkow, 2013）。</li> <li>7. 近期研究推薦以聯合顯著檢定（joint significance test，以下簡稱 JS 檢定）來檢定間接效果（Biesanz et al., 2010; Fritz, Taylor, &amp; MacKinnon, 2012; Loeys et al., 2015; MacKinnon et al., 2002; Yzerbyt et al., 2018）。Biesanz 等人（2010）更發現 JS 檢定即便在資料有遺漏值或違反常態分配的情況下，仍具穩健性。然而，該方法的侷限為無法提供間接效果的信賴區間，故仍須透過其他方法計算（Hayes &amp; Scharkow, 2013; MacKinnon et al., 2002）。例如：Biesanz 等人（2010）推薦報告百分位數拔靴法（percentile bootstrap，以下簡稱 PB 法）的信賴區間，PB 法也較無 BC 及 BCa 法的問題（見第 6 點），而且在有遺漏值或違反常態分配的情況下仍具穩健性。Hayes 與 Scharkow（2013）亦認為當樣本數偏小時，可報告 PB 法的信賴區間。Yzerbyt 等人（2018）則建議應同時報告 JS 檢定和蒙地卡羅（Monte Carlo）法的信賴區間。</li> </ol>



表 1b  $r$  類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計
<b>Cramér's <math>V</math></b> (Cramér, 1946) $V = \sqrt{\frac{\chi^2_{Pop}}{N \times \min(r-1, c-1)}}$ <ul style="list-style-type: none"> <li><math>\chi^2_{Pop}</math> = 母群卡方獨立性檢定統計量</li> <li><math>N</math> = 母群列聯表所有細格的總次數</li> <li><math>\min(r-1, c-1)</math> = 取 <math>r \times c</math> 列聯表裡列數減 1 與行數減 1 的最小值</li> </ul>	<b>Cramér's <math>\hat{V}</math></b> (Cramér, 1946) $\hat{V} = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$ <ul style="list-style-type: none"> <li><math>\chi^2</math> = 樣本卡方獨立性檢定統計量</li> <li><math>n</math> = 樣本列聯表所有細格的總次數</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>Cramér's <math>\hat{V}</math> 的值與自由度有關，當自由度為 1 時，Cohen 對 <math>\hat{V}</math> 解釋的準則 (Cohen, 1988, pp. 224-227)：小效果量 = .10、中效果量 = .30、大效果量 = .50。</li> <li>此估計值即使在大樣本的情況下仍有偏差，而且估計偏差和列聯表的大小以及樣本數有關，因此 <math>\hat{V}</math> 難以用來比較不同列聯表的關聯強度 (Bergsma, 2013)。</li> <li>Pek 與 Flora (2018) 認為在列聯表分析中報告 Cramér's <math>\hat{V}</math> 通常無法直接回答研究者背後所關心的研究問題。</li> <li>信賴區間計算方式可參考 Odgaard 與 Fowler (2010)。</li> </ol>

表 1c 其他類型的效果量：母群效果量與其樣本估計

母群效果量	樣本估計
<b>勝算比 (odds ratio, <math>OR_{Pop}</math>)</b> 一成員在其母群中被歸於某一類別的勝算 (odds)，是另一成員在其母群中被歸於同一類別之勝算的幾倍。例如在第一組母群裡，有 $A$ 個人成功、 $B$ 個人不成功，在第二組的母群裡，有 $C$ 個人成功、 $D$ 個人不成功。此時第一組母群相對於第二組母群，其成功的勝算比為： $OR_{Pop} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\frac{A}{A+B} / \frac{B}{A+B}}{\frac{C}{C+D} / \frac{D}{C+D}} = \frac{A/B}{C/D} = \frac{AD}{BC}$ <ul style="list-style-type: none"> <li><math>\pi_1</math> = 某事件（本例中的「成功」）在第一組母群發生之機率</li> <li><math>\pi_2</math> = 某事件（本例中的「成功」）在第二組母群發生之機率</li> </ul> <p>統計特性：</p> <p>勝算比介於 0 到正無窮大之間。勝算比 = 1 代表兩類別變項之間沒有關聯，勝算比的值離 1 越遠，代表關聯性越強。</p>	<b>勝算比 (odds ratio, OR)</b> 一成員在其樣本中被歸於某一類別的勝算，是另一成員在其樣本中被歸於同一類別之勝算的幾倍。例如在第一組樣本裡，有 $a$ 個人成功、 $b$ 個人不成功，在第二組樣本裡，有 $c$ 個人成功、 $d$ 個人不成功。此時第一組樣本相對於第二組樣本，其成功的勝算比為： $OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{\frac{a}{a+b} / \frac{b}{a+b}}{\frac{c}{c+d} / \frac{d}{c+d}} = \frac{a/b}{c/d} = \frac{ad}{bc}$ <ul style="list-style-type: none"> <li><math>p_1</math> = 某事件（本例中的「成功」）在第一組樣本發生之機率</li> <li><math>p_2</math> = 某事件（本例中的「成功」）在第二組樣本發生之機率</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>OR 對 <math>OR_{Pop}</math> 的估計是偏差的。為了矯正偏差可以使用校正公式，如 <math>OR_{adj} = \frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)}</math> (Grissom &amp; Kim, 2012)。對於矯正偏差的討論，讀者可以參考 Agresti (2012)、Greenland (2010) 以及 Subbiah 與 Srinivasan (2008)。</li> <li>報告 OR 與解釋列聯表分析之範例請參見 Pek 與 Flora (2018)。</li> <li>信賴區間計算方式可參考 Agresti (2012, 2018)。讀者亦可在 SAS 統計軟體的 PROC FREQ 程序中的 TABLES 指令界定 OR 選項，來計算信賴區間。</li> </ol>

**SEM 適配度指標****1. 比較性適配指標 (comparative fit index, CFI)**

$$CFI_{Pop} = 1 - \frac{F_M}{F_b}$$

- $F_M$  = 假設模型適配於母群共變數矩陣的最小函數值
- $F_b$  = 基準模型 (baseline model) 適配於母群共變數矩陣的最小函數值，在此基準模型通常為觀察變項間是零相關的模式

**統計特性：**

CFI 數值介於 0 到 1 之間，數值越高代表假設模型，相較於基準模型，所增加的適配程度越多。

$$CFI = 1 - \frac{\max(T_M - df_M, 0)}{\max(T_b - df_b, T_M - df_M, 0)}$$

- $\max(T_M - df_M, 0)$  為取  $(T_M - df_M)$  與 0 間的最大數值
- $T_M = \hat{F}_M(N-1)$ ，其中  $\hat{F}_M$  為假設模型適配於樣本共變數矩陣的最小函數值。當假設模型為正確模型時， $T_M$  統計量服從中央卡方分配；當假設模型為誤設模型時， $T_M$  統計量服從非中央 (noncentral) 卡方分配 (Yuan, 2005)
- $df_M$  = 假設模型的自由度

(續下頁)

表 1c 其他類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計
	<ul style="list-style-type: none"> <li>• <math>\max(T_b - df_b, T_M - df_M, 0)</math> 為取 <math>(T_b - df_b)</math>、<math>(T_M - df_M)</math> 與 0 間的最大數值</li> <li>• <math>T_b</math> = 基準模型適配於樣本共變數矩陣的最小函數值</li> <li>• <math>df_b</math> = 基準模型之自由度。</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. CFI 指標是在測量相較於基準模型，假設模型適配度的增加程度。</li> <li>2. Hu 與 Bentler (1999) 建議 CFI 判斷模型適配度決斷值為 .95。</li> <li>3. CFI 不易受到樣本數影響模型誤設的程度 (Ainur, Sayang, Jannoo, &amp; Yap, 2017; Fan, Thompson, &amp; Wang, 1999)。</li> <li>4. Hu 與 Bentler (1999) 建議採用多個適配度指標組合共同判斷模型適配程度，而非仰賴單一適配度指標數值來解釋。</li> </ol>
II. 比較性適配裡的 Tucker-Lewis index (TLI)	
$TLI_{Pop} = 1 - \frac{\frac{F_M}{df_M}}{\frac{F_b}{df_b}} = 1 - \frac{F_M}{F_b} \times \frac{df_b}{df_M}$	$TLI = \frac{\frac{T_b}{df_b} - \frac{T_M}{df_M}}{\frac{T_b}{df_b} - 1} = 1 - \frac{\frac{T_M - df_M}{T_b - df_b}}{\frac{df_b}{df_M}} = 1 - \frac{T_M - df_M}{T_b - df_b} \times \frac{df_b}{df_M}$
<ul style="list-style-type: none"> <li>• <math>F_M</math> = 假設模型適配於母群共變數矩陣的最小函數值</li> <li>• <math>F_b</math> = 基準模型適配於母群共變數矩陣的最小函數值，在此基準模型通常為觀察變項間是零相關的模型</li> <li>• <math>df_M</math> = 假設模型之自由度</li> <li>• <math>df_b</math> = 基準模型之自由度</li> </ul> <p>統計特性：</p> <p>TLI 數值介於 0 到 1 之間，數值越高代表假設模型，相較於基準模型，所增加的適配程度越多。</p>	<ul style="list-style-type: none"> <li>• <math>T_M</math>、<math>df_M</math>、<math>T_b</math>、<math>df_b</math> 與上欄位裡的 CFI 相同</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. 同 CFI，TLI 指標測量，相較於基準模型，假設模型適配度的增加程度。TLI 數值多介於 0 到 1 之間，但在部分情況下可能超過 1。</li> <li>2. Hu 與 Bentler (1999) 建議 TLI 判斷模型適配度決斷值為 .95。</li> <li>3. TLI 數值大多較 CFI 低，隨觀察變項個數的增加，這兩個值會越接近。</li> <li>4. TLI 不易受到樣本數影響模型誤設的程度 (Ainur et al., 2017; Balderjahn, 1988; Fan et al., 1999; Marsh, Balla, &amp; McDonald, 1988)。</li> </ol>
III. 均方根誤差的近似指標 (root mean square error of approximation, RMSEA)	
$RMSEA_{Pop} = \sqrt{\frac{F_M}{df_M}}$	$RMSEA = \max\left(\sqrt{\frac{T_M - df_M}{df_M(n-1)}}, 0\right)$
<ul style="list-style-type: none"> <li>• <math>F_M</math> = 假設模型適配於母群共變數矩陣的最小函數值</li> <li>• <math>df_M</math> = 假設模型之自由度</li> </ul> <p>統計特性：</p> <p>RMSEA 數值介於 0 到無窮大之間，數值越高表示模型的誤設程度越高。</p>	<ul style="list-style-type: none"> <li>• <math>T_M</math> 與上述的 CFI 相同</li> <li>• <math>df_M</math> = 假設模型之自由度</li> <li>• <math>n</math> = 樣本數</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. RMSEA 衡量在每個自由度下，假設模型與真實模型之間的近似程度 (Preacher, Zhang, Kim, &amp; Mels, 2013)。</li> <li>2. 根據 Browne 與 Cudeck (1993) 的建議，RMSEA 數值為 .10 時，表示模型不值得考慮；數值小於 .05 則表示高度適配 (close fit)。</li> </ol>
IV. 標準化均方根殘差 (standardized root mean square residual, SRMR)	
$SRMR_{Pop} = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^i \left( \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}} \sqrt{\hat{\sigma}_{jj}}} \right)^2}{p(p+1)/2}}$	$SRMR = \sqrt{\frac{\sum_{i=1}^p \sum_{j=1}^i \left( \frac{s_{ij}}{\sqrt{s_{ii}} \sqrt{s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}} \sqrt{\hat{\sigma}_{jj}}} \right)^2}{p(p+1)/2}}$
<ul style="list-style-type: none"> <li>• <math>\sigma_{ij}</math> = 母群的觀察變項間之共變數</li> <li>• <math>\hat{\sigma}_{ij}</math> = 假設模型下的觀察變項間之共變數</li> <li>• <math>\sigma_{ii}</math> 與 <math>\sigma_{jj}</math> = 母群的觀察變項之標準差</li> <li>• <math>\hat{\sigma}_{ii}</math> 與 <math>\hat{\sigma}_{jj}</math> = 假設模型下的觀察變項之標準差</li> <li>• <math>p</math> = 變項個數</li> </ul> <p>統計特性：</p> <p>SRMR 的最小值為 0，數值越小代表適配程度越好。</p>	<ul style="list-style-type: none"> <li>• <math>s_{ij}</math> = 樣本的觀察變項間之共變數</li> <li>• <math>\hat{\sigma}_{ij}</math> = 假設模型下的觀察變項間之共變數</li> <li>• <math>s_{ii}</math> 與 <math>s_{jj}</math> = 樣本的觀察變項之標準差</li> <li>• <math>\hat{\sigma}_{ii}</math> 與 <math>\hat{\sigma}_{jj}</math> = 假設模型下的觀察變項之標準差</li> <li>• <math>p</math> = 變項個數</li> </ul> <p>注意事項：</p> <ol style="list-style-type: none"> <li>1. SRMR 根據資料共變數與假設模型下共變數的標準化差值來量化假設模型的適配程度。</li> <li>2. Hu 與 Bentler (1999) 建議 SRMR 判斷模型適配度決斷值為 .08。</li> </ol>

