教育量化研究中的 *p* 值操弄現象初探：定義、影響與避免之道

周倩*、吳俊育

1920 年代統計學家 Fisher 之統計理論體系為機率計算和評估的概念，意指在自變項不受任何影響或操弄的前提下，計算出觀察到的結果之機率，稱為 *p* 值，並且於該研究脈絡下評估該機率值的意涵；而後 Neyman 與 Pearson 則提出了虛無假設及對立假設的概念，認為假設檢定應包含此兩種假設。兩個派別對於檢定方法雖有所不同，但後世將其整合而形成虛無假設顯著性檢定，並以 *p* 值作為統計資料分析的一個標準設定。然而，近幾年來，學界開始意識到 *p*-hacking 現象，本文稱為「*p* 值操弄」，意指研究者誤用或濫用資料分析方法，以便得到統計顯著結果，並據此宣稱得到成功的實驗，撰寫研究結果文章投稿至期刊發表。本文針對教育量化研究，首先說明 *p* 值的由來與學理，其次深入探討 *p* 值操弄方式、對教育學領域之學術研究的影響，以及期刊如何偵測與避免 *p* 值操弄。最後，本文提出研究者對 *p* 值應有的正確認知及避免操弄的實務作法，包括在提出統計顯著性的同時，也應提供實務顯著性的證據，確保研究結果的可重製性，詳細陳述分析架構與細節，學習並理性地選擇合適分析技術等。

關鍵詞：*p* 值操弄、有問題的研究行為、量化研究、虛無假設顯著性檢定

* 周倩：國立陽明交通大學教育研究所教授
（通訊作者：cchou@nycu.edu.tw）
吳俊育：國立陽明交通大學教育研究所教授

Journal of Educational Practice and Research
Vol. 37, No. 1, 107-126, Apr. 2024

A Preliminary Study on *p*-Hacking in Quantitative Educational Research: Definitions, Impacts, and Preventions

Chien Chou* & Jiun-Yu Wu

In the 1920s, the statistician Ronald Fisher introduced a statistical theory system based on probability calculations and evaluations. Fisher coined the term "p-value" to denote the probability of observing the obtained results, assuming that the independent variable is not influenced or manipulated. This calculation helps evaluate the likelihood of such results within the research context. Later, Neyman and Pearson proposed the concepts of null hypothesis and alternative hypothesis, advocating for the inclusion of both in the hypothesis testing. While the two approaches to testing statistical hypotheses were different, these approaches were later integrated into what is now known as null hypothesis significance testing, with the p-value as a standardized measure for statistical analysis. In recent years, however, the academic community has identified a concerning trend known as "p-hacking," where researchers misuse or abuse data analysis to achieve misleading statistical significance. Then, researchers claim to have speciously successful experiments and publish the implausible research results in journals. This paper focuses on p-hacking within quantitative educational research, exploring its origins, definition, technique, impact, and how journals prevent the manipulation of p-hacking. Finally, this paper reinstates the correct perception of p-value with recommendations for preventing p-hacking, including providing evidence of practical significance while presenting statistical significance, ensuring replicability of research results, stating the structure and details of statistical analysis, and judicious selection of appropriate analytical techniques.

Keywords: ***null hypothesis significance test, p-hacking, quantitative research, questionable research practice***

* Chien Chou: Professor, Institute of Education, National Yang Ming Chiao Tung University (corresponding author: cchou@nycu.edu.tw)
Jiun-Yu Wu: Professor, Institute of Education, National Yang Ming Chiao Tung University

教育量化研究中的 p 值操弄現象初探：定義、影響與避免之道

周倩、吳俊育

壹、前言

教育研究可用來評估教育與教學各面向的成效，解決各式教育問題，也能促使教育的改革與創新，藉以確保培育出優質且有良好素養的人力資源，足以面對未來挑戰，支持社會永續發展。其中，量化研究方法為教育研究中常用的研究方法，常見的研究設計方法有問卷調查、對照試驗及因果研究等，研究者常常使用統計學方法分析資料，以描繪各因素之間的關連性。

近年來，各式跨國大型教育評比與資料庫林立，再加上人工智慧與虛實科技加速教育大數據的發展及應用，藉由教育環境的數位轉化，研究者得以有系統地蒐集巨量或巢套資料，深入檢視當代學習者於多元教育現場的學習與適應樣態。這些豐富的巨量資料帶來諸多新穎的教育研究視野與機會，但同時也考驗研究者選用適當研究與統計分析方法的能力，以求更周延、更嚴謹地完成量化研究。例如為了分析線上學習者在與同學或是與機器人進行合作學習時的對話差異，研究者必須選用適合巨量時序密集且符合重複取樣結構資料的分析方法，例如使用多階層模型等方法以檢驗人人或是人機的不同配對，對於合作學習的對話歷程差異。若是忽略巨量資料規模的數量級（orders of magnitude），或是忽略了資料階層性而選用了不適當的分析方法，都將可能造成參數估計失真。最常見的是呈現錯誤的統計顯著（statistical significance）結果，導致偽陽性的統計決策。

另一方面，近年來，學術界拿來評量研究者的指標之一，常常是期刊論文的數量，教育研究界也不例外。但是在研究資源日益減少，競爭者眾多的情況下，論文作者為了要擠進期刊刊登的窄門，通常都希望能寫出引人注目的論文，以便獲得編輯和審查人的青睞（Fanelli, 2012），許多量化研究領域的研究者於是相信：期刊只喜歡刊登統

