・公共行政學報・　第六十九期
〈研究論文〉
民114年9月　頁81-113
國立政治大學公共行政學系
Doi: 10.30409/JPA.202509_(69).0003

專家異質性評估：基於臺灣公共治理主觀指標調查的檢證*

莊文忠、林美榕、張順全**

《摘要》

有鑑於許多公共議題具有高度專業性，一般民眾難以熟悉和提供具體的意見或評價，因此，許多國際組織採用的公共治理調查，多依賴菁英或專家來獲取各種評分或建議，我國行政院國家發展委員會自 2008 年起，委外執行的臺灣公共治理主觀指標調查亦是如此。由文獻檢閱可知，目前仍少有實證研究就專家異質性對調查結果的影響進行評估，使我們對專家調查的測量品質了解相對有限。因之，本研究以 2012 年和 2013 年臺灣公共治理指標專家調查結果為例，利用 Q-Q 圖、蒙地卡羅模擬，以及多種統計檢定等量化分析方法，從方法論的角度探討：1. 無反應與趨中選答的專家答題模式；2. 隨機專家假設的抽樣模擬和假設檢定；3. 不同專業領域

投稿日期：113 年 9 月 30 日；接受刊登日期：114 年 2 月 28 日。

* 論文初稿發表於 2018 年台灣公共行政與公共事務系所聯合會年會暨「變動中的公共行政：新價值、新議題與新挑戰」國際學術研討會，臺北市：東吳大學政治學系，2018/06/2-3。不過，本文在投稿前已大幅修正初稿內容，本文作者除感謝行政院國家發展委員會（前身為「行政院經濟建設委員會」與「行政院研究發展考核委員會」）長期支持本項調查計畫外，亦感謝研討會評論人及兩位期刊審查人給予諸多寶貴的修正意見，惟一切文責由本文作者自負。

** 莊文忠為國立中正大學政治學系教授，email: wenjong@ccu.edu.tw
林美榕為淡江大學國際企業學系副教授，email: 134660@mail.tku.edu.tw
張順全為馬偕醫學大學全人教育中心副教授，email: zhang@mmu.edu.tw（通訊作者）

的專家異質性比較。

研究結果顯示，適度減少專家的回答量（如七大面向中允許避答三個以內）並不影響推論結果，尤其學界代表的回覆普遍呈現較高同質性。此外，不同參與次數的專家在「無反應」與「中間值」的填答模式上亦展現一致性。本研究不僅有助於理解專家異質性對公共治理評估成效的影響，亦提出專家調查設計的改進建議，包括：針對首次參與者舉辦調查說明會，以協助理解調查目的與釐清較長問卷內容；在受訪者配置上，平衡政府內、外部專家比例，以降低偏誤；適度減少學界專家樣本並增納企業與公民社會的多元觀點，以提升推論結果的包容性；專家填答方式則可透過隨機減少部分必答面向，或依據其專長分配作答，仍可維持整體評估品質之穩定性。

[關鍵詞]：公共治理、專家異質性、隨機專家假設、蒙地卡羅模擬、Q-Q圖

壹、前言

在「世界銀行」（World Bank）提出「公共治理」（public governance）的概念，並將理想化的「善治」（good governance）定義為「具有可預測性、開放的和啟發性的決策；具有專業主義的官僚體系；政府作為可被課責；以及公民社會積極參與公共事務；所有行為遵守法治」（World Bank, 1994, pp. 7），其後，公共治理漸受矚目且引起各界廣泛討論，多個國際組織也先後投入建立評估公共治理的衡量指標體系，其中著名的除了世界銀行的「全球治理指標」（Worldwide Governance Indicators, 以下簡稱 WGI）外，還包括「國際透明組織」（Transparency International）的「政府清廉印象指數」（Corruption Perceptions Index, 以下簡稱 CPI）、「自由之家」（Freedom House）的「政治權利」（political rights）及「公民自由度」（civil liberties），瑞士洛桑國際管理學院（International Institute for Management Development, 以下簡稱 IMD）的《世界競爭力年報》（World Competitiveness Yearbook）、「世界經濟論壇」（World Economic Forum, 以下簡稱 WEF）的《全球競爭力報告》（Global Competitiveness Report）等，均成為各國

競爭力的重要評比依據。

上述指數通常是由數個面向所構成的評估體系，這些指數的評估結果不僅用於政府部門，也常常被國際援助或捐助者用於分配援助計畫的金額，並成為跨國企業投資某一國家時的風險評估參考依據（Peltier, 2010, pp. 31）。在調查方法的運用方面，對跨國比較研究來說，這些國際組織較為偏好調查專家而非民眾的意見，其主要原因在於，專家乃是學有專精之士，他們所接受的訓練是強調依證據說話，較能夠摒除個人好惡而提供客觀意見，因此，許多跨國性調查的主要研究對象大都是以產、官、學界的專家為主。於其中，各國研發衡量指標的研究團隊莫不主張應儘量讓專家的組成更具「異質性」（heterogeneity），以納入各種潛在的差異觀點（Kaufmann et al., 2007；陳文學等人，2013）。

事實上，許多領域的科學研究和政策制定都非常依賴專家所提供的專業知識和判斷，特別是在風險評估、決策科學、作業可靠度、反恐以及地緣政治等領域的決策過程中，專家意見已被證實發揮了相當重要的作用（Cederman & Weidmann, 2017; Rios & Rios Insua, 2012; Cooke & Goossens, 2008）。在這些領域中，由於可能存在成本和時間方面的考量，或是研究問題與關注事件的特性（例如破壞性），導致蒐集大數據變得不切實際；另一方面，樣本個案的資料缺失或不足也會導致風險評估和判斷不佳，從而妨礙設計可靠的決策流程或協助政策制定者做出明智的決策。因此，如何另闢蹊徑，整合有限數量的不同專家的意見，以正確描述事件的不確定性，成為一個重要的替代方案（Cooke & Goossens, 2008）。

不過，Haselswerdt 和 Rigby（2021）也指出，來自工會、商業團體、智庫以及關注低收入家庭之公民團體等領域的專家（或菁英），這些政策倡議者在連結學術研究與政策制定中發揮了關鍵作用，在一項針對這些倡議者所進行的研究，考慮了美國社會自由派與保守派不同倡議者的觀點，舉例詢問這些倡議者對美國某州最低工資政策的看法及其可能產生的影響，以檢視他們對政策研究的偏好。其研究結果發現，倡議者對於政策領域中不熟悉結果的新穎研究並不特別感興趣，而是更傾向於維持他們所從事的工作中主導相關政策問題的熟悉框架。此外，自由派和保守派的倡議者較重視能證明他們自身政策立場的研究（Haselswerdt & Rigby, 2021），這也反映了專家調查可能存在的偏頗性或偏執性。

職是之故，整合專家的評估或判斷之所以具有挑戰性，主要源自兩個方面。一方面，如何界定專家的範圍，一些外顯變數例如專家來自特定領域或行業的因素，屬於「可直接觀察到的專家異質性」（observed heterogeneity），即不同領域的專

家所反映的何種差異最令學術研究者矚目？另一方面，則來自於專家個人迥異的內隱特質所造成的專家群中個體答題「潛在異質性模式」（unobserved heterogeneity）。例如，當專家面對自己不熟悉的問題時，他們會如何回應？是否會自我界定答題的範圍，甚至選擇迴避回答？換言之，當專家們並非對所有問題都瞭解，或在面對大量題項而產生調查疲乏時，他們是否可能選擇不填答問卷中幾個自己不熟悉的題目？而這樣的結果是否與原本所有受訪者完整填答的平均結果並無差異呢？如果幾乎無差異，那麼是否還需要更多的專家參與？因之，本文的研究動機即是在好奇和疑惑中浮現，期能透過實證分析來檢證可能的答案。

再者，從方法論的觀點，一般來說，任何調查研究除了提出以堅實理論為基礎的架構外，對所欲探究的概念或現象進行有效度和信度的測量，都需要相關研究文獻的指引，也需要研究者依據操作性定義設計適當的問卷題目，找到具有代表性的受訪者，依據訪員指引或填答說明，要求受訪者提供真實、無偏誤、能完全表達心中意向的答案，才能利用此一調查品質良好的資料進行統計分析與理論檢證。過去大多數有關民意調查資料品質的研究，聚焦在問卷設計（如題型選擇、題序效應、用字遣詞、一題兩問、引導誘答等）、抽樣設計（如調查母體、樣本規模、抽樣方法等）、或是訪員效應（如訪問訓練、訪員特徵等），較少針對受訪者的回答模式進行探討。因此，當我們利用問卷調查時會發現，來自專家（或菁英）與一般民眾受訪樣本所考量的重點可能是有所不同的，這引發了本文的另一個研究問題，即民眾樣本的隨機性與是否有母體代表性一直被關注，而專家樣本通常是立意選樣，專家的隨機性又意指什麼？有何方法可以檢證？

基此，本研究的主要目的是藉由檢驗臺灣公共治理主觀指標調查中專家評估意見的代表性及不同專家之間的比較意涵，進一步探討專家之間是否存在異質性。從測量品質的角度來看，本研究旨在回答以下問題：當受訪者是接受特定專業訓練或擁有不同專業知識的專家代表時，他們如何填答問卷題目？參與調查的次數或專業差異是否會影響他們的回答？此外，面對不熟悉的問題和大量的問項，他們又會如何回應？可以只回答部分的題目嗎？這些問題不僅在應用方法論上具有研究價值，亦是在實務上值得關注的調查議題，直接關係到測量或評估的品質。

針對上述問題，本研究擬提出實證方法論上的新見解，並指出專家調查理論與實務中一些值得關注但文獻較少討論的調查議題。本文內容除了首節的研究背景與目的說明外，其他章節安排如下：首先，說明專家調查與測量品質之間的關係，並論述隨機專家假設與專家異質性；其次，交代本研究所使用的資料來源與處理；第

三，討論本研究的實證分析結果；最後，針對研究發現作出結論。

貳、文獻探討

一、專家調查與測量品質

所謂的「專家」或「專業人士」，一般係指「職場上具備專業化知識及技能的職業人士，其所擁有的專業技能通常須符合科學原理，經過長時間的學習及訓練，並有經專業認證的考試獲得的合格證書或執照，擁有自我約束行為的職業操守（或道德）及可量化的專業標準等」。[1] 舉例來說，政治學界對選舉專家的定義即是，足以證明擁有選舉相關知識的人，其標準可能是曾經發表或出版某一國家之選舉制度相關研究論文或學術論著、曾經在負責選務的政府機關工作、選舉相關研究學會或專業社群的會員、在大學的政治相關科系擔任教職等，其他社會科學領域對專家的定義也大多是如此（Martínez i Coma & van Ham, 2015, pp. 7）。