(續下頁)

表 1c 其他類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計												
邏輯斯迴歸中的類決定係數（pseudo-R <sup>2</sup> ） （文獻中無對應之母群參數）	<div><math>R_L^2</math>（McFadden, 1973）</div> <div><math display="block">R_L^2 = 1 - \frac{\ln L_M}{\ln L_0} = \frac{\ln L_0 - \ln L_M}{\ln L_0} = \frac{G_M}{-2\ln L_0}</math></div> <div><ul style="list-style-type: none"><li>• <math>L_0</math> = 虛無模型（僅包含截距之模型）的概似函數</li><li>• <math>L_M</math> = 預測模型 <math>M</math>（包含截距及其他自變項）的概似函數</li><li>• <math>G_M = -2(\ln L_0 - \ln L_M)</math>，為模型 <math>M</math> 之卡方檢定統計量</li></ul></div> <div>統計特性：</div> <div><math>R_L^2</math> 之理論值介於 0 至 1 之間。若模型 <math>M</math> 能完美預測，則 <math>\ln L_M = 0</math>，此時 <math>R_L^2 = 1</math>。反之，當模型 <math>M</math> 的表現和虛無模型一樣差時，則 <math>R_L^2 = 0</math>（Menard, 2000）。</div> <div>注意事項：</div> <div><math>R_L^2</math> 在概念上和線性迴歸的決定係數 <math>R^2</math> 相似，因此，<math>R_L^2</math> 可理解為預測模型能降低 <math>\ln L_0</math> 之百分比（Menard, 2000, 2002）。由於 <math>R_L^2</math> 不易受基準發生率（base rate，即二分依變項的其中一個類別佔整體樣本之比例）的影響，其直覺解釋與 <math>R^2</math> 相仿，而且此指標和邏輯斯迴歸中最大化概似函數的目標相同，因此 Menard（2000）推薦報告 <math>R_L^2</math>。</div>												
	<div><math>R_M^2</math>（Cox &amp; Snell, 1989; Maddala, 1983）</div> <div><math display="block">R_M^2 = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}} = 1 - e^{-\frac{2}{n}(\ln L_M - \ln L_0)}</math></div> <div><ul style="list-style-type: none"><li>• <math>L_0</math> = 虛無模型（僅包含截距之模型）的概似函數</li><li>• <math>L_M</math> = 預測模型 <math>M</math>（包含截距及其他自變項）的概似函數</li><li>• <math>n</math> = 樣本數</li></ul></div> <div>注意事項：</div> <div>即便模型完美適配，<math>R_M^2</math> 仍會小於 1，此特性違反直覺（Nagelkerke, 1991）。<math>R_M^2</math> 受基準發生率與樣本數影響（Menard, 2000, 2002; Smith &amp; McKenna, 2013）。因此，僅能用來比較不同模型從同一筆資料所導出的 <math>R_M^2</math>。不可用來比較同一模型從不同的樣本，或同一樣本但切割成不同的子樣本，所導出的 <math>R_M^2</math>（Menard, 2000）。</div>												
	<div><math>R_N^2</math>（Cragg &amp; Uhler, 1970; Nagelkerke, 1991）</div> <div><math display="block">R_N^2 = \frac{R_M^2}{\max R_M^2} = \frac{1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{n}}}{1 - (L_0)^{\frac{2}{n}}}</math></div> <div><ul style="list-style-type: none"><li>• <math>R_M^2</math> = 上一欄位的 <math>R_M^2</math>（Cox &amp; Snell, 1989; Maddala, 1983）</li><li>• <math>\max R_M^2 = R_M^2</math> 之最大值</li><li>• <math>n</math> = 樣本數</li></ul></div> <div>注意事項：</div> <div>有鑑於 <math>R_M^2</math> 之最大值會大於 1，Nagelkerke（1991）提出 <math>R_N^2</math> 作為校正，將 <math>R_M^2</math> 除以 <math>R_M^2</math> 的最大值，使 <math>R_N^2</math> 之最大值為 1。如同 <math>R_M^2</math>，<math>R_N^2</math> 易受基準發生率與樣本數影響，因此在進行模型比較時應特別注意（Menard, 2000, 2002）。</div>												
邏輯斯迴歸中的分類正確率（correct classification rate） （無對應之母群參數）	<div>敏感度（sensitivity）、特異度（specificity）及整體分類正確率</div> <div>邏輯斯迴歸的結果可用一個分類表格來表示，例如：某邏輯斯迴歸模型被用來預測大學生是否能順利畢業，順利畢業與否為一個二分依變項（0 = 未畢業；1 = 畢業）。</div> <table><tr><td></td><td>預測 0（估計機率小於切截點）</td><td>預測 1（估計機率等於或大於切截點）</td><td></td></tr><tr><td>實際 0</td><td>甲</td><td>乙</td><td>甲 + 乙</td></tr><tr><td>實際 1</td><td>丙</td><td>丁</td><td>丙 + 丁</td></tr></table>		預測 0（估計機率小於切截點）	預測 1（估計機率等於或大於切截點）		實際 0	甲	乙	甲 + 乙	實際 1	丙	丁	丙 + 丁
	預測 0（估計機率小於切截點）	預測 1（估計機率等於或大於切截點）											
實際 0	甲	乙	甲 + 乙										
實際 1	丙	丁	丙 + 丁										

（續下頁）

表 1c 其他類型的效果量：母群效果量與其樣本估計（續）

母群效果量	樣本估計
	此分類表格的橫列為實際值，直欄則為根據邏輯斯迴歸的模型所估計的機率與所設定的切截點（如 .5）相比較所估計出的值。當估計機率等於或大於切截點時，二分依變項的估計值等於 1（預測為畢業）；反之，當估計機率小於切截點時，二分依變項的估計值等於 0（預測為未畢業）。細格內的甲、乙、丙、丁則為個數。敏感度（sensitivity）、特異度（specificity）及整體分類正確率可用來作為模型好壞的指標。敏感度指的是正確預測畢業的比率 = 丁 / (丙 + 丁)。特異度指的是正確預測未畢業比率 = 甲 / (甲 + 乙)。整體分類正確率 = (甲 + 丁) / (甲 + 乙 + 丙 + 丁)。
	注意事項： 當分類是邏輯斯迴歸分析的目的時，分類表格的製作與解釋是適切的。反之，如果分類不是邏輯斯迴歸分析的目的，則分類表格可作為評估模型適配度的輔助工具（Hosmer & Lemeshow, 2000）。分類正確率跟模型的適配度所導出的結論可能一致，也可能不一致（Hosmer & Lemeshow, 2000）。分類的結果和切截點以及兩組的大小有關，人數多的組會比人數少的組有更好的分類結果（Hosmer & Lemeshow, 2000）。

ROC 曲線下面積 (area under the receiver operating characteristic curve, AUC)

$$AUC_{Pop} = P(X_1 > X_2) + .5 \times P(X_1 = X_2)$$

當隨機從第一組母群與第二組母群各選出一人時，第一組母群個體的分數（ $X_1$ ）比第二組母群個體的分數（ $X_2$ ）高的機率（Kraemer & Kupfer, 2006, p. 993）

- 當  $AUC_{Pop} = .5$  時，第一組跟第二組母群個體的分數高下是隨機的，亦即兩母群分配完全重疊
- 當  $AUC_{Pop} = 1$  時，所有第一組母群個體的分數都比第二組的分數高，亦即兩母群分配完全不重疊

ROC 曲線是根據敏感度與（1 - 特異度）在所有可能的切截點上所繪製的圖，該曲線下面積（AUC）可用來檢驗一個統計模型是否能區辨出個體是否經歷過某個事件。

$$AUC = \frac{\#(X_1 > X_2) + .5 \times \#(X_1 = X_2)}{n_1 \times n_2}$$

- $\#(X_1 > X_2)$  = 在樣本資料中，第一組個體的分數（ $X_1$ ）比第二組個體的分數（ $X_2$ ）高之次數
- $\#(X_1 = X_2)$  = 在樣本資料中，第一組個體的分數（ $X_1$ ）等於第二組個體的分數（ $X_2$ ）之次數
- $n_1$  = 第一組的樣本數
- $n_2$  = 第二組的樣本數

注意事項：  
AUC 又稱為  $A$  測量（Delaney & Vargha, 2002）。 $A$  測量可以估計兩個分配的重疊程度，不需符合常態分配或變異數同質的假設，也不受分數單調轉換的影響（invariant to monotonic transformation of data; Peng & Chen, 2014）。此指標也可以從 Mann-Whitney  $U$  求得，其公式為  $AUC = \frac{U}{n_1 n_2}$ （Kraemer & Kupfer, 2006）。

很小的可能）（AERA, 2006, p. 37）。

- 作者在詮釋效果量時，需要簡述該效果量在實務上和理論上的意義，並與過去類似的研究所報告的效果量相比較，以助讀者評估研究結果橫跨不同樣本、設計與分析之穩定性（Wilkinson & the Task Force on Statistical Inference, 1999, p. 599）。在合理的情況下，作者可將效果量的詮釋推廣至母群（Wilkinson & the Task Force on Statistical Inference, 1999, p. 602）。
- 作者論述研究結果的實務意涵時（study implications），除了應解釋其在理論上、臨床上及實務上的顯著性之外，亦須考量解釋的依據及此研究的情境脈絡（如：實驗組和控制組參與者的特性，實驗組和控制組所經歷的事件）（APA, 2010, p. 36; APA, 2020, p. 90）。

（五）美國心理與教育期刊報告效果量之回顧研究

Kirk（1996）為最早對美國心理學期刊報告效果量之回顧研究，該文回顧了 391 篇在 1995 年發表於四本 APA 期刊之論文，發現四本期刊報告效果量的平均比例為 47.6%。自從 APA 於 1999 年發表專案報告（Wilkinson & the Task Force on Statistical Inference, 1999）後，針對心理與教育各次領域報告效果量之回顧研究如雨後春筍般地發表。這些回顧研究關切的角度不盡相同，有些關注的是單一研究主題的效果量報告情形（如：Zientek, Capraro, & Capraro, 2008），另一些則是針對單一期刊（如：Fritz, Morris, & Richler, 2012）、

特定次領域的少數期刊（如：Matthews et al., 2008），或不同的統計分析方法（如：Alhija & Levy, 2009）進行回顧。近期與本研究最相關的三篇論文為：Alhija 與 Levy（2009），Sun、Pan 及 Wang（2010）及 Peng 等人（2013）。

Alhija 與 Levy（2009）回顧了 99 篇於 2003 與 2004 年發表在 10 本期刊的文章，並將期刊分為兩類：有明確要求作者報告效果量與沒有明確要求報告效果量，數量各為五本。該研究以統計分析作為分析的單位（unit of analysis），檢驗了 183 個統計分析。結果發現，這兩類期刊在報告效果量的比例上沒有統計顯著差異。當作者採用相關分析與迴歸分析時，一定會報告效果量；當作者採用變異數分析時，報告效果量的比例也相當高；然而當作者採用  $t$  檢定或是卡方（ $\chi^2$ ）檢定時，則報告效果量的比例在此兩類期刊裡皆低於 50%。就效果量的類型而言，當作者採用  $t$  檢定時，大多數作者會報告 Cohen's  $d$ ；若採用變異數分析時，則多數作者會報告  $\eta^2$  或淨  $\eta^2$ ；若採用迴歸分析時，則多數作者會報告  $R^2$ 、 $\Delta R^2$ 、 $\beta$ 、 $\Delta\beta$ （迴歸係數之改變量）。在效果量的解釋上，56% 的作者在報告效果量時，會同時對效果量的意涵提出標籤化的解釋，亦即根據某文獻既定的標準（如：Cohen, 1988），直接將效果量解釋為小、中或大的效果。當作者採用  $t$  檢定或迴歸分析時，其效果量解釋的比例高於其他統計方法。除此之外，Alhija 與 Levy 也發現當統計顯著性與效果量的強度有落差時，多數作者並未討論或解釋此落差。然而，當作者使用  $t$  檢定而且期刊未規定要報告效果量時，則作者必定解釋此落差。

Sun 等人（2010）回顧了 1,243 篇從 2005 年到 2007 年發表在 14 本期刊的文章，並把期刊分為三類：APA 的期刊、AERA 的期刊，以及不隸屬於任何學會的獨立期刊，數量分別為六本、兩本、六本。此研究以文章作為分析的單位，探討期刊所屬的學會在四個效果量面向上有何不同，這四個面向分別為：報告效果量的比例、解釋效果量的比例、統計顯著性與效果量強度有落差的比率，以及作者對前述落差提供解釋的比例。此外，Sun 等人也檢驗了報告效果量與解釋效果量的比例是否與統計方法有關。結果顯示，49% 的文章報告至少一個效果量，三類期刊在報告效果量的比例上，有統計顯著差異：AERA 期刊（73%）與獨立期刊（68%）報告效果量的比例皆高於 APA 期刊（40%）。就統計方法而言，效果量出現在複雜模型（如：SEM）裡的比例比在簡單模型裡（如：變異數分析、 $t$  檢定）高。