計有顯著結果的研究論文，當這樣的信仰成為了一種習慣，造成學界存在一種特別的研究行為，即所謂 *p*-hacking 現象，本文稱為「*p* 值操弄」。*p* 值操弄屬於「有問題的研究行為」（Questionable Research Practice, QRP）（Sijtsma, 2016；Steneck, 2006）。Banks、Rogelberg、Woznyj、Landis 與 Rupp（2016）認為，QRP 的產生是研究者為了要支持某項主張（assertion）而呈現有偏誤證據之研究行為，讓同行質疑該研究之設計、分析過程或研究成果。*p* 值操弄也是有疑義研究行為之一種，即便可能不會直接被判定是違反學術倫理的樣態，但是對於科學研究的發展、科學知識的傳播與累積，卻會造成程度不一的傷害。

而今大型資料庫與教育大數據的巨型階層資料，更加需要適當的量化研究方法，以根據取樣規劃進行周全的資料搜集，以及嚴謹的分析與結果陳報。因此，讓我們有機會來重新檢視 *p* 值操弄議題。本文目的，即是針對教育常見的量化研究，透過分析探討現有文獻，說明 *p* 值的歷史與學理，其次陳述常見 *p* 值操弄的手法、影響為何，最後說明現今教育研究者於研究執行歷程中對 *p* 值該有什麼正確的認知，並建議實務作法，尤其針對巨型階層資料分析提供避免 *p* 值操弄的可行作法。希冀當代教育研究者能正視此一有疑義的研究行為，自行警惕，並極力避免此一行為。

貳、*p* 值（*p*-value）的歷史與學理

現今教育界或跨領域的量化研究中，常用的 *p* 值是 Karl Pearson 於 1900 年為了介紹卡方檢定（Chi-square test）時所提出的概念（Pearson, 1900）。而談論 *p* 值，就不得不提到統計學 Fisher 與 Neyman-Pearson 兩派的理論體系。統計學家 Fisher（1925）提出機率計算和評估之概念，主張在自變項沒有任何影響的假設前提下，計算出所觀察到結果的機率，並且稱之為 *p* 值，這就是顯著性檢驗（significance testing）之始。而 Neyman 與 Pearson（1933）則提出了研究假設概念，強調研究假設應被分為可接受檢驗的對立假設（alternative hypothesis）以及虛無假設（null hypothesis），而假設檢定應包含此兩種假設。除此之外，Neyman 與 Pearson 亦提出 α 、 β 以及 1-β 的概念，α 代表型一錯誤，其內容為錯誤地拒絕虛無假設的機率，而 β 代表型二錯誤，內容為錯誤地拒絕對立假設，1-β 則定義為檢定力（power），即代表正確地拒絕虛無假設的機率。這兩派的理論尚有許多不同之處，Fisher 的顯著性檢驗是根據觀察之數據訂定 *p*

值，而 Neyman-Pearson 檢驗則是在測量前訂定 α 、 β 以及所需樣本大小；雖上述兩派學者存有不同的統計考驗看法，但早期教科書的作者們竭力想弄出一個看上去客觀的統計推斷分析方法，將 Fisher 與 Neyman-Pearson 的方法合在一起，形成了虛無假設顯著性檢定（Null Hypothesis Significance Test, NHST）（Gigerenzer, 2004）。

NHST 是在 1940 年代開發的，最初的倡導者認為 Fisher 與 Neyman-Pearson 的兩種方法截然不同，但隨著文獻中越來越多地使用 p 值，且大多數關於統計假設制定的現代教科書和課程都沒有區分這些方法，這可能就是導致日後 p 值與 α 容易混淆的主因（Brereton, 2019）。

目前 NHST 仍是應用的主流，其標準的虛無假設顯著性檢定流程如下：研究者先行擬定研究問題，設定虛無（H_0）與實驗假設（H_1），確立母體分布與理論模型，並預先確定好在意的錯誤決策型態（例如，型一錯誤或型二錯誤）與可允許犯錯的機率大小（例如，型一錯誤發生的機率 $\alpha = .05$ 或.01）作為臨界值（cutoff value，或稱為「門檻值」），判斷是否達到顯著水準（significance level）。研究者接著歸納性地利用檢定分析，檢驗在虛無假設為真的前提下，收集到的樣本資料（x）所呈現的統計檢定值（例如，z、t、F 或 χ^2 值）的生成機率（$P(x|H_0)$）。當 p 值小於或等於 $\alpha = .05$ 時，研究者可以在虛無假設為真的前提下得到此檢定結果的機會非常小，代表此實驗現階段資料並不支持虛無假設，因此可拒絕虛無假設，且以目前的實驗資料得以保留實驗假設。例如，新的動畫教材經過實驗法比較後，發現實驗組的學習成效顯著高於控制組，所以研究者可以宣稱此新的教材「有效」。

因此，虛無假設顯著性檢定就是利用一個可允許被犯下錯誤決策的機率大小，作為一個武斷的臨界值來評斷 p 值的大小，通常是設在 0.05。如果觀測的實驗效果大於統計上的隨機差異，也就是 p 值小於.05，稱為統計顯著（也就是「陽性」，positive），研究者就可拒絕虛無假設，保留實驗假設；反之，稱為統計不顯著（也就是「陰性」，negative）。舉例來說，實驗組在學習前述提到的新動畫教材後，學習成效並沒有比未學習該教材的控制組高，所以研究結果無法證明此新動畫教材有效。這樣的結果對於教材設計的研究與實務可能有所助益，但就因為未得到統計顯著結果，研究者可能就此放棄報告此數據成果。