由此可知，專家是經過嚴格標準的訓練和系統學習後，對某一特定領域擁有完整知識的人士，而這些知識是一般大眾所欠缺的，但這並不代表他們是全知全能，能夠通曉各領域的知識，其具體貢獻在於，專家能夠迅速地針對新的情境做出精確的判斷並採取適當的對策。換言之，當缺乏足夠的歷史及背景資料可供參考，或取得資訊的成本過高時，我們往往會轉而求助於專家的意見，由於專家意見具有高時效性和低成本的特性，並能提出重要見解以供預測和決策之用，因此，如前文所言，專家意見常被廣泛應用於經濟、科技、環境、氣象、軍事和政策分析等領域（林希偉、王國全，2009，頁 182）。

所謂的「專家判斷」（expert judgment），即是依賴一位或多位專家的經驗來對重要議題進行判斷與評估的方法（Cooke, 1991）。一般而言，專家判斷與民眾調查在效度與信度可能產生的差異，主要可以歸納成以下幾個面向：首先，由於相同領域的專家大都是接受一套嚴格的、標準的訓練，在個人特質、背景知識和資訊的同質性較高，在評估時應具有較高的效度和信度，即使參與評估的人數增加，對測量品質的影響應該不大；相對地，一般民眾的個人特質、背景知識和資訊的異質性較大，也未接受一套相同系統化訓練，因此，在評估時的效度和信度應是比較低，

[1] 資料來源：https://wiki.mbalib.com/zh-tw/%E4%B8%93%E4%B8%9A%E4%BA%BA%E5%A3%AB。檢索日期：2024 年 6 月 20 日。

但是，隨著參與評估的人數增加時，其效度與信度應會隨之提升，也更方便進行交叉驗證評估樣本結構的合理性。

其次，專家的「延展性知識」（extensive knowledge）和相關經驗較為豐富，對於長期發展趨勢的評估應較具有效度，但專家的人數畢竟是社會中的少數，因此，即使將這些參與評估的專家偏好加總，也不一定可反應當前社會的主流偏好；而民眾主要是反映當前的個人需求和偏好，將眾多個人偏好加總後便可反映出當前的主流意見，其效度較高，但他們較看不清楚未來的方向或趨勢，隨著評估的時間愈長（如三年、五年、十年），其評估的效度和信度應會下降。最後，就調查的議題而言，當議題愈是複雜和專業時，專家因具有相關的背景（領域）知識，參與評估的效度比較高；而民眾則可能因知識、興趣或是時間等因素，導致資訊較少或缺乏判斷能力，對於錯綜複雜的政策議題大多以常識判斷事物的表象（Carmines & Kuklinski, 1990; Lupia, 1994），或是較有可能傾向拒答或回答「無反應」選項。

但是，也有研究指出，過度信任專家的評估或判斷並不是完全沒有風險的，過度自信的專家判斷可能影響決策的品質，研究也發現專家之間的變異小於問題之間所造成的隨機變異（林希偉、王國全，2009）。換言之，如 Cooke（1991）和 Shlyakhter 等人（1994）的研究中指出，專家的區間估計通常存在過於狹窄的問題。另以政治學為例，專家調查普遍應用在跨國比較研究中，如用在政黨和政策立場的研究、總統或首相的權力、選舉體制的評估，而其所涉及的效度問題便很值得關注。舉例來說，西方民主國家的專家因長期生活在相似的制度環境之下，其政治競爭主要圍繞在左派、右派的路線上，在意識型態上可能持有類似的信念或價值，但是，對那些來自非西方國家的專家而言，是否擁有相同的信念則不得而知，在此環境脈絡差異如此大的情況下，國際評估仰賴專家的評估是否具有可比較性及如何比較，即是一大難題（Martínez i Coma & van Ham, 2015），甚至牽涉到專家校準和加權的議題（Cooke & Goossens, 2008）。

Budge（2000）也以專家判斷法應用在研究政黨之政策立場為例，說明透過專家來評估各國的「選舉清廉度」（electoral integrity），其可能出現的四個問題或限制：（1）專家是根據選舉的哪些面向作為判斷的基礎？（2）專家是利用哪些標準來判斷選舉的清廉度？是否考慮到競選財務或是選舉暴力的議題？是否對某些異常的事件給予更高的權重考量？（3）專家在判斷事實的同時，是否也摻入了個人的觀點？（4）要求專家在什麼時間點或哪一期間對選舉的清廉度進行評估？雖然一個好的問卷設計可以減少其中某些問題的嚴重性，但無法完全排除這些限制的存在。對此，Martínez i Coma 和 van Ham（2015）也補充指出，個別專家的特質對其

判斷也是很重要的，如在地和國際的選舉專家可能對選舉過程的某些部分有不同的意見，這很可能會出現的原因在於，對某一國家的選舉過程需要更多延展性知識或不同類型的知識，例如在地的專家較有可能藉由長期經驗累積而充分掌握選舉的過程，對其選舉的清廉度提供較為深入的見解或評價；而國際選舉專家可能僅是參與或觀察選舉過程的某一階段，僅能就選舉清廉度的某些技術層面進行評價。此外，選舉專家是否本身擁有強烈的政治意識型態可能也會影響政治中立的程度，甚至該國的政治情勢的「極化程度」（polarization） 也可能都包括在代表專家異質性的變異來源。

此外，一些國際組織對專家評估作業的重視已行之有年，例如歐洲食品安全局（European Food Safety Authority, 以下簡稱 EFSA）不但制定了完整的協議，更採用結構化和明確的流程整合專家判斷納入其食安分析的工作實踐中（Hanea et al., 2021）。由於每位專家並不一定擁有與特定專業問題相關的所有知識或資訊，因此可能需要多位專家共同參與，甚至結合多元領域的專家意見，才能做出較為完整的評估或提出完善的解決方案。換言之，在專家評估的效度方面，考慮到專業知識的有限性，在特定問題的解決或決策上，專家的判斷可能會出現很大的差異，而多元專家的參與可以減少單一專家主觀判斷的風險、多元意見的整合則能獲得較為客觀的結論。這也是為什麼多數跨國調查或比較研究不會僅依賴少數幾位專家進行判斷或評估，其目的即是為了維持評估結果的客觀性和代表性。然而，有論者進一步指出，當詢問更多專家時，他們可能很少會給出相同的答案，那麼，研究者該如何將各方意見結合起來呢？是否應該給予所有專家同等的重視，還是必須更加看重其中被公認具有較高專業知識者的意見（Hanea et al., 2021, pp. 2）？換言之，即使在同一特定領域內的專家之間，其意見仍然可能不一致且存在分群現象，例如雙峰分布，呈現專家異質性，在方法上該如何進行分析，是值得研究人員思考的議題。

二、隨機專家假設與專家異質性

以公共治理的評估為例，其所涉及的層面極為廣泛，國際組織所開發的指標體系通常也是由多元面向所建構而成，除非長期關注或實際參與政府部門的政策制定過程，否則，政策投入與產出的複雜程度往往超越一般民眾的認知或經驗，正如前述文獻檢閱得知，專家的判斷在許多政策領域中發揮了重要作用。在此值得反思的是，在政策制定過程中，當決策者缺乏明確證據來輔助判別專家意見的權重時，通常會平等對待每位專家的意見，然而，這種方法並未賦予決策者優化專家意見的能

力，因為它假設所有專家的專業程度齊一且意見貢獻相等，如此一來，可能會損害不同專家水準在科學決策中的真實貢獻度（Hanea et al., 2021, pp. 55）。如 Cooke（1991）首次提出了一種基於專家績效表現的加權機制，這種方法並不對所有專家評估意見逕行計算平均，然而，一些學者持不同看法，認為基於績效的加權並不一定優於同等加權（Clemen, 2008）。

事實上，賦予所有專家同等權重意味著，如果專家表現的差異是由隨機波動造成的，則透過抽樣隨機重新組合專家評估內容，進而創建出虛擬的隨機專家代表，在統計上應該與實際專家表現相似。這一假設被稱為「隨機專家假設」（Random Expert Hypothesis, 以下簡稱 REH）（Hanea et al., 2021, pp. 58）。REH 的基本條件具有重要意義，因為它將表現最好的專家和表現不佳的專家一視同仁。申言之，若隨機專家的表現與真實專家的表現相當，意味著他們表現既然大致都一樣，答案也具有同質性，那麼就不需要特別尋找專家小組中學經歷資深或公認表現卓倫的專家，也無須追求大量的專家取樣。實務上 REH 的有效驗證，可以透過創建隨機專家小組的「蒙地卡羅模擬過程」（Monte Carlo Simulation）進行測試，[2] 這些隨機專家小組的成員的部分答案，主要來自原始專家小組相對應的數據中隨機抽樣填補，此部分詳細操作過程將列於本文實證分析結果的二、隨機專家和真實專家的比較。

綜言之，本研究嘗試探討以下實證命題是：當受邀填答問卷的專家來自政府部門、學術界、企業界及公民團體等四大領域時，若可免除填答七大面向指標中的任何 1、2 個較不熟悉面向的題目，這也正是實務上專家所樂意而常常要求的，畢竟題目眾多、作答耗時。那麼，是否會與原始七大面向指標題目全部填答的結果有所差異？尤有進者，如果無差異，可否推廣到自由刪去 3 個以上的面向題目，甚至只回答 1、2 個最熟悉面向的題目即可嗎？

這項探討除了有助於理解專家在不同情境或答題面向的表現，以及對如何整合專家意見提供了重要啟示，也是針對專家群中個體答題「潛在異質性模式」（unobserved heterogeneity）的一項相當重要的檢證，通過檢驗 REH，我們可以評估專家自由選擇填答部分面向題目的可行性，以及確定專家在填答時可以自由選擇回答的面向題目數量之最低限度或閥值，以減輕他們的作答負擔，並為專家調查的

[2] 關於「蒙地卡羅模擬」的參考文獻，為方便讀者熟悉這種方法的操作與運用，主要可參考國外重要的期刊論文，例如：https://onlinelibrary.wiley.com/doi/full/10.1111/puar.13182，並列入參考文獻中，檢索日期：2024 年 11 月 20 日。

設計與推廣簡化作答方式提供有利的參考依據。具體而言，通過蒙地卡羅模擬方法，我們可以在不縮減問卷內容的情況下，比較原始專家小組與隨機專家小組評分的平均值和標準差方面的差異。如果兩者之間沒有顯著差異，則不會影響最終的整體評估結果，並且能夠節省大量的作答時間和成本。