針對解釋效果量的比例，Sun 等人（2010）發現 57% 報告效果量的文章也解釋了效果量，且三類期刊在比例上有統計顯著差異：AERA 的期刊與獨立期刊文章解釋效果量的比例皆比 APA 的期刊高出至少 10%。就統計方法而言，作者解釋複雜模型效果量的比例高於其他統計方法。在同時報告統計顯著性與效果量的文章裡，11% 的文章呈現兩者間有落差，三類期刊在呈現落差的比率上沒有統計顯著差異。最後，當統計顯著性與效果量強度有落差時，30% 之文章對落差提出解釋，三類期刊解釋落差的比率有統計顯著差異：獨立期刊的比例（48%）遠高於 APA（22%）或 AERA（12%）的期刊。

Peng 等人（2013）亦以文章為分析的單位，回顧了 451 篇從 2009 年到 2010 年發表在 12 本期刊的文章，並把這 12 本期刊分為三類：APA 的期刊、AERA 的期刊及隸屬於其他學會的期刊，數量分別為兩本、兩本、八本。結果顯示，整體的效果量報告比例為 65%，其中 APA 的期刊文章報告效果量的比例為 73%、AERA 為 75%，其他學會則為 54%。Peng 等人也指出，即便是 APA 或 AERA 的期刊，其報告效果量的比例也有明顯差異。就統計方法而言，因為統計軟體會自動計算迴歸分析的效果量（如： $R^2$  或  $\Delta R^2$ ），這些效果量在文章裡出現的比例很高。如同 Sun 等人（2010）的結果，Peng 等人也指出，效果量出現在複雜模型裡的比例比在簡單模型裡高。

在效果量的解釋上，Peng 等人（2013）的結果顯示效果量的解釋和該效果量所對應的統計方法有關。當作者檢驗變項間的相關係數、執行迴歸分析、 $t$  檢定，或使用複雜的統計模型時，解釋效果量的比例比使用其他統計方法更高（Alhija & Levy, 2009; Matthews et al., 2008; Smith & Honoré, 2008; Sun et al., 2010; Zientek et al., 2008）。如同 Alhija 與 Levy（2009），Peng 等人的結果也顯示，作者常用標籤化的方式來解釋效果量；此外，如同 Alhija 與 Levy（2009）與 Sun 等人（2010）的結果，多數作者並未解釋統計顯著性與效果量強度的落差。

Peng 等人（2013）除了回顧 451 篇發表在 12 本期刊的量化實徵文章外，也對 32 篇效果量回顧的文章進行二度回顧，旨在瞭解 APA/AERA 效果量報告準則對美國心理學與教育學期刊所造成的影響。此研究以 1999 年為分界，比較該年前與該年後效果量報告的狀況，結果發現自 1999 年起，效果量報告的比例明顯地增加，各期刊的比例均超過 50%；且不論統計檢定是

否達顯著，作者都會報告效果量，報告效果量之信賴區間與從實務和臨床的角度來解釋效果量的比例也增加了（Peng et al., 2013）。此外，1999 年前發表的文章多數未區分效果量的類型（如：Plucker, 1997），然而 1999 年後發表的文章，無論是個別研究或是回顧研究，都傾向描述或定義所報告的效果量。隨著年代的推進，也出現更多樣的效果量類型，如：SEM 適配度指標（見 Peng et al., 2013 的附錄表 A 與表 B），和新的效果量分類系統（Andersen, McCullagh, & Wilson, 2007; Dunleavy, Barr, Glenn, & Miller, 2006）。Peng 等人認為造成這些現象的原因與期刊編審辦法、APA/AERA 效果量報告準則之公佈、統計軟體的支援，以及特定的統計方法等有關。

## （六）研究目的

過去研究回顧了美國心理學與教育學期刊報告效果量的狀況，然而尚未有研究探討效果量在臺灣心理學與教育學門裡應用的現況。為彌補文獻之不足，本研究參考 Alhija 與 Levy（2009）、Sun 等人（2010）及 Peng 等人（2013）的研究，針對報告效果量的四個面向，提出下列的研究問題：

### 1. 效果量的報告

臺灣心理學與教育學期刊文章報告效果量的比例為何？效果量報告的比例是否與學門有關？

### 2. 效果量的類型

哪些類型的效果量出現在臺灣心理學與教育學的期刊裡？效果量的類型是否與學門有關？

### 3. 效果量的解釋

臺灣心理學與教育學的期刊文章解釋效果量的比例為何？解釋效果量的比例是否與學門有關？兩學門在三種效果量的解釋方式上之比例為何？效果量的解釋方式是否與學門有關？

### 4. 統計顯著性與效果量的落差

統計顯著性與效果量有落差的比率為何？統計顯著性與效果量有落差的比率是否與學門有關？當統計顯著性與效果量有落差時，作者是否針對落差做解釋？對落差做解釋是否與學門有關？

雖然心理學與教育學在許多研究議題上的確是環環相扣而且相輔相成，然兩學門的訴求並不相同（Egan, 2012），且過去回顧國外期刊的研究（Peng et al., 2013; Sun et al., 2010）亦發現此兩學門在效果量報告上有差異。有鑑於此，本文在回顧臺灣期刊時，亦將這兩學門的比較納入本研究關注的議題之一。

## 研究方法

### （一）期刊篩選

本研究系統性地回顧臺灣心理學與教育學界中公認為頂尖的 17 本期刊，包含八本心理學門的期刊與九本教育學門的期刊。心理學期刊包括《中華心理學刊》、《中華心理衛生學刊》、《中華輔導與諮商學報》、《本土心理學研究》、《教育心理學報》、《教育與心理研究》、《臺灣精神醫學》及《應用心理研究》。教育學期刊則涵蓋《特殊教育研究學刊》、《教育研究集刊》、《教育研究與發展期刊》、《教育政策論壇》、《教育科學研究期刊》、《教育資料與圖書館學》、《當代教育研究季刊》、《課程與教學季刊》及《臺灣教育社會學研究》。期刊之篩選與分類乃綜合考量 2017 年的「臺灣人文及社會科學期刊評比暨核心期刊收錄」（TSSCI）名單（<http://www.hss.ntu.edu.tw/model.aspx?no=354>），翁儷禎、黃怡蓉及鄭中平（2012）對臺灣心理學期刊評比之結果，以及黃毅志（2009）對臺灣教育學期刊評比之結果。

根據 2017 年 TSSCI 之名單，本研究所回顧的九本教育學期刊皆被歸類為第一級的教育學期刊，四本心理學期刊（《中華心理學刊》、《中華輔導與諮商學報》、《本土心理學研究》、《教育心理學報》）亦被歸類為第一級的心理學期刊，《應用心理研究》則被歸為第三級的心理學期刊。《中華心理衛生學刊》與《教育與心理研究》被歸為第二級的綜合類期刊。雖然《臺灣精神醫學》曾收錄在 2012 年 TSSCI 名單中，卻未收錄在 2017 年 TSSCI 名單內。翁儷禎等人（2012）根據 2006 年至 2008 年的資料，將心理學期刊分為「優」與「良」兩級，其中《中華心理學刊》、《中華心理衛生學刊》、《中華輔導與諮商學報》、《本土心理學研究》、《教育心理學報》、《教育與心理研究》及《臺灣精神醫學》的評比為「優」級，《應用心理研究》的評比為「良」級。黃毅志（2009）根據 2003 年至 2007 年的資料，將教育學期刊分為 4 級，其中《教育研究集

刊》、《教育心理學報》、《教育與心理研究》、《臺灣教育社會學研究》、《特殊教育研究學刊》及《教育科學研究期刊》的評比等級為「1」；《中華輔導與諮商學報》、《教育政策論壇》、《課程與教學季刊》、《教育研究與發展期刊》及《當代教育研究季刊》的評比等級為「2」。TSSCI 與翁儷禎等人（2012）皆把《中華輔導與諮商學報》、《教育心理學報》及《教育與心理研究》歸為心理學門的期刊，因此，本研究也把這本期刊歸於心理學門。

每一篇文章結果的呈現方式，會與該期刊所規定的撰寫格式及投稿須知有關，因此本研究也檢視期刊對此部分之要求。在 17 本期刊中，《特殊教育研究學刊》、《臺灣教育社會學研究》並未在期刊網站上提及作者須遵守的任一學會之發表手冊。《教育資料與圖書館學》要求作者遵守美國心理學會 Author-date 格式（APA format）或芝加哥 Note 格式（Chicago-Turabian Style）。《臺灣精神醫學》則要求作者遵循「Uniform Requirements for Manuscripts Submitted to Biomedical Journals」之規定（International Committee of Medical Journal Editors, 1997），該規定明確指出作者報告量化結果時須符合的原則，強調報告信賴區間或測量誤差的重要性，並提醒作者勿全然依賴統計檢定的結果，如  $p$  值，來表達研究結果之重要資訊（International Committee of Medical Journal Editors, 1997）。除此之外，《臺灣精神醫學》亦額外自訂有關統計結果呈現的規定，然而效果量一詞並未出現在此規定中。其餘 13 本期刊在本研究所回顧的年間，都要求作者依據美國心理學會發表手冊第六版（APA, 2010）的格式撰寫文章或引用文獻。

綜上所述，除《臺灣精神醫學》之外，其他 16 本期刊對統計結果的呈現方式均無特別的要求。有關這 17 本期刊之簡介、期刊所屬學會、期刊評比、收錄資料庫、撰寫格式及投稿須知，請參見本文的補充材料。

## （二）文章篩選

由於報告效果量指標之目的為反映量化統計結果之實徵意涵，因此本研究僅回顧 17 本期刊在 2017 與 2018 年間所刊登的實徵研究類文章，而且文章必須使用量化資料，並進行至少一種統計分析以回答其主要研究問題。若文章僅是論述或文獻回顧、僅用質化分析、或評論他人的文章，則不納入分析。此外，由於整合分

析缺乏合適的效果量定義，亦不納入；蒙地卡羅模擬研究（Monte Carlo simulation study）中的效果量僅有理論意涵而未有實徵意義，因此亦不納入。最後，以量表編製為其主要研究目的之文章亦被排除，因為這類文章使用效果量的方式與一般實徵研究不同。經由這一系列的篩選，本研究最後共回顧了 268 篇文章（回顧文章的引文，請參見本文的補充材料），各期刊之文章所使用的研究方法及其數量如表 2 所示。

## （三）編碼程序

本研究以文章為分析單位，每一位研究員皆負責回顧至少一本期刊，其於 2017 年與 2018 年刊登的所有量化實徵文章，並根據標準化的編碼準則進行編碼。編碼準則（coding scheme）如附錄所示，主要針對本研究所探討的四個面向收集資訊：是否報告主要研究結果的效果量？若文章報告一個以上主要研究結果的效果量，則接續以下編碼：作者報告了哪些效果量？作者是否解釋了效果量？作者若解釋了效果量，用哪一種方式解釋效果量？統計顯著性與效果量之間是否有落差？若有落差，則作者是否解釋了落差？

為確保所有研究員對編碼準則達成共識，研究員在文章編碼階段，進行定期討論，以便將編碼程序標準化，討論結果皆儲存於共享的雲端資料庫，供所有研究員參考。此外，每一期刊編碼完成後，會有另一位研究員隨機挑選出該期刊至少 30% 的文章進行編碼檢驗，若有差異或歧見，兩研究員會相互討論，直到達成一致共識才將最終的編碼定案。

## （四）資料分析

由於 2017 年與 2018 年的資料無顯著差異，本研究整合了各期刊在這兩年的數據一併分析，以便瞭解效果量在心理學與教育學門裡的應用現況。為了檢驗「效果量的報告」和「學門類別」是否相互獨立，本研究採用卡方檢定來進行獨立性檢定，所有卡方檢定的  $\alpha$  值皆設為 .05。此外，本研究計算了勝算比及其 95% 信賴區間作為效果量的指標，該指標表徵了兩類別變項間的關係強度（Fleiss, 1994），適用於回溯性研究（retrospective study）。上述分析皆以 SAS 9.4 之 PROC FREQ 程序進行。