Nuzzo（2014）在其文章中指出，當 Fisher 提出顯著性檢定或 p 值時，只是單純希望以較客觀、精確的量化統計值，能協助科學家得到更為科學性的證據以作出決策與結論，加上 1920 年代末期興起了所謂「以證據為基礎的決策」（evidence-based

decision-making）風潮，要求「嚴謹及客觀的科學證據」，當時其他學者如波蘭數學家 Jerzy Neyman 與英國統計學家 Egon Pearson 也提出不同的規則導向（rule-based）資料分析架構與概念，包含統計檢定力（statistical power）、偽陽性（false positives）與偽陰性（false negatives）等（Perezgonzalez, 2015）。這幾位學者不停爭論，也互相批評對方的分析法，但是最終被後世將其學說合併寫成操作手冊給非統計學背景的研究者使用（Gigerenzer, 2004）。

參、*p* 值操弄的定義與相關名詞

在此先針對上述提及的其他相關統計概念進行名詞釋義，包含偽陽性、偽陰性、和 *p* 值操弄。在教育領域中，偽陽性是指：如果一個統計檢驗得到了 *p* 值小於臨界值（通常設定為 0.05）的結果，拒絕了虛無假設，也就是結果宣稱支持變數間有關係或研究處理有成效，但真實的情況卻是符合虛無假設，該檢驗結果則犯了型一錯誤（Type I error，其相對應發生的機率稱為α）；反之，偽陰性為：如果檢驗結果不拒絕虛無假設，而實際上真實的情況屬於對立假設，該檢驗結果即犯了型二錯誤（Type II error，其相對應發生的機率稱為β）。

所謂的 *p* 值操弄，就是研究者有意濫用或誤用資料分析方法，以便得到統計顯著結果，因此引起學界諸多討論的問題是：現今以統計結果為主要證據的論文中，可能隱含了太多型一錯誤（Ioannidis, 2005），事實是本來沒有統計顯著性，卻被研究者操弄到好像有。造成此問題的原因並不在於統計檢驗前臨界值所設定的數值大小為何，而是這數值引發了一些研究者的迷思，以為有統計顯著結果的研究就是實驗成功，實驗處理方能被宣稱為有所成效，研究才有價值。Nuzzo（2014）即言，*p* 值已經是統計檢定效度的黃金標準，但是它並不如科學家想像的可靠。的確，*p* 值常用於醫學、心理、教育、社會科學等領域，然而即使研究結果已獲得達到顯著水準的 *p* 值，研究者還是應謹記並注意，*p* 值會受到樣本數大小以及抽樣誤差的影響。在大型資料庫與教育大數據蔚為常見的浪潮下，此一問題將會因為資料的數量級遽增或是階層性複雜化而更為加劇。因此，當代教育研究者必須更加體認，*p* 值確有其侷限性（Nelson, Wooditch, & Dario, 2015）。

2013 年美國三位統計學者 Simonsohn、Nelson 以及 Simmons（2014）正式在期刊

論文中提出 *p*-hacking 這個名詞。這三位學者早在之前就提出此問題，甚至在該篇期刊論文中稱之為「研究者的自由度」（researcher degrees of freedom）（Simmons, Nelson, & Simonsohn, 2011）。他們觀察到心理學領域的研究者在實驗前，常尚未決定要收多少樣本數據，也未事先決定收到數據後的統計分析方式、怎麼處理離群值（outlier）等。研究者會探索各式的統計分析方式，甚至還自我合理化地修改數據（例如，刪除離群值）、選擇性報告（例如，不報告共變數）、反覆計算分析，直到計算出符合個人期待的 *p* 值。在這三位學者的文章中，共進行了二個研究，並用電腦數值模擬來展現一個錯誤的假設（聽披頭四的歌《當我 64 歲》就可以年輕 1.5 歲），卻可以藉著操弄樣本數、控制變項以及選擇性報告，輕易得到具有統計顯著性的偽陽性結果。

許多學者也提出一些與 *p* 值操弄相關的名詞。例如美國統計學者 Andrew Gelman 與 Eric Loken 用「岐路花園」（the garden of forking paths）來形容研究者先有數據後再選擇各種分析方法的現象（Gelman & Loken, 2013）。另外，「數據挖掘」（data mining，這個名詞現今更常用於資料探勘技術的總稱）、「數據釣魚」（data fishing）、「釣魚探索」（fishing expedition）、「數據梳理」（data dredging）、「數據屠宰」（data butchery）、「顯著追求／尋求」（significance chasing/questing）、「選擇性推論」（selective inference）等，也是泛指 *p* 值操弄的行為（Aschwanden, 2019；Nuzzo, 2014；Smith & Ebrahim, 2002）。

肆、*p* 值是如何被操弄的

哪些行為算是 *p* 值操弄？最常見的行為是先看了數據或初步分析結果之後，再決定假設怎麼寫，直到有顯著的結果出現。例如要探討學生遊戲成癮的現象，收完資料以後進行統計分析，再依據統計結果決定研究假設為男生有遊戲成癮的比例比女生高、國中生的比例比國小生高。上述情形與近年來研究界出現的另一個現象相關：HARKing（Hypothesizing After the Results are Known）相似，此由美國社會心理學家 Norbert Kerr 於 1998 年提出。這個由首字母縮寫的新名詞指的是：研究者根據統計結果才提出的假設，被佯裝為研究前就擬定的虛無假設；或是研究者在統計分析前擬定的虛無假設，因為研究產生不樂見的統計結果而在研究報告中略而不提（Kerr, 1998）。