參、資料來源、變數處理與統計分析

本研究評估專家異質性的資料來自臺灣公共治理指標調查，該調查計畫由行政院研究考核發展委員會（後與行政院經濟建設委員會合併成為「行政院國家發展委員會」）自 2008 年開始委託臺灣公共治理研究中心執行，共計執行了 8 年，其後因委辦契約結束而終止此一調查計畫。[3] 如同其他國際上的作法，此一計畫建立了公共治理的評估構面與相對應的衡量指標，從七大面向評估政府治理績效，包含「法治化程度」、「政府效能」、「政府回應力」、「透明化程度」、「防治貪腐」、「課責程度」與「公共參與程度」，各大面向下共有 20 個次面向（莊文忠等人，2012；2013）。該計畫主要貫穿馬英九總統主政時期，每年皆會檢討修正或調整問卷題數及內容，舉例來說，公共治理指標的七大面向中，各面向均增加了一題「綜合評估」，例如在 2013 年版「政府回應力」這一面向中，會詢問：「與 100 年度相比，101 年度我國的政府回應力變得更好還是更不好？」。此一評估指標系統結合主、客觀指標，廣泛邀請不同領域的專家學者參與評比，並蒐集國內、外相關調查報告。這不但能夠充分反映出臺灣治理的特色，也能與國際接軌，透過將研究報告內容公開於網路上，更有助於國人瞭解我國當時公共治理的發展狀況。

在主觀指標調查方面（如表 1），此計畫採取問卷調查法，並成立了「專家評估委員會」，該委員會由「學術界」、「政府部門」、「產業界」和「公民社會」等四大領域的專家組成。學術界的邀請對象主要為法律、管理、經濟、公行、社會等相關系所的系所主任及期刊編輯委員；政府部門代表則包括中央部會和縣市政府的主秘及研考單位首長；產業界專家係來自天下雜誌前一千大企業的副經理級以上主管；公民社會則主要包括民間團體的執行長（或秘書長）和傳播媒體的總編輯（或社長），這些專家個別進行針對前一年度政府治理表現的主觀評估工作。

[3] 資料來源：https://www.ndc.gov.tw/cp.aspx?n=B42EEE2A1B1B5188&s=ED98AFD6DEA1C463。檢索日期：2024 年 6 月 20 日。

表 1

臺灣公共治理主觀指標概念內涵

面　向	次面向	題目數	概念內涵
法治化程度	法規管制	9	評估各類法規與制度的制定對人民保護的程度，包含政府對勞動市場補助的程度、或是政府對商業補貼的看法。
	司法體系	4	評估司法機關執行法律的手段是否符合公平、正義原則的看法。
	警政治安	2	評估我國的治安、社會安全情況與警察執法的效率等。
政府效能	個別施政項目	14	以政府服務特定利害關係人為主，針對我國政府特定業務施政成效進行評估。
	整體施政	5	評估政府的整體治理能力、政策方向以及公務機關落實政策的看法。
政府回應力	個別施政項目	14	係以政府服務特定利害關係人為主，檢視我國政府在各個主要政策之施政上，對於一般民眾及使用者需求回應程度的看法。
	整體施政	3	檢視我國政府在整體施政上，對於民意及一般民眾需求回應程度的看法。
透明化程度	資訊取得	2	檢視我國民眾取得政府資訊的容易程度的看法，作為評量政府資訊公開法是否落實的標準。
	資訊透明化	8	對我國政府機關進行評估，以檢討政府資訊透明化執行成效，落實主動公開精神。
	政治透明化	5	檢視公務機關人員財產及財務上的透明程度，包含公務人員的財產申報部分以及政治人物的政治獻金公告情形等看法。
	財政透明化	2	針對預算執行、審計稽核等項目，以評估我國財政資訊透明度的看法。
防治貪腐	貪腐印象	14	以「行政部門」、「立法部門」、「司法部門」及「其他」等四個項目，分別檢視政府各部門濫用公權力謀取私利的程度。
	倫理基礎建設	4	檢視提升倫理和防治貪腐的體制是否健全。
課責程度	審計	3	針對公務機關及人員經濟狀況及財物審計的看法，以評估整體政府的財務情形。
	預算控制	3	此次面向參考來自於瑞士洛桑國際管理學院（IMD）所建制的資料庫，屬於衡量「政府效率」，相關「財政」項目的看法。
	政府採購	5	瞭解解各政府機關在發包採購等事項上，是否有足夠公開之程序及是否有違法之情事，並進一步檢視採購稽核績效的看法。
	行為準則	2	檢視當公務人員發生失職、濫權亦或是處理不當之時，其懲戒的情形的看法。
	代議課責	1	評估立法部門對於行政部門是否發揮監督課責之功能的看法。
公共參與程度	政治參與	6	針對直接政治參與或間接政治參與的看法。
	媒體自由與言論自由	8	評估媒體自由指大眾傳播業者是否具有報導的自由的看法。言論自由則是泛指社會各界民眾是否具備表達自身意見之自由，並沒有特定對象或團體。

註：表中主觀指標係以 2013 年版本為準，各面向均會加問「綜合評估」一題，共計 121 題。

資料來源：作者自行整理。

該計畫自 2008 年啟動，但在 2011 年之前均採取紙本問卷的方式填答，2011 年首次導入線上問卷調查系統後，能更完整地記錄填答人的背景資訊。例如，在 2013 年的調查中，完成問卷填答的專家委員共計 240 位，其中學術界有 111 位，政府單位有 43 位，企業界有 43 位，公民社會有 43 位。[4] 在這 240 位完成問卷的專家委員中，有 100 位為第一次參與此一主觀評估工作；另外，有 140 位為前一年度就參與填答的專家委員，2012 年專家則共有 230 位。除非他們的職務有變動，否則每年都會受邀參加。這個調查計畫一直持續到 2017 年，但只有 2012 年和 2013 年的指標題項及內容幾乎完全一致，因此適合用來探討專家的再測情形，也具有研究價值。[5] 至於其他年度的題目，則皆有部分調整。

在資料處理方面，該計畫各項指標的評估有別於傳統五分或七分測量尺度的李克特量表（Likert scale），改採用 0 至 10 級距，0 分表示對政府表現非常不滿意或非常不同意題目陳述；反之，10 分則表示對政府表現非常滿意或非常同意題目陳

[4] 學術界的專家人數較多，主要是一開始考慮到不同學門的專業背景，因此邀請了更多元且學養經驗豐富的學者參與。這些學者包括「臺灣公共行政與公共事務系所聯合會」（TASPAA）的理監事、政治學門 TSSCI 期刊的編輯委員，以及公私立大學相關系所的主任（涵蓋法律、社會、政治、經濟和管理等領域）。雖然這樣的安排使得學者的總樣本數多於其他領域，但不同學門學者的表現也成為本研究探討的一種專家異質性評估，旨在檢驗多元學門專家在表現上是否存在差異，例如是否呈現雙峰分布。值得注意的是，在變異數分析中，通常假設各組的方差相等（同方差性）。如果分組樣本大小不均，理論上可能會影響方差的估計，導致檢定結果的不穩定，本文後續將對此問題進行實證討論。此外，這些參與的專家可能在背景或身份上存在重疊，例如從業界轉回學術界任教或從學術界轉往業界發展。然而，在邀請時是根據研究團隊先確定的專家特定現職身份來參與填答。這樣可以確保每位受訪者僅填寫單一回復樣本，避免跨領域重複填答。專家的多元性和豐富經歷實有助於降低刻板印象或狹隘專業的影響，使他們更有能力同時回答七大面向的問題。

[5] 此一資料的蒐集時間雖然較久，不足以反映當前我國政府治理表現的現況，然本研究係由方法論的觀點，探討受過「專業訓練」和擁有「領域知識（資訊）」的調查對象，在填答問卷時是否具有一致性或異質性，且由於這些受訪對象均在同一時期內完成，可排除時間和空間的「脈絡效應」（contextual effect），應較不受資料蒐集時間點的限制。此外，此一資料係屬優良的專家樣本，至今仍有研究價值的主要理由有二：1. 國內大多數公共行政領域中以專家為對象的研究主要是以質性方法為主，即使採取量化的問卷調查法，樣本數超過 200 個的研究亦不多見。2. 此一資料為二個年度的調查數據，在併檔並去識別化後，將受訪專家區分為三類：(1)2012 年第一次填答，(2)2013 年第一次填答，(3)2012 年和 2013 年同時填答，可區分出為「首次參與」和「重複參與」的樣本，並進行比較分析，具有「準實驗設計」意涵，亦屬難得。

述。5 分表示「中間」或「普通」程度，除了表示「比 4 分好、比 6 分差」外，更清楚地劃分了滿意與不滿意之間的界線。這樣的設計相較李克特量表，使得受訪者在評價時能夠更容易定位自己的感受。如果受訪者不瞭解或難以評價，則調查問卷中會建議勾選「無法評估」的無反應選項。當蒐集問卷資料完畢後，本研究首先對問卷調查中反向問法的題目進行處理，以使各個指標的好壞方向一致。其次，本研究除了分析不同年份的樣本外，還根據專家委員的參與填答次數將其分為「參與二次以上」和「第一次參與」兩類。此外，本研究也依據專家背景，將其分為「產業界」、「政府部門」、「學術界」和「公民社會」四大領域。

在統計分析方面，在大部分採用指標調查的研究中，以「變數為中心」（variable-centered）與以「受試者為中心」（subject-centered）的研究方法，是當前社會科學研究領域量化方法的兩種主流取向，這兩種方法各有其特點。以變數為中心的方法主要探究變數之間的關係（Muthén & Muthén, 2000），其前提假設是研究的母群體和回答樣本具有同質的特徵，因此可以通過隨機抽樣分析樣本的特徵來推論母體的特徵，相關分析、迴歸分析和因子分析等都屬於這一類方法。[6] 從問卷測量的角度看，每個題目即是一個變數，而因子則是用一組題目測量得到的潛在特質。因此，在以變數為中心的研究中，最關心的是測量的信度與效度，其中「收斂效度」（convergent validity）與「區別效度」（discriminate validity）也取自因子分析的結果（Hair et al., 2019）。以受試者為中心的方法則主要識別變數的相似關係或分辨受試者子群體，這類研究強調同一類型的受試者具有相似的特質和答題反應模式（Jung & Wickrama, 2008），並側重於探討個體差異性的成因和影響，進一步分析各類型受試者在各種變數上的表現差異。由於本研究的主要的目的並非評估題項設計的信度和效度，而是檢證不同領域專家在填答上是否存在異質性，以及參與填答次數是否會影響其回答模式，故本研究採用以受試者為中心的分析取向。