表 2 各期刊量化實徵研究文章所使用的研究方法及其數量

期刊名稱	實驗設計	準實驗設計	調查法	次級資料分析	實驗 + 調查 <sup>a</sup>	其他	小計
心理學門期刊							
《中華心理學刊》	9	1	5	0	0	0	15
《中華心理衛生學刊》	0	4	8	0	0	0	12
《中華輔導與諮商學報》	4	0	5	2	0	0	11
《本土心理學研究》	3	2	2	0	1	0	8
《教育心理學報》	3	6	19	8	0	1	37
《教育與心理研究》	3	1	16	5	0	0	25
《臺灣精神醫學》	3	15	7	4	0	0	29
《應用心理研究》	1	1	2	1	0	0	5
教育學門期刊							
《特殊教育研究學刊》	0	9	2	2	0	1	14
《教育研究集刊》	0	0	0	5	0	0	5
《教育研究與發展期刊》	0	3	5	1	0	0	9
《教育政策論壇》	0	0	17	4	0	0	21
《教育科學研究期刊》	2	7	16	10	0	0	35
《教育資料與圖書館學》	2	0	2	3	0	3	10
《當代教育研究季刊》	0	0	7	3	0	0	10
《課程與教學季刊》	1	7	12	0	0	0	20
《臺灣教育社會學研究》	0	0	1	1	0	0	2
小計	31	56	126	49	1	5	268

註：此表整合了 2017 年與 2018 年的文章。

<sup>a</sup> 這類文章同時採用了實驗設計與調查法來回答其研究問題（參見本文的補充材料）。

## 研究結果

研究結果分成五個部分，第一至第四部分為效果量的四個面向，分別為：效果量的報告、效果量的類型、效果量的解釋，以及統計顯著性與效果量的落差；第五部分則總結回顧的結果。

### （一）效果量的報告

針對面向 (1)，本文探討兩個研究問題，分別是：臺灣心理學與教育學期刊文章報告效果量的比例為何？效果量報告的比例是否與學門有關？

表 3 呈現在心理學門與教育學門中，量化實徵文章報告至少一個效果量之篇數與比例。如表 3 所示，量化

實徵文章在兩學門的總篇數為 268 篇，心理學與教育學門各為 142 篇與 126 篇。跨學門的效果量報告總比例為 72%，心理學期刊報告效果量之比例為 65%（92 篇），教育學期刊則為 79%（100 篇）。表 3 的卡方檢定結果達顯著（ $\chi^2(1, N = 268) = 6.98, p = .01$ ），因此，效果量報告的比例與學門之間獨立關係的虛無假設被推翻；換言之，兩學門在報告效果量的比例上有顯著差異。教育學期刊報告效果量的勝算（odds）是心理學期刊的 2.09 倍，而且在 95% 信心水準下，勝算比的最高估計值是 3.63，最低估計值是 1.20。

### （二）效果量的類型

針對面向 (2)，本文探討兩個研究問題，分別是：



表 3 效果量報告的比例與學門之間獨立關係的卡方檢定結果

學門	至少報告一個效果量	未報告效果量	文章總數 <sup>a</sup>	勝算 <sup>b</sup>	$\chi^2$	<i>df</i>	<i>p</i>
心理學	92 (65%)	50 (35%)	142 (100%)	1.84	6.98	1	.01
教育學	100 (79%)	26 (21%)	126 (100%)	3.85			
跨學門	192 (72%)	76 (28%)	268 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 為表 2 中所有量化實徵研究文章總數。<sup>b</sup> 為各學門期刊「至少報告一個效果量」的次數除以「未報告效果量」的次數。

哪些類型的效果量出現在臺灣心理學與教育學的期刊裡？效果量的類型是否與學門有關？

在報告效果量的 92 篇心理學及 100 篇教育學文章中，心理學平均報告效果量 1.42 次，教育學則為 1.87 次。對報告效果量之文章，本研究進一步分析其效果量的類型：*d* 類型、*r* 類型及其他類型（Kelley & Preacher, 2012; Kirk, 2005; Rosenthal, 1994），結果如表 4、表 5 所示。整體而言，不論在心理學門或教育學門裡，作者

最常報告的是 *r* 類型的效果量（67.0%），最不常報告的是 *d* 類型的效果量（9.1%）。如表 4 所示，心理學期刊報告 *d* 類型、*r* 類型、其他類型效果量之比例分別為 6.2%、67.7%、26.2%，各類型報告的平均次數分別為 0.09、0.97、0.37 次。教育學期刊報告 *d* 類型、*r* 類型、其他類型效果量之比例則分別為 11.2%、66.5%、22.3%，各類型報告的平均次數分別為 0.21、1.24、0.42 次。

表 4 臺灣心理學與教育學期刊報告之效果量類型、名稱及數量

效果量類型	心理學期刊		教育學期刊	
	數量	比例	數量	比例
<b><i>d</i> 類型</b>				
原始平均數差異	0	0.0%	10	5.3%
Cohen's <i>d</i>	8	6.2%	7	3.7%
Cohen's <i>q</i>	0	0.0%	1	0.5%
在階層線性模型中以學校間的平均差作為單位所算出的效果量（ES in between-school SD units in HLM）	0	0.0%	1	0.5%
百分比差	0	0.0%	2	1.1%
小計	<b>8</b>	<b>6.2%</b>	<b>21</b>	<b>11.2%</b>
<b><i>r</i> 類型</b>				
Spearman's rho	1	0.8%	1	0.5%
Pearson's <i>r</i>	14	10.8%	21	11.2%
$R^2$	24	18.5%	26	13.8%
$\Delta R^2$	10	7.7%	12	6.4%
Adjusted $R^2$	5	3.8%	10	5.3%
Negelkerke $R^2$	3	2.3%	2	1.1%
多層次（multilevel）模型中的類（pseudo）決定係數 $R^2$	0	0.0%	1	0.5%
事件歷史分析法（event history analysis）的類（pseudo）決定係數 $R^2$	0	0.0%	1	0.5%

（續下頁）

表 4 臺灣心理學與教育學期刊報告之效果量類型、名稱及數量（續）

效果量類型	心理學期刊		教育學期刊	
	數量	比例	數量	比例
$\hat{\eta}^2$ 或淨 $\hat{\eta}^2$ (partial $\hat{\eta}^2$ )	18	13.8%	14	7.4%
$\hat{\omega}^2$	1	0.8%	0	0.0%
組內相關係數	1	0.8%	2	1.1%
$\Gamma$	0	0.0%	1	0.5%
$\phi$	0	0.0%	1	0.5%
迴歸係數	0	0.0%	2	1.1%
中介分析裡的迴歸係數	8	6.2%	22	11.7%
Cramér's $\hat{V}$	1	0.8%	2	1.1%
列聯係數 (contingency coefficient)	0	0.0%	1	0.5%
Tau 非對稱關聯係數	0	0.0%	1	0.5%
在階層線性模型中可解釋的變異量比例 (% of variance explained in HLM)	1	0.8%	1	0.5%
在階層線性模型中，不同階層可解釋的變異量比例 (% of explained variance at different levels in HLM)	0	0.0%	3	1.6%
在 SEM 中所有變項對潛在依變項的可解釋的變異量比例 (% of total variance explained)	0	0.0%	1	0.5%
中介分析中可解釋的變異量比例 (% of variance accounted for 或 VAF)	1	0.8%	0	0.0%
小計	<b>88</b>	<b>67.7%</b>	<b>125</b>	<b>66.5%</b>
其他類型				
勝算比	3	2.3%	2	1.1%
對數勝算比 (log odds ratio)	1	0.8%	0	0.0%
風險比率 (relative risk)	1	0.8%	0	0.0%
SEM 適配度指標	23	17.7%	34	18.1%
邏輯斯迴歸的分類正確率	2	1.5%	1	0.5%
ROC 曲線下面積	1	0.8%	0	0.0%
潛在轉移模式 (latent transition analysis) 之熵 (entropy)	1	0.8%	0	0.0%
重疊率 (degree of overlap)	0	0.0%	1	0.5%
典型區辨分析 (canonical discriminant function analysis) 的正確分類比率	0	0.0%	1	0.5%
Moran's I	0	0.0%	1	0.5%
殘差變異量	0	0.0%	1	0.5%
決策樹 (decision tree) 中的 Kappa statistics、F-measures 與 ROC 曲線下面積	0	0.0%	1	0.5%
未明確說明 (作者僅用「效果量」來描述)	2	1.5%	0	0.0%
小計	<b>34</b>	<b>26.2%</b>	<b>42</b>	<b>22.3%</b>
總計	<b>130</b>	<b>100.0%</b>	<b>188</b>	<b>100.0%</b>

表 5 效果量的類型與學門之間獨立關係的卡方檢定結果

學門	<i>d</i> 類型	<i>r</i> 類型	其他類型	總計 <sup>a</sup>	勝算 <sup>b</sup>	$\chi^2$	<i>df</i>	<i>p</i>
心理學	8 ( 6.2%)	88 (67.7%)	34 (26.2%)	130 (100%)	2.10	2.61	2	.27
教育學	21 (11.2%)	125 (66.5%)	42 (22.3%)	188 (100%)	1.98			
跨學門	29 ( 9.1%)	213 (67.0%)	76 (23.9%)	318 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 為表 4 中各類型效果量的總計。<sup>b</sup> 為各學門期刊報告「*r* 類型」的次數除以「*d* 類型」+「其他類型」的次數。

在心理學的文章裡，*d* 類型的效果量僅報告了 Cohen's *d*；在教育學期刊文章裡，則報告了五種 *d* 類型的效果量。在心理學的文章中，*r* 類型的效果量有 13 種，最常報告的三種為： $R^2$  (18.5%)、 $\eta^2$  或淨  $\eta^2$  (13.8%)、Pearson's *r* (10.8%)；教育學文章則報告了 20 種 *r* 類型的效果量，最常報告的三種為： $R^2$  (13.8%)、中介分析裡的迴歸係數 (11.7%) 及 Pearson's *r* (11.2%)。心理學的文章共報告了八種其他類型的效果量，教育學的文章也報告了八種其他類型的效果量。無論在心理學門或教育學門，最常報告的其他類型的效果量皆是 SEM 適配度指標（心理學 17.7%、教育學 18.1%）。

如表 5 所示，效果量的類型與學門之間的卡方檢定未達顯著 ( $\chi^2(2, N=318) = 2.61, p = .27$ )，因此，效果量的類型與學門之間獨立關係的虛無假設無法被推翻；換言之，兩學門在報告效果量的類型上無顯著差異。針對每一學門，本研究計算報告 *r* 類型效果量相對於報告 *d* 類型或其他類型效果量的勝算。結果顯示，教育學期刊報告 *r* 類型效果量的勝算是心理學期刊的 0.95 倍，在 95% 信心水準下，勝算比的最高估計值是 1.52，最低估計值是 0.59。

### (三) 效果量的解釋

針對面向 (3)，本文探討四個研究問題，分別是：臺灣心理學與教育學的期刊文章解釋效果量的比例為何？解釋效果量的比例是否與學門有關？以及，兩學門在三種效果量的解釋方式上之比例為何？效果量的解釋方式是否與學門有關？

本研究將效果量的解釋方式分為三種：第一種為「標籤化」，亦即作者根據學者過去對某一效果量所定義的小、中、大的標準來標註效果量。第二種為「和過去研究比較」，亦即作者將研究結果與過去相似議題的研究結果做比較。第三種為「討論臨床和實務上的重要

性」，亦即作者從臨床或實務的觀點討論研究結果的意義、貢獻或重要性。在這三種解釋方式中，只有第二種與第三種解釋方式符合 APA 與 AERA 對效果量解釋之要求。

以第一種「標籤化」方式解釋效果量的範例取自簡馨瑩、連啓舜及張紹盈 (2017) 的〈故事提示策略、工作記憶能力對幼兒故事理解能力的影響〉。這篇論文引用了 Cohen (1988) 的標準 (見表 1b)，將效果量  $\eta^2$  標籤化：

中工作記憶能力組的幼兒在「有提示教學」及「無提示教學」時亦同， $F(1, 42) = 6.11$ ， $p = .02$ ，效果量  $\eta^2 = .13$ 。有關效果量的意義，依 Cohen (1988) 的建議標準達 .0588 至 .1379 之間，表示具有中等的效果量，效果量為 .13，顯示有提示的教學可以解釋「故事理解能力 (總分)」平均得分總變異量的 13%。  
(簡馨瑩等人，2017，頁 193)

以第二種「和過去研究比較」來解釋效果量的範例，取自 Huang 與 Chen (2018) 之「The Relationship Between Alcohol and Injury at Emergent Department in Northern Taiwan」。此文引用過去相似議題的結果，來解釋其所報告的效果量—風險比率 (relative risk, RR)：

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, p. 203, p. 206)

以第三種方式「討論臨床和實務上的重要性」來解釋效果量的範例，則取自陳淑麗、曾世杰及林慧敏（2018）的〈以讀寫合一課程提昇五年級偏鄉地區學生的寫作能力〉。該論文不僅檢驗實驗組與控制組在寫作能力各面向上的差異，也報告了淨  $\eta^2$ ；此外，陳淑麗等人進一步把實驗組與控制組學生分成低程度、中程度及高程度以便分析學生文意層次進步的階數，本研究認為此論文討論了效果量在臨床和實務上的意義：

