

大型教育調查研究實務

以TASA為例 ▶▶▶▶▶▶

Using Large-scale Assessment Datasets for Research:
Taiwan Assessment of Student Achievement (TASA)



目錄

- 001 **第一章** ——
大型教育調查與臺灣學生學習成就評量的源起
- 021 **第二章** ——
TASA 學科成就評量架構
- 049 **第三章** ——
背景問卷的心理計量特性
- 085 **第四章** ——
抽樣設計與權重
- 105 **第五章** ——
測驗設計與量尺化程序
- 129 **第六章** ——
TASA 資料庫的二次分析
- 169 **第七章** ——
應用決策樹建立學生學習模型
- 185 **第八章** ——
臺灣學生學習成就評量資料庫 (TASA) 轉型內涵
-- 邁向 108 課程之素養導向大型評量模式

第一章 大型教育調查與臺灣學生學習成就評量的源起

蕭 儒 棠

國家教育研究院助理研究員

一、緒 論

評量 (assessment) 是一種透過測驗或其它方式獲得訊息，進而針對個人、事物，或程序進行推論的系統性方法 (AERA, APA, & NCME, 1999)。而大型評量 (Large-scale assessments) 則是一種包含大量受試對象的大規模標準化評量，依照評量目的，評量的規模可能是區域性、全國性，或國際性的。在實務方面，大型評量在教育政策、績效責任，與教育規劃等面向，提供重要的參考資料；在研究方面，無論是目標的設定、評量的規劃、工具的發展、施測的方式，或是結果的運用，每個環節都是學界相當重視的研究領域。本章首先將介紹「國際教育成就比較協會」(International Association for Evaluation of Educational Achievement, 簡稱 IEA) 發展國際大型評量的緣起，接著述說美國創設「國家教育進展評量」(National Assessment of Educational Progress, 簡稱 NAEP) 的始末，最後再回到我國自行發展的「臺灣學生學習成就評量」(Taiwan Assessment of Student Achievement, 簡稱 TASA)，藉由國內外發展大型評量的經驗，認識評量目的、評量設計，與評量結果運用時應考量的面向。

二、國際大型評量誕生：1958 年

（一）理念萌發

「國際教育成就比較協會」(International Association for Evaluation of Educational Achievement, 簡稱 IEA)，於 1967 年正式以法人形式成立，是第一個為研發並進行國際性評量而成立的組織，其宗旨在提供國際性評量基準，協助教育決策者釐清可能影響教學與學習的因素，並比較各國教育制度與教育場域的優缺點，進而為試圖進行教育監測或教育改革的國家提供可行的參考策略。IEA 隸屬於「聯合國教育、科學與文化組織」(United Nations Educational, Scientific and Cultural Organization, 簡稱 UNESCO)，其成立的動機，來自 1958 年「UNESCO 教育研究所」(UNESCO Institute for Education, 簡稱 UIE) 的會議決議。會議中與會的學者們思考，應如何透過實徵的調查方式，進行跨國的教育制度比較，並評估推動大規模國際教育成就調查的可行性 (IEA, 2017a)。

為了更全面性地了解教育資源投入與學生學習成效之間的關係，會中來自教育心理學、心理計量學，以及比較教育學等領域的學者認為，各國在推動教育政策時，為因應各式不同的挑戰，必定會考量其國情，提出相應的對策。因此，可將各個參與國家的教育體系視為獨立的實驗單元，透過跨國共同組成的研究計畫，研究每個教育體系在「投入 - 成效」的關係。研究成果不只可作為各教育體系自身制定政策的依據，相關經驗也可作為其他成員的參考。

（二）評量誕生

1959 ~ 1962 年，IEA 針對 13 歲學生的數學、科學、地理，與閱讀

理解所進行的「前導性十二國研究」(Pilot Twelve-Country Study, 簡稱 PTCS), 是第一個針對學生學習成就進行跨國比較的大型國際評量。PTCS 的經驗不只證實了進行跨國大型評量的可行性, IEA 的研究者們進一步比對施測內容與學習內容後發現, 相對於科學與閱讀兩個領域, 語言對數學的學習影響較小, 因此, IEA 選擇數學作為下一波跨國大型評量的調查科目。1964 年, IEA 針對 13 歲以及大學入學前一年的學生, 進行「第一次國際數學研究」(The First International Mathematics Study, 簡稱 FIMS), 參與調查的教育體系同樣是 PTCS 包含的十二個國家 (IEA, 2017b)。

經過 PTCS 和 FIMS 兩次成功的經驗後, IEA 決定進一步擴大調查規模, 探討學生在其它學科的學習狀況, 是否也可能是影響數學學習的因素。因此, IEA 於 1970 ~ 1971 年推動「六學科研究」(Six-Subjects Study), 包含的調查研究分別是「第一次國際科學研究」(The First International Science Study, 簡稱 FISS)、「閱讀理解研究」(The Study of Reading Comprehension)、「文學教育研究」(The Study of Literature Education)、「外語研究—英文」(The Study of English as a Foreign Language)、「外語研究—法文」(The Study of French as a Foreign Language), 以及「公民教育研究」(The Study of Civic Education)。在「六學科研究」的調查中, IEA 也同時考慮了學習興趣、動機與態度、教學方法, 與學校實務 (school practice) 等, 以進一步了解影響學生學習成就的因素 (IEA, 2017c)。

除了擴大調查學科廣度的「六學科研究」, IEA 也試圖深入學生在數學與科學學習成就的研究。在數學領域方面, 「第二次國際數學研究」(The Second International Mathematics Study, 簡稱 SIMS) 於 1980 ~ 1981 年進行, 調查的重點是數學教育中, 課程、教學, 與成效三者的關係。此外, SIMS 為了蒐集更詳細的資料, 在一個學年的教學前後, 特別在研究設計中考慮了前測與後測, 以分析學生數學的學習成就變化; 在科學領域方面, IEA 於 1983 ~ 1984 年再度進行科學領域的研究, 稱為「第二次國際科學研究」(The Second International Science Study, 簡稱 SISS), 以檢視各國學生

在 FISS 和 SISS 兩次科學學習成就的變化 (IEA, 2017c)。

(三) 趨勢調查

「數學研究」(FIMS 和 SIMS) 與「科學研究」(FISS 和 SISS) 各自經歷了兩次的調查後，IEA 整合兩個領域的調查，於 1995 年進行了「第三次國際數學與科學教育成就研究」(Third International Mathematics and Science Study 1995，簡稱 TIMSS 1995)，自此已具備現今國際所熟悉 TIMSS 的雛形，其研究對象主要為四年級、八年級學生 (Beaton, 1996)。另外 IEA 也針對高中最後一年的學生進行數學與物理的評量，稱之為 TIMSS Advanced。1999 年，IEA 又舉辦「第三次國際數學與科學教育成就研究後續調查」(Third International Mathematics and Science Study Repeat，簡稱 TIMSS-R) 以了解學生在科學與數學的學習成就趨勢發展，從此，TIMSS 開始成為 4 年一個調查週期的數學與科學國際調查研究 (Ruddock, 2000)。2003 年，依往例應命名「第四次國際數學與科學研究計畫」(Fourth International Mathematics and Science Study) 的調查，由於 IEA 將原本的 Third (第三次) 改以 Trends (趨勢) 代替，將計畫的名稱改為「國際數學與科學教育成就趨勢調查」(Trends in International Mathematics and Science Study，簡稱 TIMSS)，不只兼顧原本的簡稱 TIMSS，也突顯了趨勢研究的特性。TIMSS 每次調查對象同樣是四年級與八年級的學生，但其中八年級學生與前一次調查的四年級學生來自相同的母群，研究者可透過統計方法分析學生表現的變化 (IEA, 2017d)。

除了數學與科學領域，寫作與閱讀也是 IEA 調查研究的重點。在寫作方面，1984 ~ 1985 年的「寫作研究」(Written Composition Study)，試圖了解影響寫作能力的因素，並檢視學校的寫作教學的成效，調查對象包含小學、國中和高中最後一年的學生。在閱讀方面，經過「前導性十二國研究」與「六學科研究」兩次包含閱讀能力評量的跨國調查後，為了建構一套有效的閱讀

發展評量工具，作為閱讀素養的表現基準，IEA 於 1990 ~ 1991 年推動「閱讀素養研究」(Reading Literacy Study)，調查對象有 9 歲與 14 歲的學生。此外，IEA 也於 2006 年開始推動「促進國際閱讀素養研究」(Progress in International Reading Literacy Study，簡稱 PIRLS)，進行 5 年一次的週期性調查，評量小學四年級學生的閱讀素養成就，目的在研究、比較不同教育政策、教學環境、教學方法下四年級學生的閱讀能力 (IEA, 2017e)。繼 1995 年針對數學與科學領域的 TIMSS，評量學生閱讀素養的 PIRLS，也正式於 2006 年誕生，成為 IEA 主導下，另一個週期性進行的國際大型趨勢調查。

TIMSS 和 PIRLS 是兩項由 IEA 主導，較為臺灣民眾熟知，且與學科內容直接相關的國際大型評量。另外，IEA 也針對如何培養現代公民，發展了「公民教育研究」(Civic Education Study，簡稱 CIVED)，分別於 1971 年(六學科研究)與 1994 年進行兩次調查工作。2009 年，IEA 考量週期性調查的評量趨勢，發起「國際公民教育與素養調查計畫」(International Civic and Citizenship Education Study，簡稱 ICCS)，以 7 年為一個週期，針對 13 歲半(相當於我國 8 年級)學生，進行公民學習成就調查。ICCS 評量架構評估的範圍包含公民知識、公民參與，以及態度、行為與認知等相關因素，以了解各國青少年對成為未來公民的教育準備，並進行各參與地區或國家研究成果之比較，調查對象包含受試學生的授課教師及該校校長 (IEA, 2017f)。

除了臺灣參與的 TIMSS、PIRLS，和 ICCS 等國際大型評量，1958 年以來，IEA 持續調查了學生在數學、科學和閱讀等科目的能力，對公民和公民教育進行評估，也調查了學生的電腦和資訊素養。目標是協助所有參與成員了解各國的教育成效，進而依據實徵資料制定教育政策或進行教育革新，其它由 IEA 主導的調查與研究成果，讀者可逕行參考 IEA 官方網頁的資料。

(四) 異軍突起

除了 IEA 主導的各項國際大型評量，另一項著名的國際大型評量，

則是「國際學生能力評量計畫」(The Programme for International Student Assessment, 簡稱 PISA)，它由成立於 1961 年的「經濟合作暨發展組織」(The Organization for Economic Cooperation and Development, 簡稱 OECD) 所主導的，每 3 年為一個調查週期。PISA 評量的目的在評估學生是否於義務教育階段獲得進入社會或下一教育階段應具備的素養，強調真實生活情境中，將知識用於解決問題的能力，而不是對特定課程內容的熟練程度，考量多數國家義務教育結束的年齡為 15 歲，因此，PISA 以 15 歲學生為調查對象 (OECD, 2017a)。

PISA 涵蓋三個不同素養領域，分別是數學、科學以及閱讀，每一次的調查均包含閱讀、數學與科學三個素養，但每一次調查只以一個素養為主，例如，2012 年主要測量的素養領域為數學，閱讀以及科學為次領域。2000 (或 2009) 年是閱讀，2003 (或 2012) 年是數學，2006 (或 2015) 年則為科學素養，2018 年的 PISA 則重新回到閱讀素養為主的評量。OECD 自 1997 年起開始規劃 PISA，2000 年正式推出後，參與的國家與經濟體持續增加，2000 年有 32 個、2003 年有 41 個、2006 年有 57 個、2009 以及 2012 年有 65 個，至 2015 年已經有超過 72 個國家與經濟體參與。2012 年，PISA 嘗試推出數位化評量，共有 32 個國家參與，2015 年開始，PISA 全面採取數位化評量。

除了與學科領域相關的調查，第一個針對學校學習環境與教師工作情況的國際性調查，則是另一個由 OECD 主導的「教學與學習國際調查」(Teaching and Learning International Survey, 簡稱 TALIS)。TALIS 於 2008 年第一次展開正式調查，蒐集教師與校長提供的資訊，做為政策研擬的參考，主要調查的對象是國中年段的教師和校長，國小和高中職年段則由參與調查國家自行決定是否參與。TALIS 於 2008 年舉辦第一屆，每 5 年一個調查週期，第一屆有 24 個國家參與，2018 年的第三屆則增加至 47 個國家 (OECD, 2017b)。

三、國家教育進展評量：1969 年

（一）國際競爭

美國是一個聯邦制國家，聯邦政府在對外事務享有明確的權力，但是，國內事務上應如何行使權力，尤其在教育職權的劃分，美國憲法並未做出明確的規定。在教育相關事務中，州政府究竟應聽從聯邦的命令，或者由州政府自行決策，一直是聯邦與州政府間存在的爭議。雖然美國國會於 1867 年設立教育部時，即賦予教育部進行資料統計與概況蒐集的任務，並公布不同地區的教育現況與進展，以促進全國整體的教育發展，但聯邦政府並未積極參與教育相關事務。

直到 1957 年 10 月，蘇聯成功發射 Sputnik 號人造衛星後，二次世界大戰後美國民眾的樂觀與自信開始動搖，美國各界開始檢討自己國家的科學發展與工程技術現況，各種要求教育改革的呼籲開始浮現。為此，美國國會於 1958 年通過《國防教育法案》（National Defense Education Act，簡稱 NDEA），其目的在提升教育品質，培養美國公民應付社會急遽變遷，以及面對國際競爭所需的能力（Urban, 2010）。法案授權美國聯邦政府撥款協助地方的教育需求，進行教育研究、教育實驗，以及各種具體的改革措施。聯邦政府特別在外語、數學，和科學等課程挹注大量資源，試圖在短時間內培養大量工程與科學人力。雖然聯邦政府開始依據《國防教育法案》參與教育相關事務，但並未真正落實教育部資料蒐集的任務。

（二）行政權責

1962 年，美國總統約翰·甘迺迪（John F. Kennedy）任命弗朗西斯·

凱普爾（Francis Keppel）擔任教育辦公室（Office of Education）的教育專員（Commissioner of Education），凱普爾就職時注意到，將近 100 年過去了，聯邦政府在教育事務方面，仍然缺少全面性、系統性的數據。當時蒐集的資料著重於教育資源的投入，例如，行政人員數量、教師數量、建築預算，或學生註冊率等。至於教育的成效，例如學生對知識或技能的學習成果，當時能掌握的數據並不豐富。少數可參考的數據來自「美國大學測驗」（American College Testing，簡稱 ACT）與「美國大學入學考試委員會」（College Entrance Examination Board，簡稱 CEEB），但這兩項測驗的測驗對象是申請進入大學的高中學生，其結果並無法推論全國整體的教育成效。

就在 IEA 發展「前導性十二國研究」的同時，在幾個著名學者與政策制定者的領導之下，美國也開始進行相同的工作，試圖了解「投入—成效」間的相關性。1963 年，凱普爾邀請史丹佛大學（Stanford University）行為科學高等研究中心（The Center for Advanced Study in the Behavioral Science）主任拉爾夫·泰勒（Ralph W. Tyler），針對學生的學習成效調查，發展一套定期的全國性評量計畫。凱普爾相信，國會和民眾有權了解全國學生的學習狀況，且相關的調查數據，是如何幫助弱勢學生改善學習成效的重要參考（Jones, 2003）。

然而，這項計畫卻招致了許多反對的聲浪。凱普爾和泰勒知道，進行全國性的學習調查，面臨的最大挑戰不在技術層面，來自政治上的壓力才是最大的絆腳石。反對方認為，由聯邦政府建立的全國性評量計畫，並不是單純行使國會賦予的職責，它同時顯示聯邦向州政府擴權的企圖，更意味著未來進一步建立全國統一課程的可能性，甚至有控制學校的意圖。凱普爾理解到，評量計畫由聯邦政府主導是反對方主要的疑慮來源，他一方面解釋自己只是忠實地履行國會賦予的職責，一方面也同時思考其它可行的評量計畫執行方式（Kirsch et. al., 2013）。

(三) 求同存異

由於凱普爾的父親，弗雷德里克·保羅·凱普爾（Frederick Paul Keppel）曾於 1923 ~ 1941 年，擔任紐約卡內基基金會（Carnegie Corporation of New York）的董事長，紐約卡內基基金會是一個以「促進知識和理解的進步與傳播」為宗旨的非營利組織，1911 年由美國鋼鐵大王安德魯卡內基（Andrew Carnegie）成立於紐約。透過這一層關係，評量計畫所需的資金與研發工作均由紐約卡內基基金會負責，以消除聯邦政府試圖控制學校或課程的疑慮（Finder, 2004）。該計畫由基金會 1955 ~ 1967 年擔任董事長的約翰·加德納（John W. Gardner）作為計畫負責人，他不只是專業的心理學家，1965 ~ 1968 年也在美國總統林登·詹森（Lyndon Baines Johnson）任內擔任衛生、教育和福利部（Department of Health, Education and Welfare）部長，以及 1965 年白宮教育大會（White House Conference on Education）的主席。

透過紐約卡內基基金會的財政支持，凱普爾等人於 1963 年 12 月與 1964 年 1 月召開了二次預備會議，徵詢各界對進行全國性學生評量的意見。會中與會學者指出，全國性學生評量是一個可行的計畫，但其中具有潛藏的隱憂。他們認為評量的結果是一把雙面刃，一方面可用於協助弱勢者的學習，但也可能因此而傷害受試的學生或學校。由於擔心評量結果會用來做不當或有害的比較，「美國學校行政人員協會」（The American Association of School Administrators，簡稱 AASA）等組織，在會中強烈反對進行各州評量，並公布各州的評量結果（Jones, 2003）。

為了化解實施全國性評量可能引發的疑慮，同時也為了尋求全國教育機構、政府機構、測驗專家，以及一般民眾的支持。1964 年 6 月，「教育進展評量探索委員會」（Exploratory Committee for the Assessment of Progress in Education，簡稱 ECAPE）成立，由泰勒擔任主席。此外，為了將來自各界的意見整合成具體可行的方案，1965 年「技術諮詢委員會」（Technical Advisory Committee，簡稱 TAC）成立，由普林斯頓大學統計學系主任，同

時也是 AT & T 貝爾實驗室資訊系統研究的副執行總監，約翰·圖基（John Tukey）主持。考量針對學校課程內容進行評量可能造成不當比較的疑慮，TAC 最後決議，課程內容不作為評量的目標，將由專門小組根據公民應該了解的內容設計各領域的評量目標（NAEP, 2017b）。

（四）評量實施

由於進入學校對學生施測需要學校的同意，ECAPE 特別重視 AASA 的意見，1965 年至 1967 年間，泰勒與全美各地點教育機構的代表進行了深入的討論。泰勒同意審慎思考 AASA 的憂慮，並保證不會針對個別學生、學校、學區發布評量結果，也不會特別針對州的層級發布報告。AASA 方面為此特別設立了「AASA-ECAPE 聯合委員會」（joint AASA-ECAPE Committee），進一步研究相關議題（Jones, 2003）。

1966 年 AASA 年會期間，泰勒指出全國性評量的結果並不用於州與州之間、社區與社區之間的競爭或比較，而是以美國西北、東南、西部和遠西部（far west）四個不同區域來進行報告，藉此緩和來自 AASA 的疑慮。1967 年，「國家教育協會」（National Education Association）也通過了與 ECAPE 保持合作的決議，全國性的評量上路的时间似乎已經不遠了（Finder, 2004; Jones, 2003）。1969 年，評量計畫轉由「美國各州教育委員會」（the Education Committee of the States，簡稱 ECS）負責，教育界對全國學生評量計畫的敵意才逐漸消失。今天，凱普爾與泰勒的評量計畫被稱為「國家教育進展評量」（National Assessment of Educational Progress，簡稱 NAEP）。

從 1963 年泰勒提出計畫草案，歷經 6 年的努力，NAEP 終於 1969 年正式上路，針對學生在各個學科領域學習成就和進展作全國性的調查。NAEP 系統性的蒐集並報告全國與各州的學生學習資訊，以評估全國教育的概況與進展，相關的訊息收錄於每年出版的「國家報告卡」（Nation's Report Card）中，美國聯邦政府真正落實國會於 1867 年即賦予的職責。

四、臺灣學習成就評量：2004 年

(一) 入學測驗

雖然早在 1960 年代開始，國際性或全國性的大型評量已經在歐美進行，但如同美國早期只能透過入學測驗了解學生的學習情況，臺灣的情況也是如此。臺灣的大學入學測驗，有 1954 年由國立臺灣大學、臺灣省立師範學院（今國立臺灣師範大學）、臺灣省立農學院（今國立中興大學）、臺灣省立工學院（今國立成功大學）4 所公立大學組成的「大專聯招會」，是臺灣大學聯合招生的濫觴（陳揚琳，1983）。近年來，大學招生考試又轉型為「大學入學學科能力測驗」（學測）以及「大學入學指定科目測驗」（指考）；另外，在高中入學測驗方面，1965 年臺北市 5 所公立高中仿效大學校院聯合招生的模式，成立「臺北市五省中聯合招生委員會」，以聯招方式招生。後續，為因應時代變遷與課綱的調整，又先後設立了「國民中學學生基本學力測驗」（基測）以及「國中教育會考」（會考）。

然而，這些測驗只能依據當年度測驗的結果，衡量個別學生在全體考生中的相對位置，做為學生進入各校系的依據，無法有效檢驗學生的學習成效，更遑論監控整體教育成效的演變。在這些入學測驗中，每個科目的測驗限制在一堂課左右的時間內完成，有限的時間、有限的試題，學生能作答的試題無法涵蓋學科的全部內容，再加上測驗只有學科成績，無法結合與學生、教師、學校，或是家庭相關的背景資料，在在顯示其應用的侷限，以至於行政部門在制定教育政策，或進行教育與課程改革時，始終缺乏實徵資料作為依據。一直到進入二十一世紀之前，對臺灣而言，大型評量所指的仍是各式的入學測驗。

（二）長期追蹤

「臺灣教育長期追蹤資料庫」（Taiwan Education Panel Survey，簡稱 TEPS）是 2001 年至 2007 年，由中央研究院、教育部、國立教育研究院籌備處，和國科會共同規劃的全國性長期的調查計畫，以問卷調查方式蒐集資料（TEPS, 2017）。TEPS 除了針對國中、高中、高職，及五專學生進行追蹤，計畫同時以學生為核心，將調查對象擴及可能影響學生學習經驗的家長、教師，和學校等，研究的重點在學校與家庭學習環境對學生的影響。

除了背景問卷的填寫之外，TEPS 也要求受測學生完成「綜合分析能力測驗」，測驗運用各種題材測量學生的問題解決能力，而非課程內容，因此，TEPS 並非一般針對各學科內容進行施測的學習成就測驗。透過 TEPS 的追蹤，研究者可由國中學生的樣本，得到接受國中階段的教育後，學生能力的變化情形。此外，由於 TEPS 調查的時間適逢「九年一貫課程」、「多元入學方案」等變革，以及「國民中學學生基本學力測驗」的實施，因此，TEPS 不只提供學者專家深入研究的資料，相關研究結果也是行政部門研擬政策的重要參考（張荳雲，2003）。

正如前文所述，行之有年的各項入學測驗，其測驗時間與試題內容有限，無法涵蓋全部課程內容，且因測驗目的並不相同，入學測驗也不適合用於追蹤學生的學習成效；另一方面，剛起步的 TEPS，其建置目的是蒐集全國學生在身心發展、學習活動、學習成就表現等實徵數據，雖有追蹤的功能，但其重點是在研究學校與家庭環境對學生學習的影響，並不針對學習內容進行評量，且 TEPS 資料蒐集的對象僅局限於國高中學生，並不包含國小學生。

（三）評量建置

進入二十一世紀的臺灣，在國小、國中，與高中職學生的學習成就，以及各級學校課程實施成效的監控方面，仍然缺少有效的評量工具。為了兼顧

課程內容與學力追蹤兩項需求，常態性蒐集資料，建置一套完整且客觀的全國性學生學習成就資料庫已是各界的共識。因此，教育部於 2004 年函請國立教育研究院籌備處（2011 年正式成立後，正式名稱為國家教育研究院）規劃「臺灣學生學習成就評量」（Taiwan Assessment of Student Achievement，簡稱 TASA）建置，用以發展量化指標和標準化測量工具，以檢視學生學習成就的表現及其差異，進而了解課程實施的成效，協助課程發展之進行與相關教育政策之研擬。TASA 資料庫建置的目的為（臺灣學生學習成就評量資料庫建置計畫，2017）：

1. 建立國民中小學、高中及高職學生學習成就長期資料庫，以追蹤、分析學生在學習上變遷之趨勢，進而檢視目前課程與教學實施成效。
2. 提供完整、標準化的學習成就資料，作為分析學生學習成就上差異表現變項資料，以評估學生未來在學術方面能力之發展與社會期許。
3. 了解國內學校教學及學生學習成效之現況，作為課程與教學政策改進之參考，並為縣市 政府教育局及學校推動補救教學之重要參據。
4. 提供各縣市學生學習表現資料，建立與縣市合作機制，以擴大資料庫應用效益。
5. 以資料庫的量化資料，提供國內外相關研究人員，深入探討學生學習成就方面的相關政策議題。
6. 建立本國學生學習成就評量資料庫，同時考慮與國際接軌，利於加入國際比較行列，藉以了解臺灣教育之獨特面與優缺點。

（四）評量工具

TASA 的施測對象包含國小（四、六年級）、國中（八年級），與高中職（二年級）學生，施測科目因年段不同而有調整，國小四年級，每位學生由

國語文、數學科、自然科等三科中抽測二科；國小六年級、國中八年級，及高中（職）二年級，每位學生由國語文、英語文、數學科、自然科，以及社會科等五科中抽測二科。另外，TASA 也加考國語文寫作、英語文的寫作與口說，稱之為特殊考科。除了五個考科外，每位受測學生均須填寫學生問卷，目的在蒐集學生學習、家庭等背景資料，作為資料庫變項分析使用。另外，每個受測學校也委請校長填寫學校問卷，回答關於校長年資、教學現況、學校資源、學生在學率、家長參與學校活動等問題。

（五）抽樣設計

TASA 透過全國性樣本進行抽測，藉以推估全國國小、國中及高中（職）學生，於各領域科目學習成就的表現、統計數據及分析資料作為擬訂教育政策之參考，或發展為縣市學力檢測或補救教學工具。為確保所抽取的樣本具有全國代表性，並考量臺灣各縣市人口數的差異，TASA 在國中小年段採二階段隨機抽樣，第一階段採分層叢集隨機抽樣，依據縣市、人口密度、與班級數等三個變項進行分層，第二階段則以所抽到的樣本學校，以學生個人為單位，進行簡單隨機抽樣；在高中職部分，TASA 採全國學校普測，學生部分則依比例進行抽測。此外，考量離島縣市樣本數可能不足的問題，TASA 抽樣時，學生人數最少的連江縣全縣普測，而金門縣和澎湖縣則針對抽到的學校進行全校普測，以增加離島縣市的樣本代表性。

（六）等化設計

為了有效涵蓋所有課程內容與測驗目標，國內外大型評量每個科目每次均包含大量試題，施測時要求受試者於有限時間內完成所有試題並不切實際，倘若延長施測時間，讓受試者有足夠的完成時間，多半會因為受試者逐漸疲倦而影響後續能力估計的準確性。因此，為了同時測量所有測驗目標，且考慮施測時間的限制，TASA 參考國際大型評量的經驗，施測題本採取

BIB 設計 (balanced incomplete block design)，將題庫中的試題分為若干區塊，區塊間與區塊內的試題均不重複，每位受試者僅回答部分區塊試題，後續再透過等化程序，建立可互相比較的共同量尺。

(七) 資料釋出

TASA 自 2004 年起，由教育研究院籌備處提出建置計畫報部後試行，2004 年至 2006 年期間，委託各大學進行相關試題研發，2006 年起，由國家教育研究院承接試題研發及調查施測相關工作。並於 2008 年調整施測週期為國小（四、六年級）、國中（八年級）、高中職（二年級），每年段三年為一循環，2008 年至 2016 年施測考科與年段說明如表 1。TASA 調查結果與相關資料包括公告試題（含試題品質分析）、問卷題目與學生作答反應、學生學科能力的可能值。其中，學生問卷提供學生基本人口學訊息，性別、家庭組成及父母原生國籍等變項，可歸為背景變項、家庭結構、學習歸因、學習策略、學習特質等五大類，而學校問卷委請各校校長或其職務代理人填寫，提供教學現況、學校資源、師資專長、組織氣氛等相關變項。提供大專校院或研究機構之教學及研究人員申請使用，於申請案收件之次日起，一個月內召開審議諮詢委員會，並以函文通知申請人審核結果。此外，針對本書使用之資料，可直接網路下載，網址：<http://trac.naer.edu.tw/106TASA>

表 1-1 2008 年至 2016 年施測考科與年段

	2008	2009	2010	2011	2012	2013	2014	2015	2016
國小 四年級	預試	正式 施測		預試	正式 施測		預試	正式 施測	
國小 六年級	預試	正式 施測		預試	正式 施測		預試	正式 施測	
國中 二年級		預試	正式 施測		預試	正式 施測		預試	正式 施測
高中職 二年級			預試	正式 施測		預試	正式 施測		預試

(八) 資料運用與未來方向

資料庫建置目的在提供研究用之資料，是教育研究與政策研擬的基礎工作，國際大型評量發展初期，「前導性十二國研究」與「六學科研究」的調查報告明確指出發展跨國研究的目的，並不是為了建立另一項以學習成就為評比項目的奧林匹亞競賽。發展跨國評量的動機是為了在教育資源、課程內容、教學方法等因素之外，補充學生學習成就的訊息，以研究各個教育系統內的教育「投入—成效」關係，提出教育成效優劣的可能解釋，並作為其他國家參考或仿效的對象。

國家教育研究院自 2006 年承接「臺灣學生學習成就評量資料庫」建置計畫至今已逾 10 年，常態性、系統性蒐集而來的資料，有助於追蹤與探討學生的學習成就、差異與變遷，進而檢視當前教育體制與政策之成效。資料庫必須經過使用才能彰顯其功能與重要性，建置之初即開放各界申請使用，部分資料甚至開放網路直接下載使用，然而考量大型評量資料庫的複雜性，使用者務必小心其中的限制，避免誤用或過度推論，為此，本書針對 TASA 之評量設計與資料分析做了詳盡的介紹。

在各學習領域的評量方面，為了有效達成測驗目的，檢核試題內容品質，評量架構的建立尤為重要，本書的第二章以評量架構為主題，分別介紹國語文、英語文、數學、社會與自然五科的評量架構與各科題本的信度。而測驗的設計—組卷、量尺化程序也需要參酌評量架構，後續測驗設計與量尺化程序則於第五章介紹。

大型評量除了針對學生進行表現評量，也透過背景問卷了解學生所處環境及其個人特質、心理健康等因素和學習表現間的關係，此外，同時透過學生作答反應與背景訊息得到的能力估計值也更能回應母群體的樣貌，第三章的主題則針對 TASA 的問卷設計有詳細的討論。

在大型評量中，若採用一般常見的簡單隨機抽樣，選取的學生樣本可能會散佈在非常多學校，不僅增加調查成本，且以學校中的少數樣本代表該校的整體表現也有誤差過大的考量，因此，各項大型評量往往先針對母群進行適當的分層後，再於各分層中抽出學校後，最後再抽出班級和學生。本書第四章先概覽 TASA 2016 的抽樣設計，接著再詳細說明抽樣設計和權重計算，最後再特別針對讀者可能的疑惑做出解答與建議。

大型評量建立後，資料庫的運用是學界重要的資料來源也是決策者的施政參考，然而，對許多不熟悉資料庫結構的使用者而言，資料庫因應調查需要衍生的複雜結構，不只增加使用上的難度，誤用資料或誤判數據的情形也相當常見。因此，第六章以 TASA 2016 的國中資料庫為例，介紹大型教育成就調查二階段分層叢集抽樣設計下的資料分析，包括權重的使用、統計量及其抽樣誤差的概念和計算等。此外，由於資訊科技的快速升級，由大量的數據資料萃取出有用的資訊，也是傳統統計分析之外的另一項研究方法。第七章運用資料探勘中決策樹的技術，試圖分析並建立學生的學習模型，以 TASA 2016 資料庫中國二學生數學科學生學習成就與共同問卷填答反應進行分析，探究影響國中學生數學學習的關鍵因素。

本書的第一章述說大型評量的需求於誕生，接著以 TASA 2016 資料庫

試題評量架構、問卷、抽樣、估計及其應用等層面作為主要內容的主幹，然而，在「十二年國民基本教育課程綱要」脈絡下，國家教育研究院主導的 TASA 應如何轉型，符合素養導向課程目標的試題究竟應該為何？本書的第八章再度回到評量建置的基本問題，依據課綱的願景說明 TASA 的未來發展。

大型評量在歐美已有超過半個世紀的歷史，為反映社會變遷與科技發展，調查的領域與工具也必須與時俱進。在評量領域方面，早期的研究均以學科領域為調查重點，而隨著公民民主意識逐漸抬頭，2009 年，「國際公民教育與素養調查計畫」再度成為國際調查關注的領域。而 IEA 也於 2013 年發起 5 年為一個評量週期「國際電腦與資訊素養研究」(International Computer and Information Literacy Study，簡稱 ICILS)，調查的對象以八年級學生為主，針對電腦使用、資料搜尋，網頁編寫等，以反映各界對資通訊科技發展的需求。

在評量工具方面，受限於科技的發展，早期的調查均使用紙筆測驗的方式進行，而隨著科技的發展與網路的普及，除了推動「國際電腦與資訊素養研究」，IEA 和 OECD 兩個主導國際大型評量的龍頭也開始思考數位化甚至線上評量的可能性，PISA 2015、PIRLS 2016，與 TIMSS 2019 也相繼採取數位化的施測模式。臺灣在全國性大型評量進行數位化評量的腳步略為落後，但 2017 年起，TASA 已採用線上施測方式進行預試，未來不只在命題、施測、閱卷與分析，將全面數位化，自動化命題也是國家教育研究院未來的研發重點。

參考文獻

- 陳揚琳 (1983)。大學聯招的大變革：三十年的沿革，三方案的抉擇。《聯合月刊》，21，82-87。
- 張荳雲 (2003)。台灣教育長期追蹤資料庫：第一波 (2001) 國中學生問卷。取自 <http://srda.sinica.edu.tw/group/sciitem/2/113>。
- 臺灣學生學習成就評量資料庫建置計畫 (2017)。臺灣學生學習成就評量資料庫簡介。取自 <http://www.naer.edu.tw/files/11-1000-1408-1.php?Lang=zh-tw>。
- 劉擇憲 (2011)。回首前路，展望未來－淺談臺灣百年來高中職升學制度之發展與演變。《教師天地》，172，68-72。
- American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Beaton, A. E. (1996). *Mathematics Achievement in the Middle School Years. IEA's Third International Mathematics and Science Study (TIMSS)*. Boston College, Center for the Study of Testing, Evaluation, and Educational Policy, Campion Hall 323, Chestnut Hill, MA 02167.
- Finder, M. (2004). *Educating America: How Ralph W. Tyler Taught America to Teach*. Greenwood Publishing Group.
- IEA (2017a). *Brief history of IEA: 50 years of educational research*. Retrieved from <http://www.iea.nl/brief-history-iea>.
- IEA (2017b). *Pilot Twelve-Country Study*. Retrieved from <http://www.iea.nl/pilot-twelve-country-study>.
- IEA (2017c). *Other IEA studies*. Retrieved from <http://www.iea.nl/other-iea-studies>.