另一種 *p* 值操弄的行為就是進行重複比較。Gelman 與 Loken（2014）指出，這發生在研究者針對同一個實驗的虛無假設，進行多次不同組的重複數據比較，或是把同

一筆數據集體加或減一個變數，又或者合併變數等方式再重複進行多次比較。例如，把同一個實驗內的幾組數據重複進行兩兩比較，跑完 20 次統計檢定後至少得到一個顯著結果，但相對的至少犯一次型一錯誤率達到了 64%（$1-(1-0.05)^{20}\approx .64$），此時，除了 Bonferroni correction 等族系錯誤率（familywise error rate）校正技術以外，變異數分析等整組實驗假設的整體性考驗（omnibus test）就成為適切的分析方法，確保預設可允許犯錯的機率不被重複比較所影響而膨脹。

其他 *p* 值操弄行為還包括變更離群值的刪除條件，例如，把分數過低或過高的學生資料刪掉。或者再多收一些數據，或是選用出現統計顯著結果的特定統計分析方法，例如不論各組樣本數不同，就選用 LSD 而不用 Scheffe 來做事後比較等。此外，當研究者處理大型資料庫或教育大數據時，若忽略了資料的數量級或是階層性，沒有將此考慮進數據分析框架之中，都會造成偽陽性的統計決策。如果研究者選擇性地報告這些有顯著的結果，而不報告其他未顯著的統計結果，這就成了所謂的「摘櫻桃」（cherry-picking）的論文了。

伍、操弄 *p* 值對學術研究的影響

值得注意的是，現在國際學界已經發現 *p* 值操弄的問題，國內也逐漸有相關討論與教學（林澤民，2016），大陸學者也有相關介紹（程开明、李泗娥，2019），可供參考。*p* 值操弄或許不算是嚴重的研究不當行為（Research Misconduct, RM），但是絕對屬於一種有問題的研究行為（QRP）（Banks et al., 2016）。

操弄 *p* 值到底對學術研究的進展會有什麼影響？首先，操弄 *p* 值會偏離學術求真求實的宗旨，違犯研究誠信中誠實、嚴謹、透明的基本原則（臺灣研究誠信守則起草委員會，2020）。本來做研究的目的是累積人類的知識以接近真理，滿足人類對未知的探索與好奇，解決學術上或實務上問題。以教育界來說，研究的目的是探索學習的本質，解決教育的學術或實務問題，理解所有成員（學校、教師、行政人員、學生、家長等）之間，以及他們與外在社會經濟環境之間的關係。不論如何，若是因為研究者需要發表以換取一些利益，因而不自覺或故意操弄假設、實驗、數據及統計方法，得到不可靠的結果，不但會讓社會大眾對教育研究者、教育研究成果失去信心，除了浪費研究資源外，也會讓這個不可靠的教育研究成果永遠留存在人類的知識庫中，讓

後世的研究者在一個不堅實甚至是錯誤的基礎上繼續研究下去。

其次，現今許多政策或標準的設立，都需要有系統地回顧過去所有相關的研究，以便得到整體的發展脈絡與結論，例如後設分析（meta-analyses），已經成為學界的黃金定律，用來彙整所有成效證據或變數的相關性，教育界也會進行各式的後設分析研究，以得知某一教育議題的整體結果，例如余民寧、翁雅云與張靜軒（2018）針對國小到高中職學生數理科學學習動機的性別差異所做的後設分析；又如廖遠光、陳政煥與楊永慈（2020）針對行動學習對學生學業成就影響的後設分析研究。後設分析結合多數相關研究的效果量估計（effect size estimate）以得出一個整體的效果量估計（overall effect size estimate），再加上樣本數較大的研究，常被認爲所得參數估計的標準誤較小，所以在後設分析計算整體效果量估計時會被給予較高的權重（例如，inverse variance weighting）（Borenstein, Hedges, Higgins, & Rothstein, 2009）。這是後設分析領域中一個基本原則。一旦有不可靠的研究成果被納入並再次分析，可能會膨脹後設分析研究結果的效果（Head, Holman, Lanfear, Kahn, & Jennions, 2015）。以教育研究來說，如果上述後設分析研究中的每一個別研究被操弄到有顯著結果，讓後設分析的結果也將顯示具有正向的成效，那麼針對學習動機性別差異、行動學習對於學業成就影響的後續研究，或是相對應的政策與標準設立，都將會做出失當的判斷。

陸、期刊如何偵測及避免 *p* 值操弄

如果 *p* 值操弄是現今學界的一個問題，造成此現象的原因，期刊也肩負部分責任。Fanelli（2012）的研究利用 Essential Science Indicators（ESI）資料庫分析了 4600 篇發表在 1991-2007 年各領域的論文，發現在這 17 年間研究結果為正向（陽性）的論文成長了 22%，並有領域與國家差異。社會科學及部分的生醫領域成長幅度較大；亞洲國家（特別是日本）的正向結果論文顯著高過美國，又高於歐洲（特別是英國）。亦有其他研究顯示，學術期刊也真的如研究者的懷疑，較少刊登沒有統計顯著的論文（Allen & Mehler, 2019；Ioannidis, 2005）。針對此種現象，Simmons 等人（2011）對期刊提出一些建議：