實驗組在造句商數及文意層次分別高於對照組 1.15 分及 1.83 分，達顯著差異（造句商數： $F(1, 53) = 6.69, p < .05$ ；文意層次： $F(1, 53) = 25.21, p < .001$ ）。其中，寫作造句商數的模型解釋力達到 11.0% [淨  $\eta^2$ ] <sup>4</sup>，寫作文意層次的整體模型解釋力則高達 32.0% [淨  $\eta^2$ ] <sup>4</sup>。以上結果顯示，經過一學年的讀寫合一教學介入，實驗組學生的寫作能力，在造句商數與文意層次分數的進展明顯比對照組好，介入本身有相當不錯的解釋力。…實驗組學生文意層次進步的階數，有 23 位學生（82.1%），後測 2 的文意層次進步 1 階或 2 階，僅有 5 位學生（17.9%）維持在原來的層次。這個資料顯示，隨著實驗課程的介入，

學生漸漸脫離低分群的範疇；而且，高分群學生的人數有增加的趨勢。（陳淑麗等人，2018，頁 86、89）

表 6 呈現心理學門與教育學門裡，有解釋效果量之文章篇數與比例。如表 6 所示，在報告效果量的 192 篇文章中，106 篇（55%）有解釋效果量；心理學的文章解釋效果量之比例為 52%，教育學則為 58%。卡方檢定結果未達顯著（ $\chi^2(1, N = 192) = 0.66, p = .42$ ），因此，效果量的解釋與學門之間獨立關係的虛無假設無法被推翻；換言之，兩學門在解釋效果量的比例上無顯著差異。教育學期刊解釋效果量的勝算是心理學期刊的 1.27 倍，在 95% 信心水準下，勝算比的最高估計值是 2.24，最低估計值是 0.72。

針對 106 篇解釋效果量的文章，本研究進一步檢視作者所採用的解釋方式。結果發現，使用「標籤化」的解釋方式占的比例（89%，見表 7）遠高於「和過去研究比較」（9%）與「討論臨床和實務上的重要性」（2%）。由於「和過去研究比較」與「討論臨床和實務上的重要性」兩種解釋方式的比例甚低，為了比較學門間的差異，本研究將這兩種方式合併以便進行統計分析；心理學的文章使用「標籤化」的解釋方式之比例為 81%，教育學則為 95%。效果量的解釋方式（「標籤化」或其他兩種解釋方式）與學門之間獨立性的卡方檢定達統計顯著（ $\chi^2(1, N = 106) = 4.82, p = .03$ ），因此，

表 6 效果量的解釋與學門之間獨立關係的卡方檢定結果

學門	有解釋	未解釋	文章總數 <sup>a</sup>	勝算 <sup>b</sup>	$\chi^2$	<i>df</i>	<i>p</i>
心理學	48 (52%)	44 (48%)	92 (100%)	1.09	0.66	1	.42
教育學	58 (58%)	42 (42%)	100 (100%)	1.38			
跨學門	106 (55%)	86 (45%)	192 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 為表 3 中「至少報告一個效果量」的文章總數。<sup>b</sup> 為各學門期刊「有解釋」的文章次數除以「未解釋」的文章次數。

表 7 效果量的解釋方式與學門之間獨立關係的卡方檢定結果

學門	標籤化	和過去研究比較	討論臨床和實務上的重要性	文章總數 <sup>a</sup>	勝算	$\chi^2$ <sup>b</sup>	<i>df</i> <sup>b</sup>	<i>p</i> <sup>b</sup>
心理學	39 (81%)	8 (17%)	1 (2%)	48 (100%)	4.33	4.82	1	.03
教育學	55 (95%)	2 (3%)	1 (2%)	58 (100%)	18.33			
跨學門	94 (89%)	10 (9%)	2 (2%)	106 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 為表 6 中「有解釋效果量」的文章總數。<sup>b</sup> 根據 2（學門）× 2（「標籤化」或其他兩種解釋方式）的表格計算。

效果量的解釋方式與學門之間獨立關係的虛無假設可被推翻；換言之，兩學門在效果量的解釋方式上有顯著差異。針對每一學門，本研究計算「標籤化」解釋方式相對於其他兩種方式的勝算，結果顯示，教育學期刊把效果量標籤化的勝算是心理學期刊的 4.23 倍，而且在 95% 信心水準下，其勝算比的最高估計值是 16.64，最低估計值是 1.08。

#### （四）統計顯著性與效果量的落差

針對面向 (4)，本文探討以下四個研究問題：統計檢定的顯著性與效果量有落差的比為何？該落差的比列是否與學門有關？以及，當統計顯著性與效果量有落差時，作者是否針對落差做解釋？對落差做解釋是否與學門有關？

為回答上述四個問題，本研究從報告效果量的文章中，篩選出同時有執行相對應之統計檢定的文章，若一篇文章中所有的效果量皆無對應的統計檢定（如：ROC 曲線下面積），則不納入此面向的分析。本文參考 Sun 等人（2010）的研究來判定落差是否存在：首先，檢視統計檢定是否達顯著；其次，判斷統計檢定與其對應的效果量之間是否有落差。統計檢定顯著性以 .05 或是作者自訂的  $\alpha$  值為依據，效果量的大小判斷，則是依循現行文獻所提出的效果量基準，或是該文作者自訂的準則來判斷。現行文獻所提出的效果量基準如：Cohen's  $d = 0.2$ 、 $0.5$ 、 $0.8$  分別為小、中、大效果量（Cohen, 1988），或 Hu 與 Bentler（1999）建議的 CFI、TLI  $> .95$  為模型適配度良好，其他效果量標準請參見表 1a 至 1c。儘管根據現行文獻所提出的基準來判斷效果量的大小受到部分學者的批判（e.g., Thompson, 1999, 2008），然而若要評估心理學或教育學各領域之效果量的大小，採用文獻所設立的標準可作為本研究判斷落差的依據。

本研究將兩種情況視為統計顯著性與效果量強度之間有落差：其一，在判斷 SEM 的模型適配度時，卡方檢定達顯著，表示模型不適配於資料，但其對應的適配度指標顯示適配良好；抑或卡方檢定未顯著，表示模型適配於資料，但其對應的適配度指標顯示模型適配不佳。其二，在判定其他統計方法時，若統計檢定達顯著，但其對應的效果量未達中度標準；抑或統計檢定未達顯著，但其對應的效果量等於或大於中度標準。有關統計檢定顯著性與效果量之間是否有落差，以及作者是否解釋了落差的編碼細節，請參見附錄。

「有落差」的文章如：邱倚璿（2017）的〈時間

次序與空間位置訊息於記憶運作中的運用：以失聰手語使用者為例〉。此文交互作用未達顯著，然其效果量  $\eta_G^2$  為 .24，根據 Cohen（1988）的準則， $\eta_G^2 \geq .1379$  即屬於大效果，因此兩者間有落差，但作者並未解釋該落差，故該文被歸類為「有落差未解釋」：

結果發現，參與者族群達顯著 [ $F(1, 40) = 20.35, p < .01, \eta_G^2 = .34$ ]，顯示失聰手語者的正確率比聽常非手語者顯著較低，一致性效果亦達顯著 [ $F(1, 40) = 11.52, p < .01, \eta_G^2 = .08$ ]，反映出時間和空間線索一致下有較高的正確率，而交互作用未達顯著 [ $F(1, 40) = 1.06, p = .31, \eta_G^2 = .24$ ]。（邱倚璿，2017，頁 106）

對於「有落差」的文章，本研究進一步探討作者是否解釋落差。若作者針對其中一個有落差的結果提供可能的原因，則該文章被歸類為「有落差且有解釋」。例如：葉光輝、鄭欣佩及楊永瑞（2005）的〈母親的後設情緒理念對國小子女依附傾向的影響〉一文，對卡方檢定之顯著性與適配度指標的落差解釋如下：

此測量模型的確證性因素分析展現出相當良好的整體模型適合度， $\chi^2(48, N = 538) = 77.53, p < .01$ ，SRMR = .03，NNFI = .99，IFI = .99，CFI = .99，RMSEA = .03〔以上是採用 Hu & Bentler（1999）建議的列示指標來呈現〕<sup>5</sup>。儘管卡方值仍達到顯著水準，但這應該是大樣本數令模型卡方值的統計考驗力提高（Bollen, 1989），以致十分容易偵測到模型微小差距而拒絕虛無假設的結果。因此，此時應該參考適合度指標來判斷測量模型設定的有效性，並接受此測量模型。（葉光輝等人，2005，頁 188-189）

此外，陳淑麗、曾世杰、張毓仁及蔡佩津（2017）在〈永齡國語文補救教學方案及補救教師專業背景對國小二年級學生讀寫進展之成效研究〉一文裡，發現補救教學介入的效果雖達統計顯著，但其效果量  $\Delta R^2$  甚小，兩者間有落差。陳淑麗等人分別從統計方法與專業領域知識兩個角度，提出造成落差可能的原因，進而論述介入仍有其成效。因此，這篇文章亦被歸類為「有落差且有解釋」：

階層二所投入的方案組別虛擬變項雖然達到顯著水準 ( $p < .05$ )，但在國字聽寫、閱讀理解和 ASAP<sup>6</sup> 的獨特解釋變異量則是僅只有 1%。研究者推測，本研究實驗方案介入效果雖然達到統計上的顯著水準，但是其所能增加的  $R^2$  解釋量卻是有限的，其可能原因有二：1. 階層迴歸分析統計的限制……。2. 實驗方案執行的時間不夠長。（陳淑麗等人，2017，頁 96）

本研究結果顯示，在報告效果量的 192 篇文章中，有 17 篇文章所報告的效果量無對應的統計檢定，因此將這 17 篇文章從表 8 排除。如表 8 所示，在其餘 175 篇文章中，86 篇（49%）有落差，89 篇（51%）則無。心理學的文章裡有落差的比為 43%，教育學則為 55%，卡方檢定結果未達顯著 ( $\chi^2(1, N = 175) = 2.58, p = .11$ )。因此，統計顯著性與效果量有無落差與學門之間獨立的虛無假設無法被推翻；換言之，兩學門在落差的比上無顯著差異。針對每一學門，本研究計算有落差相對於無落差的勝算，結果顯示，教育學期刊在統計顯著性與效果量上有落差的勝算是心理學期刊的 1.63 倍，在 95% 信心水準下，勝算比的最高估計值是 2.97，最低估計值是 0.90。

針對 86 篇有落差的文章，本研究進一步分析作者是否對落差提出解釋。表 9 顯示，35% 的作者解釋了落差，其中心理學文章的作者解釋落差之比例為 31%，教育學則為 37%，卡方檢定結果未達顯著 ( $\chi^2(1, N = 86) = 0.31, p = .58$ )，因此解釋落差的比與學門之間獨立的虛無假設無法被推翻；換言之，兩學門在解釋落差的比上無顯著差異。針對每一學門，本研究計算解釋落差相對於未解釋落差的勝算，結果顯示，教育學期刊的勝算是心理學期刊的 1.30 倍，在 95% 信心水準下，勝算比的最高估計值是 3.22，最低估計值是 0.52。

## （五）總結

根據以上的結果，本研究對效果量在臺灣心理與教育學期刊文章裡應用的現況，歸納出以下四點結論：

1. 整體報告效果量的比為 72%。此外，教育學期刊文章報告效果量的比顯著地高於心理學期刊文章。
2. 當把效果量歸納成  $d$  類型、 $r$  類型及其他類型時，這三個類型在心理學與教育學期刊文章內報告的比相似，超過 65% 的效果量都屬於  $r$  類型的效果量。
3. 在心理學和教育學期刊文章裡，解釋效果量的比相似（約 55%）。當把效果量的解釋方式分成「標

表 8 統計顯著性與效果量有落差的比與學門之間獨立關係的卡方檢定結果

學門	有落差	無落差	文章總數 <sup>a</sup>	勝算 <sup>b</sup>	$\chi^2$	$df$	$p$
心理學	35 (43%)	47 (57%)	82 (100%)	0.74	2.58	1	.11
教育學	51 (55%)	42 (45%)	93 (100%)	1.21			
跨學門	86 (49%)	89 (51%)	175 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 為表 3 中「至少報告一個效果量」的文章總數扣除 17 篇文章，因其所報告的效果量無對應的統計檢定。<sup>b</sup> 為各學門期刊「有落差」的次數除以「無落差」的次數。

表 9 對落差做解釋與學門之間獨立關係的卡方檢定結果

學門	有落差且有解釋 <sup>a</sup>	有落差未解釋 <sup>b</sup>	文章總數 <sup>c</sup>	勝算 <sup>d</sup>	$\chi^2$	$df$	$p$
心理學	11 (31%)	24 (69%)	35 (100%)	0.458	0.31	1	.58
教育學	19 (37%)	32 (63%)	51 (100%)	0.594			
跨學門	30 (35%)	56 (65%)	86 (100%)				