- IEA (2017d). *TIMSS. Trends in International Mathematics and Science Study*. Retrieved from <http://www.iea.nl/timss>.
- IEA (2017e). *PIRLS. Progress in International Reading Literacy Study*. Retrieved from <http://www.iea.nl/pirls>.
- IEA (2017f). *ICCS. International Civic and Citizenship Education Study*. Retrieved from <http://www.iea.nl/iccs>.
- Jones, L. V. (2003). National assessment in the United States: The evolution of a Nation's report card. In *International handbook of educational evaluation* (pp. 883-904). Springer Netherlands.
- Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1-11). Springer Netherlands.
- OECD (2017a). *PISA. Programme for International Student Assessment*. Retrieved from <http://www.oecd.org/pisa/>.
- OECD (2017b). *TALIS. Teaching and Learning International Survey*. Retrieved from <http://www.oecd.org/talis/>.
- Ruddock, G. (2000). *Third international mathematics and science study repeat (TIMSS-R): First national report*. London: Department for Education and Employment.
- Urban, W. J. (2010). *More than science and sputnik: the National Defense Education Act of 1958*. University of Alabama Press.
- NAEP (2017a). *From The NAEP Primer: A Technical History of NAEP*. Retrieved from <https://nces.ed.gov/nationsreportcard/about/newnaephistory.aspx>.
- NAEP (2017b). *Measuring Student Progress Since 1964*. Retrieved from <https://nces.ed.gov/nationsreportcard/about/naephistory.aspx>.
- TEPS (2017) *Taiwan Education Panel Survey*. Retrieved from <http://www.teps.sinica.edu.tw/main.htm>.

第二章 TASA 學科成就評量架構

曾建銘

國家教育研究院副研究員

一、緒論

為了有效達成測驗目的，檢核試題內容品質，測驗評量架構的建置是相當重要的，評量架構好比是建築的藍圖，透過評量架構可以檢核測驗目的與內容是否相符，而測驗的設計一組卷、量尺化程序也要參酌評量架構，後續測驗設計與量尺化程序將於第五章會有詳細的介紹。本章主要介紹 TASA2013、2016 兩年的評量架構。第一節至第五節將分別介紹國語文、英語文、數學、社會與自然五科的評量架構與各科的信度。TASA 釋出的各科評量結果所包含的可能值 (plausible value) 除英語外，僅有各科分數，評量架構中各科知識面向和認知面向下各主題的分數並無釋出。

二、評量架構

(一) 國語文評量架構

1. 架構

秉持階段性、可行性、延續性理念，建置國語文成就資料庫，

建置初期以「讀、寫」評量為主，待人力、資源與經驗逐漸累積後，再逐漸發展「聽、說」評量。

依照測驗內容，本測驗可分為閱讀題型評量架構及寫作題型（短文寫作）評量架構。在閱讀評量架構中，特別將基本語文表達與閱讀能力分開，並置於閱讀架構前，此項設計主要考量基本語文表達乃閱讀能力之基礎，且易使用選擇題型方式進行檢測。架構中「知識面向」之分類，於基本語文表達評量架構主要參考國內重要大型國語文考試評量內容設計，配合 TASA 之目的為檢視九年一貫課程學習成效，因此，其評量指標內容主要是根據 97 年版九年一貫能力指標，與其之對應如表 2-1；閱讀評量架構中，則主要參考國際 PIRLS（Progress in International Reading Literacy Study，以下簡稱 PIRLS）、PISA（The Programme for International Student Assessment，以下簡稱 PISA）、美國 NAEP 及本國語文相關能力測驗。而本測驗架構亦與新修訂語文（本國語）課程綱要作一對應，以達到檢視目前國家教育體制與政策實施之成效。

「認知歷程面向」乃參酌新修訂 Bloom 認知歷程面向研訂，於不同年級學生所具備的能力而有所著重區別，並依據此測驗架構為八年級學生編製題目，以形成題庫。

2013、2016 八年級國語文評量範圍參考 97 年版語文（本國語文）課程綱要「語文表達」第三階段能力指標並配合布魯姆能力面向，將國語文領域測驗內容分為選擇題和非選擇題兩部分：

（1）選擇題部分

選擇題部分依題型規畫可分為語文應用與理解之單題題型及閱讀理解之題組題。單題依布魯姆認知面向可分為記憶、應用、理解三層次，各層次之評量指標內涵及題數分布如表 2-1、2-3。

閱讀題組部份，則可分為理解及分析二層，理解層次之評量重點包含：詮釋、摘要、推論；分析層次則包含組織、歸因、及區辨。評量重點之評量指標內涵及題數分布如表 2-2、2-4。

表 2-1 2013 國語文單題評量指標 - 題數分配

認知面向	評量指標	九年一貫能力指標	合計題數	題數合計
記憶	1. 回憶字形（含字音）	能認識常用漢字 3500-4500 字。	15	15
理解	2-1 能詮釋詞義	能熟習並靈活應用語體文及文言文作品中詞語的意義。	4	13
	3-1 能詮釋句子涵義	能配合語言情境理解字詞和文意間的轉化。	5	
	3-2 詮釋句子觀點	能配合語言情境理解字詞和文意間的轉化。	4	
應用	2-2 能運用詞語	能精確的遣詞用字，並靈活運用各種句型寫作。	9	37
	3-3 能運用完整句子及段落	能精確的遣詞用字，並靈活運用各種句型寫作。掌握寫作步驟，充實作品的內容，精確的表達自己的思想。	10	
	4. 能運用標點符號	了解標點符號的功能，並適當使用。	9	
	5. 能形成結論	掌握寫作步驟，充實作品的內容，精確的表達自己的思想。	9	
	合計		65	65

表 2-2 2013 國語文題組評量指標 - 題數分配

認知面向	評量重點	評量指標	九年一貫能力指標	題數	合計題數	合計題數
理解	1. 詮釋	2-1 能詮釋詞義	同上表	4	10	40
		3-1 能詮釋句子涵義	同上表	6		
	2. 摘要	7-1 能掌握文章主旨	無(雖無能力指標對照, 為求連貫性, 仍須納入評量架構中)	7	7	
	3. 推論	6 能理解內容細節及要素	能主動思考與探索, 統整閱讀的內容, 並轉化為日常生活解決問題的能力。	15	23	
		3-2 詮釋句子觀點	同上表	8		
分析	4. 組織	8-2 能歸納文章架構及重點	能靈活運用不同的閱讀理解策略, 發展自己的讀書方法。	6	6	25
	5. 歸因	7-2 能分辨文章表述方式	能欣賞作品的寫作風格、特色及修辭技巧。	3	6	
		9-1 能推論文章寫作觀點	能主動思考與探索, 統整閱讀的內容, 並轉化為日常生活解決問題的能力。	2		
		9-2 能推論文章寫作寓意	能主動思考與探索, 統整閱讀的內容, 並轉化為日常生活解決問題的能力。	1		
	6. 區辨	8-1 能歸納文章段落重點	能主動思考與探索, 統整閱讀的內容, 並轉化為日常生活解決問題的能力。	13	13	
合 計				65	65	65

表 2-3 2016 國語文單題評量指標 - 題數分配

層 次	評 量 指 標	合 計 題 數	題 數 合 計
記 憶	1 回憶字形(含字音)	10	10
理 解	2-1 能詮釋詞義	1	5
	3-1 能詮釋句子涵義	2	
	3-2 詮釋句子觀點	2	5
應 用	2-2 能運用詞語	10	40
	3-3 能運用完整句子及段落	12	
	4 能運用標點符號	8	
	5 能形成結論	10	
	合 計	55	55

表 2-4 2016 國語文題組評量指標 - 題數分配

層 次	評量重點	評 量 指 標	題 數	合 計 題 數	合 計 題 數
理 解	1. 詮 釋	2-1 能詮釋詞義	8	21	45
		3-1 能詮釋句子涵義	13		
	2. 摘 要	7-1 能掌握文章主旨	7	7	
	3. 推 論	6 能理解內容細節及要素	16	17	
		3-2 詮釋句子觀點	1		

層次	評量重點	評量指標	題數	合計題數	合計題數
分析	4. 組織	8-2 能歸納文章架構及重點	4	4	30
	5. 歸因	7-2 能分辨文章表述方式	7	14	
		9-1 能推論文章寫作觀點	5		
		9-2 能推論文章寫作寓意	2		
	6. 區辨	8-1 能歸納文章段落重點	12	12	
合計			75	75	75

(2) 非選擇題

即寫作測驗，設計有【評論短文】1題及【創作短文】1篇，以45分鐘獨立施測，2013、2016評量指標內涵及題數分布如下表2-5。

表 2-5 寫作評量架構

層次	寫作評量指標	九年一貫能力指標	題數
創造	10. 能撰寫短文	6-4-4 掌握寫作步驟，充實作品的內容，精確的表達自己的思想。	1
評鑑	11. 能評論短文	6-4-3 練習應用各種表述方式寫作。	1

2. 計分與信度

國語文測驗之選擇題，計分方式採電腦閱卷，依正確答案由讀卡機判斷，正確給1分，錯誤給0分；非選擇題部分，評論短文採

3 級計分 (0 ~ 3)，創作短文採 6 級計分 (0 ~ 6)。

在題本信度部分是利用 cronbach's α 計算得之，2013 年 13 個題本的信度介於 0.81 ~ 0.86 之間，整體信度為 0.8；2016 年 13 個題本的信度介於 0.72~0.85 之間，整體信度為 0.77。各題本信度詳如下表 2-6。

表 2-6 2013、2016 國語文各題本信度值

2013		2016	
題本 1	0.83	題本 1	0.79
題本 2	0.85	題本 2	0.72
題本 3	0.85	題本 3	0.76
題本 4	0.82	題本 4	0.72
題本 5	0.82	題本 5	0.76
題本 6	0.83	題本 6	0.82
題本 7	0.86	題本 7	0.85
題本 8	0.81	題本 8	0.82
題本 9	0.84	題本 9	0.83
題本 10	0.81	題本 10	0.82
題本 11	0.83	題本 11	0.76
題本 12	0.84	題本 12	0.79
題本 13	0.84	題本 13	0.82
整體	0.80	整體	0.77

（二）英語文評量架構

1. 架 構

臺灣學生學習成就評量資料庫英語文科命題，以民國九十三年五月十三日所頒布「國民中小學九年一貫課程綱要」英語文學習領域能力指標為依據。

評量架構中，「知識面向」以九年一貫國民中小學能力指標為主要依據，並參考 NAEP 2004 年外國語言之評量架構（Framework for the 2004 Foreign Language National Assessment of Educational Progress，簡稱 FLNAEP）之共同能力指標，英語科基本學力測驗、指定科目考科（及學測）作為試題設計的依據。

九年一貫課程英語文學習領域旨在奠定國人英語溝通能力的基礎、提昇英語學習的動機與興趣。故「能力指標」同時注重聽、說、讀、寫四項語言能力的培養。從國小三、四年級的啟蒙階段強調聽、說的學習，讓兒童藉由豐富的英語聽、說的學習，奠定良好的英語口語溝通基礎。讀、寫活動適時融入課程，讓學生藉由接觸簡易的閱讀材料，以及適當的臨摹及填寫字詞等練習，自然體驗語言的不同形式，以收聽、說、讀、寫四者相輔相成的效果。臺灣學生學習成就評量資料庫英語文科的知識面向，即是要評量學生基本溝通的學習成就。

「認知歷程面向」乃參酌新修訂 Bloom 認知歷程面向。

記憶：語言基礎知識如字彙、片語、文法句型的知識，知道它的語意；知道它在上下文如何運用，例如：杯子跟 cup 的連結；其特性為通常是以單題試的方式出現

理解：考段落、篇章結構、克漏字、閱讀測驗，其特性為大部分已篇章出現，有上下文。其中閱讀測驗會出現考文章的 Main

Idea（主旨）、Details（故事情節）以及推論性的問題。

基礎應用：考 language skills 中的寫、說，from listening/reading to writing/speaking。學生應用所習得的語言知識，產出於回答問題、對話溝通，以及寫作中。

語言基礎知識如字彙、片語、文法句型，在認知歷程面向中屬於記憶，評量考生對各類詞彙、片語、句法掌握及運用的能力。閱讀測驗部份在認知歷程面向裡屬於了解，評量考生理解篇章結構及文意，加以分析與綜合的能力。測驗內容中寫說部份，認知歷程面向為基礎應用，評量考生理解對話或篇章文意，根據提示，加以運用綜合的能力，回答問題或作文。

惟礙於英語文的科目本質與我國英語作為第二語言的特殊條件，部分解題過程所牽涉的大腦思考無法截然區分為特定單一向度，例如單字的中翻英可能同時需要記憶與瞭解。而圖像、聲音與文字的轉換配對更可能需要大腦的多工運作。因此，在評量學生認知能力的圭臬下，TASA 英文科部份作了微觀的調整，俾能權衡各科題庫之連結與英文科之特性。於此，測驗內容的兩大範疇：語言基礎知識與閱讀，將分別朝向記憶、理解、分析、評鑑等認知範疇，但不會拘泥於單一向度的套用。

（1）取材原則

國中試題的取材原則主要是選擇通過審查的第一冊到第四冊英語教材，包含平面教材及視聽教材。選擇英語教材的原則為以落實英語課程目標，分段能力指標及符合課程大綱所列的主題、體裁及溝通功能為主。

（2）測驗內容

測驗全部都是選擇題，其中包括單題選擇題以及短篇閱讀

測驗題題組，每題正確給 1 分，錯誤給 0 分。聽力測驗題型有看圖辨義：學生聽錄音機播出題目和選項，選出與所看到的圖畫最相符的答案；問答：學生聽錄音機播出英語句子，再從試題本上三個選項中，選出一個最適合的答案；簡短對話：學生聽錄音機播出一段對話和一個相關的問題後，再從試題本選項中，選出一個最適合的答案。閱讀測驗題型包括單題選擇題：單字、文法；短篇測驗題組：克漏字、短篇閱讀測驗。口說題型：複誦經由耳機聽到的短句；朗讀句子：朗讀書面上呈現的句子；朗讀短文：朗讀書面上呈現的一段短文；看圖回答問題：畫面上呈現一個圖片及四個句子，依據圖片回答問題；回答問題：經由耳機播放問題，依據事實回答問題。寫作測驗題型包括單題及題組，單題題型如：字詞填空、句子改寫、合併句子、短句翻譯。題組題型如：克漏字以及一段段落寫作。下表 2-7、2-8 為八年級英語文測驗之評量架構：

表 2-7 2013 八年級英語文評量架構

語 言 能 力	能 力 指 標	認 知 面 向		
		記 憶	了 解	基 礎 應 用
聽	1-1-8 能聽懂簡易句型的句子。		7	
	1-2-3 能聽懂日常生活對話和簡易故事。		5	
	1-2-4 能辨識對話或訊息的情境及主旨。		1	
	5-1-5 能聽懂日常生活應對中常用語句，並能作適當的回應。		9	
讀	3-2-3 能看懂常用的英文標示和圖表。		8	
	3-2-6 能瞭解對話、短文、書信、故事及短劇等的重要內容與情節。		9	

語 言 能 力	能 力 指 標	認 知 面 向		
		記 憶	了 解	基 礎 應 用
讀	3-2-7 能從圖畫、圖示或上下文，猜測字義或推論文意。	21	15	
說	2-1-4 能以正確的語調說出簡易句型的句子。			10
	2-1-10 能朗讀和吟唱歌謠韻文。			5
	2-2-4 能以簡易的英語描述日常生活中相關的人、事、時、地、物。			5
	2-2-5 能依人、事、時、地、物作提問和回答。			5
	3-2-4 能用適切的語調、節奏朗讀短文、簡易故事等。			1
寫	4-1-6 能依圖畫、圖示填寫重要字詞。			10
	4-2-2 能依提示合併、改寫句子及造句。			10
	4-2-4 能將簡易的中文句子譯成英文。			5
	4-2-5 能依提示書寫簡短的段落。			1
合 計		21	54	52

表 2-8 2016 八年級英語文評量架構

語 言 能 力	能 力 指 標	認 知 面 向		
		記 憶	了 解	基 礎 應 用
聽	1-1-8 能聽懂簡易句型的句子。		10	
	1-2-3 能聽懂日常生活對話和簡易故事。		5	
	1-2-4 能辨識對話或訊息的情境及主旨。		3	
	5-1-5 能聽懂日常生活應對中常用語句，並能作適當的回應。		10	

語 言 能 力	能 力 指 標	認 知 面 向		
		記 憶	了 解	基 礎 應 用
讀	3-2-3 能看懂常用的英文標示和圖表。		4	
	3-2-6 能瞭解對話、短文、書信、故事及短劇等的重要內容與情節。		16	
	3-2-7 能從圖畫、圖示或上文，猜測字義或推論文意。	24	28	
說	2-1-4 能以正確的語調說出簡易句型的句子。			10
	2-1-10 能朗讀和吟唱歌謠韻文。			5
	2-2-4 能以簡易的英語描述日常生活中相關的人、事、時、地、物。			5
	2-2-5 能依人、事、時、地、物作提問和回答。			5
	3-2-4 能用適切的語調、節奏朗讀短文、簡易故事等。			1
寫	4-1-6 能依圖畫、圖示填寫重要字詞。			10
	4-2-2 能依提示合併、改寫句子及造句。			10
	4-2-4 能將簡易的中文句子譯成英文。			5
	4-2-5 能依提示書寫簡短的段落。			1
合	計	24	76	52

2. 計分與信度

2013 年在選擇題部分共設計 75 題，聽力測驗包含 22 題，閱讀測驗包含 53 題，編製成六份測驗題本，每份題本分聽力測驗和閱讀測驗兩部分共 32 題。在非選擇題部分共設計 52 題，編製成兩份測驗題本，分為口語能力測驗和書寫能力測驗兩本。口說能力測驗題

本包含 5 個大題，共有 26 個小題；書寫能力測驗題本包含 6 個大題，共有 26 個小題。

2016 年在選擇題部分共設計 100 題，聽力測驗包含 28 題，閱讀測驗包含 72 題，編製成六份測驗題本，每份題本分聽力測驗和閱讀測驗兩部分共 32 題。在非選擇題部分共設計 52 題，編製成兩份測驗題本，分為口語能力測驗和書寫能力測驗兩本。口說能力測驗題本包含 5 個大題，共有 26 個小題；書寫能力測驗題本包含 6 個大題，共有 26 個小題。

英語文測驗的測驗時間規劃，聽讀題本、寫說題本皆 45 分鐘。聽讀題本依正確答案批改，正確給 1 分，錯誤給 0 分。口說測驗各題評分標準：第一、二大題，複誦以及朗讀句子，評分向度為單音錯誤、漏音、字重音錯誤、漏字、句重音錯誤、語調、整體流利度、理解度等，評分分層定三級分制，特重理解度，其次重流利度。作答反應無法令人理解的為 0 級分 (fail)，有 3-5 項向度有誤的為 1 級分 (pass)，有 1-2 項向度有誤或完全無誤的為 3 級分屬於 high pass。第三大題朗讀短文評分向度為單音錯誤、漏音、字重音錯誤、漏字、句重音錯誤、語調、句子間的停頓、整體流利度、理解度等；分五級分，特重理解度，其次重流利度和句子間的停頓。未作答者 0 級分，1、2 級分屬於 fail，3、4 級分以上 pass，如能達到 5 級分者屬於 high pass。第四、五大題，看圖回答問題以及回答問題，評分向度為發音錯誤、語調、詞彙的正確性、文法的正確性、整體流利度、理解度、內容的正確性等，評分分層定三級分制，特重理解度及內容的正確性。

書寫測驗各題評分標準：單字層次：第一字詞填空、二大題克漏字，如未作答、作答不是一個完整的單字、作答為正答不相關的單字不給分。如有書寫錯誤、拼字錯誤、句法上的錯誤扣一

分。句子層次：第三大題句子改寫、第四大題合併句子、第五大題短句翻譯，如有未答、字詞順序錯誤嚴重影響讀者之理解、因果關係錯誤、違反邏輯，與原本文意不相符不給分。如有句法上的錯誤扣二分。如有拼字、標點、大小寫之錯誤、遺漏一字、抄錯一字等扣一分。第一至五大題百分比計算佔全部試題之 80%；第六大題段落寫作佔 20%。段落寫作就內容組織、文法、用字、結構、體例分別給分。

在題本信度部分是利用 cronbach's α 計算得之，2013 年 6 個題本的信度介於 0.93~0.94 之間，整體信度為 0.94；2016 年 6 個題本的信度介於 0.93~0.94 之間，整體信度為 0.93。各題本信度詳如下表 2-9。

表 2-9 2013、2016 英語文各題本信度值

2013		2016	
題本 1	0.94	題本 1	0.93
題本 2	0.94	題本 2	0.93
題本 3	0.94	題本 3	0.93
題本 4	0.94	題本 4	0.93
題本 5	0.94	題本 5	0.93
題本 6	0.93	題本 6	0.94
整體	0.94	整體	0.93

(三) 數學評量架構

1. 架 構

數學能力指標係依主題及階段學習能力而訂定，然因多數指標需採分年進階式教學方能達成其教學目標。因此，由階段能力指標演繹出更細緻的分年細目及詮釋，以利分年進階式教學進度目標的明確掌握。故在設計評量架構能力內涵時，參照九年一貫能力指標分年細目之內容設計為評量依據，以期能對國家課程發展提出良方及建言。

數學認知層次參酌 Anderson et al. (2001) 主編的《學習教學與評量的分類：Bloom 教育目標分類的修定》一書，可與 NAEP 之數學能力大致互相呼應：概念理解（記憶、了解）、程序知識（應用）、問題解決（分析、評鑑、創作）。

TASA 參酌 NAEP 之數學評量架構將評量架構的分成測驗之縱向「知識面向」與橫向「認知歷程面向」之設計，並結合九年一貫能力指標分年細目進行評量架構的細部規劃，並參酌各版本之教材，以期評量能使各類題型適切分佈，達到檢測的目的。

數學測驗主要以四選一單一選擇題型評量檢測八年級數學的基本能力，其雙向細目表之組成是由數學內容面向：數與量、幾何、統計與機率、代數及認知歷程面向：概念理解、程序執行、解題思考，雙向組成。

另外，為求九年一貫課程綱要的銜接性，以 92 年版本數學國中八年級之九年一貫能力指標分年細目為主，參酌 97 年版本數學國中八年級之九年一貫能力指標分年細目，來研訂本數學知識面向的評量內涵。

概念理解、程序執行和解題思考的闡釋

(1) 概念理解

係指數學內容的概念性知識，評量概念理解的試題，是測驗學生是否具備數學基本概念，要求作答者將腦中的記憶知識應用出來做判斷。能理解、指認題目中相關的符號、圖表、公式及原理，並能比較、對照並統整相關概念與原理來延伸概念與原理的性質。

(2) 程序執行

係指數學內容的程序性知識，評量程序執行的試題，是測驗學生是否「知道要如何完成數學運算」的數學知識，包括操作數與符號的運算、幾何構圖的執行及估計、正確選擇適當的程序、能判別或判斷算式或符號運用方法過程的正確性或適切性，並運用不同的數學邏輯有效解決數學問題。

(3) 解題思考

係指對數學問題的解題能力，評量解題與思考的試題，是測驗學生是否能瞭解題目及組織相關的數學知識，來進行解題思考，並瞭解、判斷題目，進而採取適當解題策略、數據及算式且使用數學語言表達解題過程，並能進行歸納、統計、演繹等推理與分析的能力。一般而言，這部分的數學能力表現通常需要數學的概念理解和程序執行的基礎來充實。

2013、2016 年 TASA- 數學科八年級各有選擇題 78 題、非選擇題 13 題，測驗內容架構如下表 2-10~13 所示。

表 2-10 2013 八年級數學科選擇題評量架構

知識面向 \ 認知面向	概念理解	程序執行	解題思考	題數
數與量	8	5	7	20
幾何	8	5	4	17
統計與機率	0	2	5	7
代數	9	12	13	34
合計	25	24	29	78

表 2-11 2013 八年級數學科非選擇題評量架構

知識面向 \ 認知面向	概念理解	程序執行	解題思考	題數
數與量	0	1	2	3
幾何	1	0	2	3
統計與機率	0	0	1	1
代數	0	2	4	6
合計	1	3	9	13

表 2-12 2016 八年級數學科選擇題評量架構

知識面向 \ 認知面向	概念理解	程序執行	解題思考	題數
數與量	8	5	7	20
幾何	8	4	5	17
統計與機率	1	4	5	10
代數	8	11	12	31
總題數	25	24	29	78

表 2-13 2016 八年級數學科非選擇題評量架構

知識面向 \ 認知面向	概念理解	程序執行	解題思考	題數
數與量	0	1	2	3
幾何	1	1	1	3
統計與機率	0	2	0	2
代數	0	1	4	5
總題數	1	5	7	13

2. 計分與信度

數學科測驗之選擇題，計分方式採電腦閱卷，依正確答案由讀卡機判斷，正確給 1 分，錯誤給 0 分；非選擇題部分，一律採 3 級計分（0~3）。

在題本信度部分是利用 cronbach's α 計算得之，2013 年 13 個題本的信度介於 0.83~0.88 之間，整體信度為 0.85；2016 年 13 個題本的信度介於 0.78~0.85 之間，整體信度為 0.78。各題本信度詳如下表 2-14。

表 2-14 2013、2016 數學科各題本信度值

2013		2016	
題本 1	0.85	題本 1	0.83
題本 2	0.87	題本 2	0.85
題本 3	0.87	題本 3	0.82

2013		2016	
題本 4	0.87	題本 4	0.82
題本 5	0.88	題本 5	0.80
題本 6	0.87	題本 6	0.82
題本 7	0.87	題本 7	0.83
題本 8	0.86	題本 8	0.80
題本 9	0.83	題本 9	0.79
題本 10	0.86	題本 10	0.80
題本 11	0.86	題本 11	0.78
題本 12	0.84	題本 12	0.81
題本 13	0.86	題本 13	0.79
整體	0.85	整體	0.78

(四) 社會評量架構

1. 架構

TASA 參考 NAEP 社會科評量架構的設置模式以及考量國內教育規範，設計規劃八年級評量架構。國中八年級的評量架構之知識面向分類按照教育部所頒訂七～九年級基本內容為縱軸；橫軸部分則是依照布魯姆的認知分類，將題目分為記憶、了解與應用、高層次思考（其下包含分析、評鑑、創作）。設計內容說明如下：

(1) 知識面向

按照教育部所訂九年一貫綱領，以學生於各階段所應具備之相關知識以及技能能力的的能力指標作為課程發展的依據。第四學習階段（七、八、九年級）除能力指標外，更多了學科基本內容作為更明確的課程範圍。TASA 八年級評量架構以主題軸含括之基本內容為縱軸，並將學科內容說明以及可參酌之能力指標附於其後，以供命題教師參考。評量架構縱軸以基本內容為主，以對應的能力指標為輔，由於部分能力指標之說明並不完全符合學科內容，故將其對應之部分以底線畫記，使命題教師能有更清楚之依據。

(2) 認知面向

NAEP 在認知程度上，按照年級的不同，題目的設計認知層次也不同，不同認知層次的題型比例也會隨之調整。低年級的學生偏重於事實性的認知與描述再生，而高年級的學生就轉趨於理解分析，並且開始學習評估各種知識內容的優劣以及做出回應，甚至進一步給予挑戰。而 TASA 評量架構的認知層次也是依照類似的概念進行，將社會科的知識內容依照 Bloom 的認知程度分類，分為記憶、理解（了解與應用）、高層次思考（其下包含分析、評鑑、創作）。記憶－以學生過去曾經學習過的知識再認為主；理解（了解與應用）－學生利用所學的知識進行詮釋、分類、摘要、推理、比較、解釋、應用則是執行和實踐學習過的知識；高層次思考－分析、評鑑、創作。

國中八年級社會科測驗內容，係依照教育部發布之九年一貫課程綱要（92 課綱），其中「社會學習領域」第三到第四學習階段九大主題軸的能力指標，並參酌課程綱要中增列之基本內容為輔，構

成本年度正式施測之評量架構。第一主題軸「人與空間」、第二主題軸「人與時間」與其餘 7 個主題軸綜合的比重分配以 1：1：1 為原則（題數分別為 47：54：55）。此外，本測驗試卷按照認知歷程面向將題目分為記憶、理解及高層次思考三類。2013、2016 社會科國中八年級評量架構如下表 2-15、表 2-16 所示。

表 2-15 2013 八年級社會科評量架構

知識面向（主題軸）	認 知 面 向			總 題 數
	記 憶	理 解	高 層 次 思 考	
一、人與空間	6	34	7	47
二、人與時間	7	36	11	54
三、演化與不變	0	6	1	7
四、意義與價值	0	1	0	1
五、自我、人際與群己	0	4	0	4
六、權力、規則與人權	3	24	3	30
七、生產、分配與消費	0	5	2	7
八、科學、技術和社會	0	2	0	2
九、全球關聯	0	4	0	4
題 數	16	116	24	156

（1）選擇題

第一主題軸「人與空間」、第二主題軸「人與時間」與其餘主題軸綜合的比重分配以 1：1：1 為原則，選擇題題數分別為

38：40：39，而認知歷程面向將題目分為記憶、理解及高層次思考三類，試題分布如表 2-16 所示。

表 2-16 2016 年八年級社會科選擇題評量架構

知識面向（主題軸）	認 知 面 向			總 題 數
	記 憶	理 解	高層次思考	
一、人與空間	0	30	8	38
二、人與時間	2	33	5	40
三、演化與不變	0	9	0	9
四、意義與價值	0	0	0	0
五、自我、人際與群己	0	8	0	8
六、權力、規則與人權	1	16	2	19
七、生產、分配與消費	0	2	0	2
八、科學、技術和社會	0	1	0	1
九、全球關聯	0	0	0	0
題 數	3	99	15	117

（2）非選擇題

此次測驗有 13 題非選擇題題組，其中 4 個題組為歷史試題（總共 10 小題），5 個題組為地理試題（總共 11 小題），4 個題組為公民（總共 4 小題）。試題分布如表 2-17 所示。

表 2-17 2016 年八年級社會科非選擇題評量架構

知識面向（主題軸）	認 知 面 向			總 題 數
	記 憶	理 解	高層次思考	
一、人與空間	0	6	5	11
二、人與時間	0	3	7	10
三、演化與不變	0	0	1	1
四、意義與價值	0	0	1	1
六、權力、規則與人權	0	0	2	2

2. 計分與信度

社會科測驗之選擇題，計分方式採電腦閱卷，依正確答案由讀卡機判斷，正確給 1 分，錯誤給 0 分；非選擇題部分，每一小題一律採 2 級計分（0~2）。

在題本信度部分是利用 cronbach's α 計算得之，2013 年 13 個題本的信度介於 0.84~0.88 之間，整體信度為 0.86；2016 年 13 個題本的信度介於 0.72~0.85 之間，整體信度為 0.81。各題本信度詳如下表 2-18。

表 2-18 2013、2016 社會科各題本信度值

2013		2016	
題 本 1	0.88	題 本 1	0.77
題 本 2	0.88	題 本 2	0.72
題 本 3	0.87	題 本 3	0.74

2013		2016	
題本 4	0.88	題本 4	0.82
題本 5	0.88	題本 5	0.83
題本 6	0.85	題本 6	0.82
題本 7	0.84	題本 7	0.84
題本 8	0.87	題本 8	0.76
題本 9	0.87	題本 9	0.83
題本 10	0.87	題本 10	0.80
題本 11	0.88	題本 11	0.77
題本 12	0.87	題本 12	0.85
題本 13	0.86	題本 13	0.83
整體	0.86	整體	0.81

(五) 自然評量架構

1. 架 構

八年級自然科的測驗評量架構參考 Anderson et al. (2001) 主編的《學習教學與評量的分類：Bloom 教育目標分類的修定》一書，以縱向「知識面向」，橫向「認知歷程面向」為評量架構依據。評量架構中，「知識面向」主要參考各領域教材細目、美國 NAEP 及國內自然科相關能力測驗的架構建置而成；「認知歷程面向」乃參酌新修訂 Bloom 認知歷程面向與鄭湧涇教授綜合分類研訂。科學學習表現之認知精熟度階層 (Cognitive Proficiency Levels)，認知學習

的表現分為四個階層：

(1) 知曉科學知識 (Knows, K)

- 記憶在學校課程或日常生活經驗中所習得之科學事實或知識。
- 區別或界定基本科學名詞、術語或科學實驗器材。
- 閱讀圖表。

(2) 了解基本科學原理法則 (Understands, U)

- 了解基本科學概念、原理、法則 (Principles)。
- 了解科學學說和定律的內容。
- 了解科學知識間的關係。

(3) 應用基本科學資訊 (Applies, A)

- 分析及解釋資料。
- 應用科學知識進行推理、推論、預測。
- 分析資料並應用資料進行推理、推論、預測。

(4) 統整科學資訊 (Integrates, I)

- 綜合各項資訊，指出各變項之間的關係。
- 統整實驗過程及數據，指出擬驗證之假說及提出結論。
- 統整科學概念，提出結論。
- 綜合各階層科學知識，以解決問題。

後來經過專家會議將上述之 2、3 層整併為一層，最後成為知道、理解應用、統整推理共三層。

自然科試題規劃係參酌教育部所公告「國民中小學九年一貫課程綱要」自然與生活科技學習領域第四階段（國中二年級）之分段能力指標與附錄一、二為主要依據，並參照學校教學實務發展而