1. 審查人一定要確保作者都有遵循科學研究的假設陳述、實驗設計、數據收集、統計分析與詮釋等規定。

2. 審查人要能多容忍不完美、不顯著的結果；但對明顯統計檢定力不足、卻還呈現統計顯著結果的研究要多加檢視。
3. 審查人應該要求作者展示其研究結果不是取決於單一、武斷的分析決策。
4. 如果認為研究中數據收集或分析的說法不具說服力，審查人應要求作者重新再做一次相同的研究。

除此之外，Head 等人（2015）建議，期刊一定要提供清楚的投稿須知，要求論文中要報告效果大小、樣本數、要求 *p* 值到小數點後第三位等。最重要的是，期刊應該要求作者完整敘述資料分析的過程，不能只列出得到夠小 *p* 值的那一次分析結果。同時，作者也應該詳實交代資料收集規劃，例如，樣本規模、取樣框架與階層性；另外，期刊也應鼓勵作者提供原始數據供審查人審查、供其他研究者使用，以上種種作法或許可以減少 *p* 值操弄的可能性。但是國內的教育類期刊，似乎較少要求提供原始數據以供審查之用，更遑論作者將資料公開給其他研究者使用。

最後，現今期刊界亦推行「註冊論文」（Registered Report, RR），鼓勵研究者寫完研究背景、文獻、方法後就先向期刊註冊並接受匿名審查，如果被接受了，期刊會要求作者將研究資訊放在一個開放平台（例如 Center for Open Science, https://www.cos.io/），接著作者確實依照審查過的研究計畫收集資料，不必擔心結果不符假設的問題，也不會讓作者有操弄 *p* 值的需要（Chambers, 2019；Power, 2016）。但是國內的教育類期刊，似乎尚未採行 RR 的措施。

柒、研究者如何處理及避免 *p* 值操弄

因為學界已經開始意識到這個可能有疑慮的研究行為，有些學者討論是否要把臨界值從 0.05 改到 0.005（Chawla, 2017），更有些學者呼籲學界不要再做什麼顯著性檢定，不要再用 *p* 值了，這種種的反思，也值得讓教育研究者參考。但是，這個已經在學界用了近百年，被多數人視為理所當然的統計方法，不可再用的話要用什麼？的確，有些學術期刊，例如 *Basic and Applied Social Psychology*，宣稱投稿文章都不可以用 *p* 值。筆者瀏覽了這本期刊最近幾期的論文，不管是實驗法或相關性研究，果真都沒有出現 *p* 值，而是提供效果量（effect size）、信賴區間（confidence interval），整合了信心水準的區間估計等訊息，給讀者參考。

不過，美國統計學會（American Statistical Association, ASA）出面表示，還不到全面捨棄 *p* 值的時代，只是研究者需要充分了解 *p* 值，不要濫用誤用即可（Wasserstein & Lazar, 2016）。ASA 所發表之聲明（ASA Statement on Statistical Significance and *p*-Values）對 *p* 值有進一步的澄清，聲明中的六點原則提到，*p* 值可顯示數據與某個特定統計模型的不相稱程度（indicate how incompatible the data are with a specified statistical model）。用來說明數據證據與虛無假設之間的關係，但並未具有測量效果的大小或研究結果的重要性（does not measure the size of an effect or the importance of a result）。結論的決定不應只是基於 *p* 值有沒有過了特定的截結值（should not be based only on whether a *p*-value passes a specific threshold）。如果 *p* 值缺乏情境資訊（contextual information）或其他證據，則所提供推論決策的訊息有限。因此我們不能、也不應該將 *p* 值當作分析或是發表的最後一步，而是應該配合完整數據資訊、統計方法的考量與選定等細節，才能得到較全面且精確的推論。ASA 所發表之聲明亦強調報告研究結果時的完整、透明性（full reporting and transparency）（Wasserstein & Lazar, 2016）。上述原則也正與 Simmons 等人（2011）對論文作者提出之建議類似：作者應在收集數據前，決定數據收集結束的準則、列出所有收集的變數、報告所有的實驗情境，以及當數據被排除或共變數被納入時，需報告未排除或未納入的統計結果等。

除了注意這些研究過程中的資料處理與統計細節，Head 等人（2015）建議，如果可能，在「盲目」（blind）的情況下做統計分析（e.g., blind analysis）（MacCoun & Perlmutter, 2015）。例如不知道哪些是控制組或實驗組的數據，或許也是一個減少 *p* 值操弄的方式，至於這種「盲目」處理能不能實施在教育研究的資料分析階段中，還需要思考。Head 等人亦建議，只要研究者並非故意涉入有問題的研究行為（QRP），只是誤用了統計方式去處理資料，則表示這些研究者對方法、統計的知能不足，需要用教育的方式去解決。而 Greenland 等人（2016）的文章，對避免誤用統計檢定、*p* 值、信心區間、統計考驗力等都有詳細說明，可以作為教育研究者學習的起點。

捌、教育研究者如何不依賴 NHST 或 *p* 值

Wasserstein 與 Lazar（2016）在導讀 ASA 聲明時，提及 Mount Holyoke College 的退休名譽教授 George Cobb 在 2014 年於 ASA 討論區貼了一篇問與答的文章，大意