註：括號裡顯示列百分比。

<sup>a</sup> 同篇文章只編碼一次，因此，若作者在所有效果量中，只針對一個有落差的效果量做解釋，則此文章編碼為「有落差且有解釋」。<sup>b</sup> 同篇文章只編碼一次，因此，若所有效果量中，僅有一個有落差且作者並未針對此落差做解釋，則此文章編碼為「有落差未解釋」。<sup>c</sup> 為表 8 中「有落差」的文章總數。<sup>d</sup> 為各學門期刊「有落差且有解釋」的次數除以「有落差未解釋」的次數。為了避免計算勝算比時有捨入誤差，計算勝算時，我們把所得的值四捨五入到小數點後第三位。

籤化」、「和過去研究比較」及「討論臨床和實務上的重要性」時，高於 80% 的文章採用「標籤化」的解釋方式。此外，教育學期刊文章使用「標籤化」的比例顯著地高於心理學期刊文章。

4. 在心理學和教育學期刊文章裡，統計顯著性與效果量強度有落差的比率相似（約 50%）。此外，兩學門對落差提出解釋的比率亦相似（約 35%）。

## 討論與建議

APA 在 1999 年所發表的專案報告中（Wilkinson & the Task Force on Statistical Inference, 1999），首度要求學者在報告統計分析的結果時，也報告效果量。此要求之後陸續出現在 APA 與 AERA 所發表的研究報告準則裡（APA, 2010, 2020; AERA, 2006）。這些報告準則確實提高了近 20 年來 APA 與 AERA 期刊文章中報告效果量的比例（如：Kirk, 1996; Peng et al., 2013）。鑑於文獻欠缺對效果量在臺灣心理學與教育學期刊應用的探討，本研究系統性地回顧於 2017 與 2018 年，發表在臺灣高評價的八本心理學與九本教育學期刊之 268 篇文章，並根據每篇文章在效果量應用的四個面向上的表現，探討臺灣心理學與教育學期刊文章報告效果量的現況。這四個面向包括：效果量報告的比例、類型、解釋方式，以及當統計顯著性與效果量強度有落差時，作者是否對該落差進行討論並提出解釋。

以下的段落將依序討論本研究結果與國外研究的異同、效果量報告之建議、效果量的誤用與錯誤解釋，以及研究限制與未來研究方向。

### （一）本研究結果與國外研究的異同

本研究結果顯示，72% 的期刊文章報告了效果量。其中，超過 65% 都屬於  $r$  類型的效果量。在報告效果量的文章中，55% 也同時解釋了效果量，但是 80% 以上的文章僅將效果量以小、中或大的標籤標註。在執行統計檢定又報告效果量的文章中，大約 50% 的文章出現兩者有落差的問題，然而僅 35% 的文章解釋或探究該落差。心理學與教育學期刊文章在效果量的類型、統計顯著性與效果量的落差，以及對落差提出解釋的比例相似。儘管心理學期刊文章報告效果量的比例顯著地低於教育學，然而教育學文章使用標籤化的方式解釋效果量之比例卻顯著地高於心理學。

相較於 Sun 等人（2010）與 Peng 等人（2013）的

研究結果，臺灣教育學期刊報告效果量的比例與 AERA 期刊相近，兩者皆高於心理學期刊。雖然 Sun 等人（2010）、Peng 等人（2013）及本研究所收集到有關心理學和教育學期刊報告效果量的數據不一，但這三個研究皆指出，即使在同一學門裡，不同的期刊報告效果量的比例亦有所差異（參見補充材料）。相似於美國期刊文章，臺灣期刊文章也同樣頻繁地報告  $r$  類型的  $R^2$  與  $\eta^2$ ，唯國內文章報告 Cohen's  $d$  的比例相對較低。此外，臺灣期刊作者比美國期刊作者更常報告 Pearson 相關係數、中介分析裡的迴歸係數、SEM 適配度指標及其他類型的效果量。

在效果量的解釋上，本研究結果與 Alhija 與 Levy（2009）、Sun 等人（2010）及 Peng 等人（2013）的研究結果相似，亦即，略高於 50% 的期刊文章在報告效果量的同時，也解釋了效果量；然而，大多數的解釋採用本研究所歸納的第一種「標籤化」方式。再者，本研究發現在臺灣心理學與教育學期刊文章裡，統計顯著性和效果量強度有落差的比率（約 50%）高於 Sun 等人（2010）的結果（11%）。臺灣的心理學期刊與教育學期刊對落差提出解釋的比例（約 35%）則介於 Sun 等人（2010）所回顧的 APA 期刊（22%）與獨立期刊（48%）之間。

本研究認為，造成效果量在臺灣心理學及教育學期刊應用之現況，可能有以下三個原因：首先，教育學期刊的作者較心理學期刊的作者更常使用複雜的統計方法（如：階層線性模型、SEM），而這些方法的執行都涉及了模型適配度指標的計算。由於適配度指標可視為效果量（Kelley & Preacher, 2012; Peng et al., 2013; Sun et al., 2010），可能因而提高教育學期刊作者報告效果量的比例。其次，一般統計軟體會例行計算  $r$  類型的效果量，作者不必刻意設定，可能因而提高此類型效果量的報告比例。最後，就效果量的解釋而言，多數作者採用標籤化的解釋方式，這些作者或認為標籤化的解釋比較簡潔，抑或認為此解釋方式達到期刊對效果量解釋的要求。

### （二）效果量報告之建議

除了 APA 與 AERA 所提供之報告準則外，本文整合 Alhija 與 Levy（2009）、Sun 等人（2010）、Peng 等人（2013）及本研究的回顧結果，歸納出五點報告效果量時所應特別注意的事項，然而這些事項不論在 1999 年前或是在 1999 年後，都常被作者所忽略（Peng

et al., 2013)。

1. 作者在報告效果量時，需明確說明效果量的計算公式或列出所引用的文獻，以及所使用的統計軟體與版本。例如：若作者僅報告標準化的兩組平均數差異為效果量，則這樣的報告並不精確，因為標準化的兩組平均數差異在文獻上至少有 Cohen's  $d$  (Cohen, 1969) 與 Hedges's  $g_u$  (Hedges, 1981) 兩種定義 (參見表 1a)。
2. 作者應選擇具良好統計特性的效果量來報告，讓讀者易於理解<sup>7</sup>，且該效果量確實能表達研究結果的實用性、理論上或臨床上的重要性。針對效果量的統計特性，作者需檢驗資料是否符合效果量指標之假設，且所選擇的效果量應是母群參數的不偏估計。當單一的效果量指標不能滿足以上所有的條件時，作者可以報告數個效果量指標。本文的表 1a 至表 1c (含補充材料) 整理了常見效果量的注意事項，可供讀者參考。
3. 當作者選定效果量指標後，應計算並報告其信賴區間，因為效果量的信賴區間能讓讀者掌握抽樣對效果量的影響程度 (Wilkinson & the Task Force on Statistical Inference, 1999)。本文補充材料的表 1a 至表 1c 整理了效果量信賴區間之相關文獻，可供讀者參考。
4. 作者需要根據過去研究的結果或是臨床上 / 實務上的重要性，對效果量提出完整的解釋，且在解釋時需要考量研究的情境，如：關注的母群、研究設計與程序、操弄的獨變項或介入方案、測量工具、控制組和實驗組經歷的事件及兩組參與者在依變項上的分數分布情形。若僅用類似 Cohen (1988) 所界定的小、中、大效果量標籤來詮釋效果量，而忽略研究的情境脈絡，將無法正確傳遞效果量所代表的意義<sup>8</sup>。
5. 當統計顯著性與效果量的強度有落差時，作者需要解釋造成此落差的可能原因與其所代表的實徵意涵<sup>9</sup>。例如：是否由於樣本數不足導致統計檢定力低，因此  $p$  值未達顯著，但效果量值卻很高。有關如何搭配  $p$  值來解釋效果量，讀者可參考 Fan (2001) 的表 2。

本文推薦 Sun 等人 (2010) 所統整的量化實徵研究之七個步驟，供讀者參考。這七個步驟涵蓋自擬訂研究計劃階段起始，學者如何根據先前研究，進行事前統計檢定力分析，以估計有效樣本數，到最後撰寫研究報

告時，應如何闡述任何可能影響研究效度或效果量數值的因素，以利學者在未來研究裡有效地控制這些影響因子。關於效果量的計算，除了由一般統計套裝軟體提供外<sup>10</sup>，也有免費的程式可供學者利用，例如：實用整合分析效果量軟體 (<https://campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD-main.php>) 或是效果量軟體 (<https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html>)。

APA 與 AERA 已對效果量的報告提出明確的要求，並鼓勵學者在報告量化研究結果時，依據效果量所提供的訊息，對研究結果提供完整的解釋。然臺灣心理學與教育學期刊尚未針對效果量的報告擬定正式的準則。在本研究所回顧的 17 本期刊裡，除了《臺灣精神醫學》對統計結果的呈現有特別的要求外，其他 16 本期刊僅明列論文格式要求。本文建議期刊主編在投稿須知裡，要求作者報告效果量並闡述其意義，以提升效果量對量化實徵研究的貢獻。本文也在研究結果的部分收錄四個效果量報告的典範，旨在示範如何參考過去的研究對效果量提出解釋 (Huang & Chen, 2018)；如何從臨床顯著性的角度來解釋效果量 (陳淑麗等人, 2018)；以及如何對統計顯著性與效果量強度之間的落差提出解釋 (陳淑麗等人, 2017；葉光輝等人, 2005)。

### (三) 效果量的誤用及錯誤解釋

儘管 APA 與 AERA 皆強調效果量在量化研究中的重要性，但近年來陸續有學者提出對效果量不同的看法。這些學者的主要論點為：以效果量的強度來判定介入方案或獨變項之有效性，未必有實際意義。

Simpson (2018, 2020) 強調，效果量的強度會受研究程序、測量工具、實驗組與對照組的定義，以及樣本性質不同而影響。相關的實徵研究也證實上述觀點，例如：(1) 受試者內設計的研究，其效果量高於受試者間設計的研究 (Bakker et al., 2019; Kraft, 2020; Schäfer & Schwarz, 2019; Simpson, 2020)；(2) 非預先註冊研究 (studies without pre-registration) 之效果量高於註冊研究 (Schäfer & Schwarz, 2019)；(3) 研究者自行設計的測量，其效果量高於標準化的測量 (Cheung & Slavin, 2016; Li & Ma, 2010)；(4) 控制組與實驗組所經歷的事件差別較大者，其效果量高於差異較小者 (Simpson, 2018; Steenbergen-Hu & Cooper, 2014)；以及 (5) 測量分數的變異較小者，其效果量高於變異較大者



(Shadish, Cook, & Campbell, 2002; Simpson, 2018)。

綜上所述，效果量的值實際上反映了整個研究的過程 (Simpson, 2018, 2020)，若學者忽略研究情境，僅以效果量的值判斷介入方案或獨變項的有效性，便容易錯誤解讀效果量所傳達的訊息 (Simpson, 2020)。

由於效果量的數值深受上述因素的影響，Simpson (2018, 2020) 更進一步批判整合分析與整合分析 (meta-meta-analysis) 的合理性。Wrigley 與 McCusker (2019) 則指出，學者不能單以效果量來判斷「什麼是有效的？」，更應重視「在什麼樣的情境脈絡下，能夠使介入方案或操弄變項有效？」。

#### (四) 研究限制與未來研究方向

本研究雖然根據過去對臺灣學術期刊評比的文獻，選出了 17 本聲望極高的心理學期刊與教育學期刊進行效果量報告的回顧，然而此結果並不能推論到臺灣所有的心理學與教育學期刊，或是同期刊裡不同年份所發表的論文，因為本研究發現，即使在同一本期刊裡，各文章在效果量的報告品質上仍參差不齊。再者，本研究雖發現效果量的報告與統計方法有關，但本文並未對此現象做進一步的探討，未來研究可以更深入地研究這兩者間的關聯。未來研究也可以針對學者們對效果量的理解與認識，進行問卷調查或訪談，以瞭解導致本研究結果的原因，例如：為何某類型的效果量鮮少被應用？為何學者選擇標籤化的方式來解釋效果量？為何學者並未針對統計顯著性與效果量強度的落差提出解釋？

### 結語

本研究結果顯示在臺灣心理學與教育學量化實徵研究裡，絕大多數的作者在研究結果裡報告了效果量。可惜的是，作者未必對效果量提供完整的解釋，以幫助讀者理解研究結果的重要性及其實徵意涵。本文根據 APA/AERA 效果量報告準則與相關文獻，探討效果量所提供的訊息，並檢視效果量在量化研究裡的重要性，以及如何完整地報告並解釋效果量。希冀推進效果量在臺灣量化實徵研究裡的廣泛應用與正確闡述，亦期本文達拋磚引玉之效，鼓勵更多臺灣學者投入效果量相關議題的探討與研究。