成。針對國中二年級學生的自然與生活科技領域學習目標，訂定測驗領域內容包含認識時空環境、認識生命世界、探索物質科學、謀求永續經營，命題範圍涵蓋國中二年級的自然科評量架構，內容詳見下表 2-19、2-20。

表 2-19 2013 八年級自然科測驗評量架構

知識面向（領域）	認識面向			題數
	知道	理解應用	統整推理	
物理	2	22	3	27
化學	6	16	5	27
生物	18	20	22	60
地球科學	4	4	1	9
九年一貫（跨年段）	10	14	9	33
合計	40	76	40	156

表 2-20 2016 八年級自然科測驗評量架構

知識面向（領域）	認識面向			題數
	知道	理解應用	統整推理	
物理	10	27	2	39
化學	14	17	8	39
生物	19	25	22	66
地球科學	5	4	3	12
合計	48	73	35	156

2. 計分與信度

自然科測驗之選擇題，計分方式採電腦閱卷，依正確答案由讀卡機判斷，正確給 1 分，錯誤給 0 分。

在題本信度部分是利用 cronbach's α 計算得之，2013 年 13 個題本的信度介於 0.85 ~ 0.90 之間，整體信度為 0.86；2016 年 13 個題本的信度介於 0.75 ~ 0.86 之間，整體信度為 0.79。各題本信度詳如下表 2-21。

表 2-21 2013、2016 自然科各題本信度值

2013		2016	
題本 1	0.85	題本 1	0.78
題本 2	0.86	題本 2	0.76
題本 3	0.86	題本 3	0.75
題本 4	0.86	題本 4	0.78
題本 5	0.87	題本 5	0.78
題本 6	0.85	題本 6	0.78
題本 7	0.86	題本 7	0.82
題本 8	0.87	題本 8	0.81
題本 9	0.87	題本 9	0.83
題本 10	0.89	題本 10	0.86
題本 11	0.89	題本 11	0.84
題本 12	0.90	題本 12	0.83
題本 13	0.87	題本 13	0.82
整體	0.86	整體	0.79

第三章 背景問卷的心理計量特性

謝 佩 蓉

國家教育研究院助理研究員

一、緒 論

大型評量除了針對學生進行表現評量，也請學生填答問卷，目的在於深入了解學生所處環境及其個人特質、心理健康等因素和學習表現間的關係，也為了透過學生作答反應和背景訊息兩者一起估計似真值（plausible values），使學生學習表現的估計值更能回應母群體的樣貌 (Wu, Tam, & Jen, 2016)。TASA 同為大型評量設計，參與評量的學生，除了作答學科試題，也要填寫一份學生問卷。

本章主要針對 TASA 2010、2013 及 2016 八年級學生問卷進行題項的心理計量特性分析，略涉獵學校層級資料的說明。學生問卷以紙本形式呈現，學生於答案卡畫記。題型為選擇題或複選題，作答時間為 45 分鐘。在一節課的時間內，每一位學生能作答的題數有限，故問卷議題需慎選，也必須隨著時代推進調整議題方向，讓問卷題項更能反映學生所處時空背景與生活型態。

問卷內容概分為一般性和學科特定性兩部分。一般性題項為該年所有接受調查的學生均須作答之題項。學科特定性題項於 2010 年和 2013 年兩屆施測時，僅由作答該學科測驗的學生填答，例如，作答國文和數學測驗的學生僅填答國文和數學學科問卷，不填答英語、社會及自然學科問卷。於 2016 年施測時，所有受測學生不論接受哪些學科的測驗，均需填答所有的學科問卷題目。

二、架 構

(一) 學生層級的資料

學生問卷主軸分為個人、家庭及學校三部分，Hattie (2009) 統合後設分析的結果發現，這三項主軸對於學習表現的效果量 (Cohen *d*) 分別為 0.40、0.31 及 0.23，具有不可忽視的影響力。TASA 問卷架構及構念題數如表 3-1 至表 3-5，分為人口學、家庭資本與教育活動、休閒活動、心理學構念及針對學科所設定的題項。另外，考量資料分析者的編碼取向，複選題的題數以「該題的選項數」列計。

表 3-1 人口學變項題數分布

變 項	2010 題 數	2013 題 數	2016 題 數
性別 (S2010Q1、S2013Q1、S2016Q1)	1	1	1
出生序 (S2010Q2_2、S2013Q2_2)	1	1	0
出生年 (S2016Q2)	0	0	1
出生月 (S2016Q3)	0	0	1
出生地 (S2010Q6)	1	0	0
家中子女數 (S2010Q2_1、S2013Q2_1)	1	1	0
家庭組成 (S2010Q3、S2013Q3、S2016Q6_1~5)	1	1	5
主要照顧者 (S2010Q4_1~9、S2013Q4_1~9)	9	9	0
父母原生國籍 (S2010Q5_1~2、S2013Q5_1~2、S2016Q4、S2016Q5)	2	2	2

註：括弧內為試題編碼，請參照編碼簿。

由表 3-1 可知，TASA 提供學生基本人口學訊息，性別、家庭組成及父母原生國籍為歷年皆有的變項，可分析不同性別、單親、雙親家庭及跨國婚姻家庭學生之心理健康與學習表現，2016 年的題項更可提供繼親家庭訊息。此外，依照國民教育法施行細則（2016）第八條規定「學齡兒童入學年齡之計算，以入學當年度九月一日滿六歲者」，TASA 2016 年提供學生出生年與月份，除可計算學生年齡，亦可分析於同一學年就學，九月出生與隔年八月出生（summer-born）學生之心理健康與學習表現。

表 3-2 家庭資本與教育活動題數分布

變 項	2010 題 數	2013 題 數	2016 題 數
父母教育程度 (S2010Q7_1~2、S2013Q6_1~2)	2	2	0
家中物品 (S2010Q8_1~11、S2013Q7_1~11、S2016Q7_1~10)	11	11	10
家中書量 (S2010Q10、S2013Q9、S2016Q8)	1	1	1
文化活動 (S2010Q11_1~3、S2013Q10_1~3)	3	3	0
自我教育期望 (S2010Q7_3、S2013Q6_3)	1	1	0
父母教育期望 (S2010Q7_4~5、S2013Q6_4~5)	2	2	0
國中畢業後規劃 (S2016Q10)	0	0	1
校外課程 (與補習) (S2010Q9_1~3、S2013Q8_1~3、S2016Q9)	3	3	1
補習時間 (S2010Q14_1~5、S2013Q13_1~5、S2016Q26_1~4)	5	5	4
小考次數 (S2016Q25_1~5)	0	0	5
語言使用 (S2010Q12_1~3、S2013Q11_1~3)	3	3	0

表 3-2 家庭資本與教育活動題數分布 (續)

變 項	2010 題 數	2013 題 數	2016 題 數
聯絡簿簽名 (S2010Q12_4~5)	2	0	0
做作業時間 (S2010Q13_1~5、S2013Q12_1~5)	5	5	0
學習寫程式經驗 (S2016Q11、S2016Q12)	0	0	2

註：括弧內為試題編碼，請參照編碼簿，可直接至 <http://trac.naer.edu.tw/106TASA> 下載。

家庭社經地位 (socioeconomic status, SES) 常透過父母之教育程度、職業、收入三項指標之組合分數來衡量 (American Psychological Association, 2017)。由於八年級學生難以精確表達父母親之職業與收入，TIMSS 八年級調查以家庭學習資源量表 (Home Resources for Learning (HRL) scale) 來蒐集，內容包括家中藏書量、家中有沒有網路與自己一人一個房間、父母親其中一人之教育程度共三項指標來測量 (Martin, Mullis, Foy, & Arora, 2013)。表 3-2 可知，TASA 歷年均有、可用於表徵家庭資本之題項為家中物品與家中書量，2016 年資料檔中，已將這兩項目之填答結果以部分計分模式 (partial credit model) 估算組合分數 (變項名稱：home_resources)，供後續資料分析者參考使用。

十二年國民教育科技領域課綱草案規劃國中與高中生必修「程式設計」(課程及教學研究中心，2016)，為反映新課綱實施前後學生學習程式撰寫經驗是否有差異，和這樣的差異和學習表現的影響，2016 年問卷特別新增寫程式經驗題項，以銜接並比較不同世代學生的學習樣態。

表 3-3 休閒活動題數分布

變 項	2010 題 數	2013 題 數	2016 題 數
課外讀物 (S2010Q22_1、S2013Q21_1)	1	1	0
幫忙做家事 (S2010Q22_2、S2013Q21_2)	1	1	0
看電視、影片 (S2010Q22_3、S2013Q21_3)	1	1	0
線上遊戲 (S2010Q22_4、S2013Q21_4)	1	1	0
上網聊天 (S2010Q22_5、S2013Q21_5)	1	1	0
和朋友玩耍或聊天 (S2010Q22_6、S2013Q21_6)	1	1	0
運動 (S2010Q22_7、S2013Q21_7)	1	1	0
工作 (S2013Q21_8)	0	1	0
電子遊戲投入 (S2016Q27_1~4)	0	0	4
電子遊戲類型 (S2016Q28_1~5)	0	0	5
電子遊戲助益 (S2016Q29_13)	0	0	1

註：括弧內為試題編碼，請參照編碼簿，可直接至 <http://trac.naer.edu.tw/106TASA> 下載。

隨著科技革新，八年級學生身邊帶著手機的現象益發普及，投入各式電子遊戲現象隨處可見，故 2016 年休閒活動題項特別著重學生們從事電子遊戲時間與類型，並由學生自評電子遊戲對其學習表現的影響。所得資料可和其他心理健康、學習表現或學校層級資料進行交叉分析。

表 3-4 不分學科的心理學構念題數分布

構	念	2010 題 數	2013 題 數	2016 題 數
家庭關係 (S2010Q15_1~6、S2013Q14_1~6)		6	6	0
同儕關係 (S2010Q16_1~4、S2013Q15_1~4、S2016Q14_1~4)		4	4	4
師生關係 (S2010Q17_1~4、S2013Q16_1~4)		4	4	0
班級常規適應 (S2010Q18_1~4、S2013Q17_1~4、S2016Q13_1~5)		4	4	5
學習策略 (S2010Q19_1~10、S2013Q18_1~10)		10	10	0
學習偏好 (S2010Q20_1~6、S2013Q19_1~6)		6	6	0
自我效能 (S2010Q21_1~2、S2013Q20_1~2)		2	2	0
趨向表現 (S2010Q21_3~4、S2013Q19_7~9)		2	3	0
自我概念 (S2016Q15_1~5)		0	0	5
恆毅力 (S2016Q16_1~8)		0	0	8
電子遊戲滿足感 (S2016Q29_1~12)		0	0	12

註：括弧內為試題編碼，請參照編碼簿，可直接至 <http://trac.naer.edu.tw/106TASA> 下載。

心理學構念分布於一般性問卷題項和學科特定性問卷題項之中。由表 3-4 可知，歷年不分學科的心理學構念內涵包括學校、家庭、同儕及個體本身的向度。由表 3-5 可知，2010 年與 2013 年學科特定性問卷大多為單題或類別題，能構成一個分量表的構念不多，著重了解學校提供的學習環境或學生自己的學習與生活經驗；2016 年增加構念分量表的比重，以期能對學習表現提供更豐富的心理學闡釋基礎。

表 3-5 學科特定性變項題數分布

變 項	2010 題 數	2013 題 數	2016 題 數
國 文			
國文學習遷移 (S2010Q23_2、S2013Q22_2)	1	1	0
國文趨向表現 (S2010Q23_3)	1	0	0
國文學習動機 (S2010Q23_4~6、S2013Q22_3~5)	3	3	0
作文學習動機 (S2010Q23_7、S2013Q22_6)	1	1	0
作文撰寫篇數 (S2010Q23_8、S2013Q22_7)	1	1	0
作文撰寫指引 (S2010Q23_9、S2013Q22_8)	1	1	0
作文擅長文體 (S2010Q23_10、S2013Q22_9)	1	1	0
國文喜歡內容 (S2010Q23_11、S2013Q22_10)	1	1	0
國文能力優勢 (S2010Q23_12、S2013Q22_11)	1	1	0
國文學習策略 (S2010Q23_13、S2013Q22_12)	1	1	0
閱讀策略教學 (S2016Q17_1~2)	0	0	2
寫作策略教學 (S2016Q17_3~5)	0	0	3
閱讀態度 (S2016Q18_1~5)	0	0	5
英 語			
英語學習遷移 (S2010Q24_2、S2013Q24_2)	1	1	0
英語趨向表現 (S2010Q24_3、S2016Q22_2、S2016Q22_4、 S2016Q22_8)	1	0	3
英語逃避表現 (S2016Q22_6、S2016Q22_10、S2016Q22_12)	0	0	3

表 3-5 學科特定性變項題數分布 (續)

變 項	2010 題 數	2013 題 數	2016 題 數
英語趨向精熟 (S2016Q22_1、S2016Q22_3、S2016Q22_7)	0	0	3
英語逃避精熟 (S2016Q22_5、S2016Q22_9、S2016Q22_11)	0	0	3
英語學習動機 (S2010Q24_4~6、S2013Q24_3~5)	3	3	0
英語學習策略 (S2010Q24_7~10、S2013Q24_6~9)	4	4	0
家庭英語環境 (S2010Q24_11~13、S2013Q24_10~12)	3	3	0
班級英語教學 (S2010Q24_14~16、S2013Q24_13~15)	3	3	0
學校英語環境 (S2010Q24_17~20、S2013Q24_16~19)	4	4	0
英語課教學方式 (S2016Q21_1~6)	0	0	6
數 學			
數學學習遷移 (S2010Q25_2、S2013Q23_2)	1	1	0
數學趨向表現 (S2010Q25_3)	1	0	0
數學學習動機 (S2010Q25_4~6、S2013Q23_3~5)	3	3	0
電腦輔助學習 (S2010Q25_7~12)	6	0	0
計算機輔助學習 (S2010Q25_13~15、S2013Q23_6~8)	3	3	0
此次測驗難度 (S2010Q25_16、S2013Q23_9)	1	1	0
再次測驗努力意願 (S2010Q25_17、S2013Q23_10)	1	1	0
數學題難度偏好 (S2010Q25_18、S2013Q23_11)	1	1	0
解題態度 (S2010Q25_19~20、S2013Q23_12~13)	2	2	0
解題策略 (S2016Q19_1~6)	0	0	6

表 3-5 學科特定性變項題數分布 (續)

變 項	2010 題 數	2013 題 數	2016 題 數
精熟策略 (S2016Q19_7~11)	0	0	5
成功期望 (S2016Q20_1~4)	0	0	4
興趣價值 (S2016Q20_5~8)	0	0	4
社 會			
曾修習社會課程 (S2013Q25_3)	0	4	0
社會學習遷移 (S2010Q26_2、S2013Q25_3、S2016Q24_11)	1	1	1
社會趨向表現 (S2010Q26_3)	1	0	0
社會學習動機 (S2010Q26_4~6、S2013Q25_4~6)	3	3	0
時事關注 (S2010Q26_7~8、S2013Q25_7、S2013Q25_11、 S2016Q24_10)	2	2	1
公民學習動機 (S2016Q24_8~9)	0	0	2
歷史輔助學習 (S2010Q26_9~10、S2013Q25_8~9、 S2016Q24_3~4)	2	2	2
歷史學習動機 (S2016Q24_1~2)	0	0	2
地圖運用 (S2010Q26_12、S2013Q25_10、S2016Q24_7)	1	1	1
地理節目 (S2010Q26_11、S2013Q25_12~15)	1	4	0
地理學習動機 (S2016Q24_5~6)	0	0	2
自 然			
曾修習自然課程 (S2013Q26_2)	0	5	0

表 3-5 學科特定性變項題數分布 (續)

變 項	2010 題 數	2013 題 數	2016 題 數
自然學習遷移 (S2010Q27_2、S2013Q26_3)	1	1	0
自然趨向表現 (S2010Q27_3)	1	0	0
自然學習動機 (S2010Q27_4~6、S2013Q26_4~6、 S2016Q23_1~5)	3	3	5
自然自我效能 (S2016Q23_6~9)	0	0	4
實驗課頻率 (S2010Q27_7、S2013Q26_7)	1	1	0
實驗課態度 (S2010Q27_8、S2013Q26_8)	1	1	0
戶外教育 (S2010Q27_9、S2013Q26_9)	1	1	0
課外讀物 (S2010Q27_10、S2013Q26_10)	1	1	0
科學節目 (S2010Q27_11~14、S2013Q26_11~14)	4	4	0

(二) 學校層級的資料

在學校層級資料方面，除了提供「哪些學生隸屬於同一校」之訊息，TASA 亦依據侯佩君、杜素豪、廖培珊、洪永泰及章英華（2008）針對臺灣鄉鎮市區類型之研究的分類，將受測學校所處之鄉鎮市區分為都會核心、工商市區、新興市鎮、傳統產業市鎮、低度發展鄉鎮、高齡化鄉鎮及偏遠鄉鎮七個集群，以供次級資料分析者進行多層次模型（multilevel model）相關研究，或了解學校所處地區不同都市化程度學生之心理健康與學習表現。

三、分析

心理學構念之計量特性分析方法如下。

(一) 題項分析

題項分析 (item analysis) 包括計算各題之平均數 (mean)、標準差 (standard deviation) 以及題分與「刪除該題後的構念分量表總分」之相關 (corrected item-total correlation, CITC)，也就是鑑別度，用來作為題項篩選之參考。若某個題項的平均數甚高或甚低，顯示受試者對於題項的陳述具有某種特定傾向；而若某個題項的「題分與刪除該題後的構念分量表總分」之正相關過低，顯示該題和分量表之間不夠緊密。

(二) 信度分析

採用 Cronbach alpha 檢驗同一構念分量表的題目是否一致地測量同一特質。

四、結果

學生問卷所包含的心理學構念，部分為一般性非指涉某學科，部分則具學科特定性，測量學生對於某一學科的主觀感受。

(一) 不分學科的心理學構念

TASA 2010、2013 及 2016 學生問卷不分學科的心理學構念合計共 11 項。負向問句的題項反向計分後，2010 與 2013 為 Likert 四點量表，計分方式為「一直如此 =4」、「經常如此 =3」、「偶爾如此 =2」、「很少如此 =1」，2016 則為 Likert 五點量表，計分方式為「非常同意 =5」、「同意

=4」、「無意見=3」、「不同意=2」、「非常不同意=1」。

1. 題項分析

(1) 家庭關係

家庭關係於 TASA 2010 和 2013 學生問卷測量，共 6 題，第五題和第六題反向計分，2010 年與 2013 年各題遺漏值比例均為 0.1% 至 0.6%。

表 3-6 家庭關係之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我會和家人談論學校發生的事	2.49	1.00	.49	2.55	1.02	.49
我可以感受到家人對我的關心	2.95	0.94	.62	3.03	0.92	.60
家人會陪我唸書或寫功課	1.60	0.85	.34	1.59	0.85	.34
我需要幫忙時，家人會來幫我	2.84	0.98	.53	2.91	0.95	.53
我覺得父母的管教嚴（反向計分）	2.82	0.94	.16	2.91	0.93	.16
我覺得父母很囉唆（反向計分）	2.67	0.95	.39	2.75	0.92	.38

由表 3-6 可知，第五題「我覺得父母的管教嚴」的題分與「刪除該題後的構念分量表總分」相關僅有 .16 明顯低於其他題項，分析構念分量表分數內部一致性時將刪除。

(2) 同儕關係

同儕關係於 TASA 2010、2013 及 2016 學生問卷測量，共 4 題。2010 與 2013 年之第一題和第二題反向計分，2010 年各題遺漏值比例為 0.3% 至 0.4%，2013 年則為 0.1% 至 0.4%，2016 年為 0.8%。

表 3-7 同儕關係 2010 與 2013 之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我會和同學們吵架（反向計分）	3.66	0.58	.30	3.63	0.57	.23
同學們會欺負我（反向計分）	3.71	0.63	.32	3.82	0.50	.30
我會和同學們一起討論功課	2.23	0.87	.31	2.21	0.88	.33
我需要幫忙時，同學們會來幫我	2.95	0.86	.44	2.93	0.86	.44

由表 3-7 可知，同儕關係 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .30 至 .44 之間、2013 介於 .23 至 .44 之間，雖然相關都不高但沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

表 3-8 同儕關係 2016 之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
我覺得我的朋友夠多	3.89	1.02	.54
我和朋友在一起的時候，覺得很快樂	4.33	0.84	.68
我的朋友能了解我	3.75	1.01	.75
我會和朋友談心事	3.80	1.11	.64

鑒於 2010 與 2013 同儕關係題項間的一致性不夠理想，2016 更換同儕關係題項。由表 3-8 可知，同儕關係 2016 各題之題分與「刪除該題後的構念分量表總分」相關介於 .54 至 .75 之間，沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

(3) 師生關係

師生關係於 TASA 2010 和 2013 學生問卷測量，共 4 題，2010 年各題遺漏值比例為 0.2% 至 0.3%，2013 年則為 0.1% 至 0.2%。

表 3-9 師生關係之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我喜歡我的導師	2.71	0.89	.80	2.76	0.85	.80
我可以感受到導師對我的關心	2.88	0.81	.81	2.91	0.77	.82
我和導師之間相處得很好	2.75	0.82	.83	2.79	0.78	.83
我需要幫忙時，導師會來幫我	2.92	0.80	.72	2.95	0.75	.73

由表 3-9 可知，師生關係 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .72 至 .83 之間、2013 介於 .73 至 .83 之間，沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

(4) 班級常規適應

班級常規適應於 TASA 2010、2013 及 2016 學生問卷測量。2010 與 2013 題項內容相同，為 4 題，第四題反向計分；2016 共 5 題，沒有需要反項計分的題項。2010 年各題遺漏值比例為 0.7% 至 0.9%，2013 年則為 0.1% 至 0.3%，2016 年為 0.8%。

表 3-10 班級常規適應 2010 與 2013 之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
上課時我能遵守老師的規定	2.85	0.83	.71	2.88	0.82	0.70
我能準時完成老師交代的作業	2.98	0.85	.59	2.97	0.84	0.58
我能遵守班上的規定	2.95	0.84	.71	3.00	0.82	0.69
我會受到懲罰（反向計分）	3.37	0.81	.41	3.24	0.87	0.32

由表 3-10 可知，班級常規適應 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .41 至 .71 之間、2013 介於 .32 至 .70 之間，分析構念分量表分數內部一致性時將全數保留。

表 3-11 班級常規適應 2016 之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
我會遵照老師的指示行事	4.02	0.86	.66
我會仔細地聆聽老師的教學	3.75	0.86	.69
我對於班級內的活動感興趣	3.84	0.93	.50
我會迅速反應老師提出的要求	3.42	0.91	.58
我會小心地使用教室內的設備 （如：置物櫃、黑板、桌椅等）	4.00	0.89	.59

為了提升班級常規適應的信效度，2016 將此構念測量題項修改自 Betts 與 Rotenberg (2007)。由表 3-11 可知，同儕關係 2016 各題

之題分與「刪除該題後的構念分量表總分」相關介於 .50 至 .69 之間，沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

(5) 學習策略

學習策略於 TASA 2010 與 2013 學生問卷測量，共 10 題，分為記憶或複述策略 3 題、控制策略 4 題及精緻化策略 3 題。2010 年各題遺漏值比例為 0.7% 至 1.0%，2013 年則為 0.1% 至 0.4%。

表 3-12 學習策略之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
記憶或複述策略						
我會溫習某些課業問題	2.17	0.92	.68	2.24	0.96	.71
我會盡力背誦必須記住的東西	2.82	0.94	.57	2.80	0.94	.59
我會重複做練習	2.05	0.88	.68	2.01	0.88	.69
控制策略						
我還記得某些已經學過的東西	2.95	0.67	.47	2.94	0.67	.49
我知道目前還不懂的地方有哪些	2.98	0.70	.45	2.97	0.70	.48
對於不懂的問題，我會想辦法解決它	2.59	0.88	.59	2.58	0.89	.62
我會找出必須先學習的重點	2.45	0.92	.57	2.43	0.92	.59
精緻化策略						
我會用新方法來解決舊問題	2.40	0.88	.59	2.42	0.89	.60
我會將學到的東西應用到日常生活上	2.52	0.86	.64	2.50	0.88	.65
我會將學到的東西應用到其他學科上	2.27	0.89	.67	2.25	0.88	.68

由表 3-12 可知，學習策略 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .45 至 .68 之間、2013 介於 .48 至 .71 之間，分析各策略構念分量表分數內部一致性時將全數保留。

(6) 學習偏好

學習偏好於 TASA 2010 與 2013 學生問卷測量，共 6 題，分為偏好合作 5 題與偏好競爭 1 題。2010 年各題遺漏值比例為 0.8% 至 1.2%，2013 年則為 0.1% 至 0.4%。

表 3-13 學習偏好之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
偏好合作						
上課時，我喜歡和其他同學分組討論問題	2.39	0.96	.63	2.46	0.97	.64
當在分組討論問題時，我會整合小組中其他同學的意見	2.37	0.96	.66	2.39	0.95	.66
當在進行小組討論時，我認為我可以展現出最好的實力	2.22	0.91	.64	2.24	0.91	.66
上課時，我喜歡幫助小組中的其他同學，使他們表現得更好	2.38	0.92	.69	2.43	0.93	.69
上課時，若以分組方式進行學習，我認為我可以學得最多	2.37	0.96	.61	2.41	0.96	.62
偏好獨立						
上課時，我覺得自己一個人學習的方式，可以學得最多	1.98	0.91	—	2.02	0.90	—

由表 3-13 可知，偏好合作 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .61 至 .69 之間、2013 介於 .62 至 .69 之間，分析偏好合作構念分量表分數內部一致性時將全數保留。

(7) 自我效能

自我效能於 TASA 2010 和 2013 學生問卷測量，共 2 題，第二題反向計分，2010 年各題遺漏值比例為 1.0%，2013 年則為 0.2% 至 0.4%。

表 3-14 自我效能之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
任何學科的學習，只要我努力，就一定可以學好	2.94	0.82	.38	2.92	0.81	.40
就某些學科而言，不管我再怎麼努力，還是學不好（反向計分）	2.56	0.93	.38	2.47	0.93	.40

由表 3-14 可知，自我效能 2010 題分與「刪除該題後的構念分量表總分」相關為 0.38、2013 為 0.40，分析構念分量表分數內部一致性時將全數保留。

(8) 趨向表現

趨向表現於 TASA 2010 和 2013 學生問卷測量，2010 年 2 題、2013 年為提升信度增加為 3 題，2010 年各題遺漏值比例為 1.0%，2013 年則為 0.4% 至 0.6%。

表 3-15 趨向表現之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我在每個學科上都下很大的功夫，因為我想成為班上功課最好的人之一	—	—	—	1.99	0.96	.76
我希望成為班上表現最好的學生	2.49	1.01	.59	2.25	1.02	.72
我很努力，因為我想在考試上表現得比其他同學好	2.64	0.98	.59	2.40	1.00	.75

由表 3-15 可知，趨向表現 2010 題分與「刪除該題後的構念分量表總分」相關為 0.59、2013 介於 .72 至 .76 之間，分析構念分量表分數內部一致性時將全數保留。

(9) 自我概念

自我概念於 TASA 2016 學生問卷測量，題項修改自 Rosenberg (1965)，共 5 題，沒有需要反向計分的題項，各題遺漏值比例為 0.8%。

表 3-16 自我概念之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
我認為自己是個有用的人，至少與別人差不多	3.83	0.95	.75
我覺得我有許多優點	3.48	0.97	.81
我做事可以做得和大多數人一樣好	3.72	0.95	.76
我覺得自己很棒	3.50	1.01	.82
整體而言，我對自己感到滿意	3.53	1.02	.78

由表 3-16 可知，自我概念各題之題分與「刪除該題後的構念分量表總分」相關介於 .75 至 .82 之間，分析構念分量表分數內部一致性時將全數保留。

(10) 恆毅力

恆毅力是指「投入一件你非常在乎的事情，在乎到你願意一直守著它」（洪慧芳譯，2016），於 TASA 2016 學生問卷測量，共 8 題，修改自 Duckworth、Peterson、Matthews 及 Kelly（2007），分為熱情 4 題和毅力 4 題，各題遺漏值比例為 0.8% 至 1.1%。

表 3-17 恆毅力之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
熱 情			
我時常立下一個目標，但一陣子過後又改追求別的目標	3.02	1.09	.59
新的計劃和想法有時候會讓我無法專心於現有的計劃	3.29	1.02	.57
當我著迷在某件計畫一陣子後，會失去興趣	2.90	1.12	.59
我很難把我的注意力集中在一個「預期要花好幾個月時間」才能完成的計畫上	2.96	1.15	.56
毅 力			
無論什麼事情，我開了頭就要完成它	3.48	0.96	.49
挫折不會使我氣餒	3.45	1.04	.48
我是個努力做事的人	3.57	0.92	.67
我是個勤勞的人	3.26	0.96	.62

由表 3-17 可知，恆毅力各題之題分與「刪除該題後的構念分量表總分」相關介於 .48 至 .67 之間，分析各構念分量表分數內部一致性時將全數保留。

(11) 電子遊戲滿足感

電子遊戲滿足感於 TASA 2016 學生問卷測量，共 12 題，依據 Selnow (1984) 及 Colwell 與 Payne (2000) 的理論發展而成，分為行動 4 題、陪伴 4 題及友誼 4 題。考量電子遊戲雖然是很多青少年生活的一部分，但卻非所有的八年級學生都會接觸，故將此構念分量表題項置於問卷最後一頁，並於指導語說明「若完全沒在打電玩，不用填答第 29 大題」，以致各題遺漏值比例為 24.6% 至 25.0%。

表 3-18 電子遊戲滿足感之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
行 動			
打電玩時讓我忘了自己是孤單一人	3.11	1.32	.62
打電玩能讓我獨立思考	3.58	1.12	.62
打電玩能讓我成為遊戲中的一份子	3.40	1.19	.73
打電玩能讓我成為自己想要的樣子	2.96	1.24	.65
陪 伴			
打電玩時讓我覺得自己有伴	3.30	1.22	.79
打電玩時就像和朋友在一起	3.20	1.26	.80
電玩是我的好夥伴	3.09	1.21	.74

表 3-18 電子遊戲滿足感之描述性統計 (續)

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
電玩能讓我紓壓	3.91	1.02	.56
友 誼			
電玩讓我和朋友們聊天的話題更豐富	3.59	1.19	.67
電玩讓我和朋友們感情更好	3.25	1.20	.74
藉由電玩我在網路上認識新朋友	3.05	1.32	.77
藉由電玩我在網路上交到好朋友	2.87	1.30	.77

由表 3-18 可知，電子遊戲滿足感各題之題分與「刪除該題後的構念分量表總分」相關介於 .62 至 .79 之間，分析各構念分量表分數內部一致性時將全數保留。

2. 信度分析

將題項分析後所保留的構念分量表試題分數計算內部一致性係數，由表 3-19 可知 TASA 學生問卷不分學科的心理學構念分量表信度介於 .53 至 .91 之間。

表 3-19 不分學科的心理學構念分量表分數之內部一致性

	Cronbach α		
	2010	2013	2016
家庭關係	.73	.72	—
同儕關係	.55	.53	.82

表 3-19 不分學科的心理學構念分量表分數之內部一致性 (續)

	Cronbach α		
	2010	2013	2016
師生關係	.91	.91	—
班級常規適應	.79	.76	.81
學習策略			
記憶或複述策略	.80	.81	—
控制策略	.73	.74	—
精緻化策略	.79	.80	—
學習偏好			
偏好合作	.84	.85	—
偏好獨立	—	—	—
自我效能	.55	.57	—
趨向表現	.74	.86	—
自我概念	—	—	.91
恆毅力			
熱情	—	—	.78
毅力	—	—	.76
電子遊戲滿足感			
行動	—	—	.83
陪伴	—	—	.87
友誼	—	—	.88

(二) 和學科有關的心理學構念

TASA 2010、2013 及 2016 具學科特定性的心理學構念以動機和策略為兩大主軸，題項均無需反向計分。2010 與 2013 為 Likert 四點量表，計分方式為「一直如此 =4」、「經常如此 =3」、「偶爾如此 =2」、「很少如此 =1」或「完全同意 =4」、「同意 =3」、「不同意 =2」、「非常不同意 =1」。2016 則為 Likert 五點量表，計分方式為「非常同意 =5」、「同意 =4」、「無意見 =3」、「不同意 =2」、「非常不同意 =1」。

1. 題項分析

(1) 特定學科綜合性學習動機

學習動機長期以來都是教育心理學的核心議題。特定學科綜合性學習動機於 TASA 2010 和 2013 學生問卷測量，各科均為 3 題。由於學科問卷於 2010 和 2013 並非必填 (optional)，有接受該科測驗的學生才填寫，致使 2010 年各題遺漏值比例為 54.2% 至 57.6%，2013 年則為 57.5% 至 58.8%。

表 3-20 特定學科綜合性學習動機之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
國 文						
我喜歡學習國語文	2.65	0.83	.77	2.64	0.80	.78
我覺得學習國語文很容易	2.45	0.82	.62	2.51	0.81	.63
我覺得國語文很有趣	2.53	0.83	.74	2.53	0.82	.75
數 學						
我喜歡學習數學	2.35	0.96	.85	2.29	0.94	.84

表 3-20 特定學科綜合性學習動機之描述性統計 (續)

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我覺得學習數學很容易	2.12	0.92	.74	2.10	0.90	.72
我覺得數學很有趣	2.31	0.98	.83	2.27	0.95	.83
英 語						
我喜歡學習英語文	2.50	0.97	.84	2.51	0.93	.84
我覺得學習英語文很容易	2.30	0.95	.72	2.33	0.92	.72
我覺得英語文很有趣	2.44	0.96	.81	2.46	0.94	.82
社 會						
我喜歡學習社會	2.73	0.86	.82	2.71	0.80	.81
我覺得學習社會很容易	2.55	0.86	.70	2.52	0.80	.66
我覺得社會很有趣	2.72	0.87	.80	2.69	0.82	.76
自 然						
我喜歡學習自然	2.42	0.92	.83	2.41	0.87	.83
我覺得學習自然很容易	2.11	0.86	.70	2.14	0.83	.70
我覺得自然很有趣	2.49	0.94	.79	2.50	0.89	.79

由表 3-20 可知，特定學科綜合性學習動機 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .62 至 .85 之間、2013 介於 .63 至 .84 之間，沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