是：為什麼大學及研究所都在教 $p = .05$？因為科學界及期刊主編都在用。為什麼這麼多人都在用 $p = .05$？因為他們在學校都學這個。此問答雖莞爾，卻讓我們在了解 p 值，深入知道 p 值操弄及其後果之後，可以更進一步的來了解可以避免操弄的實務作法。

因應教育研究的創新需求，量化研究法社群應時刻反省、精進分析與研究法的技術與規範。自 NHST 將 p 值作為一項基於分配假設的理性證據開始，反對這項作法的聲音就不曾削弱。近百年來，各方研究者依循著科學求真求實的原則，持續發出異議，提供不同做法或替代方案。目前最廣為人接受的做法，就是要求研究者在提供 p 值與 NHST 等「統計顯著性」證據的同時，也必須提供「實務顯著性」的證據，例如前述常見 p 值操弄的重複比較手法，可以透過變異數分析等整體性考驗（omnibus test）提供 F 統計值與 p 值，有效避免整組實驗虛無假設檢定 p 值因重複比較而失真。此時，再搭配效果量（effect size）（Cohen, 1988）如 η^2，*partial* η^2 *或* ω^2 等實務性顯著指標，或如 Nuzzo（2014）建議，在論文中給數據更多的描述，例如報告「效果量的信賴區間」（Thompson, 2002），不要只說 p 值有顯著或沒顯著，將可提供統計決策更完備的量化證據。效果量的定義為「差異、效果或是關聯的強度」（Snyder & Lawson, 1993），包含常用的 Cohen's d、Glass's *delta*、以及常用的判定係數 R^2、相關係數 r 或是標準化迴歸係數等。而信賴區間即為整合了信心水準的區間估計。例如研究者在執行多元迴歸分析時，可以回報標準化迴歸係數的信賴區間（Jones & Waller, 2013），作為結合效果量與信賴區間的更完備報告。這些數值較廣為研究社群熟悉，也都漸漸包含在各層級研究人員養成的系列統計教學內容裡。換言之，統計分析結果除了利用 p 值檢驗所提出的假設是否通過特定的截結值外，另外提供信賴區間與效果量等統計值，更能協助讀者有信心地全面檢視所估測結果的實際效應（practical significance）。

但是在推動使用這些數值的研究發表報告會比較容易嗎？以在教育研究領域常用的 APA（American Psychological Association）研究發表格式為例，到 1994 年第四版才第一次出現「鼓勵使用效果量」（encourage reporting effect sizes）字句，直到 2001 年第五版才增強為「建議使用效果量」（recommend to use effect sizes），並加註警語「如果沒有呈現效果量，那是一種研究設計與發表報告的缺失（defect）」，但卻因為此版手冊中缺乏完整報告例子而飽受抨擊。時至今日，報告效果量已漸成常態（Pek & Flora, 2018）；不過，這時熟悉統計的教育研究者就會發現，無論是效果量、信賴區間等等作法，本質上都還是與 p 值和 NHST 有緊密關係。所以，以上做法仍然無法完全杜絕 p 值操弄等「有問題的研究行為」。

因此，完整嚴謹地報告科學研究的研究設計與數據分析細節，確保研究結果的「可重製性」（scientific reproducibility）是唯一關鍵的誠信原則。例如分析大型資料庫資料或教育大數據時，研究者需詳細報告取樣框架（sampling scheme），說明資料檔的樣本數量級與取樣分層，也澄清排除樣本未被選取的原因；也要詳細陳述分析架構（analytical framework），切實使用取樣權重（sampling weight）、重複抽樣加權值（replicate weight）等複雜取樣分析應使用的校正技術（Lee & Wu, 2012, 2013）。研究者可進行多階層線性模型（Hierarchical Linear Modeling, HLM）（Raudenbush & Bryk, 2002）或多階層結構方程式模型（Multilevel Structural Equation Modeling）（Mehta & Neale, 2005）等建構，將不同取樣單位的變項關係明確地以不同階層方式檢驗，藉以得到不同層級的模型呈現與參數估計；或在進行多階層的模型建構時加入取樣權重，以得到更接近母體參數的不偏估計值（Asparouhov, 2006；Rabe-Hesketh & Skrondal, 2006）。如果是統計模擬（statistical simulation）研究，除了報告模擬實驗的詳細設定外，需檢驗不同資料規模的數量級對參數估計、統計決策的影響，也要呈報效果量等資訊。初始設定值（例如 seed）也會影響結果，因此建議進行重複模擬實驗，例如回報多個 seed 所產出的總整性分析結果，或是設定 random seed 選項等（Morris, White, & Crowther, 2019）。

更重要的是，所有奠基於分配假設的統計分析，都必須在符合其假設前提之下，才能產出一致、不偏頗的分析結果。因此，在未檢驗所用分析技術所需假設是否被滿足前，研究者不可盲目地使用軟體預設值逕自進行分析。研究者必須根據分析假設來檢驗結果，理性地選擇合適分析技術，以獲得一致不偏的估計值、p 值、效果量與區間估計。不然，就算是報告了多種的統計值，所得的數值可能仍然是偏誤的。例如，若線性模型分析的常態性假設被拒絕了，研究者可以採用強健式參數與標準誤估計器（robust parameter and standard error estimator）（Wu & Kwok, 2012）。如果取樣獨立性假設被違犯了，那多階層建模（multilevel modeling）（Wu, Lin, Nian, & Hsiao, 2017）或是重複取樣分析技術（repeated measure techniques）（Liang & Zeger, 1986；Wu, Kwok, & Willson, 2014）就是必須採用的分析技術。如果研究者發現手中的資料很難符合諸多分配假設，那使用無母數分析（nonparametric analysis），或是拔靴法（bootstrap method）（Efron & Tibshirani, 1993），這些都是較能得到強健統計決策結果的分析技術。