經過以上處理程序後，本文接下來的實證部分所使用統計方法計有 T 檢定、變異數分析、蒙地卡羅抽樣模擬、二項式檢定、Q-Q 圖視覺化比較法，以及 Kolmogorov-Smirnov 檢定等，再就分析結果進行討論。

[6] 相關分析用於研究兩個或多個變數之間是否存在共變關係；迴歸分析則探討某個變數的變化在多大程度上可以由其他變數的變化來解釋和預測；因子分析則根據變數之間的相關性將其分組，使同組變數之間的相關性較高，每組變數對應於一個因子。

肆、實證分析結果

一、無反應與趨中選答的專家答題模式分析

Krosnick 和 Presser（2010, pp. 265）指出，受訪者面對調查題目時，通常會先進行「認知」動作，嘗試理解題目的意涵；第二步驟是進行「解碼」，從個人的記憶中尋找合適和相關的資訊；接著，試圖整合這些資訊並做出一個「判斷」；最後，將這個「判斷」轉換成「答案」。而當受訪者不見得對所有的題目都有動機、能力和誘因去努力搜尋相關資訊時，便可能傾向提供一個「安全的」答案，例如趨中或無反應的選項，以免暴露自己的無知或立場，又不會造成無效的樣本。根據前述文獻，本研究分析在臺灣公共治理主觀指標 121 個題目中，專家填答「無反應」（即無法評估）和「中間值（5）」的次數當作專家的重要答題特徵或習性。[7] 一般專家會填答前者的答案，代表其認為這些問題可能超出個人的專業領域，或對該問題較為不瞭解；而填答後者的答案，則隱含兩種可能：一是，該位專家認為政府在這個題項上公共治理的表現是居於中間位置，即「比 4 分好、比 6 分差」；二是，該位專家對該題項並不熟悉，以趨中的答案來表達模糊的中立態度。至於是哪一種原因，因未對這些專家進行訪談，無法確認其真實想法，這也是本研究將無反應和中間值之答題數據分別討論的原因。不過，為了檢證受訪者較常在題項回答中間值 5，是否也較會有無反應的答題傾向，本研究就受訪者回答這二種選項的題目數進行相關分析，若為正相關程度越高，表示兩個答案之間應具有可互換性（interchangeable），受訪者可能只是任意選擇回答其中一個答案；反之，若為負相關程度越高，表示兩個答案之間不具有可互換性，受訪者認知到這二個選項具有不同之意涵；若越接近無關，則是越不易判斷二個選項之間的關係為何。

根據表 2 的數據可知，在同一年度的調查中，專家回答無反應題數的多寡與回答中間值 5 的題數並無明顯關聯（101 年為 -0.081，102 年為 -0.082）。相反地，兩個年度的無反應題數之間存在顯著的高度正相關（0.737），而兩個年度的中間

[7] 本文作者特別感謝其中一位審查人指出，若評分等級為是 0 至 10 分，中間值 5 分的測量意義在於「比4分好、比6分差」，並非表示「不好不壞」，本文也據以修正相關陳述。此外，本文作者增加表 2 分析並說明無反應與趨中選答的意義確實不同，可據以分別用來探討專家答題模式。

值5題數則具有顯著的中度正相關（0.451）。這表明，專家在填答時不會用中間值的評價來替代無反應選項。此外，跨年度填答的這兩種模式都顯示出相當高的一致性，這意味著部分專家習慣於填答「中間值（5）」。將這兩者分別陳列可以產生相互參照的效果，有助於理解第一次參與的專家與參與兩次以上的專家之間的意見是否具有同質性。

表 2

101 年與 102 年受訪專家填答「無反應」與「中間值 5」題目相關分析

		101 年回答 無反應題數	101 年回答 中間值 5 題數	102 年回答 無反應題數	102 年回答 中間值 5 題數
101 年回答 無反應題數	相關係數		-0.081	.737*	0.068
	個數		230	140	140
101 年回答 中間值 5 題數	相關係數			0.054	.451*
	個數			140	140
102 年回答 無反應題數	相關係數				-0.082
	個數				240

註：*P<0.05。

資料來源：作者自行整理。

本研究分析原始資料中的無反應樣本，發現在 2012 年和 2013 年參與填答的專家中，七大面向下有 3 個次面向有一成多填「無法評估」，主要分布在「課責程度」面向下的審計和預算控制（2012 年的比例達 18.4%），以及透明化程度的資訊取得部分則介於 9.6% 至 13.9% 之間，其他面向下的次面向填答無反應的比例均在一成以下。至於為何專家在審計、[8] 預算控制、[9] 資訊取得[10] 等次面向的題項填答「無法評估」比例偏高，本研究認為對專家而言，相關題項的內容並不難理解，但

[8] 「審計」次面向的題項有三：「我國的審計方式足不足夠？」、「我國審計機關可以獨立於行政及立法體系之外而行使職權？」、「我國審計機關有沒有確實查核各項屬於主管可裁量決定的支出（例如：管理費、特別費、機要費及預備金…等）」？

[9] 「預算控制」次面向的題項有三：「我國政府預算執行有沒有符合預算法的規範？」、「我國審計機關對行政部門總決算的審核報告適不適當？」、「我國政府官員的使用經費中，屬主管可裁量決定的支出（例如：管理費、特別費、機要費及預備金…等），有沒有遵守預算法規來執行？」

[10] 「資訊取得」次面向的題項有二：「我國現行的『政府資訊公開法』進行修法的必要程度為何？」、「我國『政府資訊公開法』落實的程度為何？」

一方面是政府部門的審計與預算相對較為專門，一般管理背景的專家較少有實際接觸經驗，另一方面是這些題項大多需要判斷其實際執行或落實程度，在專家的訓練過程要求「證據基礎」（evidence-based）的分析或判斷下，不會輕易給評價或下結論，以致於專家選擇避答的比例偏高。

從表 3 不同參與次數的專家分析可知，在全部 121 個評估指標中，第一次參與的專家回答「無反應」選項的平均值為 5.99 題，參與二次以上的專家的平均值為 3.78 題，顯示這些專家對於大多數的題目都有能力回答，回答無反應選項所占比例均不到總題目 5%；且 T 檢定結果也顯示，不同參與次數與其填答「無反應」選項的題數雖在 95% 的信心水準下尚未達到統計上的顯著差異，不過，第一次參與專家填答「無反應」選項題數的標準差為 11.73 題，明顯大於參與二次以上的專家（6.27 題），顯示參與二次以上可有效減少回答「無反應」的題數。

在回答「中間值（5）」的題數方面，第一次參與專家的平均值為 16.20 題，參與二次以上專家的平均值為 17.04 題，這兩個比例均遠高於回答無反應選項的比例，反映出專家們認為政府部門在這些指標的治理表現持平而已。此外，T 檢定結果也顯示，第一次參與和參與二次以上的專家，其平均值無統計顯著差異，且兩者的標準差分別為 9.84 題和 10.95 題，亦是相當接近，此一結果顯示，不同參與次數的專家對於問題的掌握程度頗為相近，並未因參與填答經驗不同而有明顯差異。

表 3

不同參與次數的專家填答「無反應」與「中間值 5」的 T 檢定

比較面向	組別	個數	平均數	標準差	T 值	P 值
回答「無反應」的題數	參與二次以上	140	3.78	6.27	-1.718	.088
	第一次參與	100	5.99	11.73		
回答「中間值 5」的題數	參與二次以上	140	17.04	10.95	.624	.533
	第一次參與	100	16.20	9.84		

資料來源：作者自行整理。

進一步由表 4 不同專業領域的專家分析可知，在全部 121 個評估指標中，公民社會領域專家回答「無反應」選項的平均值為 6.49 題，為四個領域之中最高，而產業界專家的平均值為 2.05 題，為四個領域之中最低，不過，變異數分析結果也顯示，不同專業領域與其填答「無反應」選項的題數並無統計顯著差異。值得說明的是，學術界專家填答「無反應」選項題數的標準差為 11.39 題，是四個領域之中最

高，顯示學術界專家填答此一選項的變異性遠大於其他領域，反映出學者可能只熟悉其專業領域，對其他領域的熟悉度偏低而選擇無法評估的選項，而產業界專家的標準差只有3.99題，是四個領域之中最低，或許與產業界專家平時對議題的接觸層面較為廣泛，勇於表達個人的觀點或評價。

在回答「中間值（5）」的題數方面，產業界專家的平均值為19.02題，其他三個領域專家的平均值則為16題左右，由F檢定結果顯示，四者亦無統計顯著差異，至於四者的標準差，產業界和政府部門的分別為 12.90 題和 10.97 題，略大於學術界和公民社會，此一結果同樣顯示，表示四大領域的專家對於問題的掌握程度，似乎並未因專業領域不同而有明顯差異。

表 4

不同專業領域的專家填答「無反應」與「中間值 5」的變異數分析

		個數	平均數	標準差	F 值	P 值
回答「無反應」的題數	產業界	43	2.05	3.99	1.996	.115
	政府	43	4.23	6.38		
	學術界	111	5.22	11.39		
	公民社會	43	6.49	7.31		
回答「中間值 5」的題數	產業界	43	19.02	12.90	.920	.432
	政府	43	16.09	10.97		
	學術界	111	16.42	9.76		
	公民社會	43	15.65	9.06		

註：表中主觀指標計算係以 2013 年版本為準，*P* 值> 0.05 顯示檢定結果不顯著，另以 2012 年計算結果也同樣是 *P* 值> 0.05（*P* 值依上表排列分別為 0.07 和 0.55）。

資料來源：作者自行整理。

二、隨機專家和真實專家的比較

從前述文獻探討得知，隨機專家與真實專家的表現是否相同，攸關判斷特定領域內的專家具異質或同質性，如果兩者相同，我們就不需要特別尋找專家群中是否存在更優質的專家。基此，本研究嘗試以模擬的數值分析方式處理以下的問題：當受邀填答問卷的專家來自政府、學界、企業及公民團體等四大領域時，若可免除填答七大面向指標中的任一個以上不熟悉面向的題目（本研究實證以各面向內的「綜

合評估」題代表該面向，進行表 5 的計算），[11] 是否會與七大面向指標題目全填的結果有所差異？模擬分析的過程和結果如下：

本研究使用前文提及的蒙地卡羅模擬法（Schwarz & Newman, 2020）比較了 2012 年和 2013 年的數據。假設隨機專家小組中，每位專家評估七大面向的題目。除了專家自己選擇作答的面向外，其餘面向則從原始作答的其他專家的真實回應中隨機抽取填補。假設這一過程重複 1,000 次，將生成 1,000 組隨機專家小組的答案。如果隨機創建的專家與原始真實專家之間不存在系統性差異，那麼在這 1,000 次運算中，預期原始專家群的平均評分只有約一半的情況下會高於隨機專家群。為了驗證原始專家小組和隨機專家小組的答題是否具有相同的統計分布，本研究設定了具體的驗證目標是：如果隨機專家假設（REH）為真，則假設檢定的虛無檢設有二：1. 原始小組中專家在各面向的統計平均值高於隨機專家小組的機率為 50%；2. 原始小組中專家在各面向的統計標準差高於隨機專家小組的機率為 50%。

