為研究計量指標而訂的萊登宣言

Diana Hicks, Paul Wouters 及其同儕力促採用十項原則以引導學術研究評估

近來治理科學相關事務時，越來越倚重各項數據，曾經以同儕審查為主的學術研究評估如今已慣於使用並依賴計量指標[1]。由此而衍生的問題在於，現今的評估乃由數據所主導而非判斷力。數量激增的量化指標通常經過精心設計，但卻總是無法被完整理解，甚至也經常被誤用。我們使用了原先設計來改善此現象的評估工具，但卻可能冒著傷害學術研究體系的風險，此舉就如同在缺乏對量化指標實踐及詮釋的知識與建議下，仍有越來越多機構執行學術研究評估。

在2000年之前，專家們利用 Institute for Scientific Information （ISI） 發展的 CD-ROM 形式 Science Citation Index 資料庫進行計量分析。2002年Thomson Reuters 發表整合性的網路平台，使得 Web of Science 資料庫的取用更廣泛。之後有其他引用索引競爭者成形，包括 Elsevier 的 Scopus （發表於2004年），與 Google Scholar （beta版發表於2004年）。以網路為基礎的工具日益增加，機構間研究生產力與影響力的比較也更容易進行，例如 InCites（使用 Web of Science 的資料）與 SciVal （使用 Scopus 的資料），或利用 Google Scholar 的資料分析個人引用表現的軟體 （發表於2007年的 Publish or Perish）。

2005年，美國加州州立大學聖地牙哥分校的物理學教授 Jorge Hirsch 提出*h*-index，使得被引用次數的計數廣泛應用於個別研究者的評鑑。1995年之後，對期刊影響係數的關心程度也不斷成長（見 Nature 原文 p.431 圖 ‘Impact-factor obsession’）。

晚近以來，與社會使用及網路上評論相關的量化指標也日益受到重視，例如設立於2002年的 F1000Prime、2008年的 Mendeley，Altmetric.com （由 Nature Publishing Group 所屬的 Macmillan Science and Education 所支持）則於2011年成立。

身為科學計量學者、社會科學家以及研究管理者，我們目睹量化指標在科學研究表現評估中被誤用的情形越來越普遍，以下是一些例子。全世界的大學日益執迷於自己在世界大學排名中的位置（例如上海交通大學的世界大學學術排名以及Times 的高等教育排名），即使這些排名在我們看來是建立在不精確的數據資料以及武斷的指標之上。

有些招聘者要求應徵者提供其 *h*-index 的數值，部分大學以 *h*-index 的數值以及有多少篇論文被刊登於「高影響力」期刊做為能否升等的門檻。特別在生物醫學領域，有些研究人員甚至在自己的履歷表中誇耀這些數據。指導教授們在博士生們準備好前，就要求他們在具有高影響力的期刊上發表論文與取得更多研

究經費的例子更是隨處可見。

在北歐與中國，有些大學根據數據來分配研究經費或獎金，例如計算研究人員個人的影響係數分數來分配「績效資源」，或當研究人員在影響係數超過15的期刊上發表論文時，即發給獎金。（參考文獻 2）

雖然在諸多個案中，研究人員與評估者仍然能找到相對平衡的判斷力，但對於研究計量指標的濫用已經太過普遍且不容忽視。

我們因而提出萊登宣言，這個名稱源於形成本宣言的會議（見 http://sti2014.cwts.nl）。這十個原則對科學計量學者而言並非前所未聞，但我們之中並沒有人能夠詳述所有原則，因為至今仍缺乏完整的成文論述。我們領域中的啟蒙者，例如 Eugene Garfield （ISI 的創辦人），曾提出部分原則[3,4]，但是評估者與大學的主管人員均非相關研究方法論的專家，自然不可能理解這些概念。被評估的科學家試圖搜尋相關文獻以駁斥評估結果，卻因其散落於各處常不得其門而入。

我們提出這些經過精鍊後的原則，以做為應用計量指標進行研究評量的最佳實務典範，如此一來研究人員可問責於評估者，評估者亦可規範計量指標的使用。

十個原則

1. 量化評估應支援質化評估與專家評量

量化指標能挑戰同儕審查中的偏見，並促使更謹慎的審議。量化指標能強化同儕審查機制，因為在沒有足夠相關資訊的情況下，要評論同儕總是艱鉅的任務。然而，評估者絕對不能讓數字主宰其決策，指標絕對不能取代基於充足資訊所形成的判斷力，每位評估者皆須為其評量負責。

2. 衡量績效表現需基於機構、團體或研究者的使命

在評鑑起始之際即先載明計畫的目的，且所選用的評估指標應與計畫的目的有明確關聯。指標的選擇與使用應顧及更廣泛的社會經濟與文化脈絡，科學家有其多元的研究使命，著重學術知識前沿的研究有別於關注在提供各種社會問題之解決方案的研究。評鑑有時是選用與政策、產業或公眾有關的指標，而非僅重視學術思想的卓越性。沒有任何單一的評估模式可適用於所有的情境。