(2) 數學學習動機

為了更精緻化動機成分，TASA 2016 以期望價值理論 (Eccles, 2014; Eccles et al., 1983) 之中，成功期望與興趣價值兩個成分為基礎，測量數學的學習動機。其中成功期望 4 題、興趣價值 4 題，各題遺漏值比例為 1.2% 至 1.4%。

表 3-21 數學成功期望與興趣價值之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
成功期望			
學習數學對我而言並不是什麼難事	3.05	1.28	.86
數學是我厲害的科目之一	2.68	1.39	.83
我知道我可以將數學學好	3.49	1.18	.79
我可以解決數學難題	3.07	1.22	.86
興趣價值			
學習數學對我來說非常有趣	2.94	1.26	.80
我期待每一次的數學課	2.67	1.16	.80
一想到數學課的內容很精采，就讓我覺得很棒	2.70	1.18	.85
我期待在數學課能夠發現新的觀念與挑戰性的想法	3.18	1.26	.79

由表 3-21 可知，數學成功期望與興趣價值各題之題分與「刪除該題後的構念分量表總分」相關介於 .79 至 .86 之間，分析各構念分量表分數內部一致性時將全數保留。

(3) 英語學習動機

TASA 2016 以四向度目標導向理論（程炳林，2003；Elliot & Murayama, 2008）及其量表（Achievement Goal Questionnaire, AGQ）測量英語的學習動機，共 12 題，將英語學習行為分為四種目標導向：精熟導向的學習者專注於趨向學習目標，稱之為趨向精熟目標（mastery-approach goal），3 題；精熟導向的學習者逃避其學習目標，稱之為逃避精熟目標（mastery-avoidance goal），3 題；表現導向的學習者專注於趨向學習目標，稱之為趨向表現目標（performance-approach goal），3 題；表現導向的學習者逃避其學習目標，稱之為逃避表現目標（performance-avoidance goal），3 題，各題遺漏值比例為 1.3% 至 1.5%。

表 3-22 英語四向度目標導向量表之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
趨向精熟目標			
我想要完全精熟英文課堂中所教的內容	3.72	1.13	.78
我的英文課學習目標是學得越多越好	3.54	1.11	.76
我很努力地去理解英語課的內容，而且希望理解得越透徹越好	3.75	1.06	.78
逃避精熟目標			
我想要避免「英文課沒有盡力學習」	3.58	1.03	.78
我的學習目標是要避免「英文課沒有全力以赴」	3.44	1.04	.80
我很努力地避免「對英文課內容只有一知半解」	3.63	1.05	.77

表 3-22 英語四向度目標導向量表之描述性統計（續）

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
趨向表現目標			
我很努力地要在英語課中表現得比其他學生來的好	3.37	1.10	.80
我想要在英文課表現得比其他學生好	3.47	1.08	.82
我的學習目標是要在英語課中表現得比其他學生好	3.34	1.08	.74
逃避表現目標			
我的學習目標是要避免「英文課表現得比其他同學差」	3.48	1.09	.76
我很努力地要避免「英文課表現得比其他同學差」	3.48	1.05	.86
我想要避免「在英文課中表現得比其他學生差」	3.55	1.05	.85

由表 3-22 可知，英語四向度目標導向量表各題之題分與「刪除該題後的構念分量表總分」相關介於 .76 至 .85 之間，分析各構念分量表分數內部一致性時將全數保留。

(4) 國文學習策略教學

學生若能具備閱讀策略能力，可能透過閱讀學習新知；若能具備寫作策略能力，更可能透過文字表達或論述自己的想法。TASA 2016 測量教師教學時採用閱讀策略與寫作策略的情形，共 5 題，各題遺漏值比例為 0.8% 至 1.0%。

表 3-23 國文學習策略之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
閱讀策略教學			
閱讀時，老師會教我們找出段落的重點	4.22	0.87	.88
閱讀時，老師會教我們找出段落或全文的主旨	4.26	0.83	.88
寫作策略教學			
寫作文時，老師會教我們如何分段	3.92	0.99	.81
寫作文時，老師會教我們從題目去想可以寫作的內容有哪些	4.02	0.95	.80
寫作文時，老師會教我們組織結構	3.88	1.01	.83

由表 3-23 可知，國文閱讀策略各題之題分與「刪除該題後的構念分量表總分」相關為 .88、寫作策略則介於 .80 至 .83 之間，分析各構念分量表分數內部一致性時將全數保留。

(5) 數學學習策略

學生若能善用某些學習策略，學習數學可能事半功倍。TASA 2016 測量解題策略與精熟策略，共 11 題，各題遺漏值比例為 1.2% 至 1.3%。

表 3-24 數學學習策略之描述性統計

題 項 內 容	2016		
	<i>M</i>	<i>SD</i>	CITC
解題策略			
老師會教我們找出數學題目的關鍵字句，例如：「未知數」是什麼	4.13	0.90	.76
老師會教我們如何找出題目的條件，並根據題目的條件列式	4.22	0.83	.77
遇到題目時老師會教我們先從簡單例子著手思考	4.03	0.93	.76
遇到題目時老師會教我們先從特殊例子著手思考（例如：先想想看如果題目是正三角形、正方形、或特殊幾何圖形，會有什麼結果）	3.90	0.97	.70
老師會請我們在求出問題解答之後，再思考有沒有其他可能的解決辦法	4.00	0.94	.74
老師會請我們在求出問題解答之後，教我們進行驗算	3.56	1.05	.63
精熟策略			
老師會鼓勵我們課前預習數學內容	3.67	1.03	.68
老師會鼓勵我們課後複習學過的數學內容	3.99	0.93	.79
老師會要我們多多練習題目	4.19	0.87	.72
老師會鼓勵我們彼此討論數學	3.91	0.99	.74
老師會鼓勵我們堅持不懈地解決數學難題	3.82	0.99	.74

由表 3-24 可知，數學解題策略各題之題分與「刪除該題後的構念分量表總分」相關介於 .63 至 .77 之間、精熟策略則介於 .68 至 .79 之間，分析各構念分量表分數內部一致性時將全數保留。

(6) 英語學習策略

英語學習策略於 TASA 2010 和 2013 學生問卷測量，均為 4 題。由於學科問卷於 2010 和 2013 並非必填，有接受該科測驗的學生才填寫，致使 2010 年各題遺漏值比例為 55.7%，2013 年則為 58.4% 至 58.5%。

表 3-25 英語學習策略之描述性統計

題 項 內 容	2010			2013		
	<i>M</i>	<i>SD</i>	CITC	<i>M</i>	<i>SD</i>	CITC
我會閱讀英文圖書、報紙、雜誌和小說等課外讀物	1.62	0.84	.66	1.62	0.84	.67
我會聽英語廣播節目、英語歌曲或者觀賞英語電視節目、電影等	2.04	1.03	.58	2.10	1.05	.58
我會上英文網站搜尋所需要的資料	1.53	0.80	.63	1.56	0.82	.63
我會勇於嘗試用英語與人交談	1.66	0.87	.61	1.71	0.89	.62

由表 3-25 可知，英語學習策略 2010 各題之題分與「刪除該題後的構念分量表總分」相關介於 .58 至 .66 之間、2013 介於 .58 至 .67 之間，沒有特別低的題項，分析構念分量表分數內部一致性時將全數保留。

2. 信度分析

將題項分析後所保留的構念分量表試題分數計算內部一致性係數，由表 3-26 可知 TASA 學生問卷和學科有關的心理學構念分量表信度介於 .80 至 .94 之間。

表 3-26 和學科有關的心理學構念分量表分數之內部一致性

	Cronbach α		
	2010	2013	2016
特定學科綜合性學習動機			
國文	.84	.85	—
數學	.90	.90	—
英語	.89	.90	—
社會	.88	.86	—
自然	.88	.88	—
數學成功期望與興趣價值			
成功期望	—	—	.93
興趣價值	—	—	.92
英語四向度目標導向量表			
趨向精熟	—	—	.88
逃避精熟	—	—	.89
趨向表現	—	—	.89
逃避表現	—	—	.91

表 3-26 和學科有關的心理學構念分量表分數之內部一致性（續）

	Cronbach α		
	2010	2013	2016
國文學習策略			
閱讀策略	—	—	.94
寫作策略	—	—	.91
數學學習策略			
解題策略	—	—	.90
精熟策略	—	—	.89
英語學習策略	.80	.80	—

五、結 論

TASA 2010、2013 及 2016 評量的對象均為八年級學生，學生問卷題項亦以八年級學習脈絡進行設計，涵蓋學生自身的評估，和學生對於所處家庭、班級及學校的知覺。妥善利用學生層級和學校層級資料進行分析，可延伸出甚多研究議題。尤其 2014 年 5 月為第一屆國中教育會考，TASA 2013 和 2016 資料恰好能反映國中教育會考前後學生學習環境與學習表現的變化。期透過多年度、多向度的豐富資料，為臺灣教育政策擬定與教育方案規劃提供實證證據。

參考文獻

- 侯佩君、杜素豪、廖培珊、洪永泰、章英華（2008）。臺灣鄉鎮市區類型之研究：「臺灣社會變遷基本調查」第五期計畫之抽樣分層效果分析。《調查研究：方法與應用》，23，7-32。
- 洪慧芳譯（2016）。A. Duckworth 著。恆毅力：人生成功的究極能力。臺北市：天下雜誌。
- 國民教育法施行細則（2016）。
- 程炳林（2003）。四向度目標導向模式之研究。《師大學報：教育類》，48（1），15-40。
- 課程及教學研究中心（2016）。新課綱「程式設計」，學邏輯解問題。《國家教育研究院電子報》，134，取自 http://epaper.naer.edu.tw/index.php?edm_no=134&content_no=2672
- American Psychological Association (2017). Education and socioeconomic status. Retrieved from <http://www.apa.org/pi/ses/resources/publications/education.aspx>
- Betts, L. R., & Rotenberg, K. J. (2007). A short form of the teacher rating scale of school adjustment. *Journal of Psychoeducational Assessment*, 25(2), 150-164. doi:10.1177/0734282906296406
- Colwell, J., & Payne, J. (2000). Negative correlates of computer game play in adolescents. *British Journal of Psychology*, 91(3), 295-310. doi:10.1348/000712600161844
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), 1087-1101. doi:10.1037/0022-3514.92.6.1087
- Eccles, J. S. (2014). Expectancy-value theory. In R. C. Eklund & G. Tenenbaum (Eds.), *Encyclopedia of sport and exercise psychology* (pp. 269-273).

- Thousand Oaks, CA: Sage.
- Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75-146). San Francisco, CA: W. H. Freeman.
- Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*(3), 613-628. doi:10.1037/0022-0663.100.3.613
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Oxford, UK: Routledge.
- Martin, M. O., Mullis, I. V. S., Foy, P., & Arora, A. (2013). Context questionnaire scales details. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Selnow, G. W. (1984). Playing videogames: The electronic friend. *Journal of Communication, 34*(2), 148-156. doi:10.1111/j.1460-2466.1984.tb02166.x
- Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer Nature.

第四章 抽樣設計與權重

黃馨瑩

國家教育研究院助理研究員

講到抽樣的方式，大家最常想到的便是簡單隨機抽樣，這個抽樣方法大家直觀上覺得「最公平」，也就是每個人有相同的機率被抽到，且從研究的角度來看，簡單隨機抽樣計算抽樣誤差也很方便，有既定的公式可以帶入計算，然而，儘管簡單隨機抽樣在直觀上有很多優點，但在不同情境脈絡下，卻未必是最方便的方法，例如在大型教育調查中，如果用簡單隨機抽樣選取學生進行調查，樣本可能會散佈在非常多學校，不但大量增加調查成本，且一校只抽到少許幾個人，若要進一步瞭解班級氣氛、班級效應對學生學習的影響，樣本大小並不充足，因此在許多大型教育調查中，無論是PISA（Programme for International Student Assessment）、TIMSS（Trends in International Mathematics and Science Study）、或是美國知名的NAEP（The National Assessment of Educational Progress），在進行抽樣調查時，均不採取簡單隨機抽樣，而採取複雜抽樣設計（complex sampling），往往先針對母群進行適當的分層後，各層先抽出學校後再抽出班級和學生，此種方式較能事先掌握一校被抽中的樣本大小，學校也不至於過於分散，但缺點是估計抽樣誤差的難度會大大提高。無論抽樣方式為何，只要是機率抽樣便可以代表母群，只是抽樣誤差會隨著抽樣人數和抽樣方式不同有差異，舉例來說：若從臺灣隨機抽取10人，這10人可以代表全臺灣母群，它仍具代表性，但用這10人描述母群，誤差會很大。因此，大型教育調查會將誤差控制在有意

義的範圍內。

TASA 2016 為國內針對國中八年級的大型教育調查，其抽樣設計比照國際規格，利用複雜抽樣抽取出樣本，本章將先概覽 TASA 2016 的抽樣設計，再詳細說明抽樣設計和權重計算，最後進行討論。

一、TASA 2016 抽樣概覽

2016 年施測科目為國語文、英語（含聽力與閱讀）、數學、自然、社會五學科，考量英語包含聽力測驗，較難和其他科目一起施測，故將施測「國數社自」的學生一起抽樣，而施測英語的學生則另外自己抽樣，特別提醒的是，兩組樣本皆是回應相同的母群，讀者可以想成是兩個獨立的研究，所以兩者可能抽到相同的學校。為確保所抽取之樣本具有全國代表性，TASA 2016 採取先抽學校再抽班級的兩階段分層叢集抽樣（two-stage stratified cluster analysis），第一階段抽取單位（primary sampling unit, PSU）為學校，第二階段抽取單位（secondary sampling unit, SSU）為班級，讀者可以把 PSU 想成是一個群集（cluster）。在 TASA2016 中，第一階段依據行政區域分為北、中、南、東與離島 4 層，採用系統機率比例抽樣（systematic probability proportional to size, PPS）抽取樣本學校，第二階段再從樣本學校裡利用簡單隨機抽樣選取班級，原則上各校抽 1 個班，若那間學校被抽到 2 次，則抽到 2 個班，施測完之後，會將全部題目未作答的學生剔除，剩下的即為有效樣本。關於「國數社自」的抽樣，抽取 10182 人，共 304 間學校、333 個班，有效樣本則為 9955 人，共 304 間學校 333 個班；關於「英語」的抽樣，抽取 4225 人，共 141 間學校、141 個班，有效樣本則為 4096 人，共 140 間學校 140 個班，少一間學校一個班，是因為施測當天發生突發狀況，導致無法施測。

（一）調查母群

標的母群為 105 學年度在學的八年級學生，但考量施測上的地理因素、某些學校的獨特性，將海外中文學校、特殊教育學校排除在外，以及將臺灣本島 19 縣市中，八年級學生數 10 人以下的學校，以及離島 3 縣市八年級總學生數 5 人以下的學校排除在外，這種人數較少的學校共排除 9 校 29 人。另外，抽到學校後，進一步排除資源班學生、分校學生和在家教育的學生後（特殊學生則包括在家教育與 12 類障礙類別，其中，12 類障礙類別包括：智能障礙、視覺障礙、聽覺障礙、語言障礙、肢體障礙、身體病弱、嚴重情障、學習障礙、多重障礙、自閉症、發展遲緩及其他顯著障礙），所有公私立國民中學、有八年級附設國中部學校的學生，即為 TASA 2016 的八年級調查母群。

（二）樣本大小與精確性

樣本大小和推估母群的精確度有關，若要使母群估計值的 95% 信賴區間，能夠小於 $\pm 10\%$ 標準差（亦即標誤差要小於 5% 標準差），以隨機抽樣方式至少需 400 人，但 TASA 為先決定抽樣學校，再從中抽取班級的兩階段分層叢集抽樣，因此從中抽取出來的學生，可能會因為來自同一個班、受到同一個老師的薰陶、受到同樣學校氛圍的感染，而受到相近的影響，故樣本代表性會不如簡單隨機抽樣，導致 30 人代表的可能只有簡單隨機抽樣下的 15 人，又因為有時會進行子群體的比較（例如：性別差異比較），需讓各個子群體有足夠的樣本大小，加上 TASA 需要提供報告給各縣市，每個縣市、每個學科均需要足夠的樣本大小，故樣本大小遠高於臺灣參加 TIMSS、PISA 時的數目，「國數社自」和「英語」兩個調查，預計各抽 10000、4000 人。

（三）分層變項

抽樣設計時使用分層變項有幾個目的（PISA, 2012）：

1. 使抽樣設計更有效率、更可靠。
2. 確保母群中各類型樣本都能被抽到。
3. 確保母群中某些特定樣本的抽樣可以具代表性。

一般而言分層變項可以分為兩種：顯性分層（explicit stratification）、隱性分層（implicit stratification）：

1. 顯性分層

若研究目的需要推論到各子群的表現，就得透過顯性分層確保抽樣的人數和方法能反應各子群的情況，例如：若需要瞭解各縣市的表現情形，則可將縣市作為顯性分層變項，以確保參加調查的各個縣市，有足夠的樣本反應母群，也有正確的抽樣架構讓各縣市的樣本可以更精準的勾勒該縣市情況。因此，顯性分層往往是顯性名義變項，可能是地理區、公私立校別…等等。

2. 隱性分層

主要目的在於減少抽樣誤差，故隱性分層變項若能跟調查結果有高相關，對提升抽樣精準度將大有助益，例如：臺灣參加 TIMSS 2011 調查時，以學生的基本學力成績當成輔助變項，其抽樣誤差大幅降至 3.2 個量尺分數（TIMSS 量尺分數平均數 500，標準差 100）。

TASA 在抽樣時，會考量教育行政機關的需求，審慎估計樣本大小，並選取合宜的分層變項，但每年也會檢討修正，根據歷屆的目的和結果調整抽樣方法。

（四）班級抽樣

第一階段決定受測學校後，第二階段將抽出受測班級，從受測學校中隨機抽出數個班級，以 TASA 2016 年為例，原則上一校抽一班。而因為同樣班級的學生可能受到相同背景變項影響，導致抽出來的學生代表性不及簡

單隨機抽樣的人數，班級內的相關程度越高，表示學生同質性較高，能代表的訊息就比隨機抽樣下少很多，故在估計樣本大小時要考慮因為組內相關係數（intraclass correlation）而衍伸出的叢集抽樣設計效果（design effect），叢集抽樣的設計效果如下（kish, 1995）：

$$D_{effect} = 1 + \rho(n_c - 1), \rho : \text{組內相關係數}, n_c : \text{平均一校抽樣多少人}$$

根據 TIMSS 2011 臺灣報告可知，臺灣 8 年級學生組內相關係數約界在 0.2~0.3 之間，國中平均一班約 20 人，則叢集抽樣的設計效果約在 4.8~6.7 之間，表示以叢集抽樣方式抽取 1920~2680 人，才相當以簡單隨機抽樣抽取 400 人，而考量還要進行子群體的差異分析（例如：性別、城鄉的差異分析），故 TASA 在估算時至少會抽取 8000 人以上。

二、TASA 2016 抽樣步驟

TASA 2016 年和 TASA 2013 年不同的是，2016 年施測科目為國語文、英語（含聽力）、數學、自然、社會五學科，考量英語包含聽力測驗，較難和其他科目一起施測，故將施測「國數社自」的學生一起抽樣，而施測英語的學生則另外自己抽樣，要提醒的是，兩組樣本皆是回應相同的母群，讀者可以想成是兩個獨立的研究，所以兩者的學校會有部分重疊。

TASA 2016 年為先抽取學校再抽取班級的兩階段分層叢集抽樣，第一個抽樣單位為學校，在抽取學校時，學校只依照行政區分層後，決定出抽樣學校，在從學校中以簡單隨機抽樣隨機抽取 1~2 個班級，現說明如下：

（一）決定抽樣學校

1. 蒐集學校名冊：TASA 2016 在五月進行施測，此時為八年級一整年的學期末，八年級的學習內容幾乎學完，可以測得較完整的學習情形。

在 2015 年 12 月先根據教育部統計處¹提供的學校資料（包含學校八年級人數、班級數），先決定學校的抽樣架構（sampling frame），抽完學校之後，再請抽出的學校提供學生名單，我們依此名單挑選班級後將整班施測。

2. 分層變項：TASA 2016 年的分層變項為行政區，將臺灣分成北、中、南、東部與離島四區²，在 2016 年將此變項當成隱性分層，在 TASA 成果報告不會呈現四個區域的學生表現，此分層變項的目的是試圖減少抽樣時的誤差。
3. 依照各層人數佔總人數的比例，算出各層應抽出多少人，然後決定抽樣學校數，接著以 PPS 決定抽樣學校，母群和預估抽樣人數如表 4-1 和表 4-2 所示。

表 4-1 國數社自四科調查的母群和預估抽樣人數

分 層	母 群 學 生 數	母 群 學 校 數	預 估 抽 樣 人 數
北	107288	348	4419
中	65197	252	2685
南	62811	257	2587
東 與 離 島	7539	74	311
小 計	242835	931	10002

¹ 學校資料可由教育部統計處下載：<https://depart.moe.edu.tw/ED4500/News.aspx?n=5A930C32CC6C3818&sms=91B3AAE8C6388B96>

² 北區：新北市、宜蘭縣、桃園市、新竹縣、基隆市、新竹市、臺北市；中區：苗栗縣、彰化縣、南投縣、雲林縣、臺中市；南區：嘉義縣、屏東縣、嘉義市、臺南市、高雄市；東部及離島：臺東縣、花蓮縣、澎湖縣、金門縣、連江縣

表 4-2 英語調查的母群和預估抽樣人數

分 層	母 群 學 生 數	母 群 學 校 數	預 估 抽 樣 人 數
北	107288	348	1768
中	65197	252	1074
南	62811	257	1035
東 與 離 島	7539	74	125
小 計	242835	931	4002

4. PPS 決定抽樣學校的步驟如下：

- 步驟 1：先將學校按照八年人數由多到小排序，計算八年級累加人數。假設共有 N 間學校，最後累加人數為 X ，每校平均人數為 N/X 。
- 步驟 2：根據所需樣本大小（令為 S ），估算出需要多少學校數（令為 n ）。
- 步驟 3：算出抽樣間距（令為 I ），也就是每多少人就抽一間學校： $I = (X/S) * (N/X)$ 。
- 步驟 4：產生起始值（令為 s ），服從 Uniform $(1, I)$ ，第一個包含起始值的累加人數，就是第一間被抽到的學校。
- 步驟 5：第一個起始值加上抽樣間距，得到第二個值（即 $s+I$ ），第一個包含 $s+I$ 的累加人數就是第二間抽到的學校…以此邏輯繼續抽取學校。

PPS 抽樣舉例

假設母群有 2500 人，平均一班 25 人，若需從中抽取 100 位學生，先抽取學校再抽學生，每校各抽一班，抽學校的步驟如下：

步驟 1：先將學校按照八年人數由多到小排序，計算累加人數。

→ 累加人數在第三欄，分別是 500、920、1240……

步驟 2：根據所需樣本大小，估算出需要多少學校數。

→ $n=100/25=4$ ，需要 4 間學校。

步驟 3：算出抽樣間距，也就是每多少人就抽一間學校。

→ $I = (2500/100) * 25 = 625$

步驟 4：產生起始值，服從 Uniform (1, 625)，第一個包含起始值的累加人數，就是第一間被抽到的學校。

→ 在這邊產生的起始值 s 為 310，故 S1 為第一間抽到的學校。

步驟 5：第一個起始值加上抽樣間距，得到第二個值，第一個包含這個值的累加人數就是第二間學校……以此邏輯抽學校。

→ 因為要抽 4 間學校，起始值為 310，之後分別 $s+I=310+625$ 、 $s+2*I=310+325*2$ 、 $s+3*I=310+625*3$ ，分別為 935、1560、2185（第四欄）。

學 校	人 數	累 加 人 數	間 距	是否被抽樣
S1	500	500	310	是
S2	420	920		
S3	320	1240	935	是
S4	315	1555		
S5	250	1805	1560	是
S6	210	2015		
S7	190	2205	2185	是
S8	135	2340		
S9	100	2440		
S10	60	2500		

5. 若學校因故無法施測，會視情況採用替代學校，也就是原抽到學校的後一間，例如以上述的例子，若 S7 學校無法參加，則由 S8 參與施測。TASA 2016 施測英語當天，有一所學校有狀況無法施測，但來不及替換學校，故該年度少一間學校，而無替代學校。

（二）決定抽樣班級

第二階段抽樣單位為班級，決定抽樣學校後，再從該學校中以簡單隨機抽樣方式抽取 1 個受測班級。原則上，從抽樣學校抽取一個班級，若該學校被抽到兩次，則抽取 2 個班。在「國數社自」調查的抽樣，一間學校最多被抽到 2 次即 2 個班，在「英語」的調查下，一間學校最多被抽到 1 次即 1 個班。決定受測班級後，被抽到受測班級的學生均接受測驗，惟身心障礙類別、分校學生、在家教育三大類別不施測（這部分已被排除在母群之外），當天請假者也不另進行補測。

（三）資料清理（data clean）

施測完畢後，會將全部未作答的學生剔除、缺考的學生剔除，剩下的為有效樣本，在國數社自四科調查中缺考人數為 116 人，英文缺考人數為 116 人，表 4-3、4-4 則為兩個調查的母群和樣本大小。

表 4-3 國數社自四科調查的抽樣人數和有效樣本

分 層	實際抽樣 學生數	實際抽樣 學校數	有效樣本 大小	有 效 學校數	有 效 班級數
北	4501	137	4374	137	147
中	2713	75	2646	75	86
南	2667	82	2634	82	89
東與離島	301	10	301	10	11
小 計	10182	304	9955	304	333

表 4-4 英語科調查的抽樣人數和有效樣本

分 層	實際抽樣 學生數	實際抽樣 學校數	有效樣本 大 小	有 效 學 校 數	有 效 班 級 數
北	1822	61	1735	60	60
中	1151	38	1129	38	38
南	1112	37	1095	37	37
東與離島	140	5	137	5	5
小 計	4225	141	4096	140	140

三、權重計算

權重是指一個樣本在母群會代表多少人，也就是抽樣機率的倒數，舉例來說，若母群有 100 人，以簡單隨機抽樣從中抽取 20 人，表示要用 20 人代表 100 人，一個人要代表母群中的 5 個人（ $100/20$ ），一個人的權重就是 5，而從機率來看，一個人被抽中的機率是 0.2（ $20/100$ ），抽樣機率和權重互為倒數。在一般情況下，希望每個人被抽中的機率相同，也就是權重相同，但實際運作上，各大型調查中可能因為對某些族群加重抽樣、缺考等因素，使得每個學生被抽到的機率（權重）很難相同，TASA 2016 因為幾個原因促使每個人的權重不同：

1. 母群名冊和實際可能有出入：儘管母群名冊是當屆才蒐集，但仍有可能因為學生轉學等情形，我們無法收到最新的名冊，導致影響每個人被抽到的機率。
2. 班級數的差異：TASA 2016 以學校八年級累加人數進行 PPS 挑選學校，學校被抽到的機率和八年級人數呈正比，但在第二階段抽樣，從學校中抽取班級時，學生被抽到的機率是和班級數有關，而非和各校人數多寡有

關，又因為各校班級數不同，所以每個學生被抽到的機率會不同，如果第二階段抽樣是從學校隨機挑選學生，則每個學生被抽到的機率則會相同。

3. 缺考：考試當天因故缺考，使得該校要用較少學生代表一整校的學生，所以權重將調大。

以 TASA 的抽樣架構來看，先抽學校再抽班級後才施測學生，以 PPS 抽學校時，即可知道一間學校被抽到的機率，便知道學校權重，抽班級時也可得到班級被抽到的機率，便知道班級權重，故在沒有人缺考的情況下，學生權重即為學校權重乘上班級權重，現說明如下。

(一) 學生權重

1. 學校權重

學校權重就是學校被抽到的機率倒數，TASA 以 PPS 方式決定抽樣學校，因此第 i 所學校權重的計算為 W_{sc}^i ：

$$W_{sc}^i = \frac{X}{n \cdot x_i}, \quad X = \sum_{i=1}^N x_i$$

其中， n 為預計抽樣校數， x_i 為第 i 所學校的母群學生數，而 N 為該分層的學校總數， X 為該分層的所有學校的學生總數。

2. 班級權重

班級權重就是班級在校內被抽到的機率倒數，TASA 以簡單隨機抽樣方式抽取該校班級，因此在第 i 所學校中，班級權重的計算如 W_{class}^i ：

$$W_{class}^i = \frac{C^i}{c^i}$$

其中， C^i 為第 i 學校總班級數， c^i 為第 i 所學校預計抽樣班級數。

3. 學生權重

學生權重為學生在班級內被抽到的機率倒數，TASA 抽到的班級內學生將全數施測，因此在第 i 所學校，第 j 個班級內學生的權重如 $W_{stu}^{i,j}$ ：

$$W_{stu}^{i,j} = 1$$

4. 班上學生調整權重

若施測當日如有學生請假等學生無法施測的情況，或是學生出席但成就測驗和問卷完全未作答等無效樣本，這些將被剔除，故樣本權重將經過調整。例如若班上應該有 30 人施測，但當天因故卻只有 29 人施測或 29 個有效樣本，表示該班要用 29 人來代表 30 人，學生權重須經調整要再乘上 30/29。校正後的學生權重計算如 $FW_{stu}^{i,j}$ ：

$$FW_{stu}^{i,j} = A_{stu}^{i,j} \cdot W_{stu}^{i,j}, \quad A_{stu}^{i,j} = \frac{S_{rs}^{i,j} + S_{nrs}^{i,j}}{S_{rs}^{i,j}}$$

其中， $A_{stu}^{i,j}$ 為第 i 所學校，第 j 個班級的學生權重校正係數， $S_{rs}^{i,j}$ 為第 i 所學校，第 j 個班級中作答反應有效的學生數，表示缺席的或無效樣本。

5. 學生最後的權重

學生最終的權重為學校權重、班級權重、班上學生權重三者相乘，公式如下：

$$TW_{stu}^{i,j} = W_{sc}^i \cdot W_{class}^i \cdot FW_{stu}^{i,j}$$

(二) 釋出資料權重

TASA 釋出兩種權重，分別是 Totwgt (total student weight)、Houwgt (student house weight)，以上公式 $TW_{stu}^{i,j}$ 算出來即為學生的 Totwgt，所

有學生的 Totwgt 加總後，會很接近母群的人數。而 Houwgt 是將 Totwgt 除以推估的母群人數再乘以樣本大小，故學生的 Houwgt 加總後會接近樣本大小。

兩者關係用公式表示如下：

$$\text{Houwgt} = \frac{\text{Totwgt}}{\text{母群人數}} \times \text{實際抽樣人數}$$

進行描述性分析時，建議使用 Totwgt，但要進行假設檢定分析、跨屆分析的時候，建議用 Houwgt，以避免樣本膨脹而影響假設檢定結果（表 4-5）。

另外，要提醒讀者在分析大型教育調查時要使用正確的權重，以避免低估抽樣標準誤。表 6 顯示有無使用權重時，國語第一個似真值的估計標準誤，不管有沒有使用權重，平均數都是一樣的，但沒有使用權重時，標準誤會被低估，這會造成假設檢定時，較容易達到顯著。

表 4-5 Totwgt 和 Houwgt 比較表

變項名稱	全名	加總結果	使用時機
Totwgt	Total student weight	等於母群人數	描述性統計
Houwgt	Student house weight	等於樣本大小	顯著性檢定

表 4-6 不同加權方式對國語第一個似真值的影響

加權方式	人數	平均值	標準誤
不加權	246153	499.88	1.00
Totwgt	9959	499.88	2.07
Houwgt	9959	499.88	2.07

四、抽樣誤差

在許多大型教育調查裡，多為先抽學校再抽班級（或學生）的兩階段抽樣，故同一個班級或同一個學校的學生彼此有一定程度的相關性而非獨立的，已經違反簡單隨機抽樣的假設，故沒有現成簡單的公式可以計算抽樣誤差，比較常見的計算方式為利用平衡重複取樣法 (balanced repeated replication, BRR) (Plackett & Burman, 1946)、刀切重複取樣法 (jackknife repeated replication technique, JRR) (Frankel, 1971)、拔靴法 (bootstrap method) (Efron & Tibshirani, 1993) 來進行抽樣誤差的計算，像 PISA 採用 BRR、TIMSS 採用 JRR 來計算抽樣誤差。這些方法的概念是利用重複取樣的方式計算抽樣誤差，亦即利用抽出來的樣本，模擬在研究設計的抽樣架構下，重複抽出這些樣本所得的統計量，這些統計量分配的標準差即是抽樣誤差。要注意的是，讀者在分析大型教育調查的資料時，如果沒用正確的方法估計抽樣誤差，而是用 SPSS 傳統的處理方式，則會將樣本當成來自簡單隨機抽樣之下去處理，抽樣誤差往往會被低估，假設檢定就會很容易顯著，而誤用了分析結果。

不管是用 BRR、JRR、拔靴法…哪一種方式估算抽樣誤差，其概念均為從抽出的樣本中，仿照抽樣架構重複抽取樣本，以得到統計量的抽樣誤差，為便利讀者重複取樣，本資料提供「JKZONE」、「JKREP」兩個變項，現說明如下：

(一) JKZONE

TASA 2016 將行政區當作同一個隱性分層，將所有學校依照八年級總人數由大到小排序後，將相近的兩個學校配對成同一組，稱為「JKZONE」。而因為行政區是隱性分層變項，所以兩個學校來自不同行政區也沒關係，對抽樣誤差估計影響不大。TASA 2016 國數自社四科調查中，因抽取 304 間學

校，故有 152 組 JKZONE，而英語有 140 間學校，固有 70 組 JKZONE。

(二) JKREP

在估算抽樣誤差時，是從樣本中模擬，在相同的抽樣架構下，進行重複抽樣，在這邊的模擬方式，是每一次重複抽樣時，變動某一組 JKZONE 的學校，會將其中一間學校當成沒抽到權重為 0，另一間被抽到兩次權重變兩倍。讀者可利用 JKZONE、JKREP 來計算抽樣誤差，詳情請見第六章。

五、結 論

TASA 2016 八年級母群人數為 242835 人，國數社自四科的抽樣人數為 9955 人，Houwgt 最小為 0.14 最大為 2.95，權重加總回推母群為 246153 人，英語文抽樣人數為 4096 人，Houwgt 最小為 0.46，最大為 2.42，加總回推母群為 245461 人，兩個調查回推母群的狀況還算良好。考量權重後，各科誤差如表 7，由結果可知，誤差大概界在 2.02~3.49 之間，小於 0.05 個標準差 ($0.05 \times 100 = 5$)，表示在 TASA 2016 的抽樣架構下，有將學生表現估計值控制在我們希望的誤差之內。