玖、結語

的確，多數研究者都習得以 *p* 值判讀教育量化研究結果是否具備統計顯著性。然而，除非是統計專業的研究者，很少有人會去懷疑其研究結果的侷限性與潛藏的危機。Aschwanden（2019）提醒我們，hacking 這個字聽起來有點像壞人做的不道德的事，以至於研究者可能會覺得自己沒有在欺騙他人，不會做這種壞人才做的事。操弄 *p* 值可視為人類行為的一種：我們傾向去找證據支持我們相信的東西是存在的，然後對不支持的證據視而不見。但是，在教育研究的範疇中，研究者即便是無意地操弄 *p* 值，也不能免除犯下數據資料不當分析的行為瑕疵。

現今教育系統與現場的數位轉化，研究者取得大型資料庫或教育大數據等資料的機會大幅增加，並使用量化方式數據化再現當代學習者於多元教育現場的複雜學習與適應樣態。本文藉此機會，在簡述了 *p* 值及 NHST 的歷史與學理之後，詳細地呈現了 *p* 值操弄的定義、相關名詞、手法，更針對資料數量級與階層性等衍生問題進行討論。同時，本文也提出了教育類期刊出版界與研究者自身可以努力避免的方向，更提醒了研究社群應有的正確認知及實務作法。本文認為，如果我們暫時無法拋棄 NHST、還是要靠顯著性檢定去分析資料，就必須秉持科學研究的嚴謹態度，遵從本文所列的諸項專家建議，並對 *p* 值應用有更深一層的素養與反思：理解它、善用它，不要誤用它，更不要操弄它。如前所述，即便 *p* 值操弄不見得就違反了學術倫理，但這是一種必須盡力避免的有問題研究行為。

如本文之介紹，不用 NHST 或 *p* 值，或是在呈現「沒有效果」假設為否、「效果存在」的統計顯著證據時，研究者也需一併報告「實務效果重要與否」的實務顯著證據，並詳實呈現研究設計、數據分析完整細節與結果。此外，如同前文所述，除了 NHST 之外，我們還有很多統計技術可以多加學習並且合理運用。身為研究者，我們應終生記住，研究是為了求真求實，讓人類文明能夠奠基於具可重製性的嚴謹科學研究之上向前邁進，而操弄 *p* 值，恰好是反其道而行。

本文期收拋磚引玉之效，期待更多教育研究學者研發更多統計技術，以及更具科學客觀性與嚴謹性的分析判讀指標，以增進教育量化研究的適用性與精準性。也期待更多統計教育者開發相關課程，推廣並提升研究者有關 *p* 值操弄的基礎認知，反思背

後潛藏的問題，精進複雜取樣分析可使用的校正作法，不但讓教育研究品質提升，更具有學術或實務價值，若成果能貢獻給社會國家，必定能讓我們的教育更美好。

誌　謝

本研究感謝國家科學及技術委員會專題研究計畫補助。計畫名稱：邁向研究者之路：臺灣學生學術倫理與研究誠信之虛實整合課程的研發與實施（MOST110-2511-H-A49-008-MY4）、探研數位轉化之混成式學習場境內的教與學（MOST111-2410-H-A49-066-MY3），謹誌謝忱。

參考文獻

余民寧、翁雅芸、張靜軒（2018）。數理科學的學習動機有性別差異嗎？一個來自後設分析的證據。**當代教育研究季刊，26**（1），45-75。doi: 10.6151/CERQ.201803_26(1).0002

[Yu, M. -N., Weng, Y. -Y., & Chang, C. -H. (2018). Students' learning motivation to math and science: Using the meta-analysis to find the gender difference in Taiwan. *Contemporary Education Research Quarterly, 26*(1), 45-75. doi: 10.6151/CERQ.201803_26(1).0002]

林澤民（2016）。看電影學統計：*p* 值的陷阱。**社會科學論叢，10**（2），I-XXV。

[Lin, T. -M. (2016). The pitfalls of *p*-values. *Review of Social Sciences, 10*(2), I-XXV.]

程开明、李泗娥（2019）。科學研究中的 *p* 值：誤解、操縱及改進。**數量經濟技術經濟研究，7，**117-136。

[Cheng, K., & Li, S. (2019). *p*-value in scientific research: Misunderstanding, *p*-hacking and improvement strategy. *The Journal of Quantitative & Technical Economics, 7,* 117-136.]

廖遠光、陳政煥、楊永慈（2020）。行動學習對臺灣學生學業成就影響之後設分析。**當代教育研究季刊，28**（3），67-102。doi: 10.6151/CERQ.202009_28(3).0003

[Liao, Y. -K., Chen, C. -H., & Yang, Y. -C. (2020). A meta-analysis of the effects of mobile learning on students' academic achievement in Taiwan. *Contemporary Educational Research Quarterly, 28*(3), 67-102. doi: 10.6151/CERQ.202009_28(3).0003]

臺灣研究誠信守則起草委員會（2020）。**臺灣研究誠信守則**。台北：台灣聯合大學系統。

[The Drafting Committee of the Taiwan Code of Conduct for Research Integrity. (2020). *Taiwan code of conduct for research integrity.* Taipei: University System of Taiwan.]