3. 保護與在地相關的卓越研究

在這個世界的許多地方，卓越的研究等同於發表英文論文。舉例而言，西班牙的法律明文鼓勵西班牙學者在高影響係數的期刊發表論文，而計算影響係數的數據來源為收錄於 Web of Science 且以美國與英文為主的期刊，這些偏誤尤其對著重區域與本國研究的社會科學及人文學科造成嚴重的影響。許多其他學科領域也有國家或區域層面的議題，例如在撒哈拉以南的非洲愛滋

病流行病學研究。

這種多元化與社會相關的議題往往不為擁有高影響係數的英文期刊之編輯或審查者所青睞，在 Web of Science 中取得高被引次數的西班牙社會學學者多以抽象的模型或美國的數據做研究。特定的社會學家撰寫與當地的勞動法、高齡者家庭健康照護或移工就業[5]等議題相關的論文，這些真正有高影響力的西班牙文論文卻被忽略。使用基於高品質且非英文文獻為衡量基礎的指標，才能有助於識別並獎勵與本地相關的卓越研究。

4. 數據蒐集與分析的過程應為公開、透明與簡潔

用於評估的資料庫，在建置過程中應遵守明確的規則，而這些規則也應在研究評估之前即清楚地闡述。這是在過去數十年中，學術界與商業機構在建構書目計量學評估方法論時的慣例，而這些處理數據的準則與流程亦發表於經同儕審查的期刊文獻中。這樣透明的過程可確保被再度檢驗的可行性。舉例來說，在2010年，荷蘭萊登大學科學技術研究中心（CWTS）所創建的一項指標，其技術特性引發了公開的辯論，導致該項指標的計算方式隨後被修改[6]。該領域的新進商業機構也應遵照此項標準；任何人都無法接受黑箱作業的學術研究評估。

對指標而言，簡潔就是美，因為可保有其透明性。但簡單的指標可能會導致偏頗的結果（詳見原則7）。評估者必須在選用簡潔的指標與忠於複雜的研究程序中取得平衡。

5. 允許被評估者檢驗數據與分析

為確保數據的品質，所有被列入於書目計量研究中的研究人員都必須能查驗其研究產出是否均被正確收錄。任何人在進行或管理研究評估的過程時，應透過自我的檢驗或是第三方團體的檢核，確保數據的正確性。大學在建置其研究資訊系統時，可參考此項指導原則以做為系統的選擇依據。精確、高品質的數據往往需要花時間與金錢蒐集與處理，故應編列足夠的預算以支應開銷。

6. 留意論文發表數量與被引用次數在不同領域上的差異

最佳實踐應該是挑選出一套合適的指標讓各領域可各取所需。幾年前，一群歐洲的歷史學家們在全國性的同儕審查評量中，得到相對較差的評比結果，原因在於他們的主要研究產出為書籍，而非在被 Web of Science 索引的期刊上發表論文。另外的原因則是這群歷史學家很不幸地被歸類在心理學科系。因此，在研究評估中，歷史學家與社會科學家往往會要求在計算產出時將書籍與使用本國語言撰寫的文獻列入；電腦科學家則會要求列入研討會論文。

引用率也會因為領域而有所差異：在數學領域，頂級期刊的影響係數在 3 左右，細胞生物學領域的頂級期刊影響係數則可高達30。因此，在計算過程中，有必要將指標標準化，而最常使用的標準化方法則是將其轉換為百分比：例如

每一篇論文依據於其所屬領域的被引用次數分布計算權重，檢視該篇論文的被引用次數位於該領域的前百分之幾（例如前百分之一、前百分之十、前百分之二十等）。使用以百分比為計算基礎的指標時，一篇高被引論文可能會略微提高該大學在排名系統的位置，但在以平均被引用次數的指標進行排名時，卻可能使該大學的排名從中等位置躍升至頂尖排名上[7]。

7. 個人層級的研究評量基礎應是對其研究成果歷程檔案進行質化的判斷

即便在沒有新論文產出的情況下，當研究者的年紀愈大，其 *h*-index 的數值可能也會持續提高。*h*-index 也會因領域不同而產生差異：例如生命科學家的 *h*-index 可高達200、物理學家最高可達100，但社會科學家卻只有20~30（參考文獻 8）。另外，該數值的差異也取決於資料庫的選擇，有些電腦科學領域的學者，若使用 Web of Science 進行計算，其 *h*-index 約為10，但若選用 Google Scholar 進行計算，其 *h*-index 卻變為20~30之間[9]。解讀與評斷研究人員的論文會遠比僅使用單一指標分析後的數值來得更為適當。縱使在比較多位研究人員的表現時，若能綜合考量他們的專長、經驗、活動與影響力等，亦會較為適當。