表 4-7 TASA 2016 各科學生表現標準誤

	<i>n</i>	<i>M (SD)</i>	SE
國 文	9955	500 (100)	2.07
數 學	9955	500 (100)	2.14
自 然	9955	500 (100)	2.05
社 會	9955	500 (100)	2.02
英 語	4096	500 (100)	3.49

另外，讀者可能有幾個問題，現說明如下：

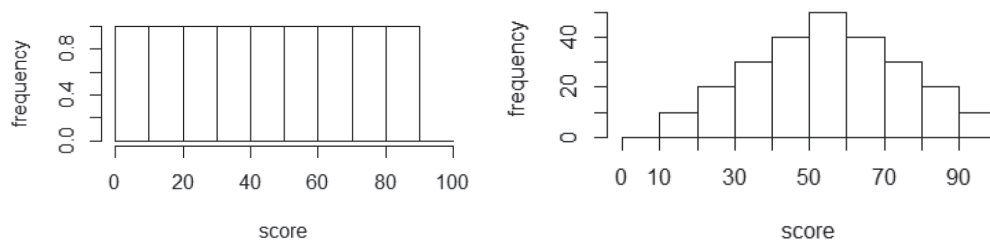
（一）TASA 抽樣架構不同，歷屆八年級的成績可以互相比較嗎？

在回答這問題之前，大家可以先思考無論是 PISA 或 TIMSS，皆會呈現各國的分數，儘管各國的抽樣方式不同，但各國間的分數是可以比較的，而一個國家內，也可以看出自身的表現較前一次進步或退步，這其實意味著每年的成績都是在相同的量尺上，各次成績可以互相比較。其實每一次評量的抽樣設計，在分層變項選取上、抽樣架構均未必相同，以臺灣參加 TIMSS 為例，在 TIMSS 2015 才開始以學校基本學力測驗當成分層變項，在這之前並未以此分層過。其實，只要是機率抽樣，樣本皆可以回推母群，樣本均具有代表性，而不斷修正抽樣設計，主要是為了減少抽樣誤差，使樣本在描繪母群樣貌時更精準。抽樣方式主要決定抽樣誤差，但大型教育調查還有等化誤差與測量誤差（Wu, Tam & Jen, 2016），若要使各次成績可以比較，更大的關鍵在於各次成績間的等化要準確，只要是機率抽樣將誤差控制在一定範圍內、等化設計良好，基本上各次的學生表現可以互相比較。

（二）分析 TASA 資料非要使用權重不可嗎？

讀者可能有個疑惑：一般的調查研究沒有估算學生個人權重，分析時也沒有把權重納入考量，在分析大型教育調查資料，非要放入權重才可以嗎？若只要知道描述性統計量，仍需要把權重放入嗎？大型教育調查的目的是要推論母群的整體表現，回答和母群表現相關的研究議題，不是推論到個人，故樣本能代表母群樣貌的程度就非常重要。如本章前面所述，放入權重才能正確估算抽樣誤差，能避免估計量有偏誤（bias），也可避免低估誤差導致檢設檢定時，容易達到統計上的顯著，即便不做假設檢定，但學生權重的意義為這個人代表母群中幾個人，放入權重才能較準確的描述母群的樣貌。以下用一個簡單的例子說明有沒有放權重，對母群描述的影響。

若抽了 9 個樣本，得到這 10 個學生的分數分別為 10、20、30...90，若只用樣本資料畫出的直方圖為圖 4-1 左邊的圖，用這個圖描繪母群，我們會以為母群分數的分佈類似均勻分配，每種分數出現的機率很接近。若 9 個學生的權重分別為 10、20、30、40、50、40、30、20、10，將權重一併考慮進來後畫出的直方圖為圖 1 右邊的圖，可以看出中間分數的學生居多，兩端分數的學生較少。由這個簡單的例子可知，儘管不做假設檢定，只是想利用次數分配推論母群的分佈情形，有無考量權重所得到的結果有些不同，其他例子還可以參見 Lohr（2009）。故強烈建議讀者使用 TASA 2016 資料進行分析時，務必放入權重進行分析。



R 語法

```
install.packages("plotrix")
library(plotrix)
score<-c(10,20,30,40,50,60,70,80,90)
w<-c(10,20,30,40,50,40,30,20,10)
hist(score,breaks = seq(0,100,by=10),xlab="score",ylab="frequency") # 左圖
weighted.hist(score,w,breaks = seq(0,100,by=10),xlab="score",ylab="frequency") # 右圖
```

圖 4-1 有無放入權重的直方圖（左圖沒放權重、右圖有放權重）

（三）抽樣誤差的算法要用哪一種比較好？

TASA 2016 提供 JKZONE、JKREP 兩個變數，這兩個變數提供了學生分層的訊息，有這兩個變數便可運用平衡重複抽樣法、費氏改良平衡重複抽樣法（Fay's BRR）（Dippo, Fay, & Morganstein, 1984; Judkins, 1990）、刀切重複取樣法（jackknife method）、拔靴法…等重複取樣的方式進行抽樣誤差的計算，不同計算抽樣誤差的方式有他的優點和缺點，例如平衡重複抽樣法較刀切重複取樣法運算簡單，但有些抽樣設計情況，例如一層不只包含兩個 PSU 時，恐無法利用平衡重複抽樣法計算（Lohr, 2009），且刀切重複取樣法的誤差有機會變成誤差變異的不偏估計量，也有機會用在更多地方（Rao & Shao, 1992）。而 TASA 2016 一層均為兩個 PSU，基本上使用平衡重複取樣法（BRR）、費氏改良平衡重複抽樣法、刀切重複取樣法、以及拔靴法四種方法估計對比例統計量（ratio estimators）的誤差估計並無顯著差異（Lehtonen & Pahkinen, 2004）。抽樣誤差詳情請見第六章。

（四）若從中擷取某特定樣本進行分析，權重要如何處理？

若有特別需要，需要抽取某部分樣本，則樣本大小不同了，例如本來「國數自社」以 9955 人代表母群，若經過處理後變成以 8000 人代表母群，或者是研究者可能只對新移民子女、原住民等特定族群感興趣時，若只抽出某一群人出來分析，權重並不用特別調整，但要注意的是，因為樣本大小減少，所以抽樣誤差會增加。

參考文獻

- Dippo, C. S., Fay, R. E., & Morganstein, D. H. (1984). *Computing Variances from Complex Samples with Replicate Weights*. Proceedings of the American Statistical Association, Section on Survey Research Methods, 489-494.
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the bootstrap: Monographs on statistics and applied probability*. New York and London: Chapman and Hall/CRC.
- Frankel, M. R. (1971). *Inference from survey samples*. Ann Arbor: Institute of Social Research, University of Michigan.
- Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6 (3), 223-239.
- Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.
- Krewski, D., & Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 1010-1019.
- Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys, statistics in practice* (2nd ed.). New York: John Wiley & Sons.
- Lohr, S. (2009). *Sampling: design and analysis*. Nelson Education.
- OECD. (2016). *PISA 2015 technical report*. Paris: OECD.
- Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33(4), 305-325.
- Rao, J. N., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4), 811-822.
- Wu, M., Tam, H. P., and Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer Nature.

第五章 測驗設計與量尺化程序

吳 慧 珉

國家教育研究院副研究員

一、前 言

TASA 是以課程綱要為主要評量內容，評量的內容包含學科科目和問卷，科目是國語文、英語文、數學、社會和自然科學；問卷包含背景問卷和學科問卷，詳細的評量內容讀者可參閱「第二章、TASA 評量架構」和「第三章、背景問卷的心理計量特性」。為了完整涵蓋評量內容，測驗的題目數量可多達 1 百多題，含括較廣的內容領域，一位學生要完成 1 百多題的測驗，需要很長的考試時間且有疲勞因素之影響，再加上其他的科目，則實務運作會有困難。為了解決這個問題，TASA 是透過特殊的組卷方式，稱為平衡不完全區塊（Balanced Incomplete Block, BIB）之設計，讓一位學生只需要考到一部份試題，另一位學生考到另一部份試題，試題之間有部分重複作為連結之用，最後再將試題全部集合起來分析，即可完整涵蓋評量內容。

TASA 使用的 BIB 組卷方式，讓學生只需要施測部分的試題即可，如此一來將造成極大的測量誤差，不適合推論個別學生的能力，但卻可透過特殊的測驗統計模式，也就是可能值方法（plausible values methodology），有效的推論群體的能力表現。以下將以 2016 年的資料為例，第二部分主要介紹 TASA 的測驗設計，說明 TASA 如何安排不同的考科題本；第三部份是

說明 TASA 的量尺化程序，包含如何刪除不良試題以及如何進行群體能力的估計。

二、測驗設計

TASA 有數種不同形式之測驗題本，每一位學生接受不同的測驗題本，其測驗的分數可以藉由統計方法轉換至共同量尺，使測驗分數能夠比較，這樣的程序就是等化。TASA 目前是採用平衡不完全區塊（BIB）等化設計，學生只需接受若干試題區塊的試題，且不同學生可能接受部分相同、完全相同、或完全不同的試題區塊。其優點在於能進行大量的施測試題（如 TASA2016 的自然科學科高達 156 題），可包含較廣的內容領域，內容領域的部分請參閱「貳、TASA 評量架構」，每位受試者僅接受約 40 題，較不會造成學生之精神負荷。

平衡不完全區塊（BIB）等化設計最早是由 Yates（1936）提出，1992 年 Rust & Johnson 兩位學者將之應用於大型測驗之題庫設計（Rust & Johnson, 1992）。所謂的「平衡」是由於成對試題區塊（block）出現於題本中的次數是相同的，因此在成對試題區塊間之有比較相同的精準度。各題本中的試題區塊可能部分相同或完全不同，但是每一個試題區塊在所有題本中出現的次數是一樣的（Kuehl, 2000；曾玉琳、王暄博、郭伯臣、許天維，2005）。平衡不完全區塊（BIB）會讓題庫中的每個試題所受測的學生約為相同的。

2016 年 TASA 國語文、數學、社會和自然科學使用 BIB 設計，英語文包含聽力測驗與閱讀兩部分，如同一班級，如果部分學生接受接力測驗，部分學生考其他學科，將會形成干擾，施測上有其限制，故在題本排列上較不適用於 BIB 設計，而是藉由各題本間重複出現之不同試題區塊來連結各題本。2016 年 TASA 將每一科試題平分成 13 個試題區塊，分別給予試題區塊

編碼，如國語文是 C01~C13，C01 有 10 題題目，C02 有 10 題題目，以此類推；數學是 78 題，平分成 13 個試題區塊，每一個試題區塊是 6 題，各科的試題數量與區塊編碼如表 5-1 所示。

表 5-1 2016 年 TASA 各科的試題數量與區塊編碼

科 目	題 數	試 題 區 塊 編 碼	每一試題區塊之題目數量
國 語 文	130	C01~C13	10
數 學	78	M01~M13	6
社 會	117	S01~S13	9
自然科學	156	N01~N13	12

TASA 2016 年的國語文、數學、社會和自然科學 BIB 等化設計總共有 78 本題本，每一題本是由四個區塊組合成，每一學生測驗一個題本，如表 5-2。題本 1 是由國語文和數學組成，包含 C01、C02、M01、M02 四個試題區塊，題數是 32 題；題本 2 是由數學和社會組成，包含 M01、M02、S01、S02 四個試題區塊，題數是 30 題；題本 3 是由社會和自然科學組成，包含 S01、S02、N01、N02 四個試題區塊，題數是 42 題。每一個試題區塊在 78 本題本出現的次數是一樣的，如 C01 是出現 6 次，M01 是出現 6 次。分配題本時，同一個班級內，一號學生考題本 1，二號學生考題本 2，以此類推，如此一來，同一班級的學生接受不同的題本，題本內有相同的試題區塊和不同的試題區塊，如 M01 即為一號學生和二號學生相同的試題區塊，如此一來，可藉由相同的試題區塊之題目，將接受不同題本測驗的學生的能力連結至同一量尺。

表 5-2 TASA2016 年國語文、數學、社會和自然科學 BIB 等化設計

題 本	科目組合	區 塊 1	區 塊 2	區 塊 3	區 塊 4
1	國 數	C01	C02	M01	M02
2	數 社	M01	M02	S01	S02
3	社 自	S01	S02	N01	N02
4	自 國	N01	N02	C01	C02
5	國 社	C01	C02	S01	S02
6	數 自	M01	M02	N01	N02
7	國 數	M02	M03	C02	C03
8	數 社	S02	S03	M02	M03
9	社 自	N02	N03	S02	S03
10	自 國	C02	C03	N02	N03
11	國 社	S02	S03	C02	C03
12	數 自	N02	N03	M02	M03
13	國 數	C03	C04	M03	M04
14	數 社	M03	M04	S03	S04
15	社 自	S03	S04	N03	N04
16	自 國	N03	N04	C03	C04
17	國 社	C03	C04	S03	S04
18	數 自	M03	M04	N03	N04
19	國 數	M04	M05	C04	C05
20	數 社	S04	S05	M04	M05

題本	科目組合	區塊 1	區塊 2	區塊 3	區塊 4
21	社 自	N04	N05	S04	S05
22	自 國	C04	C05	N04	N05
23	國 社	S04	S05	C04	C05
24	數 自	N04	N05	M04	M05
25	國 數	C05	C06	M05	M06
26	數 社	M05	M06	S05	S06
27	社 自	S05	S06	N05	N06
28	自 國	N05	N06	C05	C06
29	國 社	C05	C06	S05	S06
30	數 自	M05	M06	N05	N06
31	國 數	M06	M07	C06	C07
32	數 社	S06	S07	M06	M07
33	社 自	N06	N07	S06	S07
34	自 國	C06	C07	N06	N07
35	國 社	S06	S07	C06	C07
36	數 自	N06	N07	M06	M07
37	國 數	C07	C08	M07	M08
38	數 社	M07	M08	S07	S08
39	社 自	S07	S08	N07	N08
40	自 國	N07	N08	C07	C08

表 5-2 TASA2016 年國語文、數學、社會和自然科學 BIB 等化設計 (續)

題本	科目組合	區塊 1	區塊 2	區塊 3	區塊 4
41	國 社	C07	C08	S07	S08
42	數 自	M07	M08	N07	N08
43	國 數	M08	M09	C08	C09
44	數 社	S08	S09	M08	M09
45	社 自	N08	N09	S08	S09
46	自 國	C08	C09	N08	N09
47	國 社	S08	S09	C08	C09
48	數 自	N08	N09	M08	M09
49	國 數	C09	C10	M09	M10
50	數 社	M09	M10	S09	S10
51	社 自	S09	S10	N09	N10
52	自 國	N09	N10	C09	C10
53	國 社	C09	C10	S09	S10
54	數 自	M09	M10	N09	N10
55	國 數	M10	M11	C10	C11
56	數 社	S10	S11	M10	M11
57	社 自	N10	N11	S10	S11
58	自 國	C10	C11	N10	N11
59	國 社	S10	S11	C10	C11

題本	科目組合	區塊 1	區塊 2	區塊 3	區塊 4
60	數 自	N10	N11	M10	M11
61	國 數	C11	C12	M11	M12
62	數 社	M11	M12	S11	S12
63	社 自	S11	S12	N11	N12
64	自 國	N11	N12	C11	C12
65	國 社	C11	C12	S11	S12
66	數 自	M11	M12	N11	N12
67	國 數	M12	M13	C12	C13
68	數 社	S12	S13	M12	M13
69	社 自	N12	N13	S12	S13
70	自 國	C12	C13	N12	N13
71	國 社	S12	S13	C12	C13
72	數 自	N12	N13	M12	M13
73	國 數	C13	C01	M13	M01
74	數 社	M13	M01	S13	S01
75	社 自	S13	S01	N13	N01
76	自 國	N13	N01	C13	C01
77	國 社	C13	C01	S13	S01
78	數 自	M13	M01	N13	N01

TASA 2016 年的英語文總共有 20 個試題區塊 E01~ E20，其中 E01 和 E02 為英語聽力測驗，統一安排於區塊 1，E03~ E20 是英語閱讀，按順序分散於區塊 2~ 區塊 6，英語文總共編排六本題本、各個題本之間藉由相同的試題區塊之題目，如題本 1 和題本 2 藉由 E01 和 E09，將接受不同題本測驗的學生的英語文能力連結至同一量尺。

表 5-3 TASA2016 年英語文等化設計

題 本	區 塊 1	區 塊 2	區 塊 3	區 塊 4	區 塊 5
1	E01	E03	E09	E11	E15
2	E01	E04	E09	E12	E16
3	E01	E05	E09	E11	E17
4	E02	E06	E10	E12	E18
5	E02	E07	E10	E13	E19
6	E02	E08	E10	E14	E20

三、量尺化程序

本節將分兩大部分：資料清理和量尺化程序為讀者進行說明。假設每一位學生作答 40 題，可以收集到學生在這 40 題的對錯資料，即 (1, 1, 1, 0, 1, ..., 1)，此即為作答反應。量尺化程序是指將學生的作答反應資料，透過統計公式轉為至某一量尺分數的過程。如國人比較熟知的學生能力國際評量計劃 (Programme for International Student Assessment, PISA)，主要是評量一個國家 15 歲學生的數學素養、科學素養和閱讀素養。2015 年臺灣在數學素養的平均量尺分數是 542 分、科學素養的平均量尺分數是 532 分、閱讀素養的平均

量尺分數是 497 分，這些分數就是透過量尺化程序而來的。

在進行量尺化程序的過程中，為避免不良的資料（data），例如受試者亂答或多題未答，對於量尺化的結果產生不良的影響，可能導致量尺分數估計不佳，必須對資料進行資料清理（data clean）。資料清理涵蓋受試者部分與試題部分，以下分別就這兩部分說明 TASA2016 資料清理的標準訂定，讀者進行資料清理時，可以參考這些原則，將不良的資料刪除，以使用較乾淨的資料進行量尺化程序，TASA 2016 刪除作答反應之學生人數比例，讀者可參閱第四章抽樣設計與權重。

（一）資料清理

1. 受試者部分

TASA2016 訂定之受試者作答反應刪除標準如下：

- （1）受試者於整份題本中，連續五題試題未作答（含以上）；
- （2）受試者於整份題本中，連續四題（含以上）試題為相同反應，後續四題（含以上）試題為相同的另一種反應，連續部分的題數和為總題數 1/3 以上；
- （3）受試者於整份題本中，連續四題（含以上）試題相同反應，且此種反應題數和為總題數 1/2 以上。

2. 試題部分

TASA2016 的試題難度參數分布圖如圖 5-1，為維持測驗品質，訂定之不良試題標準如下：

- （1）依據古典測驗理論（classical test theory, CTT）訂定標準如下：

試題通過率低於 0.25， $P_{CTT} < 25\%$ ；

試題鑑別度指數小於 0.2，CTT 的鑑別度 < 0.2 。

試題鑑別度指數使用點二系列相關，公式如下：

$$r_j = \frac{(\mu_j - \mu_x)}{\sigma_x} \sqrt{\frac{P_j}{1 - P_j}}$$

其中， r_j 指第 j 題鑑別度；

μ_j 指答對第 j 題之受試者之測驗總分的平均數；

μ_x 指所有受試者之測驗總分的平均數；

σ_x 指所有受試者之測驗總分的標準差。

(2) 依據試題反應理論 (item response theory, IRT) 訂定標準如下：

試題難度參數小於或等於 -3.5， $b \leq -3.5$ ；

試題難度參數大於或等於 3， $3 \leq b$ ；

估計後的試題參數未達收斂。

圖 5-1 是試題參數的分布圖，以供讀者判斷測驗的品質

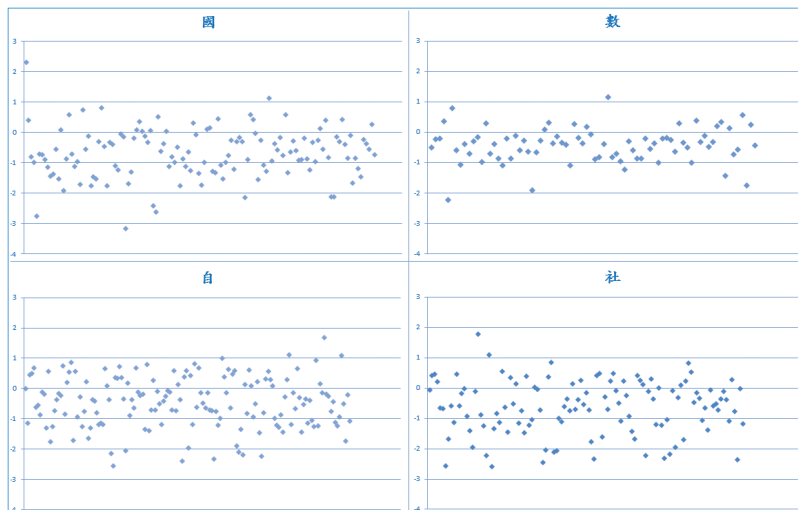


圖 5-1 TASA2016 的試題難度參數分布圖

(二) 量尺化程序

臺灣學生在 2015 年國際學生評量計畫 (The Programme for International Student Assessment, PISA) 測驗中，閱讀素養的成績是 497 分、數學素養成績是 542 分、科學素養是 532 分。這些國家的成績是怎麼來的呢？是透過可能值方法 (plausible value method) 估計而得的。除了 PISA 之外，目前許多國際大型評比調查，如國際數學與科學教育成就趨勢調查 (Trends in International Mathematics and Science Study, TIMSS)、促進國際閱讀素養研究 (Progress in International Reading Literacy Study, PIRLS) 等，都是使用可能值方法估計國家或區域的平均能力 (Lee, Grigg & Dion, 2007; Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992; OECD, 2009; 林陳涌, 2014)。

可能值方法是在估計學生的能力時，除了考慮學生的學科答題反應外，更加入了和學生學習相關的背景變項，如性別、社經背景等，估計每位學生能力值的機率分佈，再從此分佈中隨機抽取學生的能力值，呈現學生「可能合理」的能力值範圍，故國際大型評比釋出資料型態，如 PISA、TIMSS、PIRLS 是以可能值的方式釋出學生的資料，如某一位學生的可能值是 498 分、500 分、503 分、502 分、505 分，代表這五個分數都有可能是學生的分數，經由每一位學生的可能值可以推論國家的整體成績。

有些研究者會將國際評比所釋出的每一位學生的五個可能值，直接平均代表每位學生的分數，或是隨機選一個可能值代表某一位學生的分數，這些都是不適當的作法，會造成較大的估計誤差 (von Davier, Gonzalez, & Mislevy, 2009)，更進一步的訊息可參閱本書第六章。國際評比時，由於每位學生只被施測少量試題，以某一個可能值代表學生的分數，會造成較大的估計誤差，可能值主要是計算群體的統計量數，如國家層級的統計量數 (Adams, Wilson & Wu, 1997)。讀者如需要更詳細的可能值使用方法，可參閱 PISA 2003 Data Analysis Manual，提供詳細的使用範例和 SPSS 程式巨集，可讓資料分析者正確使用可能值。

TASA2016 由於學生只接受某些試題區塊測量，個體能力的測量會伴隨著相當程度的測量誤差，且 TASA 也有背景變項之問卷，使用可能值方法計算可能值，除了能得到群體統計量的良好估計值外，藉由可能值資料的釋出，能提供給次級資料分析者進行學生學習成效相關因素之探討，促進相關教育議題的討論與連結。在可能值方法中，主要是運用學生的學科作答反應資料與背景問卷資料，透過測驗統計模式估計學生的能力分布，以下將以測驗理論模式為基礎，說明 TASA2016 在使用可能值方法時所定義的能力向度評量架構，再說明可能值方法之理論基礎。

1. TASA 2016 能力向度架構

發展測驗的第一個步驟就是先決定測量的目標，也就是此份測驗想要測量的學生能力是什麼？這個學生能力可以是單一向度的，例如數學能力；也可以是多向度的，例如是數與計算、量與實測、幾何等。決定評量目標後，就需決定測驗的題型和計分方式，如選擇題、填充題和簡答題等是常見的測驗題型，針對不同的題型會搭配不同的計分方式，如使用答錯 0 分，答對 1 分，這種計分模式在測驗理論中稱為二元（dichotomous）計分，而簡答題則可以是答錯 0 分、部分答對 1 分、全對 2 分，這種計分模式則被稱為多點（polytomous）計分。

針對不同的測量目標和計分方式可搭配不同的測驗理論，稱為試題反應理論，透過試題反應理論可以估計受試者的能力，如測量的目標是單向度，目前常見的試題反應理論是單向度試題反應理論，如單參數對數模式（one-parameter logistic model, 1PL）、二參數對數模式（two-parameter logistic model, 2PL）及三參數對數模式（three-parameter logistic model, 3PL）；多點計分模式包含部分給分模式（partial credit model, PCM）和廣義部分給分模式（generalized partial credit model, GPCM），此部分有興趣的讀者可以參閱（郭伯臣、吳慧珉、陳俊華，

2012) 的文章。

如果評量的目標是多向度的，又可分為兩種模式：題間多向度測驗（within-item multidimensional test）和題內多向度測驗（within-item multidimensional test）（Adams, Wilson, & Wang, 1997）。題間多向度測驗是指測驗中每題試題只測量單一能力，如圖 5-1 所示，每一題僅測量國語文或是數學。題內多向度測驗是測驗中有些試題測量了一種以上的能力，如圖 5-2 所示，有些題目同時測量數學和自然科學。TASA2016 的設計是使用圖 5-1 題間多向度的架構，英語科獨自分析，亦是採用題間多向度的架構，將英語文分為聽力和閱讀兩個向度的能力。TASA 2016 更細緻的評量架構說明，請讀者參閱「貳、TASA 評量架構」。

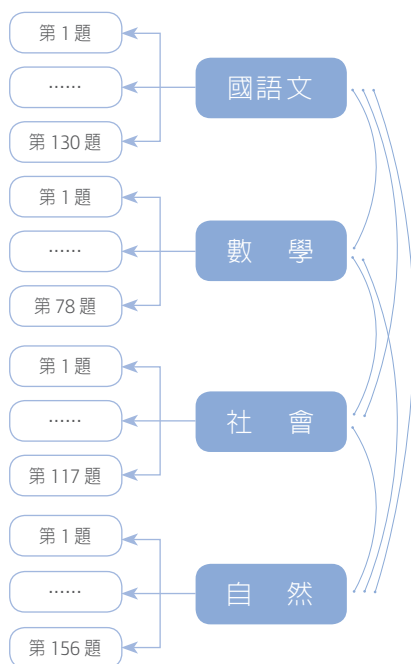


圖 5-2 題間多向度架構

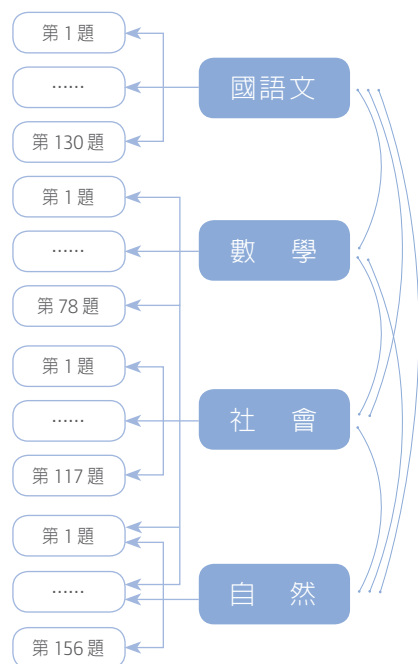


圖 5-3 題內多向度架構

2. 多向度試題反應理論

依照 TASA2016 的評量架構，是屬於能力值為多向度的評量架構，使用多向度的測驗理論模式有一個優勢，就是可以藉由各向度間的相關性來提高能力估計的精確性，且相較於將這些試題拆開成獨立的單向度能力分析，多向度的分析方法能得到較高的信度（Wang, Chen, & Cheng, 2004；余民寧，2009）。目前關於多向度的測驗理論主要有多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM）、多向度二參數模式（multidimensional two parameters model, M2PL）與多向度三參數模式（multidimensional three parameters model, M3PL）三種模式。TASA2016 使用可能值方法進行量尺化程序，是使用多向度隨機係數多項洛基模式（MRCMLM）再結合背景變項進行分析，因此以下將說明 MRCMLM，其他的模式有興趣的讀者可以參閱（郭伯臣、吳慧珉、陳俊華，2012）的文章。

Adams、Wilson 與 Wang（1997）等學者提出多向度隨機係數多項洛基模式（MRCMLM），是單向度試題反應理論所推廣而成的多向度試題反應理論模式，如公式（5-1）所示（郭伯臣、吳慧珉、陳俊華，2012）。

$$P(\mathbf{X}_{jk} = 1; \mathbf{A}, \mathbf{B}, \hat{\boldsymbol{\theta}} | \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}'_{jk} \boldsymbol{\theta} + \mathbf{a}'_{jk} \hat{\boldsymbol{\theta}})}{\sum_{k=1}^{K_j} \exp(\mathbf{b}'_{jk} \boldsymbol{\theta} + \mathbf{a}'_{jk} \hat{\boldsymbol{\theta}})} \quad (5-1)$$

$\mathbf{X}_{jk} = (X_{j1}, X_{j2}, \dots, X_{jk_j})'$ ， $(k = 0, 1, \dots, K_j + 1)$ ：受試者反應類別，如 TASA2016 是二元計分，0 分和 1 分，則有二個類別，0 分是第 1 個類別，1 分是第 2 個類別。

$$X_{jk} = \begin{cases} 1 & \text{表第 } j \text{ 題作答第 } k \text{ 個反應類別} \\ 0 & \text{表其他} \end{cases}$$

$\xi' = (\xi_1, \xi_2, \dots, \xi_p)$ ：試題參數向量（ p 個參數），如二個計分類別，則 $p=1$ 。

$\theta' = (\theta_1, \theta_2, \dots, \theta_D)$ ：受試者的能力向量（ D 個向度），如 TASA2016 的評量架構，則有 4 個向度，分別是國語文、數學、自然科學、社會。

$A' = (a_1, a_{12}, \dots, a_{1K_1}, a_{12}, a_2, \dots, a_{2K_2}, \dots, a_{nK_n})$ ：整份測驗的設計矩陣，控制試題所對應的反應類別。

a_{jk} ($j=1, \dots, n$ and $k=1, \dots, K_j$)：第 j 題中第 k 個反應類別的設計向量，每個向量長度為 p 。

$B = (B'_1, B'_2, \dots, B'_n)'$ ：整份測驗的計分矩陣。

$B_j = (b_{j1}, b_{j2}, \dots, b_{jD})'$ ：第 j 題的計分子矩陣。

$b_{jk} = (b_{jk1}, b_{jk2}, \dots, b_{jkD})'$ ：在 D 個向度中，第 j 題回答第 k 個反應類別的計分向量。

3. 可能值方法理論

在試題反應理論中，受試者的能力向量是我們無法直接觀察到的，有時會被稱為潛在特質，受試者的能力向量必須透過估計而得，也就是受試者的能力的測量含有不確定性，在計算群體統計量和相關連的標準誤時，應考量這些不確定性（Allen, Donoghue, & Schoeps, 2001; Mullis, Martin, & Foy, 2008; OECD, 2009）。可能值方法是從估計出來的受試者的能力分布隨機抽取學生的可能的分數，就是考量受試者能力測量的不確定性。可能值方法是透過迴歸模式（regression model），加入學生背景變項，估計每一位學生的能力分布，再抽取可能值，以利於次級資料分析者使用。事實上透過可能值方法中的迴歸模式就可以直接估計

群體的參數，如群體能力的平均數，如此可以使參數的估計更為精準（Mislevy & Sheehan, 1989）。

在可能值方法理論中主要包含兩個重要的模式：一個為條件式的試題反應模式 $f_x(X; \xi | \theta)$ ，一個是母群體分佈模式 $f_\theta(\theta; \alpha)$ ，其中 X 表示為在每一題上表現的觀察值， ξ 表示各題的試題參數， θ 代表能力值， α 為表示母群體特徵的參數。假設學生的作答反應矩陣為隨機的從母群體分佈抽取，其分佈如下（郭伯臣、曾建銘、吳慧珉，2012）：

$$f_x(X; \xi, \alpha) = \int_{\theta} f_x(X; \xi | \theta) f_{\theta}(\theta; \alpha) d\theta \quad (5-2)$$

其概似函數為

$$\Lambda = \prod_{n=1}^N f_x(X_n; \xi, \alpha) \quad (5-3)$$

其中 N 為全部所抽取的學生總數。

若 μ 為母群體分佈的平均數，則 f_{θ} 可以表示為平均數加上誤差變數

$$\theta_n = \mu + E_n \quad (5-4)$$

學者使用迴歸模式 $Y_n' \beta$ 取代公式 (5-4) 平均數 μ ，其中 Y_n 為 u 的矩陣，為學生 n 固定且已知的值（Adams et al., 1997），例如 Y_n 可以被視為學生的性別、社經地位或者所住的區域，也就是背景變項，而 β 為迴歸係數。母群體的模式公式 (5-4) 可以被更改如下：

$$\theta_n = Y_n' \beta + E_n \quad (5-5)$$

假設 E_n 為一標準常態分配，能力值是單向度的情況下，則公式 (5-4) 等同於

$$f_n(\theta_n; Y_n, \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\theta_n - Y_n'\beta)'(\theta_n - Y_n'\beta)\right] \quad (5-6)$$

如果是多向度的情況，模式如下：

$$f_\theta(\theta_n; \omega_n, \gamma, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\theta_n - \gamma\omega_n)^T \Sigma^{-1}(\theta_n - \gamma\omega_n)\right] \quad (5-7)$$

其中， γ 是一個 $u \times d$ 的迴歸係數矩陣， Σ 是一個 $d \times d$ 的， ω_n ： $u \times 1$ 的背景變數，如上述的學生性別、社經地位等，在考慮能力值是多向度的情況下，公式 (5-8) 可以修改為結合試題反應模式與多向度母群體模式的公式：

$$f_x(X; \xi, \gamma, \Sigma) = \int_{\theta} f_x(X; \xi | \theta) f_\theta(\theta; \gamma, \Sigma) d\theta \quad (5-8)$$

公式 (5-8) 需要估計的參數是試題參數 ξ 、迴歸係數 γ 和變異數共變數矩陣 Σ 。每一位受試者的能力值之分布如下，可能值就是從這個分布抽取出來的。

$$\begin{aligned} h_\theta(\theta_n; \xi, \gamma, \Sigma | x_n) &= \frac{f_n(x_n; \xi | \theta_n) f_\theta(\theta_n; \gamma, \Sigma)}{f_x(x_n; \xi, \gamma, \Sigma)} \\ &= \frac{f_n(x_n; \xi | \theta_n) f_\theta(\theta_n; \gamma, \Sigma)}{\int_{\theta} f_n(x_n; \xi | \theta_n) f_\theta(\theta_n; \gamma, \Sigma)} \quad (5-9) \end{aligned}$$