如表 5-1、表 5-2 所示，在模擬過程中，本研究模擬專家隨機去掉 1 個面向、2 個面向，…，直到去掉 6 個面向的題目，並比較隨機專家與真實專家的答案分布。具體來說，我們使用二項式檢定統計 1,000 次的模擬結果如下：[12]

[11] 此處假設了特定面向內的「綜合評估」題是「代理變數」（proxy variable），因為可表現該面向的整體表現評價，故以該面向綜合評估的題目作為代表該面向進行計算。

[12] 二項式檢定適用於只有兩種結果的情況。在本案例中，可以將真實專家的評分表現分為「高於隨機專家小組」和「不高於隨機專家小組」兩類。因此，這符合二項式檢定的基本要求。

表 5-1

隨機專家小組的蒙地卡羅模擬過程（2012）

專家領域	2012年各面向模擬結果						
	法治化程度	政府效能	政府回應力	透明化程度	防治貪腐	課責程度	公共參與程度
產業界	0.195	0.268	0.071	0.776	0.924	0.874	0.174
	0.015*	0.268	0.054	0.046*	0.121	0.107	0.040*
	0.359	0.018*	0.924	0.137	0.825	0.924	0.429
	0.359	0.155	0.002**	<0.001***	<0.001***	<0.001***	0.046*
	0.393	0.393	0.094	0.825	0.121	0.825	0.137
	0.107	<0.001***	<0.001***	<0.001***	0.009**	<0.001***	<0.001***
	0.548	0.155	0.975	0.062	0.174	0.728	0.825
	0.034*	0.001**	<0.001***	<0.001***	<0.001***	<0.001***	0.003**
	0.393	0.467	0.174	0.359	0.071	0.635	0.776
	0.001**	<0.001***	<0.001***	<0.001***	0.002**	<0.001***	0.010*
	0.681	0.776	0.874	0.591	0.071	0.040*	0.681
	0.107	0.015*	<0.001***	<0.001***	<0.001***	0.005**	0.003**
政府部門	0.429	0.359	0.874	0.195	0.029*	0.268	0.297
	0.327	0.924	0.094	0.217	0.040*	0.025*	<0.001***
	0.635	0.548	0.681	0.467	0.467	0.975	0.825
	0.082	0.094	0.195	<0.001***	0.001**	0.001**	0.003**
	0.681	0.591	0.429	0.591	0.025*	0.071	0.635
	0.025*	<0.001***	<0.001***	<0.001***	0.009**	0.003**	0.195
	0.137	0.975	0.874	0.776	0.155	0.924	0.393
	0.004**	<0.001***	0.001**	0.009**	0.003**	<0.001***	0.010*
	0.681	0.018*	0.137	0.776	0.359	0.635	0.217
	<0.001***	<0.001***	<0.001***	0.001**	0.054	<0.001***	<0.001***
	0.195	0.467	0.728	0.635	0.467	0.776	0.359
	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	0.001**
學術界	0.327	0.507	0.507	0.029*	0.635	0.107	0.268
	0.071	0.107	0.429	0.107	0.681	0.174	0.082
	0.825	0.429	0.924	0.776	0.268	0.874	0.825
	0.029*	0.975	0.217	0.121	0.021*	0.429	0.137
	0.728	0.297	0.467	0.429	0.217	0.359	0.242
	0.012*	0.002**	0.012*	0.046*	0.174	0.327	0.776
	0.467	0.507	0.825	0.297	0.728	0.327	0.359
	0.393	0.467	0.003**	0.054	0.155	0.874	<0.001***
	0.429	0.094	0.195	0.174	0.728	0.327	0.107
	0.009**	0.001**	0.018*	<0.001***	0.002**	0.046*	<0.001***
	0.507	0.268	0.327	0.507	0.548	0.359	0.155
	0.507	0.082	0.010*	0.268	<0.001***	0.137	0.054
公民社會	0.297	0.924	0.217	0.029*	0.467	0.467	0.297
	0.002**	0.107	0.297	0.062	0.054	0.776	0.034*
	0.874	0.591	0.591	0.635	0.297	0.054	0.635
	0.002**	0.040*	0.034*	0.003**	0.003**	0.006**	0.681
	0.825	0.728	0.062	0.548	0.046*	0.591	0.029*
	0.007**	0.015*	0.010*	0.195	<0.001***	0.001**	<0.001***
	0.217	0.327	0.975	0.591	0.825	0.776	0.018*
	0.004**	0.001**	0.029*	0.015*	<0.001***	<0.001***	<0.001***
	0.548	0.034*	0.681	0.217	0.054	0.548	0.021*
	0.018*	<0.001***	0.005**	0.021*	<0.001***	0.004**	<0.001***
	0.825	0.242	0.393	0.242	0.467	0.046*	0.825
	0.015*	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***	<0.001***

註：*$P < 0.05$, **$P < 0.005$, ***$P < 0.001$；灰色標示各面向全部和原始專家小組差異顯著。

資料來源：作者自行整理。

表 5-2

隨機專家小組的蒙地卡羅模擬過程（2013）

專家領域	隨機避答面向數	測量項目	2013年各面向模擬結果						
			法治化程度	政府效能	政府回應力	透明化程度	防治貪腐	課責程度	公共參與程度
產業界	1	平均值 p值	0.467	0.268	0.467	0.327	0.591	0.217	0.107
		標準差 p值	0.005**	0.174	0.082	0.174	0.548	0.548	0.174
	2	平均值 p值	0.728	0.874	0.728	0.393	0.776	0.874	0.728
		標準差 p值	0.040*	0.015*	0.071	0.010*	0.018*	0.034*	0.007**
	3	平均值 p值	0.107	0.548	0.681	0.635	0.467	0.121	0.874
		標準差 p值	0.006**	0.054	0.174	0.029*	0.040*	0.007**	<0.001***
	4	平均值 p值	0.975	0.728	0.874	0.548	0.467	0.155	0.217
		標準差 p值	<0.001***	0.003**	<0.001***	<0.001***	<0.001***	<0.001***	0.003**
	5	平均值 p值	0.728	0.467	0.217	0.507	0.591	0.359	0.591
		標準差 p值	<0.001***	<0.001***	0.004**	0.009**	<0.001***	<0.001***	0.009**
	6	平均值 p值	0.975	0.975	0.776	0.393	0.591	0.217	0.155
		標準差 p值	<0.001***	0.002**	<0.001***	0.006**	<0.001***	<0.001***	0.015*
政府部門	1	平均值 p值	0.635	0.137	0.467	0.924	0.062	0.094	0.021*
		標準差 p值	0.025*	0.924	0.015*	0.021*	0.021*	0.001**	0.062
	2	平均值 p值	0.429	0.924	0.548	0.776	0.297	0.548	0.155
		標準差 p值	0.054	0.297	0.018*	<0.001***	0.004**	0.174	0.082
	3	平均值 p值	0.242	0.825	0.728	0.054	0.874	0.467	0.874
		標準差 p值	0.015*	0.005**	<0.001***	0.025*	<0.001***	0.155	0.003**
	4	平均值 p值	0.635	0.548	0.429	0.635	0.776	0.975	0.327
		標準差 p值	<0.001***	<0.001***	<0.001***	0.034*	<0.001***	0.003**	0.174
	5	平均值 p值	0.155	0.242	0.082	0.071	0.874	0.297	0.635
		標準差 p值	0.001**	<0.001***	0.002**	0.003**	<0.001***	<0.001***	<0.001***
	6	平均值 p值	0.776	0.874	0.728	0.548	0.874	0.776	0.591
		標準差 p值	<0.001***	<0.001***	<0.001***	0.054	<0.001***	<0.001***	0.107
學術界	1	平均值 p值	0.728	0.155	0.507	0.467	0.825	0.327	0.467
		標準差 p值	0.776	0.327	0.082	0.776	0.924	0.507	<0.001***
	2	平均值 p值	0.728	0.874	0.681	0.054	0.681	0.029*	0.728
		標準差 p值	0.003**	0.681	0.009**	0.121	0.217	0.507	<0.001***
	3	平均值 p值	0.728	0.359	0.874	0.137	0.094	0.728	0.268
		標準差 p值	0.001**	0.010*	0.429	0.155	0.429	0.268	<0.001***
	4	平均值 p值	0.268	0.728	0.825	0.015*	0.268	0.082	0.393
		標準差 p值	0.029*	0.217	0.003**	0.021*	0.029*	0.007**	<0.001***
	5	平均值 p值	0.467	0.635	0.137	0.393	0.393	0.094	0.242
		標準差 p值	<0.001***	0.012*	0.040*	<0.001***	0.054	0.137	<0.001***
	6	平均值 p值	0.327	0.467	0.297	0.268	0.297	0.327	0.297
		標準差 p值	<0.001***	0.004**	0.082	<0.001***	0.040*	0.046*	<0.001***
公民社會	1	平均值 p值	0.548	0.825	0.507	0.393	0.924	0.591	0.728
		標準差 p值	0.005**	0.874	0.548	0.005**	0.006**	0.195	<0.001***
	2	平均值 p值	0.825	0.924	0.507	0.548	0.062	0.012*	0.874
		標準差 p值	<0.001***	0.018*	0.015*	<0.001***	0.359	0.007**	<0.001***
	3	平均值 p值	0.635	0.548	0.429	0.874	0.924	0.874	0.635
		標準差 p值	<0.001***	0.195	0.009**	<0.001***	<0.001***	0.046*	<0.001***
	4	平均值 p值	0.548	0.681	0.681	0.094	0.591	0.121	0.174
		標準差 p值	<0.001***	<0.001***	0.001**	0.094	<0.001***	0.012*	<0.001***
	5	平均值 p值	0.507	0.467	0.548	0.094	0.507	0.924	0.040*
		標準差 p值	0.007**	0.025*	0.018*	<0.001***	<0.001***	<0.001***	<0.001***
	6	平均值 p值	0.015*	0.217	0.010*	0.681	0.242	0.924	0.467
		標準差 p值	<0.001***	<0.001***	0.002**	<0.001***	<0.001***	0.002**	<0.001***