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology, 17*(5), e3000246. doi: 10.1371/journal.pbio.3000246

Aschwanden, C. (2019). *We're all "p-hacking" now.* Wired. Retrieved from https://www.wired.com/story/were-all-p-hacking-now/

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics-Theory and Methods, 35*(3), 439-460. doi: 10.1080/03610920500476598

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31,* 323-338. doi: 10.1007/s10869-016-9456-7

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. West Sussex, UK: John Wiley & Sons.

Brereton, R. G. (2019). The use and misuse of *p* values and related concepts. *Chemometrics and Intelligent Laboratory Systems, 195,* 103884. doi: 10.1016/j.chemolab.2019.103884

Chambers, C. (2019). What's next for registered reports? *Nature, 573,* 187-189. doi: 10.1038/d41586-019-02674-6

Chawla, D. S. (2017). "One-size-fits-all" threshold for *p* values under fire. *Nature*. doi: 10.1038/nature.2017.22625

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap.* Boca Raton, FL: Chapman and Hall/CRC.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891-904. doi: 10.1007/s11192-011-0494-7

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver & Boyd.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University, 348*. Retrieved from https://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102,* 460-465.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587-606. doi: 10.1016/j.socec.2004.09.033

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31,* 337-350. doi: 10.1007/s10654-016-0149-3

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*(3), e1002106. doi: 10.1371/journal.pbio.1002106

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi: 10.1371/journal.pmed.0020124

Jones, J. A., & Waller, N. G. (2013). Computing confidence intervals for standardized regression coefficients. *Psychological Methods, 18*(4), 435-453. doi: 10.1037/a0033269

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. doi: 10.1207/s15327957pspr0203_4

Lee, Y. -H., & Wu, J. -Y. (2012). The effect of individual differences in the inner and outer states of ICT on engagement in online reading activities and PISA 2009 reading literacy: Exploring the relationship between the old and new reading literacy. *Learning and Individual Differences, 22*(3), 336-342. doi: 10.1016/j.lindif.2012.01.007

Lee, Y. -H., & Wu, J. -Y. (2013). The indirect effects of online social entertainment and information seeking activities on reading literacy. *Computers & Education, 67,* 168-177. doi: 10.1016/j.compedu.2013.03.001

Liang, K. -Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22. doi: 10.1093/biomet/73.1.13

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature, 526*(7572), 187-189. doi: 10.1038/526187a

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259-284. doi: 10.1037/1082-989X.10.3.259

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074-2102. doi: 10.1002/sim.8086

Nelson, M. S., Wooditch, A., & Dario, L. M. (2015). Sample size, effect size, and statistical power: A replication study of Weisburd's paradox. *Journal of Experimental Criminology, 11*(1), 141-163. doi: 10.1007/s11292-014-9212-9

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231,* 289-337. doi: 10.1098/rsta.1933.0009

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506,* 150-152. doi: 10.1038/506150a

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5, 50*(302), 157-175. doi: 10.1080/14786440009463897

Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208-225. doi: 10.1037/met0000126

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6,* 223. doi: 10.3389/fpsyg.2015.00223

Power, A. (2016). Registered reports: What are they and why are they important? *The Royal Society.* Retrieved from https://royalsociety.org/blog/2016/11/registered-reports-what-are-they-and-why-are-they-important/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(4), 805-827. doi: 10.1111/j.1467-985X.2006.00426.x

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Sijtsma, K. (2016). Playing with data-or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*(1), 1-15. doi: 10.1007/s11336-015-9446-0

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366. doi: 10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. doi: 10.1037/a0033242

Smith, G. D., & Ebrahim, S. (2002). Data dredging, bias, or confounding [Editorial]. *British Medical Journal, 325,* 1437-1438. doi: 10.1136/bmj.325.7378.1437

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education, 61,* 334-349. doi: 10.1080/00220973.1993.10806594

Steneck, N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics, 12,* 53-74. doi: 10.1007/PL00022268

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31. doi: 10.3102/0013189X031003025

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. doi: 10.1080/00031305.2016.1154108

Wu, J. -Y., & Kwok, O. (2012). Using structural equation modeling to analyze complex survey data: A comparison between design-based single-level and model-based multi-level approaches. *Structural Equation Modeling-A Multidisciplinary Journal, 19*(1), 16-35. doi: 10.1080/10705511.2012.634703

Wu, J. -Y., Kwok, O., & Willson, V. L. (2014). Using design-based latent growth curve modeling with cluster-level predictor to address dependency. *The Journal of Experimental Education, 82*(4), 431-454. doi: 10.1080/00220973.2013.876226

Wu, J. -Y., Lin, J. J. H., Nian, M. -W., & Hsiao, Y. -C. (2017). A solution to modeling multilevel confirmatory factor analysis with data obtained from complex survey sampling to avoid conflated parameter estimates. *Frontiers in Psychology, 8,* 1464. doi: 10.3389/fpsyg.2017.01464

投稿收件日： 2023 年 02 月 24 日
第 1 次修改日期： 2023 年 05 月 22 日
第 2 次修改日期： 2023 年 12 月 19 日
接受日： 2023 年 12 月 23 日