4. 背景變數的設定

TASA2016 除了學科問卷，亦包含背景變項問卷，詳細的背景變項

問卷訊息請參閱「參、背景問卷的心理劑量特性」。各國際大型評比在可能值方法中所使用的背景變項設定並不相同，如 PISA 是將一些背景變數，如性別、母親職業、父親職業等視為主要背景變數，其餘的背景變數則透過主成分分析，解釋原始資料 95% 變異的條件下，計算每一位學生的主成分分數，納入公式 (5-7) 的背景變數中。TIMSS 則是在一些主要的背景變數之外，如性別，將其餘的背景變數取能解釋原始資料 90% 變異的條件下，計算每一位學生的主成分分數，納入公式 (5-7) 的背景變數中（郭伯臣、曾建銘、吳慧珉主編，2012）。主成分分析的主要目的是要縮減背景變數的維度，TASA2016 背景變數的設定如下：

步驟一：將學生問卷中的所有的背景變數虛擬編碼，編碼的方式請參閱附錄。

步驟二：使用主成分分析分析虛擬編碼的變數，並且計算每一位學生的主成分分數，主成份的數量必須要能解釋原始資料 90% 的變異。

步驟三：使用試題參數和經由主成分分析得到的條件變數估計群體參數分佈。

步驟四：使用上述的方法抽取十個可能值向量。

5. 可能值的抽取步驟

可能值是代表學生最有可能的能力值的值，TASA2016 是使用 R package TAM (Robitzsch, Kiefer, Wu M., 2017)，理論基礎是來自於 PISA 所使用的可能值的抽取步驟 (OECD, 2009)：

假設試題參數已知。

對於每一位學生，從能力值的邊際後驗機率 (5-9) 隨機抽取可能值。

對於每一個受試者 n ， M vector-valued random deviates, $\{\varphi_{mn}\}_{m=1}^M$ ，從多變量常態分佈， $f_{\theta}(\theta_n; \mathbf{W}_n, \gamma, \Sigma)$ 。使用蒙地卡羅積分法逼近式子 (5-9) 的分母。

$$\int_{\theta} f_x(x; \xi | \theta) f_{\theta}(x, \gamma, \Sigma) d\theta \approx \frac{1}{M} \sum_{m=1}^M f_x(x; \xi | \varphi_m) \equiv \mathfrak{Z} \quad (5-10)$$

同時，計算

$$P_{mn} = f_x(X_n; \xi | \varphi_{mn}) f_{\theta}(\varphi_{mn}; \mathbf{W}_n, \gamma, \Sigma) \quad (5-11)$$

$\{\varphi_{mn}, P_{mn} / \mathfrak{Z}\}_{m=1}^M$ 的集合可視為式子 (5-9) 的後驗機率函數之近似；且機率值 φ_{nj} 可藉由以下公式求得：

$$q_{nj} = \frac{P_{mn}}{\sum_{m=1}^M P_{mn}} \quad (5-12)$$

隨機產生 L 個服從均勻分佈的值 $\{\eta_i\}_{i=1}^L$ ；對於每一次隨機抽取，若 φ_{ni_0} 滿足下列條件則選取當作一可能值向量 (plausible vector)：

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i < \sum_{s=1}^{i_0} q_{sn} \quad (5-13)$$

TASA2016 的學生檔案中包含 70 個可能值，前 10 個數字是國語文的可能值，接著是數學、自然科學和社會。在英語文的部分，則是用多向度的英語文聽力和英語文閱讀之可能值，以及英語文之可能值。

表 5-4 TASA2016 釋出的學生檔案之可能值和其代碼說明

代 碼	說 明
PV1.C~PV10.C	第 1~10 個國語文可能值
PV1.M~PV10.M	第 1~10 個數學可能值
PV1.N~PV10.N	第 1~10 個自然科學可能值
PV1.S~PV10.S	第 1~10 個社會可能值
PV1.L~PV10.L	第 1~10 個英語文聽力可能值
PV1.R~PV10.R	第 1~10 個英語文閱讀可能值
PV1.E~PV10.E	第 1~10 個英語文可能值

6. 群體參數變異數的計算

知道群體參數估計之變異數，就可以知道估計之標準誤，也就是可以推估估計之參數的正確範圍。可能值可以提供群體參數的一致性估計，是群體能力參數的最佳估計值，如要估計群體能力參數 A ，方法如下：

步驟一：使用第一組可能值向量計算 A ，記為 A_1 。

步驟二：計算 A_1 的抽樣變異，記為 VAR_1 ，抽樣變異的計算公式請參閱第四章。

步驟三：重複步驟一和步驟二，分別計算第二組至第十組的 A_t 和 VAR_t ， $t = 2 \sim 10$ 。

A 的最佳估計值是將這十組所計算的值平均而得，也就是

$$\hat{A} = \frac{\sum_{t=1}^{10} A_t}{10} \circ$$

\hat{A} 的變異數估計包含抽樣變異 $\bar{S} = \frac{\sum_{t=1}^{10} VAR_t}{P}$ 和測量變異

$$B_M = \frac{\sum_{t=1}^{10} (A_t - \hat{A})^2}{P-1}, \text{ 加總而得 } Var(\hat{A}) = \bar{S} + (1 + P^{-1})B_M, \text{ P 是抽取的可能}$$

值之組數，TASA2016 中，P 是 10。

參考文獻

- 余民寧 (2009)。試題反應理論 (IRT) 及其應用。台北：心理出版社。
- 林陳涌主編 (2014)。國際數學與科學教育成就趨勢調查國家報告。台北市：國立臺灣師範大學科學教育中心。
- 郭伯臣、吳慧珉、陳俊華 (2012)。試題反應理論在教育測驗上之應用。新竹縣教育研究集刊。民 101，第十二期，頁 05 ~ 40。
- 郭伯臣、曾建銘、吳慧珉主編 (2012)。大型標準化測驗建置流程應用於 TASA 之研究。新北市：國家教育研究院。
- 曾玉琳、王暄博、郭伯臣、許天維 (2005)。不同 BIB 設計對測驗等化的影響。測驗統計年刊，第十三輯下期，209-229。臺中市：國立臺中教育大學。
- Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47-76.
- Allen, N. A., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report (NCES 2001-452)*. Washington DC: United States Department of Education, Institute of Education Sciences, Department of Education, Office for Educational Research and Improvement.
- Lee, J., Grigg, W., & Dion, G. (2007). *The Nation's Report Card: Mathematics 2007*. National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education, Washington, D. C.
- Kuehl, R. O. (2000). *Design of Experiments: Statistical Principles of Research Design and Analysis*. CA: Duxbury Press.

- Mislevy, R. J. (1991). Randomization-based inference about latent variable from complex samples. *Psychometrika*, 56(2), 177-196.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics form sparse matrix samples of item response. *Journal of Educational Measurement*, 29, 133-161.
- Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- OECD (2009). PISA 2003 Technical Report. OCED, Paris.
- Rust, K. F., & Johnson, E.G. (1992). Sampling and weighting in the national assessment. *Journal of Educational Measurement*, 17, 111-129.
- Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *The Journal of Agricultural Science*, 26(3), 424-455.
- Robitzsch A., Kiefer T., Wu M.(2017). TAM: Test Analysis Modules . <https://cran.r-project.org/web/packages/TAM/index.html>
- Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116-136.

第六章 TASA 資料庫的二次分析

陳冠銘

國家教育研究院助理研究員

一、緒論

多數的調查研究為抽樣調查而非普測，因此得藉由樣本「統計量」(statistic)來推估「母群參數」(parameter)。推估過程必然產生誤差，影響該誤差大小的其中一個因素為該調查研究抽取樣本的過程，亦即抽樣設計。TASA、TIMSS、PIRLS 和 PISA 等大型教育成就調查的抽樣均非學生層級的「簡單隨機抽樣」(simple random selection, SRS)，而是採用「二階段分層叢集抽樣」(two-stage stratified cluster sampling, TSCS)設計。因此，針對這些大型教育成就調查資料庫的分析，須特別考量二階段分層叢集抽樣設計所產生的誤差效應。本章以 TASA 2016 的國中資料庫為例，介紹大型教育成就調查二階段分層叢集抽樣設計下的資料分析和誤差估計，包括權重使用、統計量、抽樣誤差(sampling error)、測量誤差(measurement error)、整體誤差估計的概念和計算，最後輔以 R 程式語法實際用於 TASA 2016 資料庫計算範例。在進入本章所介紹的資料分析方法前，讀者可先了解大型教育成就調查資料庫的特性，建議閱讀本書第四章以了解 TASA 的抽樣設計，及第五章介紹 TASA 學生成就量尺化的程序。

（一） 權重使用

一般調查抽樣設計，原則上會希望符合母群定義的每位學生被抽取為樣本的機率應該相同。按本書第四章之「系統機率比例」(systematic probability proportional to size, PPS)抽樣原則，在「分層變項」(stratification)下某「分層」(strata)內的「樣本人數」(sample size)，應正比於該分層內人數對母群人數的比例，而某校被抽取為樣本的機率應正比於該校學生人數對母群人數的比例。先抽出學校樣本，再從每所樣本學校中以等機率隨機抽取相同數量的學生樣本。當抽樣程序符合上述原則時，每位樣本學生應具有相同權重。然而，大型教育成就調查常因下列情形導致樣本間的權重不同：

1. 超額抽樣 (over sampling) 或不足額抽樣 (under sampling)

大型調查有時因特定的研究考量，例如某些子群（如：原住民族、新住民等）的人數偏少但研究者又想了解該特定子群的表現時，若僅以該子群相對於母群之人數比例做為其抽樣機率，便會導致該子群的樣本人數偏少，所得之統計量誤差偏高，而減少其參考價值。因此，可針對這些特定子群採取超額抽樣，期使提高該子群統計量估計的精確性。相反地，當子群人數夠多時，亦可在不影響統計量誤差的考量下，採取不足額抽樣。但不論超額抽樣或不足額抽樣，當用以推估母群參數時，因應抽樣機率的改變，都必須適當調整權重，以得到較正確的誤差估計。

2. 樣本流失導致權重校正

多數大型教育成就調查的抽樣方式，如 TASA 2016 會先抽學校樣本，再由樣本學校抽取學生樣本進行調查施測。若原本抽取的樣本學校因故必須使用替代學校，則學校層級的權重便需調整以校正。同樣地，若某樣本學校內有一名或數名樣本學生因故無法參與調查，便會抽取同校其他替代學生來涵蓋流失樣本學生的表現，此時亦須對學生層級權重進行調整校正。

上述兩種情形常導致大型教育成就調查中，每位學生的權重有所不同。圖 6-1 為 TASA 2016 所有受測學生的 HOUWGT 權重分布，全距從最小值 0.14 到最大值 2.95。分析大型教育成就調查資料時，若未正確使用權重，便會導致統計量估計的偏誤。以 TASA 2016 八年級學生數學平均成就為例，採用正確權重的平均量尺分數應為 500 分，若未考慮權重，得出的數學平均成就則為 507 分。

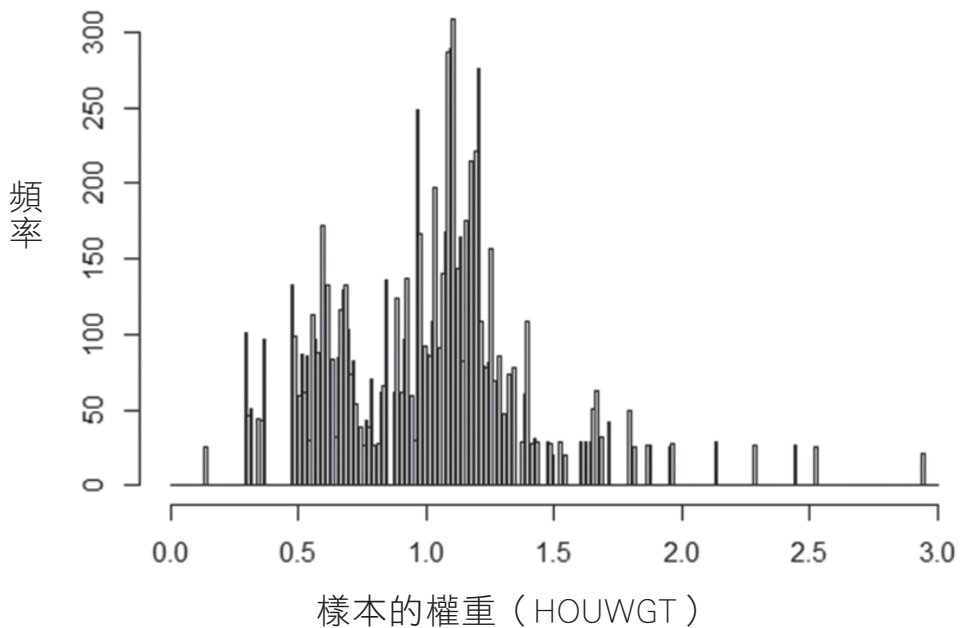


圖 6-1 TASA 2016 資料庫樣本權重的分布

(二) 抽樣誤差的估計

進行統計檢驗時，研究者需要由樣本計算統計量及其誤差。當同一母群抽取不同樣本時，每次計算的統計量皆會有差異，此即抽樣誤差。抽樣誤差的估計，可以經由相同的抽樣程序，重複多次從母群中抽取固定大小的樣本人數，並在每次抽樣後計算其統計量，最後求其重複抽樣所得到統計量分布的標準差，即為該統計量推估母群參數之抽樣誤差。

一般統計套裝軟體如 SAS、SPSS，預設所分析的樣本是從母群中以簡單隨機抽樣抽取。然而，大型教育成就調查多採用兩階段分層叢集抽樣設計，其抽樣誤差通常需特別計算。以平均數為例，以樣本平均數 (\bar{x}) 推估母群平均數時，其抽樣誤差 ($SE(\bar{x})$) 為

$$SE(\bar{x}) \approx \sqrt{D_{eff} \frac{\sigma^2}{n}}, \quad (1)$$

其中 D_{eff} 為「抽樣設計效應」(design effect)， σ 為標準差， n 為樣本大小。當抽樣設計採簡單隨機抽樣時，抽樣設計效應為 1。若為叢集抽樣，抽樣設計效應則為 (Hansen, Hurwitz, & Madow, 1953; Kish, 1965)

$$D_{eff} = 1 + \rho(m - 1), \quad (2)$$

其中 m 為叢集內平均樣本大小， ρ 為「叢集內相關」(intra-cluster correlation)。若為二階段分層叢集抽樣設計，抽樣設計效應則為 (任宗浩、譚克平和張立民，2011)

$$D_{eff} = [1 + \rho(m - 1)] - \rho\phi m, \quad (3)$$

其中 φ 為叢集分層變項解釋「叢集間變異量」(between-cluster variance) 的比例。

複雜抽樣設計其它統計量抽樣誤差之估計，因資訊科技及計算機演算速度的進步，常採用以樣本分布推估母群分布之「重複抽樣」(sampling replication) 技術。常見的重複抽樣技術包括「拔靴法」(bootstrap method) (Efron & Tibshirani, 1993)、「刀切法」(jackknife method) (Frankel, 1971)、「平衡重複抽樣法」(Balanced Repeated Replication, BRR) (Plackett & Burman, 1946)，及「費氏平衡重複抽樣法」(Fay's BRR) (Dippo, Fay, & Morganstein, 1984; Judkins, 1990)。

1. 拔靴法

拔靴法模擬由母群反覆抽取相同樣本人數的過程，利用原樣本進行抽樣、每次抽出並放回至抽出相同樣本人數後，重複數次形成重複抽樣樣本統計量之分布，並以該重複樣本分布作為母群分布之最佳估計 (Efron & Tibshirani, 1993)。母群通常人數很多，所以不會因為抽取部分樣本人數而改變其分布；然而單一樣本卻人數有限，因此從中抽取任何一個人，剩下其他人表現的分布可能會和未抽取前的分布有很大不同。為能更準確地模擬母群反覆抽樣過程又不致改變其分布，採用拔靴法重複抽樣時，必須將前次抽取的樣本放回，再抽取下一個樣本，重複此步驟直到抽出與原樣本人數相同之重複樣本，便為 1 次重複抽樣。然後，計算這些重複樣本的統計量（可為平均數、相關係數等）後，由統計量分布的變異數，作為原樣本統計量推估母群參數的抽樣誤差變異量 ($V_{bootstrap}$)。其計算公式為

$$V_{bootstrap} = \frac{1}{B-1} \sum_{b=1}^B (\theta_b - \hat{\theta})^2, \quad (4)$$

其中 $\hat{\theta}$ 為使用原樣本得出的統計量， θ_b 為第 b 次重複樣本計算之統計量， B 為重複抽樣總次數。

圖 6-2 範例採用 R 的 bootstrap 模組，先於平均數為 0、標準差為 1、呈常態分配的母群中，以簡單隨機方式抽樣一組樣本人數為 100 的樣本 X 後，再用拔靴法對樣本 X 重複抽樣 1000 次，並據以計算利用樣本 X 平均數推估母群平均數的抽樣誤差。與公式 (1) 計算得到的抽樣誤差比較後，顯示拔靴法與公式推導結果非常接近。此例雖為簡單隨機抽樣設計，但當複雜抽樣設計配合拔靴法以估計統計量的抽樣誤差時，亦不失其便利、直覺和正確性 (Chen, Jen, & Wu, 2014)。

```
R 語法範例1
1. library(bootstrap) #呼叫bootstrap語法資料庫
2. x <- rnorm(100, 0, 1) #從平均值為0標準差為1的常態分布母群中抽出一組大小為100的樣本x
3. Boot <- bootstrap(x,1000,mean) #利用拔靴法利用樣本x進行重複抽樣方式產生1000組反覆抽樣樣本，每組樣本大小都是100，計算每一組反覆抽樣樣本的平均值，將結果存成Boot物件
4. SE_boot <- sd(Boot$thetastar) #計算1000組反覆抽樣樣本的平均值之標準差（模擬從母群抽樣推估平均值的標準誤）
5. SE_boot #寫出用拔靴法估計的標準誤
6. sd(x)/(100^0.5) #寫出用公式計算簡單隨機抽樣的標準誤  $\frac{\sigma}{\sqrt{n}}$ 
7. Hist(Boot$thetastar, breaks = seq(-0.5,0.8,by = 0.01)) #繪出1000組重複抽樣樣本平均值的分布
```

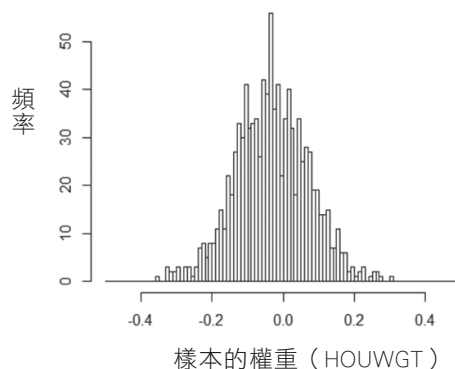


圖 6-2 以拔靴法估計樣本平均數之抽樣誤差

2. 刀切法

Quenouille (1949) 最早將「刀切重複抽樣法」(jackknife replicate method, 或簡稱刀切法) 用於估計「叢集抽樣」(cluster sampling) 設計之抽樣誤差, Frankel (1971) 則用此法估計二階層分層叢集抽樣設計之抽樣誤差。TASA 2016 的「第一階抽樣單位」(primary sampling unit, PSU) 為學校, 當使用刀切法以估計抽樣誤差時, 會先根據學校原始分層變項及學校大小加以排序, 使最相似的學校兩兩配對為同一「刀切抽樣區」(jackknife sampling zone) (如表 6-1)。然後, 每次針對一個刀切抽樣區, 隨機抽取其中一所學校將權重設為 2, 另一所權重則為 0, 藉以模擬重新抽樣時抽到的兩所學校皆近似原樣本學校的其中一所。每次重複抽樣後所得之統計量與原樣本統計量差值的平方, 即為該刀切抽樣區的抽樣變異。刀切法雖然每次只針對一個刀切抽樣區進行重複抽樣, 但實際從母群的每個學校分層重抽一組學校樣本時, 每個分層的樣本學校都會與原本的樣本學校不同, 所以抽樣誤差變異量 ($V_{jackknife}$) 必須將每次刀切重複抽樣所造成的抽樣變異加總:

$$V_{jackknife} = \sum_{j=1}^J (\theta_j - \hat{\theta})^2, \quad (5)$$

其中 $\hat{\theta}$ 為使用原樣本計算所得之統計量, θ_j 為利用第 j 次重複抽樣樣本計算之統計量, J 為刀切抽樣區的數目 (在此亦等於重複抽樣次數)。

表 6-1 二階段分層叢集抽樣的刀切重複抽樣法加權值

學校分層	學校大小	學校	刀切法抽樣區	重 複 抽 樣 次 數 與 權 重									
				第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
A	大 ↓ 小	1	1	0	1	1	1	1	1	1	1	1	1
		2		2	1	1	1	1	1	1	1	1	
		3	2	1	2	1	1	1	1	1	1	1	1
		4		1	0	1	1	1	1	1	1	1	
		5	3	1	1	2	1	1	1	1	1	1	1
	6	1		1	0	1	1	1	1	1	1		
	小 ↓ 大	7	4	1	1	1	0	1	1	1	1	1	1
		8		1	1	1	2	1	1	1	1	1	
		9	5	1	1	1	1	2	1	1	1	1	
		10		1	1	1	1	0	1	1	1	1	
B		大 ↓ 小	11	6	1	1	1	1	1	0	1	1	1
	12		1		1	1	1	1	2	1	1	1	
	13		7	1	1	1	1	1	1	0	1	1	
	14			1	1	1	1	1	1	2	1	1	
	15		8	1	1	1	1	1	1	1	2	1	
	16	1		1	1	1	1	1	1	0	1		
	小 ↓ 大	17	9	1	1	1	1	1	1	1	1	2	1
		18		1	1	1	1	1	1	1	1	0	1
		19	10	1	1	1	1	1	1	1	1	1	0
		20		1	1	1	1	1	1	1	1	1	2

3. 平衡重複抽樣法 (BRR) 和費氏平衡重複抽樣法 (Fay's BRR)

「平衡重複抽樣法」(BRR) 又稱為「對半抽樣法」(half-sampling method) (表 6-2)。上述刀切法每次只針對一個刀切抽樣區進行重複抽樣，並以兩所原樣本學校中的其中一所替代另一所。平衡重複抽樣法則是每次重複抽樣時，對所有的重複抽樣分層區（概念近似於刀切法的刀切抽樣區）同時進行權重設定為 2、0 的重複抽樣。平衡重複抽樣法每次重複抽樣類似從母群重新抽取學校的過程，進行多次重覆抽樣後，便可計算這些重複抽樣的樣本統計量，並以其分布的變異數作為原樣本統計量推估母群參數的抽樣誤差變異。若為窮盡每一種重複抽樣情況，理論上應重複抽樣 2^H 次（H 為重覆抽樣分層區的總數）。然而，平衡重複抽樣法為精簡重複抽樣次數，採用正交 (orthogonal) 方式對各重複抽樣分層區內的兩所學校進行權重設定 (Plackett & Burman, 1946)。以表 6-2 為例，當任選兩次重複抽樣結果，有一半的重複抽樣分層區（10 區中的 5 區）所抽取的學校是相同的，另一半則是抽到不同的學校。

表 6-2 二階段分層叢集抽樣的 BRR 之重複抽樣加權值

學校分層	學校大小	學校	刀切法抽樣區	重 複 抽 樣 次 數 與 權 重											
				第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次	第 11 次	第 12 次
A	大 ↓ 小	1	1	2	0	0	2	0	0	0	2	2	2	0	2
		2		0	2	2	0	2	2	2	0	0	0	2	0
		3	2	2	2	0	0	2	0	0	0	2	2	2	0
		4		0	0	2	2	0	2	2	2	0	0	0	2
		5	3	2	0	2	0	0	2	0	0	0	2	2	2
	6	0		2	0	2	2	0	2	2	2	0	0	0	
	小 ↓ 大	7	4	2	2	0	2	0	0	2	0	0	0	2	2
		8		0	0	2	0	2	2	0	2	2	2	0	0
		9	5	2	2	2	0	2	0	0	2	0	0	0	2
		10		0	0	0	2	0	2	2	0	2	2	2	0
B		大 ↓ 小	11	6	2	2	2	2	0	2	0	0	2	0	0
	12		0		0	0	0	2	0	2	2	0	2	2	2
	13		7	2	0	2	2	2	0	2	0	0	2	0	0
	14			0	2	0	0	0	2	0	2	2	0	2	2
	15		8	2	0	0	2	2	2	0	2	0	0	2	0
	16	0		2	2	0	0	0	2	0	2	2	0	2	
	小 ↓ 大	17	9	2	0	0	0	2	2	2	0	2	0	0	2
		18		0	2	2	2	0	0	0	2	0	2	2	0
		19	10	2	2	0	0	0	2	2	2	0	2	0	0
		20		0	0	2	2	2	0	0	0	2	0	2	2

Fay's BRR (Dippo, Fay, & Morganstein, 1984; Judkins, 1990) 則修改 BRR 進行同一重複抽樣分層區兩個 PSUs 的重複抽樣時，應根據取代率（介於 0 到 1 之間）而非完全取代的方式設定權重。表 6-3 範例以 70% 的取代率為例，將其中一所學校權重設為 1.7，另一所權重則為 0.3，費氏係數 $\varepsilon = 1 - 0.7 = 0.3$ 。BRR 或 Fay's BRR 重複抽樣方式估計統計量 ($\hat{\theta}$) 之抽樣誤差變異量計算方式為

$$V_{Fay's_BRR} = \frac{1}{G(1-\varepsilon)^2} \sum_{g=1}^G (\theta_g - \hat{\theta})^2 \quad (6)$$

其中， $\hat{\theta}$ 為原樣本所得之統計量， θ_g 為第 g 次重複抽樣樣本所得之統計量， G 為重複抽樣次數， ε 為費氏係數。當 ε 等於 0（取代率為 1）時，公式（6）即為 BRR 計算抽樣誤差變異量之公式。

表 6-3 二階段分層叢集抽樣的 Fay's BRR ($\epsilon = 0.3$) 之重複抽樣加權值

學校分層	學校大小	學校	刀切法抽樣區	重複抽樣次數與權重											
				第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次	第 11 次	第 12 次
A	大 ↓ 小	1	1	1.7	0.3	0.3	1.7	0.3	0.3	0.3	1.7	1.7	1.7	0.3	1.7
		2		0.3	1.7	1.7	0.3	1.7	1.7	1.7	0.3	0.3	0.3	1.7	0.3
		3	2	1.7	1.7	0.3	0.3	1.7	0.3	0.3	0.3	1.7	1.7	1.7	0.3
		4		0.3	0.3	1.7	1.7	0.3	1.7	1.7	1.7	0.3	0.3	0.3	1.7
		5	3	1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3	0.3	1.7	1.7	1.7
	6	0.3		1.7	0.3	1.7	1.7	0.3	1.7	1.7	1.7	0.3	0.3	0.3	
	小 ↓ 大	7	4	1.7	1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3	0.3	1.7	1.7
		8		0.3	0.3	1.7	0.3	1.7	1.7	0.3	1.7	1.7	1.7	0.3	0.3
		9	5	1.7	1.7	1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3	0.3	1.7
		10		0.3	0.3	0.3	1.7	0.3	1.7	1.7	0.3	1.7	1.7	1.7	0.3
B		大 ↓ 小	11	6	1.7	1.7	1.7	1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3
	12		0.3		0.3	0.3	0.3	1.7	0.3	1.7	1.7	0.3	1.7	1.7	
	13		7	1.7	0.3	1.7	1.7	1.7	0.3	1.7	0.3	0.3	1.7	0.3	0.3
	14			0.3	1.7	0.3	0.3	0.3	1.7	0.3	1.7	1.7	0.3	1.7	1.7
	15		8	1.7	0.3	0.3	1.7	1.7	1.7	0.3	1.7	0.3	0.3	1.7	0.3
	16	0.3		1.7	1.7	0.3	0.3	0.3	1.7	0.3	1.7	1.7	0.3	1.7	
	小 ↓ 大	17	9	1.7	0.3	0.3	0.3	1.7	1.7	1.7	0.3	1.7	0.3	0.3	1.7
		18		0.3	1.7	1.7	1.7	0.3	0.3	0.3	1.7	0.3	1.7	1.7	0.3
		19	10	1.7	1.7	0.3	0.3	0.3	1.7	1.7	1.7	0.3	1.7	0.3	0.3
		20		0.3	0.3	1.7	1.7	1.7	0.3	0.3	0.3	1.7	0.3	1.7	1.7

4. 重覆抽樣法的抽樣誤差變異量

上述的幾種重複抽樣方法，都可用來估計因抽樣過程所導致的統計量誤差。綜合公式（4）、（5）、（6），可將抽樣誤差變異量（U）之計算彙整為通式

$$U(\hat{\theta}) = c \sum_{r=1}^R (\theta_r - \hat{\theta})^2 \quad (7)$$

其中 $\hat{\theta}$ 為原樣本所計算之統計量， θ_r 為利用第 r 次重複抽樣樣本所得之統計量， R 為重複抽樣次數。當採用不同的重覆抽樣法時， c 為不同的常數。

1. 拔靴法時， $c = 1 / (R - 1)$ ；
2. 刀切法時， $c = 1$ ；
3. BRR 時， $c = 1/R$ ；
4. Fay's BRR 時， $c = 1/[R \cdot (1 - \varepsilon)^2]$ ，其中 ε 為費氏係數。

（三）測量誤差的估計

古典測驗理論常經由檢視測驗工具的效度和信度，認為在可接受的效度和信度水準下，以該工具測量學生在某變項的表現即是穩定的，進而忽略測量誤差存在的事實。然而，忽略測量誤差於變項的效應，會低估變項之間的相關，降低「統計檢定力」（statistical power）（Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007）。為避免因測量誤差影響統計檢定力，大型教育成就調查其一特色便採用「可能值」（plausible values）（Mislevy, 1991）還原測量誤差，取代傳統以單一學生對應單一成就分數的方式。

古典測驗理論「真分數模型」（true score model）認為學生每次的觀測分數（ x_i ），會等於其真分數（ t_i ）加上隨機誤差（ e_i ）

$$x_i = t_i + e_i \quad (8)$$

其中 e_i 符合 $cov(t_i, e_i) = cov(t_j, e_i) = cov(e_i, e_j) = 0$ 特性。因此，觀測分數 (X) 的變異量會等於真分數 (T) 變異量加上隨機誤差 (E) 變異量

$$\begin{aligned} Var(X) &= Var(T + E) \\ &= Var(T) + Var(E) + 2Cov(T, E) \\ &= Var(T) + Var(E), \end{aligned} \quad (9)$$

亦即考量隨機誤差後，觀測分數的變異量會恆大於母群變異量。

將觀測分數變異量乘以該測量工具的信度係數 $R (= Var(T)/Var(X))$ 可加以校正 (Wu et al., 2016, p. 81)

$$Var(X) \cdot R = Var(T) \quad (10)$$

此外，兩個變項真分數的相關會等於其觀測分數的相關除以兩變項測量工具信度乘積的平方根 (Wu et al., 2016, p. 82)。因此，若考量隨機誤差則觀測分數的相關會恆小於真分數的相關。

資料庫在進行二次分析時常會使用「結構方程式模型」(Structural Equation Modeling, 簡稱 SEM)，SEM 有將誤差概念納入潛在構念的測量指標中，而在估計路徑係數時則考量構念指標間的相關 (構念信度)，因此參數估計可以較準確。然而，大型教育成就調查為提高構念 (如數學成就、國文成就、科學成就構念等) 效度，通常會包含數十題甚或上百題題目 (測量指標) 以減少選題誤差。而避免學生做太多題目產生疲勞等效應影響作答反應，通常每位學生僅需回答部分試題 (請參考本書第五章)。題數太多會造成單一構念對應過多測量指標，而回答部分題目則會使得單一指標有過

多遺漏值，導致結構方程式模型適配指標不被接受（MacCallum, Browne, & Sugawara, 1996; Mannetti, Pierro, Kruglanski, Taris, & Bezinovi, 2002）。

為解決上述困境，Mislevy（1991）利用 Little 和 Rubin（1987）針對遺漏值的「多重差補方法」（multiple imputation），在同時考量學生背景變項與測驗作答結果的情況下，先行推估每位學生能力值的後驗機率分布（posterior distribution），再根據該分布隨機抽出數個可能的能力值，作為該學生能力的「可能值」（plausible value，簡稱 PV）。因此，每位學生會得出一組可能值，該組可能值的分布，既代表該學生能力值的機率分布，也代表與該學生背景變項及測驗作答結果完全一樣的一群學生其能力值的分布（任宗浩，2011，第 88 頁）。TIMSS、PISA、PIRLS 等國際大型教育成就調查資料庫，均提供每位學生單一特定領域能力 5 個可能值，TASA 2016 則對每位學生每一領域能力提供 10 個可能值。當研究者需推估母群參數 θ 時，必須先以第 i 個可能值跨不同學生計算統計量 $\hat{\theta}^{(i)}(i=1,2,\dots,10)$ ，再將得出的 i 個估計值平均，才能得到母群參數 θ 的不偏估計值 $\hat{\theta}$ （Mislevy, 1991）

$$\hat{\theta} = \frac{1}{M} \sum_{i=1}^M \hat{\theta}^{(i)} \quad (11)$$

其中 M 為對每位學生抽取可能值的次數，而其測驗誤差變異量為

$$B_M(\hat{\theta}) = \frac{1}{M-1} \sum_{i=1}^M (\hat{\theta}^{(i)} - \hat{\theta})^2 \quad (12)$$

TASA 2016 因採用 10 個可能值，所以 $M=10$ 。

（四）標準誤的估計

在利用 TASA 或其他大型教育成就調查資料庫進行統計分析時，統計量的誤差估計必須同時考量公式（7）所得之抽樣誤差和公式（12）之測量

誤差。綜合抽樣誤差與測量誤差，便可得到統計量（ $\hat{\theta}$ ）用以推估參數之誤差變異量（ V ）為（Mislevy, 1991）

$$V(\hat{\theta}) = U(\hat{\theta}) + \frac{M+1}{M} B_M(\hat{\theta}) \quad (13)$$