註：*$P < 0.05$, **$P < 0.005$, ***$P < 0.001$；灰色標示各面向全部和原始專家小組差異顯著。

資料來源：作者自行整理。

在分析中，本文將四大專業領域的專家（產業界、政府部門、學術界、公民社會）分開考量。研究結果顯示，參與填答的專家們在避答部分題目後，平均評分的趨勢仍然相似，這意味著隨機專家與真實專家的答案分布的平均評分差異並不明顯。進一步來說，雖然檢定平均值差異的 *P* 值較大，大抵顯示出兩者的分布平均值無顯著差異，但是，另一方面，隨著避答題目數量的增加，大部分隨機專家在七大面向的標準差逐漸呈現全面下降的趨勢，致隨機專家與真實專家比較的 *P* 值顯著（詳見表 5-1、表 5-2，灰色標示各面向全部和原始專家小組差異顯著的情況）。然而，這一趨勢在學術界專家的變化較小（詳見表 5-1、表 5-2，未出現灰色標示中，學術界專家即使隨機避答高達六個面向，部分 *P* 值仍有 0.08 或 0.13），推測可能是因為學術界的多數學者們答題有共識，出現如前述 Cooke（1991）指出專家評量區間較狹窄的現象，導致產生隨機專家數據後，和原始專家相比不會有太大變化，而其他專業領域中，隨機專家的標準差則較原始專家，縮小非常明顯，尤其是在去掉四個領域以上的題目後，各專業領域的隨機專家標準差大部分均會縮小，惟學術界即使在改變大量避答數目的情況下，結果仍然類似於原始的真實專家評估，這意味著增加學者人數也未必能提高結果的多樣性。

至於是否允許專家可以減少填答題目的數量，答案是肯定的，因為表 5-1、表 5-2 的模擬結果也顯示，各類專家普遍在缺答較少數三個面向以下的題目時，二項式檢定統計結果均符合隨機專家假設；而當專家缺答四個面向以上時，如果用其他人的相關答題數據隨機填補，則不易捕捉到原始數據中的少數極端值，此時隨機專家的標準差較容易內縮，這可能會大大影響隨機專家和真實專家標準差的差異擴大。

換言之，全部的專家都會有隨著回答問題數愈少，隨機插補題數上升，呈現標準差下降的趨勢，也就是當隨機插補題數上升，專家真正回答的題目數下降時，極端值會越來越不明顯，故隨機專家的標準差會下降。不過，這個現象在學術界專家群中比較不明顯，也就是說他們隨著隨機插補題數上升，標準差下降速度比較慢。可能的解釋理由是，原本學者就比較屬於中間立場，持公允中立態度並不輕易給出極端值，故分數範圍會較為集中分布在中上跟中下的區段，但其他領域專家的立場較為鮮明，也比較容易出現極端值，如產業界、公民社會的代表較為集中在低分區段，政府部門則集中在高分區段，以上理由也可從下表 6 獲得佐證。是以，隨著隨機插補題數增加，原先數據比較集中特定區間的專家，其結果高機率更容易抽中集中區段的數值，致標準差快速下降，而學者組答題較一致的特性，使得隨機學者的標準差仍變化不大。

三、不同專業領域的專家異質性檢證

從表 6 的 F 檢定結果可知，不同專業領域的專家在七大面向的評價平均數均呈現顯著差異，進一步來看，事後多重比較採「雪費法」（Scheffe test），在「法治化程度」面向上，政府領域專家的平均分數高於公民社會；在「政府效能」面向上，政府與學術領域專家的平均高於公民社會；在「政府回應力」面向上，政府領域專家的平均分數高於公民社會；在「透明化程度」、「防治貪腐」與「課責程度」面向上，政府領域專家的平均分數高於其他三個領域；在「公共參與程度」面向上，政府領域專家的平均分數高於公民社會；倘若整體治理計算來自七個面向綜合評估題的加總平均，在「整體治理」上，政府領域專家的平均數會高於產業界及學術界。

表 6

不同專業領域的專家評價平均分數的變異數分析

		個數	平均數	標準差	F 值	P 值	事後多重比較
「法治化程度」平均分數	產業界	43	4.99	1.33	4.464	.005	政府>公民社會
	政府	43	5.47	1.27			
	學術界	111	5.08	1.34			
	公民社會	43	4.45	1.29			
「政府效能」平均分數	產業界	43	4.86	1.49	8.262	.000	政府>公民社會 學術界>公民社會
	政府	43	5.63	1.51			
	學術界	111	5.02	1.43			
	公民社會	43	4.05	1.62			
「政府回應力」平均分數	產業界	43	5.04	1.52	5.557	.001	政府>公民社會
	政府	43	5.96	1.69			
	學術界	110	5.26	1.58			
	公民社會	43	4.54	1.86			
「透明化程度」平均分數	產業界	43	4.75	1.60	6.063	.001	政府>產業界 政府>學術界 政府>公民社會
	政府	43	5.84	1.44			
	學術界	110	4.98	1.49			
	公民社會	43	4.53	1.65			

表 6（續）

		個數	平均數	標準差	F 值	P 值	事後多重比較
「防治貪腐」平均分數	產業界	43	4.53	1.55	7.043	.000	政府>產業界 政府>學術界 政府>公民社會
	政府	43	5.43	1.21			
	學術界	110	4.58	1.26			
	公民社會	43	4.14	1.46			
「課責程度」平均分數	產業界	43	5.19	1.79	6.441	.000	政府>產業界 政府>學術界 政府>公民社會
	政府	43	6.39	1.39			
	學術界	110	5.58	1.56			
	公民社會	43	5.03	1.55			
「公共參與程度」平均分數	產業界	43	5.48	1.40	4.435	.005	政府>公民社會
	政府	43	5.96	0.84			
	學術界	110	5.69	1.31			
	公民社會	43	5.03	1.33			
「整體治理」平均分數	產業界	43	5.15	1.63	6.325	.000	政府>產業界 政府>公民社會
	政府	43	5.92	1.33			
	學術界	111	5.31	1.40			
	公民社會	43	4.56	1.57			

註：表中主觀指標計算係以 2013 年版本為準，另以 2012 年計算結果各個面向也同樣是 P 值< 0.05。

資料來源：作者自行整理。

此一結果顯示，政府領域的專家可能自身在公部門服務，包括中央部會和縣市政府的主秘及研考單位首長，會因較清楚和熟悉公務機關的實際運作過程，因此對政府的各項治理表現給予較高的分數，或是不吝對「自家人」的表現給予肯定，而公民社會領域的專家因扮演外部監督者，可能是以較為嚴格的標準審視政府表現，因此對於政府各項公共治理表現普遍給予偏低的評價，而學術界和產業界的專家所給的分數則是介於政府部門和公民社會的專家之間，也可能是較為缺乏系絡知識或實際接觸經驗，因資訊或證據不夠充分而未給予過高的評價。由此觀之，從不同專業背景的專家異質性初步分析，呈現政府部門專家單獨的一群，迥異於其他專業領域的專家評價。

本研究接著再利用其他統計方法加以細部觀察與檢證不同專業背景的專家分組內與分組間的異質性。

四、Q-Q 圖視覺化比較法及 Kolmogorov-Smirnov 檢定的應用

Q-Q 圖（Quantile-Quantile Plots）是一種用於比較兩組統計分布的圖形工具。通過這種圖，可以直觀地評估資料集是否服從某一理論分布，或者比較兩個資料集的分布特徵是否相似。在 Q-Q 圖中，一個資料集的「分位數」（quantiles）會與另一個資料集的分位數進行比較，分位數是用資料量來分割資料，而不是用數值本身來分割資料，簡單地講就是名次。通常，其中一個資料集是樣本資料，而另一個是理論分布，如常態分布等。本研究使用 Q-Q 圖來判斷兩組專家間答題資料的分布是否相同，如果是，那麼這些點將落在一條 45 度角的直線上，這條直線稱為「等分位數線」（line of equality）。如果點偏離直線，表明分布之間有差異。有別於表 6 的變異數分析，Q-Q 圖還可以比較出會出現差異主要是在高分組即「滿意」（高於 5 分）評價，還是低分組即「不滿意」（低於 5 分）評價的分歧。

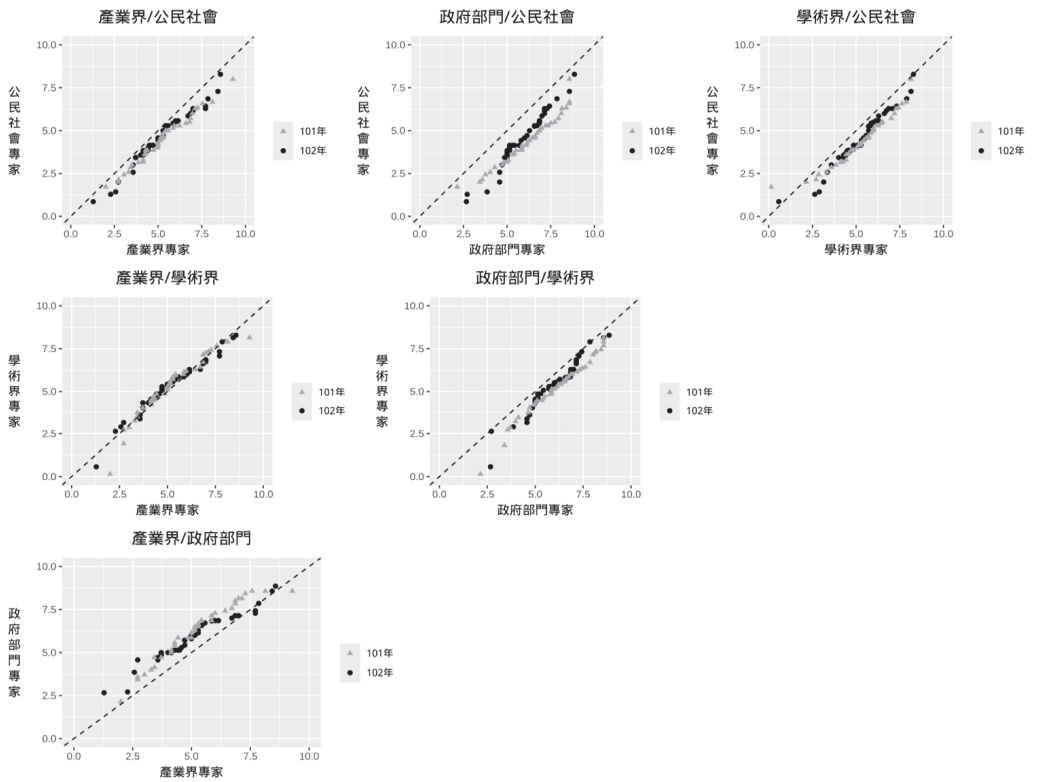
圖 1 中有三個重要發現：首先，（a）圖的分析中，除了將四大領域專家之間的兩兩比較，（b）圖也提供了 2012 年、2013 年四大領域的專家評價各年度整體治理表現的直方圖。於其中，學術界專家跨年度不論高分組或低分組幾乎分布一致，表示學者們答題確實有共識；[13] 公民社會領域專家在 2013 年高分群和低分群各有增加；產業界專家亦是如此；政府部門專家在 2013 年整體呈現上升趨勢。其次，所有的點分布大致呈現平行落在等分位數線上或下，表示四大領域的專家互比之間的差異，不僅反映在高分群，也一樣出現在低分群。例如，產業界和政府部門的比較，大部分點分布落在等分位數線上方，表示政府部門專家在低分組評分也比產業界高。政府部門與其餘三大領域的專家均有差距存在，尤其表現在低分群評分都比較高，（a）圖更顯示公民社會跟政府部門有最大之差異。第三，代表產業界、學術界、公民社會的專家點分布貼近等分位數線，目視三者分布相仿，且產業界和學術界專家的評價最為接近。綜上所述，從不同專業領域的專家 Q-Q 圖視覺化比較分析，確實也呈現政府部門專家代表是單獨的一群，迥異於其他專業領域的專家評價。有鑑於利用圖形的視覺化判斷可能失之客觀，本研究以下再使用統計檢定方法加以檢驗。