### 參考文獻

- 邱倚璿 (2017)：〈時間次序與空間位置訊息於記憶運作中的運用：以失聰手語使用者為例〉。《特殊教育研究學刊》，42 (3)，93-117。[Chiu, Y.-S. (2017). Temporal order and spatial information in short-term memory: Evidence from deaf Taiwanese signers. *Bulletin of Special Education*, 42(3), 93-117.] doi: 10.6172/BSE.201711\_42(3).0004
- 陳淑麗、曾世杰、林慧敏 (2018)：〈以讀寫合一課程提昇五年級偏鄉地區學生的寫作能力〉。《教育研究與發展期刊》，14 (4)，71-100。[Chen, S.-L., Tzeng, S.-J., & Lin, H.-M. (2018). Raising the writing ability of grade 5 students from schools in rural Taiwan using a reading and writing bi-focal curriculum. *Journal of Educational Research and Development*, 14(4), 71-100.] doi: 10.3966/181665042018121404003
- 陳淑麗、曾世杰、張毓仁、蔡佩津 (2017)：〈永齡國語文補救教學方案及補救教師專業背景對國小二年級學生讀寫進展之成效研究〉。《特殊教育研究學刊》，42 (2)，85-111。[Chen, S.-L., Tzeng, S.-J., Chang, Y.-J., & Tsai, P.-C. (2017). Effects of the Young-Ling remedial program and teachers' professional background on the literacy progress of second graders. *Bulletin of Special Education*, 42(2), 85-111.] doi: 10.6172/BSE.2017.07.4202004
- 翁麗鎮、黃怡蓉、鄭中平 (2012)：〈2008 年台灣心理學門學術期刊評比研究〉。《中華心理學刊》，54，413-431。[Weng, L.-J., Huang, Y.-J., & Cheng, C.-P. (2012). An evaluation of psychology journals in Taiwan in 2008. *Chinese Journal of Psychology*, 54, 413-431.] doi: 10.6129/CJP.20120329
- 黃毅志 (2009)：〈2008 年國內教育學術期刊評比研究〉。《教育研究集刊》，55 (2)，1-33。[Hwang, Y.-J. (2009). An evaluation of educational journals of Taiwan in 2008. *Bulletin of Educational Research*, 55(2), 1-33.]
- 葉光輝、鄭欣佩、楊永瑞 (2005)：〈母親的後設情緒理念對國小子女依附傾向的影響〉。《中華心理學刊》，47，181-195。[Yeh, K.-H., Cheng, S.-P., & Yang, Y.-J. (2005). The influence of maternal meta-emotion philosophy on children's attachment inclination.

- Chinese Journal of Psychology*, 47, 181-195.] doi: 10.6129/CJP.2005.4702.06
- 簡馨瑩、連啓舜、張紹盈 (2017) : 〈故事提示策略、工作記憶能力對幼兒故事理解能力的影響〉。《教育科學研究期刊》, 62 (4) , 181-207。[Chien, H.-Y., Lien, C.-S., & Chang, S.-Y. (2017). Effects of story prompting and working memory ability on young children's story comprehension. *Journal of Research in Education Sciences*, 62(4), 181-207.] doi: 10.6209/JORIES.2017.62(4).07
- Agresti, A. (2012). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Ainur, A. K., Sayang, M. D., Jannoo, Z., & Yap, B. W. (2017). Sample size and non-normality effects on goodness of fit measures in structural equation models. *Pertanika Journal of Science & Technology*, 25, 575-586. Retrieved from [http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JST%20Vol.%2025%20\(2\)%20Apr.%202017/16%20JST%20Vol%2025%20\(2\)%20Apr%202017\\_JST-0057-2016\\_pg575-586.pdf](http://www.pertanika.upm.edu.my/Pertanika%20PAPERS/JST%20Vol.%2025%20(2)%20Apr.%202017/16%20JST%20Vol%2025%20(2)%20Apr%202017_JST-0057-2016_pg575-586.pdf)
- Alhija, F. N.-A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement*, 69, 245-265. doi: 10.1177/0013164408315266
- American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, 35(6), 33-40. doi: 10.3102/0013189X035006033
- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington, DC: Author.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567, 305-307.
- Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology*, 29, 664-672. doi: 10.1123/jsep.29.5.664
- Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102, 1-8. doi: 10.1007/s10649-019-09908-4
- Balderjahn, I. (1988). A note to Bollen's alternative fit measure. *Psychometrika*, 53, 283-285. doi: 10.1007/BF02294138
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182. doi: 10.1037/0022-3514.51.6.1173
- Bergsma, W. (2013). A bias-correction for Cramér's  $V$  and Tschuprow's  $T$ . *Journal of the Korean Statistical Society*, 42, 323-328. doi: 10.1016/j.jkss.2012.10.002
- Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-536. doi: 10.2307/2279690
- Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, 45, 661-701. doi: 10.1080/00273171.2010.498292
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Carlson, R. (1976). The logic of tests of significance. *Philosophy of Science*, 43, 116-128. Retrieved from <https://www.jstor.org/stable/187338>
- Cheung, A., & Slavin, R. E. (2016). How methodological

- features affect effect sizes in education. *Educational Researcher*, 45, 283-292. doi: 10.3102/0013189X16656615
- Citrome, L., & Ketter, T. A. (2013). When does a difference make a difference? Interpretation of number needed to treat, number needed to harm, and likelihood to be helped or harmed. *The International Journal of Clinical Practice*, 67, 407-411. doi: 10.1111/ijcp.12142
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509. doi: 10.1037/0033-2909.114.3.494
- Cochran, W. (1954). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics*, 10, 417-451. doi: 10.2307/3001616
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65, 145-153. doi: 10.1037/h0045186
- Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York, NY: McGraw-Hill.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997-1003. doi: 10.1037/0003-066X.49.12.997
- Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (2nd ed.). Boca Raton, FL: CRC Press.
- Cragg, J. G., & Uhler, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics*, 3, 386-406. doi: 10.2307/133656
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.
- Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods*, 7, 485-503. doi: 10.1037/1082-989X.7.4.485
- Dunleavy, E. M., Barr, C. D., Glenn, D. M., & Miller, K. R. (2006). Effect size reporting in applied psychology: How are we doing? *The Industrial-Organizational Psychologist*, 43(4), 29-37. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.594.8929&rep=rep1&type=pdf>
- Egan, K. (2012). *Education and psychology: Plato, Piaget and scientific psychology* (2nd ed.). New York, NY: Routledge.
- Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, 94, 275-282. doi: 10.1080/00220670109598763
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6, 56-83. doi: 10.1080/10705519909540119
- Faraway, J. J. (2014). *Linear models with R* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Fisher, R. A. (1925a). *Statistical methods for research workers*. London, England: Oliver and Boyd.
- Fisher, R. A. (1925b). The influence of rainfall on the yield of wheat at Rothamstead. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 213, 89-142. doi: 10.1098/rstb.1925.0003
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York, NY: Russell Sage Foundation.
- Finsaas, M. C., & Goldstein, B. L. (2020). Do simple slopes follow-up tests lead us astray? Advancements in the visualization and reporting of interactions. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000266
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and

- interpretation. *Journal of Experimental Psychology: General*, 141, 2-18. doi: 10.1037/a0024338
- Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47, 61-87. doi: 10.1080/00273171.2012.640596
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2, 156-168. doi: 10.1177/2515245919847202
- Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries: R. A. Fisher memorial lecture. *Journal of the American Statistical Association*, 63, 1091-1131. doi: 10.1080/01621459.1968.10480916
- Goodman, L. A. (1969). On partitioning  $\chi^2$  and detecting partial association in three way contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31, 486-498. doi: 10.1111/j.2517-6161.1969.tb00808.x
- Goodman, L. A. (1971). Partitioning of chi-square, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables. *Journal of the American Statistical Association*, 66, 339-344. doi: 10.1080/01621459.1971.10482265
- Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's *d* family. *The Quantitative Methods for Psychology*, 14, 242-265. doi: 10.20982/tqmp.14.4.p242
- Greenland, S. (2010). Simpson's paradox from adding constants in contingency tables as an example of Bayesian noncollapsibility. *The American Statistician*, 64, 340-344. doi: 10.1198/tast.2010.10006
- Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6, 135-146. doi: 10.1037/1082-989X.6.2.135
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.
- Hays, W. L. (1963). *Statistics for psychologists*. New York, NY: Holt, Rinehard & Winston.
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67, 451-470. doi: 10.1111/bmsp.12028
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, 24, 1918-1927. doi: 10.1177/0956797613480187
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107-128. doi: 10.2307/1164588
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York, NY: John Wiley & Sons.
- Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55. doi: 10.1080/10705519909540118
- Huang, C.-C., & Chen, Y.-C. (2018). The relationship between alcohol and injury at emergent department in northern Taiwan. *Taiwanese Journal of Psychiatry* (Taipei), 32, 200-210.
- Huang, Y.-T. (2018). Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *The Annals of Applied Statistics*, 12, 1535-1557. doi: 10.1214/17-AOAS1120
- Huang, Y.-T. (2019). Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics*, 13, 60-84. doi: 10.1214/18-AOAS1181
- Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, 16(8), 4-9. doi: 10.3102/0013189X016008004
- Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal*

- of Experimental Education*, 61, 317-333. doi: 10.1080/00220973.1993.10806593
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15, 309-334. doi: 10.1037/a0020761
- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25, 51-71. doi: 10.1214/10-STS321
- International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *The New England Journal of Medicine*, 336, 309-315. doi: 10.1056/NEJM199701233360422
- Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, 17, 137-152. doi: 10.1037/a0028086
- Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology*, 67, 283-284. doi: 10.1037/0022-006X.67.3.283
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759. doi: 10.1177/0013164496056005002
- Kirk, R. E. (2005). Effect size measures. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 532-542). Hoboken, NJ: Wiley.
- Kirk, R. E. (2013). *Experimental design: Procedures for behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.
- Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, 59, 990-996. doi: 10.1016/j.biopsych.2005.09.014
- Kraft, M. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49, 241-253. doi: 10.3102/0013189X20912798
- Lancaster, H. O. (1949). The derivation and partition of  $\chi^2$  in certain discrete distributions. *Biometrika*, 36, 117-129. doi: 10.1093/biomet/36.1-2.117
- Lancaster, H. O. (1969). *The chi-squared distribution*. New York, NY: Wiley.
- Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *The British Journal for the Philosophy of Science*, 57, 69-91. doi: 10.1093/bjps/axi152
- Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, 22, 215-243. doi: 10.1007/s10648-010-9125-8
- Liao, J. G., & McGee, D. (2003). Adjusted coefficients of determination for logistic regression. *The American Statistician*, 57, 161-165. doi: 10.1198/0003130031964
- Loeys, T., Moerkerke, B., & Vansteelandt, S. (2015). A cautionary note on the power of the test for the indirect effect in mediation analysis. *Frontiers in Psychology*, 5, 1549. doi: 10.3389/fpsyg.2014.01549
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83-104. doi: 10.1037/1082-989X.7.1.83
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99-128. doi: 10.1207/s15327906mbr3901\_4
- Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*. New York, NY: Cambridge University Press.
- Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology*, 53, 372-378. doi: 10.1037/0022-0167.53.3.372
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391-410. doi: 10.1037/0033-2909.103.3.391

- Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F. C., Matthews, D., & Dixon, F. (2008). Evaluating the state of a field: Effect size reporting in gifted education. *The Journal of Experimental Education*, 77, 55-68. doi: 10.3200/JEXE.77.1.55-68
- Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology*, 66, 525-534. doi: 10.1037/0021-9010.66.5.525
- McCabe, C. J., Kim, D. S., & King, K. M. (2018). Improving present practices in the visual display of interactions. *Advances in Methods and Practices in Psychological Science*, 1, 147-165. doi: 10.1177/2515245917746792
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (pp. 105-142). New York, NY: Academic Press.
- Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115. doi: 10.1086/288135
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17-24. doi: 10.1080/00031305.2000.10474502
- Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Mittlböck, M., & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15, 1987-1997. doi: 10.1002/(SICI)10970258(19961015)15:19<1987::AID-SIM318>3.0.CO;2-9
- Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691-692. doi: 10.1093/biomet/78.3.691
- Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A, 175-240. doi: 10.1093/biomet/20A.1-2.175
- Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2020). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000299
- Odgaard, E. C., & Fowler, R. L. (2010). Confidence intervals for effect sizes: Compliance and clinical significance in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology*, 78, 287-297. doi: 10.1037/a0019294
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25, 241-286. doi: 10.1006/ceps.2000.1040
- Pearson, K. (1905). *Mathematical contributions to the theory of evolution. XIV. On the general theory of skew correlations and nonlinear regression* (Draper's Company Research Memoirs, Biometric Series II). London, England: Dulau & Co.
- Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, 23, 208-225. doi: 10.1037/met0000126
- Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's *d*: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education*, 82, 22-50. doi: 10.1080/00220973.2012.745471
- Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, 25, 157-209. doi: 10.1007/s10648-013-9218-2
- Plucker, J. A. (1997). Debunking the myth of the "highly significant" result: Effect sizes in gifted education research. *Roeper Review*, 20, 122-126. doi: 10.1080/02783199709553873
- Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66, 825-852. doi: 10.1146/annurev-psych-010814-015258
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior*