8. 避免錯誤的具體性與虛假的精確性

科學與技術指標容易在概念上讓人感到模糊且充滿不確定性，需要仰賴強而有力的假設，但卻不見得能獲得普世認同。舉例來說，對於被引用次數的定義，各方長久以來一直爭論不休。有鑑於此，最佳實務典範應為使用多項指標以提供更穩健可靠且更為多元的呈現方式，以說明研究評估後的結果。如果不確定性與誤差可以被量化，則應在公佈評估結果時，同時公佈誤差線作為附加資訊以供參考。若無法測量不確定性與誤差，則計算指標時至少應避免追求虛假的精確性。舉例來說，期刊的影響係數應計算至小數點後第三位，以避免期刊之間形成平手局面。然而考慮到被引用次數的概念模糊性與隨機誤差，實則不必要在很小的影響係數差異上區分期刊之間的優劣。為避免虛假的精確性，計算時僅需採用小數點後第一位。

9. 體認評量和指標所造成的系統性影響

評估指標的改變，影響了研究人員的動機，進而造成研究系統的改變。對於這樣的影響應該是可預期的，而這也意味著整套評估指標總是較佳的選擇，單一指標更易被操弄，且會促使原先欲作為衡量的項目轉變成為研究者追逐的目標。舉例來說，在1990年代，澳洲政府採用以論文發表數量為基礎的計算公式，作為分配經費的依據，故各大學即可估算出在期刊上發表一篇論文的「價值」：在2000年，一篇論文可取得澳幣800元（當時相當於美金480元）的研究補助經費。在此情況下，可預期的是澳洲研究論文數量大幅成長，但這些論文卻是發表在較少被人引用的期刊上，此現象也意味著論文品質的下降[10]。

10. 定期審視與更新指標

研究的使命與評量的目標會隨著研究系統的演進而一起改變。曾經為合適的指標也有可能會變得不妥當，同時，新的指標也會應運而生。因此，應隨時檢視且修正指標系統。舉例來說，在體認到簡單公式所帶來的影響後，澳洲政府於2010年推出更為複雜的卓越研究評估系統，且該系統更強調質化的評估方式。

下一階段

遵循上述十項原則，學術研究評估可在科學發展與社會互動上扮演更為重要的角色，研究指標可以提供那些難以取得或需透過個人專業知識方可判斷的關鍵訊息。但必須謹記，不可讓這些量化資訊從評估工具轉變為評估的目標。

最好的決定往往是結合穩健且對目標敏感的統計結果，與瞭解被評估的研究本質之下的產物。量化與質化的證據皆有其必要，且二者必須皆可客觀呈現。有關科學事務的決策必須基於高品質的評估過程，高品質的評估過程則是建立在最高品質的數據之上。

作者介紹

Diana Hicks是位於美國喬治亞州亞特蘭大的喬治亞理工學院下公共政策學院的教授。Paul Wouters為荷蘭萊登大學科學與技術研究中心（CWTS）的教授兼主任，Ludo Waltman亦是該中心的研究員，Sarah de Rijcke則是該中心的助理教授。Ismael Rafols是西班牙國家研究委會會和瓦倫西亞理工大學的科學政策研究員。通訊作者之聯絡 e-mail：diana.hicks@pubpolicy.gatech.edu

翻譯者介紹

林雯瑤，淡江大學資訊與圖書館學系，副教授。（Wen-Yau Cathy Lin is an associate professor of Department of Information and Library Science in Tamkang University, Taiwan）

陳明俐，財團法人國家實驗研究院科技政策研究與資訊中心，助理研究員。（Carey Ming-Li Chen is an assistant researcher of Science & Technology Policy Research and Information Center, National Applied Research Laboratories, Taiwan）

參考文獻

1. Wouters, P. in Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact (eds Cronin, B. & Sugimoto, C.) 47–66 (MIT Press, 2014).
2. Shao, J. & Shen, H. Learned Publ. 24, 95–97 (2011).
3. Seglen, P. O. Br. Med. J. 314, 498–502 (1997).
4. Garfield, E. J. Am. Med. Assoc. 295, 90–93 (2006).
5. López Piñeiro, C. & Hicks, D. Res. Eval. 24, 78–89 (2015).

6. van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J. & Waltman, L. J. Informetrics 4, 431–435 (2010).
7. Waltman, L. et al. J. Am. Soc. Inf. Sci. Technol. 63, 2419–2432 (2012).
8. Hirsch, J. E. Proc. Natl Acad. Sci. USA 102, 16569–16572 (2005).
9. Bar-Ilan, J. Scientometrics 74, 257–271 (2008).
10. Butler, L. Res. Policy 32, 143–155 (2003).