[13] 本文作者特別感謝其中一位審查人建議，可以隨機抽取等量（n=43）的學術界專家進行檢測，本文作者也發現隨機抽樣的結果，兩年小樣本和原本百人以上學者專家樣本的整體治理表現評價分布是類似的，這也正顯示學術界專家答題的特性，即不同學門的學者專家答題其實具有同質性，學者專家樣本數不會影響本文不同各個專業領域的專家異質性的比較。

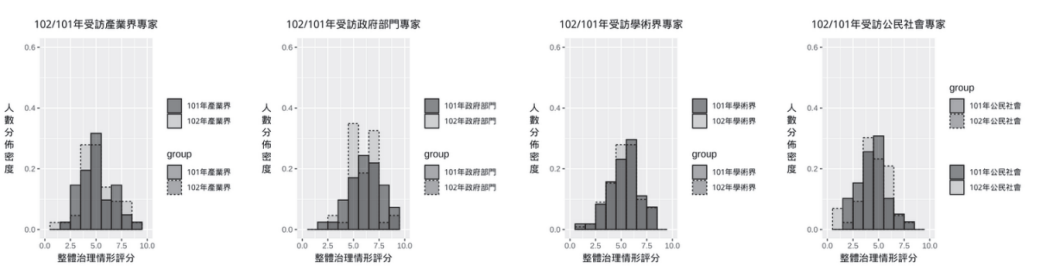
圖 1

不同專業領域的專家評價各年度整體治理表現

(a) Q-Q 圖視覺化比較



(b) 直方圖視覺化比較



資料來源：本研究。

簡單來說，Kolmogorov-Smirnov 檢定就是一種基於累計分佈函數的無母數檢定，用以檢定兩個不同領域的專家評價分布是否不同。表 7 提供了 2012 年、2013 年四大不同領域的專家評價在各個面向的比較，可以看到學術界、產業界專家的評

價十分接近（*P* 值均大於 0.05）；公民社會跟政府部門專家評價有極大之差異（*P* 值均小於 0.05，甚至出現小於 0.001），而與產業界專家相近（*P* 值較多出現大於 0.05）；同理，政府部門專家評價與其餘三組均有差異；產業界專家評價跟其餘三組差距較小。事實上，臺灣公共治理指標調查在 2017 年的最終結案報告提到了有趣的一段說明：焦點團體座談有與會者在檢視主觀問卷評估後，更進一步對各領域專家的填答狀況作出「政府部門自我肯定、產業界有期待、學術界要求高、公民社會百花齊放」之評語（蘇彩足等人，2017，頁 189），本研究對不同領域的專家異質性檢證也呼應了此一評語。綜言之，本研究的資料分析亦顯示，政府部門專家評價偏向滿意，學術界學者評價跨年度最一致，公民社會專家跨年度分數較會向兩極波動，產業界專家評價則界在公民社會跟政府部門專家評價之間。

表 7

不同專業領域的專家評價各個面向的 Kolmogorov-Smirnov 檢定

Kolmogorov-Smirnov 檢定		2013 年			2012 年		
面向	專業領域	產業界	政府部門	學術界	產業界	政府部門	學術界
法治化程度	公民社會	0.430	0.009	0.117	0.052	0.000	0.048
	學術界	0.930	0.030		0.818	0.024	
	政府部門	0.177			0.147		
政府效能	公民社會	0.111	0.000	0.009	0.010	0.000	0.000
	學術界	0.174	0.041		0.430	0.010	
	政府部門	0.017			0.007		
政府回應力	公民社會	0.286	0.002	0.069	0.782	0.001	0.205
	學術界	0.168	0.022		0.304	0.005	
	政府部門	0.177			0.098		
透明化程度	公民社會	0.782	0.001	0.205	0.002	0.000	0.001
	學術界	0.304	0.005		0.902	0.002	
	政府部門	0.004			0.014		
防治貪腐	公民社會	0.775	0.000	0.076	0.029	0.000	0.006
	學術界	0.408	0.001		0.295	0.000	
	政府部門	0.004			0.001		
課責程度	公民社會	0.777	0.000	0.011	0.234	0.000	0.021
	學術界	0.014	0.008		0.760	0.011	
	政府部門	0.000			0.014		
公共參與程度	公民社會	0.286	0.004	0.085	0.018	0.000	0.000
	學術界	0.222	0.062		0.313	0.036	
	政府部門	0.060			0.003		
整體治理評分	公民社會	0.367	0.000	0.003	0.103	0.000	0.004
	學術界	0.428	0.052		0.262	0.010	
	政府部門	0.014			0.005		

註：表中數值皆為 *P* 值，$*P < 0.05$, $**P < 0.01$, $***P < 0.001$。

資料來源：本研究。

伍、結論

在全球化與在地化的雙元發展趨勢下，當前公共治理的論述除了強調政府施政的效率、效能與回應外，也重視建立能夠因應政治、經濟、社會與文化等環境脈絡的制度，以有效解決層出不窮且錯綜複雜的公共問題，滿足人民的需求與期待。為了促使政府部門落實此治理理念，需要建構一套具體測量治理成效的方法與指標，包括如何評估及評估的內容，以系統性且全覽式地呈現評估結果，有效呼應全球治理指標的發展，助力邁向「善治」的境界。在此背景下，臺灣公共治理指標設計了一套結合主觀指標與客觀指標的評估體系。其中，本研究所採用 2012-2013 年的主觀指標調查，係針對我國政府在馬英九總統主政當時的治理狀況設計了七大面向及 20 個次面向，共計 121 個題目的主觀評估問卷。為確保評估的專業性，該計畫成立了「專家評估委員會」，邀請政府部門、學術界、產業界和公民社會等四個領域的專家組成，依據其專業背景與實務經驗給予適當的評價，這是國內公共治理研究有關專家取樣的重要里程碑。

然而，檢視國內過去公部門調查相關文獻可知，仍較少有實證研究針對專家評估是否存在異質性進行檢證，導致我們對國內專家評估的測量品質了解相對有限。因此，本研究藉由檢驗 2012 年和 2013 年的臺灣公共治理主觀指標調查中，特定專業領域的專家評分的分布，並比較不同專家之間的評分差異，深入探討專家異質性內涵中值得關注的議題。本研究主要是以受試者（專家）為中心，針對臺灣公共治理主觀指標調查中專家評估的異質性進行了深入分析，結果顯示不同專業領域的專家在評估過程中存在顯著差異，這對於理解公共治理的評價機制具有重要意義。

本文的實證分析結果進一步得到幾個研究發現，以下說明其在理論與實務上的意涵。首先，第一次參與和參與二次以上的專家，在填答「中間值（5）」的平均題數和標準差表現差不多，顯見他們對問卷題目的理解程度相近，不過，這兩群專家雖然在填答「無反應」選項的表現未達到統計上的顯著差異，然從回答無反應的平均數題數和標準差可以看到，第一次參與評估的專家不僅填答此選項的題數較多，標準差亦較大，反映出他們對某些評估指標的不熟悉感，未貿然給一個具體評價分數，且彼此之間填答無反應的題數差異亦甚大。其可能的解釋原因是，由於該項調查係以電子郵件邀請各個領域的專家參與線上調查，僅在邀請信和問卷開端說明調查目的和填答方式，並未面對面向專家說明和解釋各個面向和指標的意義，第

一次參與填答的專家若對某些指標不諳評估重點或判斷依據，便傾向選擇無反應的答案。因此，未來在專家調查實際操作上，建議可以針對第一次參與調查的專家舉辦調查填答說明會，協助他們熟悉調查內容；或是建立「預備專家」的制度，第一次參與填答的專家不納入分數計算，待其再參與填答時，對相關評估項目已經具有更充分的資訊，再納入計算。

其次，不同領域之專家對政府治理表現給予不同之評價，其中，政府部門的代表傾向給予較高評價，此一結果可能反映了他們對於政府施政的深刻理解與實際經驗，不過，相較於其他領域專家的評估結果，這種自我肯定也引發了對於評估結果是否存在角色偏見或自我感覺良好的質疑，因之，邀請外部專家代表參與評估應有助於降低此一偏誤，使評估結果更具有代表性和客觀性。另外，雖然在一般認知中，專家的專業背景不同可能影響他們對同一政策或治理現象的看法，不過，本研究發現，或許是學術訓練嚴謹所致，多數學者們答題有一致共識，學者人數雖多但仍呈現學者專家之間答題具有同質性。這一發現突顯了在此一公共治理評估中，過度仰賴學術界代表的意見，建議可以考慮專家的多元性與代表性，適度減少學術界的代表人數，擴充其他領域的專家代表，以確保評估結果即使是各個專家主觀的評價，整合後也能夠反映出符合科學的客觀價值。

最後，檢驗隨機專家假設的蒙地卡羅模擬過程與推論是本研究在研究方法創新上的主要貢獻之一，透過蒙地卡羅模擬的結果，本研究發現專家評估的隨機性在不同專業領域中存在顯著差異，這為未來的研究提供了新的視角。值得指出的是，除了學術界以外的專業領域中，模擬的隨機專家的標準差均會明顯縮小，尤其是在隨機去掉四個面向以上題目無須作答的情境。然而，百位以上學術界專家在改變大量避答數目的情況下，結果卻仍然類似於原始的真實專家評估結果，殊值注意。