- Research Methods*, 40, 879-891. doi: 10.3758/BRM.40.3.879
- Preacher, K., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93-115. doi: 10.1037/a0022658
- Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research*, 48, 28-56. doi: 10.1080/00273171.2012.710386
- Rosenthal, R. (1994). *Parametric measures of effect size*. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York, NY: Russell Sage Foundation.
- Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110, Article 104038. doi: 10.1016/j.jml.2019.104038
- Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, Article 813. doi: 10.3389/fpsyg.2019.00813
- Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1, 103-113. doi: 10.1111/j.2041-210X.2010.00012.x
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth, Cengage Learning.
- Shieh, G. (2008). Improved shrinkage estimation of squared multiple correlation coefficient and squared cross-validity coefficient. *Organizational Research Methods*, 11, 387-407. doi: 10.1177/1094428106292901
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422-445. doi: 10.1037/1082-989X.7.4.422
- Simpson, A. (2018). Princesses are bigger than elephants: Effect size as a category error in evidence-based education. *British Educational Research Journal*, 44, 897-913. doi: 10.1002/berj.3474
- Simpson, A. (2019). Separating arguments from conclusions: The mistaken role of effect size in educational policy research. *Educational Research and Evaluation*, 25, 99-109. doi: 10.1080/13803611.2019.1617170
- Simpson, A. (2020). On the misinterpretation of effect size. *Educational Studies in Mathematics*, 103, 125-133. doi: 10.1007/s10649-019-09924-4
- Smith, M. L., & Honoré, H. H. (2008). Effect size reporting in current health education literature. *American Journal of Health Studies*, 23, 130-135.
- Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo  $R^2$  indices. *Multiple Linear Regression Viewpoints*, 39(2), 17-26. Retrieved from [http://www.glmj.org/archives/articles/Smith\\_v39n2.pdf](http://www.glmj.org/archives/articles/Smith_v39n2.pdf)
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290-312. doi: 10.2307/270723
- Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological Methodology*, 16, 159-186. doi: 10.2307/270922
- Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, 106, 331-347. doi: 10.1037/a0034752
- Subbiah, M., & Srinivasan, M. R. (2008). Classification of  $2 \times 2$  sparse data sets with zero cells. *Statistics & Probability Letters*, 78, 3212-3215. doi: 10.1016/j.spl.2008.06.023
- Sun, S. Y., 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. doi: 10.1037/a0019507
- Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology*, 9, 165-181. doi: 10.1177/095935439992006
- Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development*, 80, 64-71. doi: 10.1002/j.1556-6678.2002.tb00167.x
- Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 246-262). Thousand Oaks, CA: Sage.
- VanderWeele, T. J. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37, 17-32. doi: 10.1146/annurev-publhealth-032315-021402
- VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2, 457-468. doi: 10.4310/SII.2009.v2.n4.a7
- VanderWeele, T. J., & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, 172, 1339-1348. doi: 10.1093/aje/kwq332
- Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the *CL* common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25, 101-132. doi: 10.2307/1165329
- Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of the multiple correlation. *Annals of Mathematical Statistics*, 2, 440-457. doi: 10.1214/aoms/1177732951
- Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604. doi: 10.1037/0003-066X.54.8.594
- Wrigley, T., & McCusker, S. (2019). Evidence-based teaching: A simple view of "science." *Educational Research and Evaluation*, 25, 110-126. doi: 10.1080/13803611.2019.1617992
- Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, 40, 115-148. doi: 10.1207/s15327906mbr4001\_5
- Yzerbyt, V., Muller, D., Batailler, C., & Judd, C. M. (2018). New recommendations for testing indirect effects in mediational models: The need to report and test component paths. *Journal of Personality and Social Psychology*, 115, 929-943. doi: 10.1037/pspa0000132
- 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 report. *Educational Researcher*, 37, 208-216. doi: 10.3102/0013189X08319762



## 註釋

1. 因為效果量的發展目的之一為補充統計檢定的不足，Kelley 與 Preacher (2012)、Peng 等人 (2013) 及 Sun 等人 (2010) 皆把 SEM 適配度指標視為效果量。在統計檢定中，當樣本數夠大時，即便是微小的效果也能達統計顯著，但此微小的效果未必具有實質的意涵。有鑑於此，學者們藉由發展不受樣本數影響的效果量，來評估觀察效果的強度，SEM 適配度指標的發展亦是依循此一邏輯。例如：RMSEA 表徵平均每個自由度下，假設模型不適配於母群共變數矩陣的程度，CFI 則反映假設模型相較於基準模型之適配程度。
2. 此外，該相乘積作為間接效果 (indirect effect) 的估計可延伸到其他的變項類型或統計模型，唯在詮釋上會取決於變項本身的尺度與模型假設 (Huang, 2018, 2019)。例如當依變項是低發生率的二分變項，且中介變項與獨變項都是連續變項時，若獨變項與中介變項沒有交互作用，則間接效果可用勝算比表示 ( $OR_{\text{間接效果}} \approx e^{\hat{a}\hat{b}}$ )，在此  $e$  為歐拉數， $\hat{a}$  為獨變項對中介變項的線性迴歸係數， $\hat{b}$  為在控制獨變項下中介變項對依變項的邏輯斯迴歸係數。該值代表當獨變項增加一單位時，透過中介變項，讓依變項的勝算比相較於之前的勝算比所增加的倍數 (VanderWeele & Vansteelandt, 2010)。有興趣的讀者可參考 Preacher (2015) 與 VanderWeele (2016) 的回顧。
3. 在進行二維列聯表 (two-way contingency table) 的獨立性檢定時，若  $p$  值達顯著且自由度超過 1，所得的資訊非常有限。過去已有文獻建議研究者應瞭解變項關聯的本質，而非單單仰賴整體統計檢定的結果 (Berkson, 1938; Cochran, 1954)。若要進一步瞭解變項關聯度的本質與強度 (nature or strength of the association)，可以透過拆解檢定統計量的方式進行。具體而言，研究者須將一列聯表拆解成數個獨立的子表 (subtable)，分別對每個子表進行概似比檢定 (likelihood ratio test)，則所得各個概似比檢定統計量  $G^2$  的總和應等於整體列聯表的  $G^2$ 。Goodman (1968, 1969, 1971) 與 Lancaster (1949, 1969) 提出拆解子表的原則以確保子表之間是獨立的 (Agresti, 2012, p.84)。透過此種列聯表的拆解，研究者便能更細緻地檢驗不同類別組合間的獨立性，甚至計算出每個子表的效果量，如勝算比，以便直接估計變項間的關聯強度。另外，Kelley 與 Preacher (2012) 則提出在多元迴歸分析中，計算每一個獨變項對依變項的關聯強度，此一做法能增進讀者對結果的理解。
4. 陳淑麗等人 (2018) 並未將淨  $\hat{\eta}^2$  加註在本文所節錄的段落中，然而，從該原文表 4 可得知，其所謂的「模型解釋力」即為淨  $\hat{\eta}^2$ 。因此，本文把「淨  $\hat{\eta}^2$ 」額外加在括號〔〕裡，以助讀者的理解。
5. 葉光輝等人 (2005) 僅在第一次報告模型適配度時，引用 Hu 與 Bentler (1999) 為列示適配度指標的依據。由於本文所摘錄的段落，並非原文第一次報告模型適配度指標，因此，本文在括號〔〕裡額外加註「以上是採用 Hu & Bentler (1999) 建議的列示指標來呈現」，以助讀者閱讀。
6. ASAP 在原文指的是攜手計畫課後扶助方案科技化評量系統 (After School Alternative Program technology-based testing)。
7. 例如：當作者認為風險差異 (risk differences) 比勝算比更能表達研究結果的重要性時，應選擇風險差異作為效果量。
8. 若作者選擇以標籤化方式來解釋效果量，應解釋為何此一解釋方式可以用來解釋該研究的效果量，例如：此效果量的標籤是建立在與該研究相似的研究情境下。同時，作者應該解釋在所引用的標籤化準則裡，何謂小、中、大效果。
9. 針對每個統計檢定的結果，不論其  $p$  值是否達顯著水準，作者都應該報告效果量。
10. IBM SPSS Statistics 27.0 在作者採用  $t$  檢定時，會例行計算 Cohen's  $d$  與 Hedges's  $g_u$ 。

## 附錄：回顧文章的編碼表

- 
- A. 該篇文章是否符合本研究之篩選標準（若符合本研究之篩選標準，則繼續回答 B 以後的問題；若不符合本研究之篩選標準，則不必回答以下的問題）？  
☐ 0. 否 ☐ 1. 是
- B. 該篇文章的研究方法？  
☐ 1. 實驗設計 ☐ 2. 準實驗設計 ☐ 3. 調查法 ☐ 4. 次級資料分析 ☐ 5. 實驗設計 + 調查法  
☐ 6. 其他\_\_\_\_\_
- C. 該篇文章是否針對主要研究發現報告效果量？  
☐ 0. 否 ☐ 1. 是  
 （當該篇文章有針對主要研究發現報告效果量時，研究員對每一個在該篇文章所報告的效果量，回答效果量的編碼表裡所有的問題）
- 

### 效果量的編碼表

- 
- C1. 文章裡報告了哪個效果量指標？\_\_\_\_\_
- C2. 呈現效果量指標的頁數？\_\_\_\_\_
- C3. 是否說明效果量的計算方式（例如：公式、統計軟體、引用文獻）？  
☐ 0. 否 ☐ 1. 是\_\_\_\_\_
- C4. 針對所報告的效果量指標，是否對該指標的意涵提出解釋？  
☐ 0. 未解釋  
☐ 1. 給予小、中或大的標籤  
☐ 2. 透過和過去研究比較來進行解釋  
☐ 3. 針對效果量數值之實務或臨床意義來解釋
- C5. 判斷效果量大小的方式？  
☐ 1. 文章作者主觀判定，或作者引用現行文獻對小、中、大效果量基準來判定（在 SEM 則為適配度指標是可接受或不可接受）  
☐ 2. 文章作者並未對效果量做解釋，本研究研究員引用現行文獻對小、中、大效果量的基準來判定（在 SEM 則為適配度指標是可接受或不可接受）  
☐ 3. 效果量尚未有文獻劃定小、中、大效果量基準值，本研究團隊經討論後判定該效果量的大小
- C6. 是否有出現統計顯著性和效果量大小「有落差」的現象？  
☐ 0. 效果量未有相對應統計檢定，無法判斷兩者是否「有落差」  
☐ 1. 兩者「有落差」，包含兩種情況：一、在判斷 SEM 的模型適配度時，卡方檢定達顯著，表示模型不適配於資料，但其對應適配度指標顯示適配良好（或是卡方檢定未顯著，表示模型適配於資料，但其對應適配度指標顯示適配不佳）；二、在判斷其他統計方法時，若統計檢定達顯著，但其相對應的效果量未達中度標準（或統計檢定未達顯著，但其相對應的效果量等於或大於中度標準）  
☐ 2. 兩者結論「無落差」
- C7. 是否對「有落差」作解釋？  
☐ 0. 不適用（即 C6 題，選答 0 或 2 者）  
☐ 1. 沒有解釋  
☐ 2. 有解釋
-

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

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

Counseling and Educational Psychology, University of Nevada, Reno, USA<sup>1</sup>

Department of Psychology, National Taiwan University<sup>2</sup>

Graduate Institute of Learning and Instruction, National Central University, Taiwan<sup>3</sup>

Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology<sup>4</sup>

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  $\eta^2$ . 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  $\alpha$  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  $\alpha$  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  $\eta^2$  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/effectsizefaq/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)?	<input type="checkbox"/> 0. No. <input type="checkbox"/> 1. Yes _____
4. What was the interpretation of the reported ES?	<input type="checkbox"/> 0. No interpretation. <input type="checkbox"/> 1. The ES was labeled as small, medium, or large. <input type="checkbox"/> 2. The ES was compared with those of similar studies. <input type="checkbox"/> 3. The ES was interpreted in terms of its practical implications and clinical significance.
5. How did the author(s) assess the ES magnitude?	<input type="checkbox"/> 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). <input type="checkbox"/> 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). <input type="checkbox"/> 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?	<input type="checkbox"/> 0. The reported ES did not have a corresponding significance test. <input type="checkbox"/> 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. <input type="checkbox"/> 2. No discrepancy.
7. Was the discrepancy explained or resolved?	<input type="checkbox"/> 0. Not applicable (the response to Item 6 was 0 or 2). <input type="checkbox"/> 1. Neither explained nor resolved. <input type="checkbox"/> 2. Explained or resolved.