標準誤（ $SE(\hat{\theta})$ ）為：

$$SE(\hat{\theta}) = [V(\hat{\theta})]^{1/2} = \sqrt{U(\hat{\theta}) + \frac{M+1}{M} B_M(\hat{\theta})} \quad (14)$$

二、TASA 資料庫的二次分析方法介紹

本節所介紹之分析範例皆為 R 程式語法。R 屬開放共享資源（R Core Team, 2017），讀者可從官網（<https://www.r-project.org/>）下載最新版程式安裝使用。本章會用到的 R 套件包括「boot」（Canty & Ripley, 2017; Davison & Hinkley, 1997）、「foreign」（R Core Team, 2016）、「survey」（Lumley, 2004; 2016），並會說明相關套件中使用到的功能或函數。然而個別套件的詳細使用語法及參數設定，請有需要的讀者自行由官網下載使用手冊參考。

（一）複雜抽樣的平均數與標準誤估計

```
1. install.packages("foreign")
2. install.packages("survey ")
3. library(foreign)
4. library(survey)
5. setwd("C:/.../TASA_2016") #set working directory
6. tasa2016 <- read.spss("TASA2016.sav", use.value.labels = FALSE, to.data.frame
= TRUE)
```

```

7. head(tasa2016)
8. sam <- svydesign(ids = ~JKREP, strata = ~JKZONE, nest = TRUE, data=tasa2016,
  weights = ~Totwgt)
9. # Bootstrap
10. samboot <- as.svrepdesign(sam, type="bootstrap", replicates= 1500)
11. TWN_Math_boot <- data.frame(svymean(~PV1.M + ~PV2.M + ~PV3.M + ~PV4.
  M + ~PV5.M + ~PV6.M + ~PV7.M + ~PV8.M + ~PV9.M + ~PV10.M, samboot))
12. TWN_Math_boot
13. U_boot <- mean((TWN_Math_boot$SE)^2)
14. Bm_boot <- var(TWN_Math_boot$mean)
15. SE_Math_boot <- sqrt(U+(1.1*Bm_boot))
16. TASAMATH_boot <- matrix(c(TWN_Math_boot$mean, mean(TWN_Math_
  boot$mean), TWN_Math_boot$SE, SE_Math_boot), nrow = 11, ncol = 2, byrow
  = FALSE, dimnames = list(c("TASA2016$PV1Math", "TASA2016$PV2Math", "T
  ASA2016$PV3Math", "TASA2016$PV4Math", "TASA2016$PV5Math", "TASA20
  16$PV6Math", "TASA2016$PV7Math", "TASA2016$PV8Math", "TASA2016$PV
  9Math", "TASA2016$PV10Math", "TASAMATH_boot"), c("mean", "SE")))
17. TASAMATH_boot
18. # JK replication
19. samjkn <- as.svrepdesign(sam, type="JKn")
20. TWN_Math_jkn <- data.frame(svymean(~PV1.M + ~PV2.M + ~PV3.M + ~PV4.
  M + ~PV5.M + ~PV6.M + ~PV7.M + ~PV8.M + ~PV9.M + ~PV10.M, samjkn))
21. U_jkn <- mean((TWN_Math_jkn$SE)^2)
22. Bm_jkn <- var(TWN_Math_jkn$mean)
23. SE_Math_jkn <- sqrt(U_jkn +(1.1*Bm_jkn))
24. TASAMATH_jkn <- matrix(c(TWN_Math_jkn$mean, mean(TWN_Math_
  jkn$mean), TWN_Math_jkn$SE, SE_Math_jkn), nrow = 11, ncol = 2, byrow
  = FALSE, dimnames = list(c("TASA2016$PV1Math", "TASA2016$PV2Mat
  h", "TASA2016$PV3Math", "TASA2016$PV4Math", "TASA2016$PV5Math",
  "TASA2016$PV6Math", "TASA2016$PV7Math", "TASA2016$PV8Math", "TA
  SA2016$PV9Math", "TASA2016$PV10Math", "TASAMATH_jkn"), c("mean",
  "SE")))
25. TASAMATH_jkn

```

```

26. # Fay's BRR
27. samfay <- as.svrepdesign(sam, type="Fay",fay.rho=0.5)
28. TWN_Math_fay <- data.frame(svymean(~PV1.M + ~PV2.M + ~PV3.M + ~PV4.
    M + ~PV5.M + ~PV6.M + ~PV7.M + ~PV8.M + ~PV9.M + ~PV10.M, samfay))
29. U_fay <- mean((TWN_Math_fay$SE)^2)
30. Bm_fay <- var(TWN_Math_fay$mean)
31. SE_Math_fay <- sqrt(U_fay+(1.1*Bm_fay))
32. TASAMATH_fay <- matrix(c(TWN_Math_fay$mean, mean(TWN_Math_
    fay$mean), TWN_Math_fay$SE, SE_Math_fay), nrow = 11, ncol = 2, byrow =
    FALSE, dimnames = list(c("TASA2016$PV1Math", "TASA2016$PV2Math","TA
    SA2016$PV3Math","TASA2016$PV4Math","TASA2016$PV5Math","TASA201
    6$PV6Math", "TASA2016$PV7Math", "TASA2016$PV8Math","TASA2016$PV
    9Math","TASA2016$PV10Math","TASAMATH_Fay"), c("mean", "SE")))
33. TASAMATH_fay

```

本例以 TASA 2016 資料庫推估我國八年級數學成就平均數。指令 1、2 列分別安裝 R 套件「foreign」和「survey」，其中「foreign」可用以讀取 SPSS 的資料格式 (*.sav) 並轉換為 R 語法環境的資料格式，「survey」則專為分析複雜抽樣資料所開發的套件。3、4 列以「library」指令將已安裝的套件提取至工作區，接下來的程式語法就可直接呼叫「foreign」和「survey」套件中的功能函數。第 5 列「setwd」指令用以指定預設工作資料夾路徑，讀者可自行設定，接下來語法若無特別指定其它路徑，存取動作都會以該資料夾作為預設值，並請注意 R 語法路徑符號採用的是「/」或「\」，Mac OS 版本並請以「~」代替根目錄磁碟「C:」。

第 6 列指令利用「foreign」套件中的「read.spss」功能函數以讀取工作資料夾中的 SPSS 格式資料檔「TASA2016.sav」後，無需使用該 SPSS 資料檔裡的數值標記 (use.value.labels = FALSE)，並將之轉換成列聯表結構 (to.data.frame = TRUE)，指定為物件「tasa2016」。經由「head (tasa2016)」可呈現 tasa2016 物件的檔頭及前幾筆資料 (圖 6-3)，用以確認資料是否存取成功。

	STUD	SCHOOL	Totwgt	Houwgt	Urbanizationrate	JKZONE	JKREP	S2016Q1	S2016Q2
1	160800001	403	20.74	0.8383995		2	27	1	2
2	160800002	403	20.74	0.8383995		2	27	1	3
	S2016Q3	S2016Q4	S2016Q5	S2016Q6_1	S2016Q6_2	S2016Q6_3	S2016Q6_4	S2016Q6_5	
1	5	1	1	1	1	2	2	2	
2	4	1	1	1	1	2	2	2	
	S2016Q7_1	S2016Q7_2	S2016Q7_3	S2016Q7_4	S2016Q7_5	S2016Q7_6	S2016Q7_7	S2016Q7_8	
1	1	1	1	1	1	1	1	2	
2	1	1	1	1	1	1	1	1	
	S2016Q7_9	S2016Q7_10	S2016Q8	S2016Q9	S2016Q10	S2016Q11	S2016Q12	S2016Q13_1	
1	1	1	4	1	1	2	4		
2	2	1	3	2	1	1	4		
	S2016Q13_2	S2016Q13_3	S2016Q13_4	S2016Q13_5	S2016Q14_1	S2016Q14_2	S2016Q14_3		
1	3	4	4	4	4	5	4		
2	4	3	3	4	4	4	4		

...後略。

圖 6-3 「tasa2016」資料物件頭兩筆資料顯示結果

第 8 列利用「survey」套件裡「svydesign」功能界定「tasa2016」資料物件的抽樣架構，並將之指定為物件「sam」。其中「ids = ~JKREP」宣告初階抽樣單位的權重劃分根據欄位 JKREP，「strata = ~JKZONE」宣告根據欄位 JKZONE 劃分刀切法抽樣區或重複抽樣分層區，「nest = TRUE」表示初階抽樣單位巢套在分層架構下，故先分層後，再由每一分層抽取初階抽樣單位（反之，若先將初階抽樣單位劃分後再分層抽出樣本，則應設定為「nest = FALSE」），「data = tasa2016」是宣告指定的資料物件為「tasa2016」，「weights = ~Totwgt」則宣告樣本的權重變項使用 Totwgt（total weight）。

第 10 列指令利用「as.svrepdesign」功能針對資料物件「sam」以拔靴法（type = "bootstrap"）進行 1500 次重複抽樣（replicates = 1500）後，指定為物件「samboot」。第 11 列以「survey」套件「svymean」功能得出「samboot」內 10 組數學可能值的平均數（~PV1.M + ~PV2.M + ~PV3.M + ~PV4.M + ~PV5.M + ~PV6.M + ~PV7.M + ~PV8.M + ~PV9.M + ~PV10.M），以及每一組可能值的標準誤，指定為資料物件「TWN_Math_boot」。第 12 列列出

「TWN_Math_boot」數值（圖 6-4），由於重複抽樣允許每次抽樣有所變動，因此數值會近似但不完全相同。

	mean	SE
PV1.M	500.1852	2.066928
PV2.M	500.4033	2.101328
PV3.M	500.3813	2.081947
PV4.M	499.3658	2.084907
PV5.M	500.5568	2.136751
PV6.M	499.9472	2.100418
PV7.M	500.2027	2.110107
PV8.M	499.8402	2.084168
PV9.M	499.4869	2.074084
PV10.M	499.6304	2.045997

圖 6-4 10 組數學可能值的平均數與利用拔靴法估計的標準誤

第 13 列根據公式（11）取 10 組可能值估計出來的抽樣誤差變異量之平均數，作為最後估計的抽樣誤差變異量。第 14 列則根據公式（12）計算測量誤差變異量。第 15 列按公式（14）結合抽樣誤差變異量與測量誤差變異量後，開根號得出以該樣本平均數推估母群平均數之標準誤。第 16 列指令取 10 組可能值平均數的平均（`mean(TWN_Math_boot$mean)`）作為母群平均數的估計值，新增在「TWN_Math_boot\$mean」最後一列，並將第 15 列指令得出的標準誤（`SE_Math_boot`）新增在「TWN_Math_boot\$SE」最後

一列，重新命名可能值後指定為資料物件「TASAMATH_boot」。第 17 列直接觀看 TASAMATH_boot 資料（圖 6-5），由圖判讀 TASA 2016 八年級生的數學平均分數為 500，標準誤為 2.1。因為本屆量尺分數用線性轉換方式將平均值設定為 500，表示該平均能力的估計是正確的。

第 18 列至第 25 列改以刀切抽樣法、第 26 列至第 34 列則以 Fay's BRR ($\epsilon = 0.5$) 估計各組可能值的抽樣誤差，結果各如圖 6-6、圖 6-7。經比較圖 6-5、圖 6-6、圖 6-7，讀者不難發現這三種估計標準誤的方式，所得到的結果幾乎相同，這與相關文獻的觀點一致（Lehtonen & Pahkinen, 2004; Rust & Rao, 1996）。

	mean	SE
TASA2016\$PV1Math	500.1852	2.066928
TASA2016\$PV2Math	500.4033	2.101328
TASA2016\$PV3Math	500.3813	2.081947
TASA2016\$PV4Math	499.3658	2.084907
TASA2016\$PV5Math	500.5568	2.136751
TASA2016\$PV6Math	499.9472	2.100418
TASA2016\$PV7Math	500.2027	2.110107
TASA2016\$PV8Math	499.8402	2.084168
TASA2016\$PV9Math	499.4869	2.074084
TASA2016\$PV10Math	499.6304	2.045997
TASAMATH_boot	500.0000	2.132918

圖 6-5 整合 10 組數學可能值的平均數與利用拔靴法計算的標準誤

	mean	SE
TASA2016\$PV1Math	500.1852	2.090948
TASA2016\$PV2Math	500.4033	2.126732
TASA2016\$PV3Math	500.3813	2.091505
TASA2016\$PV4Math	499.3658	2.093076
TASA2016\$PV5Math	500.5568	2.134837
TASA2016\$PV6Math	499.9472	2.095522
TASA2016\$PV7Math	500.2027	2.143535
TASA2016\$PV8Math	499.8402	2.080618
TASA2016\$PV9Math	499.4869	2.070508
TASA2016\$PV10Math	499.6304	2.051832
T A S A M A T H _ j k n	500.0000	2.142020

圖 6-6 整合 10 組數學可能值的平均數與利用刀切抽樣法計算的標準誤

	mean	SE
TASA2016\$PV1Math	500.1852	2.090171
TASA2016\$PV2Math	500.4033	2.126332
TASA2016\$PV3Math	500.3813	2.090527
TASA2016\$PV4Math	499.3658	2.092385
TASA2016\$PV5Math	500.5568	2.133801
TASA2016\$PV6Math	499.9472	2.094384
TASA2016\$PV7Math	500.2027	2.143006
TASA2016\$PV8Math	499.8402	2.079565
TASA2016\$PV9Math	499.4869	2.069140
TASA2016\$PV10Math	499.6304	2.050650
T A S A M A T H _ F a y	500.0000	2.141126

圖 6-7 整合 10 組數學可能值的平均數與利用 Fay's BRR 計算的標準誤

表 6-4 錯誤的資料分析方式對母群平均數以及其標準誤估計之影響

分析	錯誤類型				平均值	標準誤
	未使用 權重	將樣本視 為 SRS	未考慮測量誤差			
			僅用 PV1	PVs 平均		
1	✓	✓	✓		507.15	1.00
2		✓	✓		500.18	1.06
3	✓		✓		507.15	2.34
4			✓		500.19	2.09
5	✓	✓		✓	507.03	0.93
6		✓		✓	500.00	0.99
7	✓			✓	507.03	2.36
8				✓	500.00	2.06
9	✓	✓			507.03	1.09
10		✓			500.00	1.13
11	✓				507.03	2.43
12	正確分析				500.00	2.14

表 6-4 以 TASA 2016 資料庫中八年級數學成就的平均數估計為例，示範各種錯誤的資料分析方法對於估計母群平均數及其標準誤之影響。讀者仔細比對各種錯誤類型之後，應該可以歸納出下列幾點結論：

1. 未使用權重會導致平均數與標準誤估計的偏誤；
2. 未考慮複雜抽樣而將樣本視為簡單隨機抽樣會導致標準誤的低估；
3. 未考慮測量誤差則有兩種情況：較常見的錯誤是將 10 組可能值的平均當作學生的數學成就分數，這種方法對於成就平均數的估計是正確的，但是會低估標準誤；另一種錯誤是只採用其中一組可能值（例

如 PV1.M)，這麼做也會導致標準誤的低估，但低估的情形不若採用 10 組可能值的平均當成學生能力來得嚴重。

(二) 複雜抽樣資料的相關值和標準誤估計與其顯著性檢定

```

1. install.packages ("foreign")
2. install.packages ("survey")
3. install.packages ("boot")
4. install.packages ("jtools")
5. library ( foreign )
6. library ( survey )
7. library ( boot )
8. library ( jtools )
9. setwd ("C:/.../TASA_2016") #set working directory
10. data1 <- read.spss ("TASA2016_1.sav", use.value.labels = FALSE,to.data.frame
    = TRUE)
11. data0 <- read.spss ("TASA2016_0.sav", use.value.labels = FALSE,to.data.frame
    = TRUE)
12. B=500
13. p=0.05
14. sam <- svydesign ( ids = ~JKREP,strata = ~JKZONE, nest = TRUE, data=data1,
    weights = ~Totwgt )
15. samboot <- as.svrepdesign ( sam, type="bootstrap", replicates= B )
16. boot1 <- samboot$repweights$weights
17. dim ( boot1 )
18. boot1[, 1:4]
19. RMN <- matrix ( 0, nrow=10, ncol=B+1 )
20. #bootstrapping of dataset
21. for ( j in 1:B )
22. {
23. bt1 <- as.data.frame ( data0 )
24. for ( i in 1:nrow ( boot1 ) )
25. {

```

```
26. DA <- subset ( data1, PSU1 == i )
27. resamp <- sample ( c ( 1:nrow ( DA ) ) ,boot1[i,j]*nrow ( DA ) ,replace=TRUE )
28. bt2 <- as.data.frame ( DA[resamp,] )
29. bt1 <- rbind ( bt1,bt2 )
30. }
31. D1 <- cbind ( bt1$PV1.M,bt1$PV1.N )
32. D2 <- cbind ( bt1$PV2.M,bt1$PV2.N )
33. D3 <- cbind ( bt1$PV3.M,bt1$PV3.N )
34. D4 <- cbind ( bt1$PV4.M,bt1$PV4.N )
35. D5 <- cbind ( bt1$PV5.M,bt1$PV5.N )
36. D6 <- cbind ( bt1$PV6.M,bt1$PV6.N )
37. D7 <- cbind ( bt1$PV7.M,bt1$PV7.N )
38. D8 <- cbind ( bt1$PV8.M,bt1$PV8.N )
39. D9 <- cbind ( bt1$PV9.M,bt1$PV9.N )
40. D10 <- cbind ( bt1$PV10.M,bt1$PV10.N )
41. RMN[1,j] <- corr ( D1, w=bt1$Houwgt )
42. RMN[2,j] <- corr ( D2, w=bt1$Houwgt )
43. RMN[3,j] <- corr ( D3, w=bt1$Houwgt )
44. RMN[4,j] <- corr ( D4, w=bt1$Houwgt )
45. RMN[5,j] <- corr ( D5, w=bt1$Houwgt )
46. RMN[6,j] <- corr ( D6, w=bt1$Houwgt )
47. RMN[7,j] <- corr ( D7, w=bt1$Houwgt )
48. RMN[8,j] <- corr ( D8, w=bt1$Houwgt )
49. RMN[9,j] <- corr ( D9, w=bt1$Houwgt )
50. RMN[10,j] <- corr ( D10, w=bt1$Houwgt )
51. }
52. #caculation of the correlation
53. R1 <- cbind ( data1$PV1.M,data1$PV1.N )
54. R2 <- cbind ( data1$PV2.M,data1$PV2.N )
55. R3 <- cbind ( data1$PV3.M,data1$PV3.N )
56. R4 <- cbind ( data1$PV4.M,data1$PV4.N )
57. R5 <- cbind ( data1$PV5.M,data1$PV5.N )
58. R6 <- cbind ( data1$PV6.M,data1$PV6.N )
```

```

59. R7 <- cbind ( data1$PV7.M,data1$PV7.N )
60. R8 <- cbind ( data1$PV8.M,data1$PV8.N )
61. R9 <- cbind ( data1$PV9.M,data1$PV9.N )
62. R10 <- cbind ( data1$PV10.M,data1$PV10.N )
63. RMN[1,B+1] <- corr ( R1, w=data1$Houwgt )
64. RMN[2,B+1] <- corr ( R2, w=data1$Houwgt )
65. RMN[3,B+1] <- corr ( R3, w=data1$Houwgt )
66. RMN[4,B+1] <- corr ( R4, w=data1$Houwgt )
67. RMN[5,B+1] <- corr ( R5, w=data1$Houwgt )
68. RMN[6,B+1] <- corr ( R6, w=data1$Houwgt )
69. RMN[7,B+1] <- corr ( R7, w=data1$Houwgt )
70. RMN[8,B+1] <- corr ( R8, w=data1$Houwgt )
71. RMN[9,B+1] <- corr ( R9, w=data1$Houwgt )
72. RMN[10,B+1] <- corr ( R10, w=data1$Houwgt )
73. RMN_mean <- apply ( RMN,2,mean )
74. hist ( RMN_mean[1:500],breaks= seq ( min ( RMN_mean ) -0.005,max ( RMN_
    mean ) +0.005,by=0.0002 ) )
75. r_MN <- RMN_mean[501]
76. r_MN
77. mean ( RMN_mean[1:500] )
78. RMN2 <- sort ( RMN_mean[1:500] )
79. CI <- c ( RMN2[1+as.integer ( B*0.5*p ) ],RMN2[1+as.integer ( B* ( 1-
    ( 0.5*p ) ) ) ] )
80. CI

```

對許多研究者來說，變項間相關的顯著性檢定經常是進行更複雜統計模型擬合的第一步。此節將以重複抽樣技術並推估相關值的信賴區間，來替代推論統計傳統以「第一型錯誤」(Type I error) 機率值為顯著性判斷的檢定方式。該節的語法利用「survey」以及「boot」套件，計算複雜抽樣設計 TASA 2016 的數學成就與科學成就之相關值及其 95% 信賴區間。

第 1 至 9 列安裝並呼叫所需要的 R 套件，並指定工作資料夾路徑。第

10 列讀取工作資料夾中的 SPSS 檔案「TASA2016_1.sav」，並將讀進來的資料指定為列聯表結構資料物件「data1」。第 11 列利用讀取和「TASA2016_1.sav」具相同資料結構但是沒有任何資料的檔案「TASA2016_0.sav」，藉以產生一個存放資料的空間，指定為資料物件「data0」，稍後可以將從「data1」中抽樣的數據寫入「data0」中。第 12 列指定拔靴法重複抽樣次數（B = ）為 500 次。第 13 列則是決定進行統計顯著性檢定時違犯第一型錯誤的機率閾值，依慣例設為 .05。第 14 列宣告「data1」資料的抽樣設計架構。

第 15 列利用「as.svrepdesign」指定在「sam」的資料架構下進行拔靴法（type = "bootstrap"）重複抽樣（B = ）500 次後，指定為資料物件「samboot」。實際上，「samboot」儲存了許多重複抽樣後的屬性可供後續分析取用，包括每一次重複抽樣後各初階抽樣單位（PSU，此為學校）被抽取為樣本的情形，利用第 16 列指令「samboot\$repweights\$weight」呼叫每次重複抽樣原樣本學校被抽取為新樣本的情形後，指定為物件「boot1」。第 17 列指令顯示「boot1」為 304 列 *500 行的資料矩陣。



圖 6-8 利用 svydesign 功能中拔靴法產生的叢集（學校）重複抽樣矩陣示例

第 18 列指令列出「boot1」第 1 到 4 行資料為例（圖 6-8）。每一行代表每一次重複抽樣後，原來的 304 所學校樣本以抽出後放回的方式重複抽出新的 304 所學校樣本的情況。由於兩階段分層叢集抽樣設計加上叢集抽樣機率必須根據系統機率比例（PPS），這使得實際的分層變項成為原分層變項加上學校大小，所以同一個 JKZONE 的兩所學校會是分層變項和大小最近似的學校。data1 的學校排序會先按 JKZONE 由小到大，相同 JKZONE 的兩所學校則按大小順序排列。讀者可以看出 boot1 的每一行，同一個 JKZONE 內的兩所學校的重複抽樣結果一定是 (2, 0)、(1, 1) 或是 (0, 2)，因為拔靴法會控制在每一 JKZONE 中以抽出放回的方式重新抽出相同樣本大小的樣本。以第一次學校重複抽樣結果的 JKZONE1 至 3 為例，JKZONE1 的兩所學校在第一次的重複抽樣中，兩次都抽中了第 2 所，用以模擬如果回到母群中，重新在第一個叢集分層中抽兩間學校，可能都較接近第二所學校的特質。而 JKZONE2 中則是各抽了一次，模擬重抽的兩所學校特質分別相似於兩所原本被抽出的學校。JKZONE3 重抽的兩所學校則較類似 JKZONE3 原本抽出的第 1 所學校。每一次重複抽樣後的總數仍會維持在 304，即模擬從母群按原始的抽樣分層架構重新抽出 304 所學校的可能狀況。

第 19 列的指令先指定 1 個 10 列、B + 1 行、元素為 0 的矩陣 RMN。該矩陣前 B 行將用以存放 B 次重複抽樣、每次抽樣有 10 組對應的數學和科學可能值相關，最後第 B+1 行則為原樣本計算出的相關值。

拔靴法是利用樣本分布模擬母群分布，反覆以對母群的抽樣架構從原樣本抽出新的重複抽樣樣本，並計算每次重複抽樣樣本的估計值，最後以這些重複抽樣樣本估計值分布的變異量來代表抽樣誤差。語法 21 ~ 51 列設計了兩個迴圈，for (j in 1:B) {...} 可依序取出 boot1 中第 j 次、總計 B 次重複抽樣後的學校樣本結果，for (i in 1:nrow (boot1) {...}) 再於個別學校的校內學生樣本中，以抽出放回的方式重新抽出一組與該校原來學生樣本人數相同的新學生樣本。每一所學校內需重複抽出的學生樣本人數，為該校原樣

本人數乘以拔靴法第 j 次重複抽樣中該校被重複抽到的次數。將以上兩個步驟重複 B 次，便得到完整的 B 組重複抽樣樣本。

第 26 列指令讀取原樣本資料 `data1` 中學校 i 的學生樣本人數並存為物件 `DA`，第 27-28 列針對 `DA` 中學校 i ，以抽出放回的方式重複抽樣，樣本人數則為 `boot1` 中第 j 次重複抽樣學校 i 被抽到的次數（「`boot1[i,j]`」）乘以該校原樣本大小（`nrow(DA)`），之後將重複抽樣後的學校樣本資料存成資料物件「`bt2`」。第 29 列合併目前已經做完重複抽樣的樣本資料，並存放在「`bt1`」。所以 `for (j in 1:B) {...` 迴圈取出「`boot1`」第 j 次重複抽樣學校被抽取為樣本的情形，24-30 列則產生執行拔靴法後第 j 次重複抽樣的虛擬樣本資料「`bt1`」。第 31 列利用「`cbind`」合併行（`column`），將第 j 次重複抽樣的資料「`bt1`」中的第 1 組數學和第 1 組科學可能值欄位單獨取出組成資料物件 `D1`，第 32-40 列的 `D2` 至 `D10` 則分別存放第 2 至第 10 組相對應的數學和科學可能值數據。第 41-50 行利用「`boot`」套件的「`corr`」功能計算數學與科學 10 組可能值個別的相關值，並寫入 `RMN` 矩陣第 j 行的 1-10 列，即為拔靴法第 j 次重複抽樣結果所得的 10 組相關值。如此反覆執行 21-51 列迴圈 500 次後，`RMN` 矩陣就記錄了根據 500 次重複抽樣結果算出來的 10 組相關值。

第 53-72 列的目的是為了計算原樣本資料估計的數學與科學成就的相關值，其語法和 31-50 列幾乎完全相同，唯一的不同在此處是根據原樣本資料 `data1` 去計算十組對應的數學和科學可能值間的相關，並將這十組數值記錄在 `RMN` 矩陣的第 501 行。

根據公式 (11)，針對每次抽樣的相關值估計，應該是根據 10 組可能值估計出相關值的平均。`RMN` 為 10 列 * 501 行矩陣，其中第 1 到 500 行（`RMN[,1:500]`）存放利用拔靴法進行 500 次重複抽樣所估計出來的 10 組數學和科學成就相關，最後一行（`RMN[,501]`）則是利用原樣本估計出來數學和科學的 10 組相關值。第 73 列指令利用 R 語法矩陣運算「`apply`」功能將 `RMN` 矩

陣每一行進行 1-10 列相關值的平均，總共有 501 行的平均值，並將這 501 個估計值存成向量物件「RMN_mean」。語法第 74 列利用「hist」功能繪製 RMN_mean 前 500 筆相關值 (RMN_mean[1:500]) 的頻率分布長條圖 (圖 6-9)。

語法第 75 列將原樣本資料估計出來的數學與科學相關 (RMN_mean [501])，存入物件 r_MN；第 76 列語法顯示 TASA 2016 資料庫估計我國八年級生數學與科學相關值的估計值為 0.9042157。第 77 列語法計算根據 500 次重複抽樣結果得到的相關值為 0.9040803，與原樣本的估計結果非常接近，可用以佐證重複抽樣的過程是否正確。至此，我們計算拔靴法 500 次重複抽樣樣本數學成就和科學成就間相關的期望值，並和原樣本所得的相關值比較後證實差異不大。

第 78-80 列則用以取得相關估計值的信賴區間。讀者可依據圖 6-9 相關值的分布取得最小和最大兩臨界值，使得該臨界值包含分布中間 95% 的相關估計值，該值即為 95% 信賴區間臨界值。第 78 列利用「sort」功能將 500 次重複抽樣樣本估計的相關值 (RMN_mean[1:500]) 由小到大排序。第 79 列則是找到根據語法第 13 列指定第一型的機率閾值 ($p = .05$)，依雙尾檢定，頭尾各找到最接近的 2.5% ($= 0.5 * p$) 的整數序位後，再根據該序位讀出相關值估計結果作為 95% 信賴區間的上、下臨界值。執行第 80 列，讀者會看到 TASA 2016 數學和科學成就相關的 95% 信賴區間為 (0.901, 0.907)。由於拔靴法每次的抽樣結果都不一樣，讀者利用所釋出的資料和本章提供的語法，會獲得極接近的數值，但可能不會完全相同。

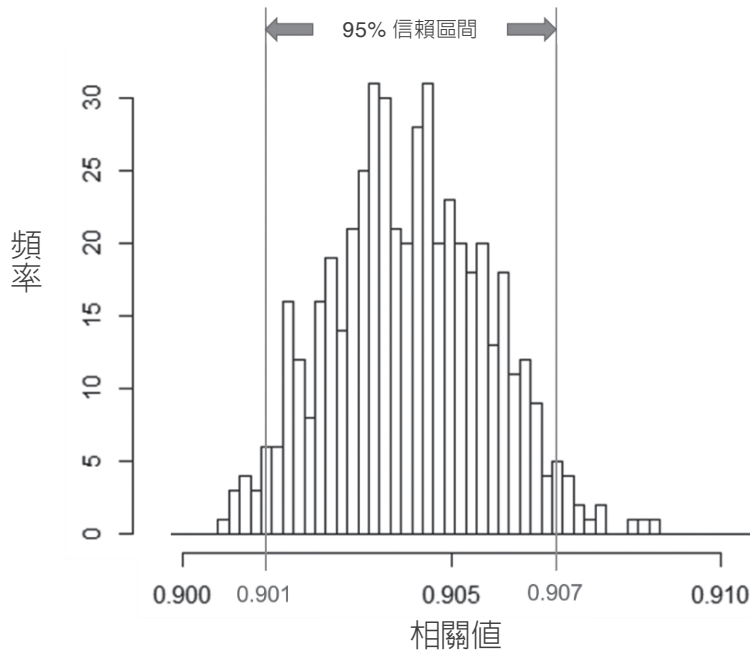


圖 6-9 利用拔靴法進行 500 次二階段重複抽樣估計 TASA 2016 八年級生之數學與科學成就相關之分布

表 6-5 顯示幾種常見的錯誤類型對於相關值估計的影響。該表的結果顯示，資料未正確加權可能導致相關值估計的偏誤。其次，利用可能值求取相關一定要做正確的配對（第 1 組數學可能值配對第 1 組科學可能值後求相關，第 2 組數學可能值配對第 2 組科學可能值...），若配對錯誤（如第 1 組數學可能值配對第 2 組科學可能值取相關）則會低估兩個變項的相關。最後如果直接先取可能值的平均後，再取變項間的相關，則會高估了兩個變項間的相關。

表 6-5. 不同錯誤類型之 TASA 2016 資料數學與科學成就相關值估計結果

錯 誤 類 型	相 關 值
無（正確估計）	0.904
未加權	0.906
可能值未正確配對	約 0.800
利用 10 組可能值之平均值取相關	0.959

（三）複雜抽樣資料的 t 檢定與迴歸分析

```

1. library ( foreign )
2. library ( survey )
3. setwd ( "C:/.../TASA_2016" ) #set working directory
4. tasa2016 <- read.spss ( "TASA2016.sav", use.value.labels = FALSE,to.data.frame
   = TRUE )
5. sam <- svydesign ( ids = ~JKREP,strata = ~JKZONE, nest = TRUE,
   data=tasa2016, weights = ~tasa2016$Totwgt )
6. A <- matrix ( 0,nrow=11,ncol=4 )
7. Reg1 <- summary ( svyglm ( PV1.M~gender, design=sam ) )
8. A[1,] <- Reg1$coefficients[2,1:4]
9. Reg2 <- summary ( svyglm ( PV2.M~gender, design=sam ) )
10. A[2,] <- Reg2$coefficients[2,1:4]
11. Reg3 <- summary ( svyglm ( PV3.M~gender, design=sam ) )
12. A[3,] <- Reg3$coefficients[2,1:4]
13. Reg4 <- summary ( svyglm ( PV4.M~gender, design=sam ) )
14. A[4,] <- Reg4$coefficients[2,1:4]
15. Reg5 <- summary ( svyglm ( PV5.M~gender, design=sam ) )
16. A[5,] <- Reg5$coefficients[2,1:4]
17. Reg6 <- summary ( svyglm ( PV6.M~gender, design=sam ) )
18. A[6,] <- Reg6$coefficients[2,1:4]

```

```

19. Reg7 <- summary (svyglm (PV7.M~gender, design=sam) )
20. A[7,] <- Reg7$coefficients[2,1:4]
21. Reg8 <- summary (svyglm (PV8.M~gender, design=sam) )
22. A[8,] <- Reg8$coefficients[2,1:4]
23. Reg9 <- summary (svyglm (PV9.M~gender, design=sam) )
24. A[9,] <- Reg9$coefficients[2,1:4]
25. Reg10 <- summary (svyglm (PV10.M~gender, design=sam) )
26. A[10,] <- Reg10$coefficients[2,1:4]
27. A[11,1] <- mean (A[1:10,1])
28. A[11,2] <- ( mean (A[1:10,2]^2) +1.1*var (A[1:10,1]) ) ^0.5
29. A[11,3] <- A[11,1]/A[11,2]
30. A[11,4] <- ( 1-pt (A[11,3],151) ) *2
31. colnames (A) <- c ("Estimate","Std. Error","t value","Pr (>|t|) ")
32. rownames (A) <- c ("PV1.M~gender","PV2.M~gender","PV3.M~gender","PV4.
    M~gender","PV5.M~gender","PV6.M~gender","PV7.M~gender","PV8.
    M~gender","PV9.M~gender","PV10.M~gender","Math~gender")
33. A