據此，在方法論方面，有鑑於模擬專家在避答超過四個面向題目時，如果用其他人的答題數據來隨機補充，則容易錯過原本數據中的極端值，這會讓標準差變小，進而影響各面向答題結果的比較和統計檢定的結果。因之，未來也可以進一步利用蒙地卡羅模擬，探討不同專家樣本數的可能影響，並且考慮如何在不同領域中調整參與專家的數量及評估其答題質量。在實務方面，有鑑於評估的面向和指標數量甚多，考量各領域的專家不一定熟悉，本研究透過模擬方式發現，少量減少專家填答的面向並不會影響整體填答結果，尤其是學術界的專家，因受過嚴謹的學術訓練，或是立場較為中立超然，其填答的一致性較高、評分區間較窄，致隨機學者的模擬結果和原始學者改變不大。將來從事此類專家調查，或許就可以考慮根據每一

位專家的專業背景、經驗和個人意願，填答部分面向和題目即可，此作法不僅可以減輕專家的負擔和提高專家參與的意願，透過專注在熟悉的評估面向，亦應有助於提高填答的品質。

在本文的研究貢獻與限制方面，從調查方法論的角度來看，問卷題目過多可能導致受訪者產生疲乏效應，進而陷入制式回應或選擇終止作答。因此，在調查實務中，簡化問卷題項是一種較為理想的做法。在過去執行該調查的審議專家會議中，修正問卷內容和減少問題數量一直是主要推動方向。然而，文獻分析顯示，公共治理涉及非常廣泛的面向，例如全球治理指標的評量也是多元的，因此若一味減少面向題目，可能會導致重要資訊的遺漏，使得問卷題項難以精簡至極少數。儘管如此，研究結果顯示，在不縮減問卷內容的情況下，減少部分回答量（例如，本研究建議在七大面向中允許避答三個以下）並不會影響整體推論結果。這也呼應了受測專家的普遍期望，即希望能僅回答部分熟悉的面向。換言之，研究團隊可以根據各領域專家的專業背景指派其填答的面向，或者簡化問卷題項讓所有專家自行勾選填寫，同時要求專家至少回答一定數量的面向和題目。由於本研究使用次級資料，因此無法具體檢驗設計分類問卷讓不同類別專家填寫是否可行，或是簡化問卷題項讓所有專家填寫是否更有效。這一點是本研究的限制，同時也是未來值得深入探討的方法議題。

此外，本研究針對專家填答的異質性進行檢證，可用來探索專家評分的代表性與調查成本（包括執行時間和人力投入）之間的平衡。如果專家的填答具有同質性，則適度調整或減少專家的人數應不會影響評分的代表性。本研究結果顯示，學者專家的回答普遍較其他領域更具同質性，因此未來在取樣時無需選取大量學者樣本，這樣可以有效節省調查成本。再者，兩年間參與次數不同的專家在填答「無反應」與「中間值」方面也展現出一致性。雖然，本研究對專家調查作出若干貢獻，但不同領域專家的最適規模和問卷填答的最適題數仍然缺乏具體答案，且可能因「根據個案具體情況」而有所不同，因此仍需更多實證研究來進一步解答。然而，本文在研究方法與測量方面在國內公共行政類研究具有創新意涵。未來相關的專家調查議題建議可採用本文所介紹的研究方法進行探索，除了檢證隨機專家假設外，不同的受訪者可能對問卷題目的中立選項有不同的解讀，有些人可能將其視為「普通」或「一般」，而另一些人則可能將其理解為「不確定」或「不願意表達意見」，為了減少中立選項的不明確性，未來調查設計者可以根據本研究對此議題的討論，考慮提供更具體的答題指導。總而言之，本研究不僅有助於理解專家異質性

對公共治理評估結果的影響，還能為改進專家調查的研究設計及推廣簡化作答方式提供重要參考。

參考文獻

林希偉、王國全（2009）。專家區間估計中過度自信現象之研究。**管理與系統，16**（2），181-200。[Lin, S.-W., & Wang, K.-C. (2009). Expert overconfidence in probability interval estimations. *Journal of Management and Systems*, *16*(2), 181-200.]

莊文忠、洪永泰、陳俊明（2012）。**臺灣公共治理指標調查**（編號：PG10104-0128）。行政院研究發展考核委員會。[Juang, W.-J., Hung, Y.-T., & Chen, C.-M. (2012). *Taiwan public governance indicators* (Project number: PG10104-0128). Research, Development and Evaluation Commission.]

莊文忠、洪永泰、陳俊明（2013）。**臺灣公共治理指標調查**（編號：PG10204-0228）。行政院研究發展考核委員會。[Juang, W.-J., Hung, Y.-T., & Chen, C.-M. (2013). *Taiwan public governance indicators* (Project number: PG10204-0228). Research, Development and Evaluation Commission.]

陳文學、孫同文、史美強（2013）。國際組織對公共治理之運用與研究，**公共治理季刊，1**（1），37-51。[Chen, W.-H., Sun, M. T.-W., & Shih, M.-Q. Roles of international organizations in public governance: Application and research perspectives, *public governance quarterly*, *1*(1), 37-51.]

蘇彩足、莊文忠、郭銘峰（2017）。**臺灣公共治理指標調查及公共治理相關議題研究**（編號：PG10606-0040）。行政院國家發展委員會。[Su, T.-T., Juang, W.-J., & Kuo, M.-F. (2017). *Investigation of public governance metrics and research on public governance topics in Taiwan* (Project number: PG10606-0040). National Development Council.]

Budge, I. (2000). Expert judgements of party policy positions: Uses and limitations in political research. *European Journal of Political Research*, *35*(1), 103-113.

Carmines, E. G., & Kuklinski, J. H. (1990). Incentives, opportunities, and the logic of public opinion in American political representation. *Information and democratic processes*, 240-268.

Cederman, L. E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, *355*(6324), 474-476.

Clemen, R. T. (2008). Comment on Cooke's Classical Method. *Reliability Engineering & System Safety*, *93*(5), 760-765.

Cooke, R. M. (1991). Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press.

Cooke, R. M., & Goossens, L. L. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety*, *93*(5), 657-674.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis: A global perspective* (8th ed.). Cengage Learning.

Hanea, A. M., Nane, G. F., Bedford, T., & French, S. (Eds.). (2021). *Expert judgement in risk and decision analysis.* Springer.

Haselswerdt, J., & Rigby, E. (2021). What do advocates want from policy research? Evidence from elite surveys. *Evidence & Policy*, *17*(3), 385-403.

Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302-317.

Kaufmann, D., & Kraay, A. (2007). Governance indicators: Where are we, Where should we be going? *The World Bank Research Observer*, *23*(1), 1-30.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of Survey Research* (pp. 263-313). Emerald Group.

Lupia, A. (1994). The effect of information on voting behavior and electoral outcomes: An experimental study of direct legislation. *Public Choice*, *78*(1), 65-86.

Martinez i Coma, F., & Van Ham, C. (2015). Can experts judge elections? Testing the validity of expert judgments for measuring election integrity. *European Journal of Political Research*, *54*(2), 305-325.

Muthén, B. O., & Muthén, L. K. (2000). Integrating Person-Centered and Variable-Centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*(6), 882-891.

Peltier, T. R. (2010). *Information Security Risk Analysis* (3rd ed.). Auerbach Publications.

Rios, J., & Rios, Insua, D. (2012). Adversarial risk analysis for counterterrorism modelling. *Risk Analysis*, *32*(5), 894-915.

Schwarz, G., Eva, N., & Newman, A. (2020). Can public leadership increase public service motivation and job performance? *Public administration review*, *80*(4), 543-554.

Shlyakhter, A. I., Kammen, D. M., Broido, C. L., & Wilson, R. (1994). Quantifying the credibility of energy projections from trends in past data: The United States

energy sector. *Energy Policy*, *22*(2), 119-130.

World Bank, (1994). *Governance: The World Bank's experience*. World Bank Group. 檢索日期 2025 年 10 月 6 日，取自 https://documents.worldbank.org/en/publication/documents-reports/documentdetail/711471468765285964/governance-the-world-banks-experience

Expert Heterogeneity Assessment: Verification Based on the Subjective Indicators Survey of Public Governance in Taiwan

Wen-Jong Juang*, Mei-Rong Lin**, Shun-Chuan Chang***

Abstract

Due to the high level of expertise required in many public issues, the general public often finds it difficult to fully understand these topics and provide concrete evaluations. Consequently, many international organizations rely on elite or expert surveys to collect scoring data and professional opinions for governance assessments. Since 2008, Taiwan's National Development Council has commissioned the Taiwan Public Governance Indicator Survey, conducted by the Taiwan Public Governance Research Center, for this purpose. However, existing domestic literature reveals a lack of empirical studies assessing the impact of expert heterogeneity on survey outcomes, which constrains our understanding of the measurement quality of expert surveys. This study draws on expert survey data from the 2012 and 2013 Taiwan Public Governance Indicators and applies Q-Q plots, Monte Carlo simulations, and a series of statistical tests to examine three methodological issues: (1) patterns of expert

* Professor, Department and Graduate Institute of Political Science, National Chung Cheng University. email: wenjong@ccu.edu.tw

** Associate Professor, Department of International Business, Tamkang University. email: 134660@mail.tku.edu.tw

*** Associate Professor, Center of Holistic Education, Mackay Medical University. email: zhang@mmu.edu.tw (corresponding author, supported by MMC-RD-113-1B-P001)

responses involving non-response and central tendency bias, (2) sampling simulations and hypothesis testing under the random expert assumption, and (3) heterogeneity among experts from different professional fields.

The results indicate that moderately reducing expert response requirements (e.g., allowing respondents to skip up to three of the seven major dimensions) does not affect inferential validity, particularly as academic experts tend to provide more homogeneous responses than experts from other fields. Moreover, experts with different participation frequencies over the two years demonstrated consistent patterns in terms of non-response and midpoint selection. Beyond contributing to the understanding of expert heterogeneity in governance evaluations, this study offers practical recommendations for improving the design of expert surveys: organizing briefings for first-time participants to clarify survey objectives and questionnaire content, balancing the ratio of internal government and external experts to reduce bias, limiting highly homogeneous academic samples while incorporating more diverse perspectives from business and civil society, and adopting simplified or specialized response assignments to ensure stability in overall assessment quality.

Keywords: public governance, expert heterogeneity, random expert hypothesis, Monte Carlo Simulation, Quantile-Quantile Plots