```

本章一開始介紹了幾種不同的重複抽樣法，用以估計複雜抽樣設計中統計量的標準誤。前節有關兩變項相關值的統計推論，主要採取信賴區間臨界值的估計。若讀者欲採用傳統的統計檢定以進行群組間差異的顯著性考驗，其標準誤的估計仍應採用相同程序，但對於 t 分佈或 F 分佈的自由度 (degree of freedom) 計算，則不同於簡單隨機抽樣設計。當採用重複抽樣法進行誤差估計並進行變異量分析時，相關研究建議自由度應為重複抽樣次數減 1 (Rust, 1986; Rust & Rao, 1996)。以 TASA 2016 資料為例，當採用 500 次的拔靴法進行分析時，其自由度為 499；當採用 JK2、BRR 或是 Fay' s BRR 重複抽樣法進行分析時，建議的自由度則為 151。因為重複抽樣次數通常都很大，這使得減 1 後的自由度也很大 (> 100)，因此 t 分布會近似 z 分布，而以 z 分布的機率作為顯著性考驗依據。

該節以「survey」套件的「一般線性模型」(generalized linear model, GLM)「glm」功能，示範 TASA 2016 調查中男、女學生數學成就差異的統計檢定。語法第 6 列創造一個 11 列 4 行、元素為 0 的矩陣空間，用以存放稍後所得數據。第 7 列以「glm」功能函數，直接取用「sam」物件中的抽樣架構，並以資料中的性別變項「gender」為自變項，對所有學生的第一組數學成就可能值「PV1.M」進行迴歸分析後，將結果指定為物件「Reg1」。由於性別變項為二元虛擬變項(dummy variable)，一般線性模型中當迴歸分析的自變項為二元虛擬變項時，可視為該自變項兩水準間的 t 檢定，而迴歸係數即為兩變項水準間(男性與女性間)平均分數的差異。指令「Reg1\$coefficients」可列出物件「Reg1」中與迴歸係數有關的部分資訊(圖 6-10)。由圖中判讀性別變項的迴歸係數為 5.126631，由於 TASA 2016 資料中男性虛擬變項代碼為「1」、女性為「2」，故表示以第一組數學可能值得出的女性平均分數比男性高 5.126631，而考慮複雜抽樣設計的標準誤為 2.310224，t 值為 2.219105，大於該臨界值以上的機率为 2.797093e-02。第 8 列以指令「Reg1\$coefficients」特定讀取資料表中第 2 列的第 1 至 4 行與迴歸係數有關的資料，寫入 A 矩陣中的第 1 列。第 9 至 26 列則是針對第 2 到第 10 組數學可能值進行男、女性差異的迴歸分析後，將迴歸係數等資訊寫入 A 矩陣中的第 2 到第 10 列，藉以彙整十組可能值的迴歸分析結果。

	Estimate	Std. Error	t value	Pr (> t)
(Intercept)	492.611513	4.254395	115.788858	2.218082e-149
gender	5.126631	2.310224	2.219105	2.797093e-02

圖 6-10 利用「survey」套件中「glm」功能產生的迴歸係數報表格式

第 27 列指令將 10 組可能值算出的迴歸係數取平均值，寫入 A 矩陣的第 11 列中的第 1 行 (A[11,1])；第 28 列則根據公式 (14) 合併抽樣誤差與測量誤差變異量後，求出正確的迴歸係數標準誤並寫入 A[11,2]；第 29 列則計算迴歸係數的 t 值 (迴歸係數除以標準誤) 後，寫入 A[11,3]；最後，我們利用 R 語法內建的「stats」套件中「pt」功能，設定自由度 151 以計算 t 分佈累積機率後，求出臨界值以上區域的機率分布，寫入 A[11,4]。第 31 至 32 列指定 A 矩陣的行、列名稱。33 列則列出 A 矩陣的內容如圖 6-11。

	Estimate	Std. Error	t value	Pr (> t)
PV1.M~gender	5.126631	2.310224	2.219105	0.027970927
PV2.M~gender	6.196357	2.436088	2.543569	0.011978089
PV3.M~gender	6.806467	2.281943	2.982750	0.003332004
PV4.M~gender	5.779807	2.443291	2.365582	0.019271925
PV5.M~gender	6.175825	2.351836	2.625959	0.009530341
PV6.M~gender	4.755177	2.392538	1.987503	0.048675133
PV7.M~gender	5.794654	2.332756	2.484038	0.014082460
PV8.M~gender	5.393130	2.219646	2.429726	0.016283473
PV9.M~gender	4.194058	2.387061	1.756997	0.080945173
PV10.M~gender	5.389077	2.408779	2.237265	0.026733351
Math~gender	5.561118	2.488532	2.234698	0.026905335

圖 6-11 利用該節語法進行 TASA 2016 八年級男女生數學成就差異的 t 檢定

在圖 6-11 中最下方一列的結果即為合併 10 組可能值的迴歸分析後，得到的性別對於數學成就的迴歸分析結果。結果顯示女生的平均分數比男生高 5.56 分，標準誤為 2.49 分， t 值為 2.23， $p < .05$ 故達顯著水準。

此節以性別（二元變項）與學生數學成就間的線性迴歸分析作為例子，當預測變項為連續變項或多元迴歸時，其分析方法亦同。惟當預測變項與被預測變項均為成就變項（可能值），或是預測變項有兩個以上的成就變項（如英文和國文成就）時，需特別注意可能值的對應配對，例如英文的第 1 組可能值必須配對國文的第 1 組可能值；每 1 組可能值都必須要利用重複抽樣的方式估計抽樣誤差變異量後，再按公式（11）至公式（14）統整 10 組可能值的迴歸分析結果，方能得到正確的標準誤估計。

三、結論與建議

本章雖以介紹 TASA 2016 資料庫統計量、抽樣誤差、測量誤差的概念、原理及計算方式為主，但亦適用於 TIMSS、PISA、PIRLS 等國際大型教育成就調查資料庫。並經由免費的 R 模組相關語法範例，進行二次分析的實際操作，希望以範例幫助讀者了解大型教育成就調查在權重使用、複雜抽樣設計之抽樣誤差與測量誤差等議題的處理方法。範例一特別使用三種不同的重複抽樣技術：拔靴法、刀切抽樣法、Fay's BRR，用以估計複雜抽樣設計資料的平均數與標準誤，可觀察到結果彼此差異不大，因此讀者可依資料庫特性或偏好，選擇其中一種方式進行。

本章亦說明分析大型教育成就調查資料庫常見的錯誤類型及其影響，包括未使用權重、將樣本視為簡單隨機抽樣設計，以及可能值的誤用（例如：直接以可能值平均數作為學生能力估計值、估計學科成就間的相關係數未正確配對可能值等錯誤）。過去數十年來，各種國際大型教育成就調查方法及

結果愈來愈受世界各國重視，相關統計分析技術也漸趨成熟。若研究者能利用這些珍貴且數量龐大的資料庫進行研究分析，了解分析方法的進展，勢將對教育政策、測驗理論、統計技術等進步卓有貢獻。

參考文獻

- 任宗浩 (2011)。研究設計與資料分析 (第三章)。張俊彥主編, **TIMSS 2007 國際數學與科學教育成就趨勢調查國家報告**。臺北市: 國立臺灣師範大學科學教育中心。
- 任宗浩、譚克平、張立民 (2011)。二階段分層叢集抽樣的設計效應估計: 以 TIMSS 2007 調查研究為例, **教育科學研究期刊**, **56** (1), 33-65。
- Chen, K.-M., Jen, T.-H., and Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics*, *8*(1), 181-196. doi: 10.1016/j.joi.2013.11.007
- Canty, A. and Ripley, B. (2017). *boot: Bootstrap R (S-Plus) functions*. R package version 1.3-19.
- Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. ISBN 0-521-57391-2
- Dippo, C. S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Proceedings of the American Statistical Association, Section on Survey Research Methods, 489-494.
- Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: Wiley.
- Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, *6* (3), 223-239.
- Kish, L. (1965). *Survey sampling*. London: John Wiley & Sons.
- Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*(1), 81-90. doi: 10.22237/jmasm/1177992480

- Lehtonen, R., and Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys, statistics in practice* (2nd ed.). New York: John Wiley & Sons.
- Lumley, T. (2016). *survey: Analysis of complex survey samples*. R package version 3.31-5.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19.
- MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130–149.
- Mannetti, L., Pierro, A., Kruglanski, A., Taris, T. and Bezinović, P. (2002). A cross-structural study of the need for cognitive closure scale: Comparing its structure in Croatia, Italy, USA and the Netherlands. *British Journal of Social Psychology*, 41, 139-156.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177–196.
- Mislevy, R. J. (1993). Should “multiple imputations” be treated as “multiple indicators”? *Psychometrika*, 58(1), 79–85.
- Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20 (3), 355–375. doi:10.1214/aoms/1177729989.
- R Core Team (2016). *foreign: Read data stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase,* R package version 0.8-67. <https://CRAN.R-project.org/package=foreign>
- R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rust, K. (1986). *Efficient Replicated Variance Estimation*. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 81-87.

Rust, K., and Rao, J.N.K. (1996). Variance estimation for complex estimators in sample surveys. *Statistics in Medical Research*, 5, 381-397.

Wu, M., Tam, H. P., and Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer Nature.

第七章 應用決策樹建立學生學習模型

蔡明學

國家教育研究院副研究員

一、分析方式選擇的參考

決策樹（Decision Tree）是具有監督式特徵萃取與描述的功能，將輸入變數根據目標設定分支，以樹狀圖的層級架構呈現的一種採礦技術，決策樹模型可做為資料探索、分類及預測，找出依變數與自變數之間的層級關係（簡禎富、許嘉裕，2014）。雖然決策樹分析屬於資料探勘較早期的技術，但本文為何還是選擇這種技術呢？我們根據謝邦昌、鄭宇庭（2016）的研究指出，在預測信用卡違約戶的目標下，建立五種不同方法的預測模型，分別是羅吉斯迴歸、類神經網路、貝氏機率、決策樹與集群分析等，結果發現類神經網路與決策樹分析兩種採礦技術，雖然潛藏過度學習的可能，但還是具有較佳的總體預測率（詳見表 7-1）。再以類神經網路與決策樹分析兩種採礦技術進行比較，類神經網路分析有不易解釋的缺點，故本文選用資料探勘技術中的決策樹分析，作為建立學生學習模型之方法。

表 7-1 五種採礦模型建立之模型對母體資料的預測能力比較表

單位：%

	非 違 約 戶 預 測 率	違 約 戶 預 測 率	預 測 違 約 戶 正 確 率	總 體 預 測 率
羅 吉 斯 迴 歸	86.43	59.11	68.22	77.41
類 神 經 網 路	99.98	100.00	99.97	99.99
貝 氏 機 率	87.19	51.15	66.32	75.29
決 策 樹 分 析	100.00	99.88	100.00	99.96
集 群 分 析	95.63	23.44	72.53	71.79

資料來源：謝邦昌等（2016）。

本文以 SQL Server Data Mining 2014 版建立學生學習分類模型，以了解影響學生學習成效的關鍵因素，並區辨不同學生學習成就高低之解釋路徑。決策樹的主要功能為探索及預測（Berry & Linoff, 2000），其概念來自於認為每一現象背後的成因，可能是由兩個或多個事件影響，進而形成不同結果。利用資料探勘演算法的方式，自動找出現象背後所隱微與影響決策因素（Steinberg & Cardell, 2016）。

本文是藉由分類已知的「利益值」（效標變項）來建立一個樹狀結構，並從中歸納出類別欄位與其它欄位間隱藏規則，以遞歸階級結構作為分類的依據原則。在使用 CART 演算法的條件下，最大的特色除了進行二元分支演算法外，同時能處理連續變項與名目變項的分類（簡禎富等，2014）。

本文分析時主要使用 k 疊交互驗證法（k-fold cross validation），乃經過以下三個步驟：（一）以全體樣本產生分類模型，建構出最大樹狀結構：使用 Gini 為分割準則，在每個分支節點進行資料分隔，建立一個二分式的決策樹，已決定最佳分支變數。本文考量樣本數與分類情況，將分支母節點（parent node）的最少樣本數設為 200 人，最後產生的子節點（child node）

的最少樣本數設為 100 人；(二) 採用 k 疊交互驗證法評估分類模型的正確率，將其值設為 10；(三) 事後修剪規則是採用最大風險差異法 (maximum difference in risk)，是使生長完的樹，如果想要避免樹狀結構過於龐大，會進行修剪，修剪的方式視效標變項及不純度量測的方式有所不同，主要是以考量分類正確與否的風險值 (risk) 大小進行設定 (江羿臻、林正昌，2014；Witten & Frank, 2005)，本文將其值設為 0，以產生最小風險值的樹狀結構。最後 CART 在樹狀結構長成以後，會對長成的樹進行分類正確率的評估 (Breiman, Friedman, Olshen, & Stone, 1984)。

操作軟體主要以 Excel 2013，首先點選「進階」，再點取「建立採礦結構」。再將所需分析的「資料表匯入」並完成命名。回到「進階」，點取「將模型加入結構」，在演算法中選擇「決策樹」。在到資料行中，將「利益值」(效標變項) 進行設定，以本文而言，以 5 個數學 plausible value (學生學習成就) 平均後即是效標變項，家庭資本變項、學科特定性變項 (教師教學與學習策略) 與心理學構念 (學生學習自我效能、趨向表現與自我概念) 則為預測變項。

二、自變項的選擇

有關於自變項的選擇，須回到探討現象的本質。以本文為例，主要在於探討學生學習的影響因素的分類建構學生學習模型。黃建翔、蔡明學 (2016) 曾於高中學生學習表現因素進行探究，該文指出教師教學是影響高中學生學習表現的主要因素，眾多，但大致上影響學生學習的因素包含三大層面，家庭文化資源、教師教學方法與學生學習態度。本文取 2016 年臺灣學生學習成就資料庫 (TASA) 國二學生數學科學生學習成就與共同問卷填答反應進行分析，探究影響國中學生數學學習的關鍵因素。學生學習因素三大層面內涵以及與共同問卷的連結，相關內容分述如下：

（一）家庭資本變項

Coleman（1988）進行的研究指出，家庭社會經濟背景愈佳，愈有助於提升子女之教育成就。易言之，家庭背景控制學生的學習成就。但以學習資本論的觀點來看，家庭社經背景並非直接影響學生的學習表現因素，家庭社會經濟背景是透過家庭所能使用的財務資本（financial capital）、文化資本（cultural capital）及社會資本（social capital）等家庭資源，對子女的學習成就產生中介效果的影響（林俊瑩，2007；蔡毓智，2008）。根據上述內容，2016 共同問卷可供分析的內容包含（詳見表 7-2）：

1. 你家中有沒有電腦或筆記型電腦
2. 你家中有沒有 iPad 等平板電腦
3. 你家中有沒有智慧型手機
4. 你家中有沒有 Wii、XBOX、PS4、PSVita、PSP 或 NDS、3DS 等遊戲機
5. 你家中有沒有鋼琴、小提琴等西洋樂器（直笛不算），或中國樂器
6. 你家中有沒有西洋畫、水墨畫、中國字畫等掛在牆壁的字畫
7. 你家中有沒有你的個人書桌

（二）學科特定性變項（教師教學與學習策略）

學校是學生學習知識和技能歷程中最主要的教育場域，而教師是知識傳遞的主要媒介。Ottmar 等人（2015）探究教學法有效性，瞭解教師課堂投入與學生數學成就之間的關係。該研究指出，透過創新教學的支持性，除提供教師增強教學效能與改善學生學習實踐。黃建翔等（2016）的研究亦指出，教師教學對於學生學習成就有關鍵性的影響。故本文在教師教學上，2016 共同問卷可供分析的內容包含：下列關於這學期學校數學老師教學的描述，你是否同意？

1. 老師會教我們找出數學題目的關鍵字句，例如：「未知數」是什麼
2. 老師會教我們如何找出題目的條件，並根據題目的條件列式
3. 遇到題目時老師會教我們先從簡單例子著手思考
4. 遇到題目時老師會教我們先從特殊例子著手思考
5. 老師會請我們在求出問題解答之後，再思考有沒有其他可能的解決辦法
6. 老師會請我們在求出問題解答之後，教我們進行驗算
7. 老師會鼓勵我們課前預習數學內容
8. 老師會鼓勵我們課後複習學過的數學內容
9. 老師會要我們多多練習題目
10. 老師會鼓勵我們彼此討論數學
11. 老師會鼓勵我們堅持不懈地解決數學難題

（三）心理學構念（學生學習自我效能、趨向表現與自我概念）

學生是學習的主體，教育的開展是植基於以「人」為中心的基本理念上，故當談論到學生學習成就表現時，就不得不談到學生學習的策略與方法。Wang、Peng、Huang、Hou 與 Wang（2008）便認為學習者基於他們對於學習內容的渴望與興趣，藉由獲取知識或深入瞭解知識的方式，去滿足此需求。在學生學習層面上，影響學習成就的因素非常多，例如：健康、智力、性向、動機、身心發展、人格特質、學習態度、學習滿意度等。故本文在學生學習上，2016 共同問卷可供分析的內容包含：下列關於學習數學的描述，你是否同意？

1. 學習數學對我而言並不是什麼難事
2. 數學是我厲害的科目之一

3. 我知道我可以將數學學好
4. 我可以解決數學難題
5. 學習數學對我來說非常有趣
6. 我期待每一次的數學課
7. 一想到數學課的內容很精采，就讓我覺得很棒
8. 我期待在數學課能夠發現新的觀念與挑戰性的想法

表 7-2 分析內容代碼表

層 面	內 容	代 碼
家庭資本變項	你家中有沒有電腦或筆記型電腦	A1
	你家中有沒有 iPad 等平板電腦	A2
	你家中有沒有智慧型手機	A3
	你家中有沒有 Wii、XBOX、PS4、PSVita、PSP 或 NDS、3DS 等遊戲機	A4
	你家中有沒有下列這些東西？鋼琴、小提琴等西洋樂器（直笛不算），或中國樂器	A5
	你家中有沒有西洋畫、水墨畫、中國字畫等掛在牆壁的字畫	A6
	你家中有沒有你的個人書桌	A7
學科特定性變項 (教師教學與學習策略)	老師會教我們找出數學題目的關鍵字句，例如：「未知數」是什麼	B1
	老師會教我們如何找出題目的條件，並根據題目的條件列式	B2

層面	內容	代碼	
學科特定性變項 (教師教學與學習策略)	遇到題目時老師會教我們先從簡單例子著手思考	B3	
	遇到題目時老師會教我們先從特殊例子著手思考	B4	
	老師會請我們在求出問題解答之後，再思考有沒有其他可能的解決辦法	B5	
	老師會請我們求出問題解答之後，教我們進行驗算	B6	
	老師會鼓勵我們課前預習數學內容	B7	
	老師會鼓勵我們課後複習學過的數學內容	B8	
	老師會要我們多多練習題目	B9	
	老師會鼓勵我們彼此討論數學	B10	
	老師會鼓勵我們堅持不懈地解決數學難題	B11	
	心理學構念 (學生學習自我效能、 趨向表現與自我概念)	學習數學對我而言並不是什麼難事	C1
		數學是我厲害的科目之一	C2
我知道我可以將數學學好		C3	
我可以解決數學難題		C4	
學習數學對我來說非常有趣		C5	
我期待每一次的數學課		C6	
一想到數學課的內容很精采，就讓我覺得很棒		C7	
我期待在數學課能夠發現新觀念與挑戰性的想法		C8	

三、分析結果的解讀

過去論及學生學習成就多進行學習成就與影響學習二變項之相關或差異分析，或以迴歸分析判斷學習成就的影響程度等統計分析方法。但上述相關方法除無法辨識哪項因素為影響學生學習的關鍵因素外，亦無法描述各因素間的階層關係。為避免上述狀況，本文將以決策樹分析進行探討，並建立決策樹分析模型。

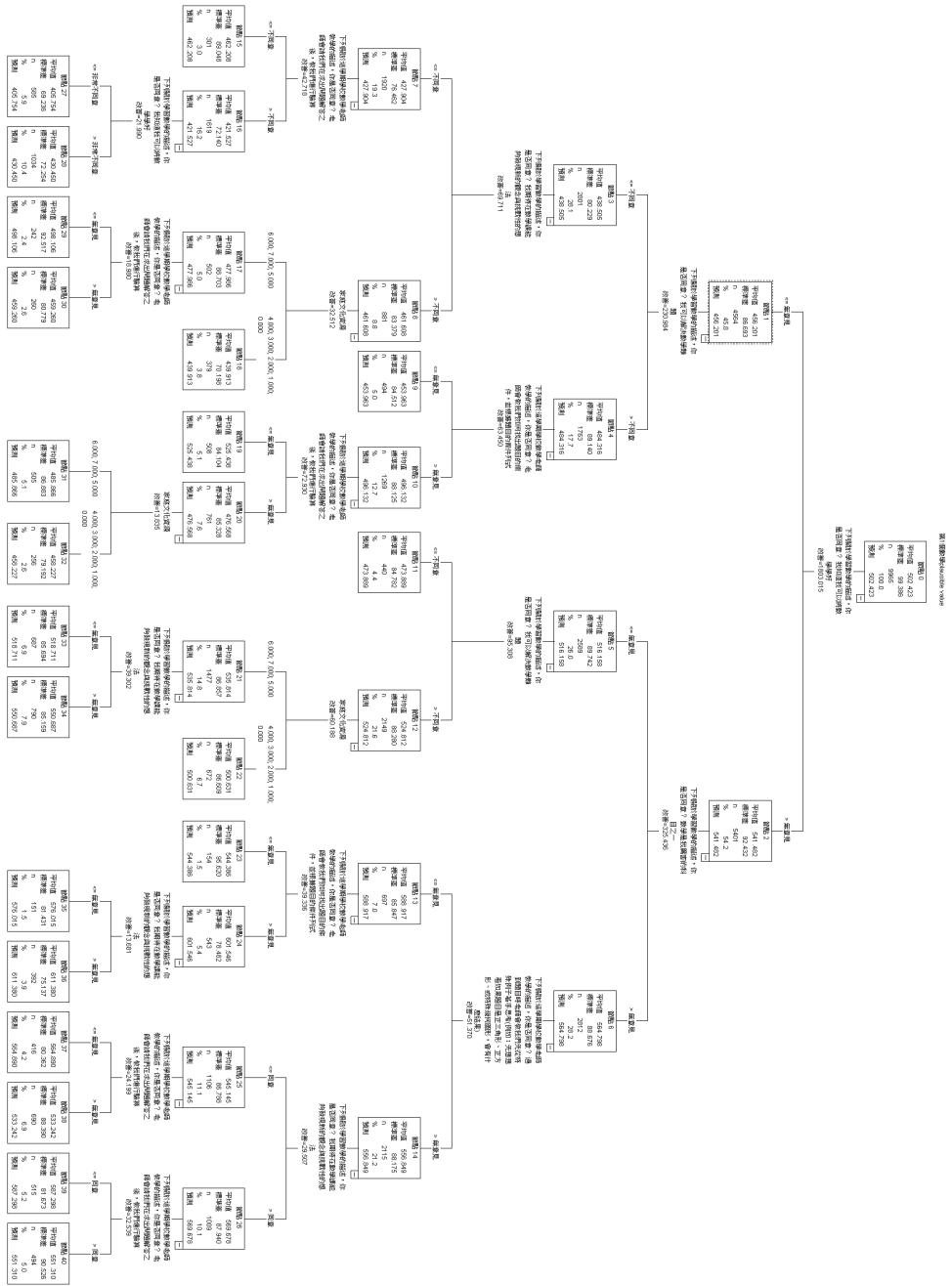
本文主要以 2016 年 TASA 數學科 PV 值 (plausible value) 作為效標變項，以家庭文化資源 (家中軟硬體設備有無，合併計算 A1~A7)、教師教學方法 (B1~B11) 與學生學習態度 (C1~C8) 等做為預測變項，用決策樹 CART 演算法進行分析。依據決策樹探勘結果，終端節點數為 21，深度為 5 (詳見表 7-3 與圖 7-1)。

表 7-3 2016 年 TASA 國二數學學生學習影響模式摘要

規 格	成長方法	CART
	依變數	第 1 個數學 plausible value
	自變數	A1, A2, A3, A4, A5, A6, A7, B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, C1, C2, C3, C4, C5, C6, C7, C8 (家庭資本變項、學科特定性變項 (教師教學與學習策略)、心理學構念 (學生學習自我效能、趨向表現與自我概念))
	確認	無
	最大樹狀結構深度	5

	父節點中最少觀察值個數	300
	子節點中最少觀察值個數	150
風 險	估計	6682.513
	標準誤	99.173
結 果	所包含的自變數	A1, A2, A3, A4, A5, A6, A7, B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, B11, C1, C2, C3, C4, C5, C6, C7, C8 (家庭資本變項、學科特定性變項(教師教學與學習策略)、心理學構念(學生學習自我效能、趨向表現與自我概念))
	節點數量	40
	終端節點數量	21
	深 度	5

圖 7-1 2016 年 TASA 國二數學學生學習影響因素模型圖



四、模型的發展脈絡與現象解釋

本次分析結果顯示，影響國二學生數學學習成效最重要的因素在學生學習態度調查中的「我知道我可以將數學學好」。除此之外，「我可以解決數學難題」、「數學是我厲害的科目之一」、「我期待在數學課能夠發現新觀念與挑戰性的想法」、「老師會教我們如何找出題目的條件，並根據題目的條件列式」、「遇到題目時老師會教我們先從特殊例子著手思考」、「老師會請我們求出問題解答之後，教我們進行驗算」，以及「家庭文化資源」（合併計算 A1~A7）皆是影響國二學生數學學習成就的因素。

數學學習成就表現較佳的族群，學習的特徵包含幾項重要因素：1. 學習自信心較強，2. 較能融入數學老師的教學方法，以及 3. 家庭文化資源相對豐富。反觀學習成就較弱的族群，1. 學習自信心相對薄弱與 2. 對於數學老師課堂教學方法較無法融入，以及 3. 家庭文化資源較缺乏。整體來說，國二學生數學學習成就、家庭文化資源、數學教師教學方法與學生學習態度有其相關性，學習成就較佳者，學生學習態度較佳、對於教師教學方法接受度較高、且家中文化資源也較豐富，學習成就較弱則反之。

從分析結構圖節點 1、2、3、4、5、6、7、8 可以發現，影響國二學生數學學習成就表現最明顯的因素為學生學習態度調查中有關學習自信心「我知道我可以將數學學好」、「我可以解決數學難題」、「數學是我厲害的科目之一」及「我期待在數學課能夠發現新觀念與挑戰性的想法」4 個題項，其次為教師教學方法中有關「老師會教我們如何找出題目的條件，並根據題目的條件列式」、「遇到題目時老師會教我們先從特殊例子著手思考」、「老師會請我們求出問題解答之後，教我們進行驗算」3 個題項，以及家庭文化資源多寡（節點 9、10、13、14、15、16、17、18）。觀察國二數學學習成就表現最佳的族群（節點 36: 平均數 611.38），其學習特徵包含：對「我知道

我可以將數學學好」選擇同意或非常同意、「數學是我厲害的科目之一」選擇同意或非常同意、「遇到題目時老師會教我們先從特殊例子著手思考」選擇非常不同意、不同意或沒意見、「老師會教我們如何找出題目的條件，並根據題目的條件列式」選擇同意或非常同意，以及「我期待在數學課能夠發現新觀念與挑戰性的想法」選擇同意或非常同意者，其數學學習成就表現最佳。而國二數學學習成就表現最弱的族群（節點 27: 平均數 405.75），其學習特徵包含：對「我知道我可以將數學學好」選擇非常不同意、不同意或沒意見、「我可以解決數學難題」選擇非常不同意或不同意、「老師會請我們求出問題解答之後，教我們進行驗算」選擇不同意、沒意見、同意或非常同意，以及「我知道我可以將數學學好」選擇非常不同意者，其數學學習成就表現最弱。

綜上所述，相較於學生學習態度，家庭文化資源對於現階段國二學生數學學習成就表現雖然存在正相關（節點 17、18、21、22），但似乎並非主要影響因素，影響學生數學學習的關鍵因素還是在於學習自信心與學習興趣等學習正向態度的建立，詳見表 4 及表 5。

表 7-4 節點增益摘要

節 點	個 數	百 分 比	平 均 數
36	392	3.9%	611.38
39	515	5.2%	587.30
35	151	1.5%	576.02
37	416	4.2%	564.89
40	494	5.0%	551.31
34	790	7.9%	550.69
23	154	1.5%	544.39

節 點	個 數	百 分 比	平 均 數
38	690	6.9%	533.24
19	508	5.1%	525.44
33	687	6.9%	518.71
22	672	6.7%	500.63
29	242	2.4%	498.11
31	505	5.1%	485.87
11	440	4.4%	473.89
15	301	3.0%	462.21
30	260	2.6%	459.26
32	256	2.6%	458.23
9	494	5.0%	453.96
18	379	3.8%	439.91
28	1034	10.4%	430.45
27	585	5.9%	405.75

成長方法：CRT

依變數：第 1 個數學 plausible value

表 7-5 自變數正規化重要性摘要表

自 變 數	正 規 化 重 要 性
我可以解決數學難題	100.0 %
我知道我可以將數學學好	88.6 %
學習數學對我而言並不是什麼難事	84.2 %

表 7-5 自變數正規化重要性摘要表 (續)

自 變 數	正 規 化 重 要 性
我期待在數學課能夠發現新的觀念與挑戰性的想法	76.7 %
數學是我厲害的科目之一	73.3 %
學習數學對我來說非常有趣	55.8 %
一想到數學課的內容很精采，就讓我覺得很棒	30.8 %
家庭文化資源	24.7 %
我期待每一次的數學課	24.6 %
老師會教我們如何找出題目的條件，並根據題目的條件列式	24.0 %
老師會要我們多多練習題目	18.8 %
老師會請我們在求出問題解答之後，教我們進行驗算	14.0 %
老師會教我們找出數學題目關鍵字句，例如：「未知數」是什麼	12.9 %
老師會鼓勵我們彼此討論數學	7.2 %
老師會請我們在求出問題解答之後，再思考有沒有其他可能的解決辦法	6.9 %
老師會鼓勵我們課後複習學過的數學內容	6.8 %
老師會鼓勵我們課前預習數學內容	6.6 %
遇到題目時老師會教我們先從簡單例子著手思考	5.7 %
遇到題目時老師會教我們先從特殊例子著手思考（例如：先想想看如果題目是正三角形、正方形、或特殊幾何圖形，會有什麼結果）	4.6 %
老師會鼓勵我們堅持不懈地解決數學難題	4.3 %

成長方法 : CRT

依變數 : 5 個數學 plausible value 平均值

參考文獻

- 江羿臻、林正昌 (2014)。應用決策樹探討中學生學習成就的相關因素。 *教育心理學報*, 45 (3), 303-327。
- 林俊瑩 (2007)。檢視個人與家庭因素、學校因素對學生學業成就的影響：以 SEM 與 HLM 分析我國國中教育階段機會均等及相關問題 (未出版之博士論文)。國立高雄師範大學，高雄市。
- 黃建翔、蔡明學 (2016)。影響高中職學生學習成就關鍵因素之研究。 *教育行政與評鑑學刊*, 18, 73-98。
- 蔡毓智 (2008)。臺灣地區國中生家庭教育資源結構之探究及其與學業表現之關連 (未出版之博士論文)。國立政治大學，臺北市。
- 謝邦昌、鄭宇庭 (2016)。資料探礦之技術及應用 -Excel 實例演練。臺北市：新陸。
- 簡禎富、許嘉裕 (2014)。資料挖礦與大數據分析。新北市：前程文化。
- Berry, M., & Linoff, G. (2000). *Mastering data mining: The art & science of customer relationship management*. NY: John Wiley and Sons.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth international group.
- Coleman, J. S. (1988) Social capital in the creation of human capital. *American Journal of Sociology*, 94, 95-120.
- Steinberg, D., & Cardell, N. S. (2016). *Methods and systems for automatic selection of preferred size classification and regression trees*. Washington, DC: U.S. Patent and Trademark Office.
- Wang, Y., Peng, H., Huang, R., Hou, Y., & Wang, J. (2008). Characteristics of distance learners research on relationships of learning motivation, learning strategy, self-efficacy, attribution and learning results. *Open Learning*, 23(1), 17 -28.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques (2nd ed.)*. San Francisco, CA: Morgan Kaufmann.

第八章

臺灣學生學習成就評量資料庫 (TASA) 轉型內涵 - 邁向 108 課程 之素養導向大型評量模式

謝名娟、謝進昌

國家教育研究院副研究員

一、課程改革脈絡下之素養導向大型評量需求

臺灣國民教育歷經兩次重大的課程改革，其一是 2000 年實施、2001 年修訂的國民中小學九年一貫課程綱要，其二乃是預計於 2019 年公布及逐年推展的十二年國民基本教育課程，其中，為因應九年一貫課程的推動，教育部頒布了九年一貫課程綱要以取代 1993 年適行的課程標準，而此課程綱要的推動也象徵著教育部引導課程教學邁入新紀元的決心，其內涵是有別於重視學科本位的精神，改以十大基本能力為編纂綱要的導引架構，而在九年一貫課程改革的背景脈絡下，教育部於 2004 年函請國立教育研究院籌備處（2011 年正式成立後，更名為國家教育研究院）規劃「臺灣學生學習成就評量資料庫」（Taiwan Assessment of Student Achievement, TASA）建置，用

於發展量化指標和標準化測量工具，以檢視學生學習成就的表現及其差異，進而了解九年一貫課程實施的成效，協助課程發展之進行與相關教育政策之研擬，在過去近十年內（2008年-2017年），逐年針對四、六、八、十一年級學生進行大規模評量與調查，完整建立了九年一貫課程脈絡下臺灣學生學習成就表現資料庫，而本書內容則是在說明九年一貫課程脈絡下，TASA 資料庫試題評量架構、問卷、抽樣、估計及其應用等層面，然而，對於未來預計推動的十二年國民基本教育課程綱要脈絡下，對於臺灣學生表現評估，乃是未來逐步規劃發展之標的，也是本書最後一個章節所欲著重的焦點。

十二年國民基本教育課程綱要總綱是由教育部於2014年年11月28日正式發布，預計於108學年度起，依序自國民小學、國民中學及高級中學一年級逐年實施（教育部，2014），以漸進推動方式，以取代目前適行的國民中小學九年一貫課程綱要。其中，根據國家教育研究院目前公布的綱要，108課程綱要是在九年一貫課程綱要的基礎上，進一步延伸，其願景是「重視成就每一個孩子，以適性揚才、終身學習」為培養目標，並結合「自發、互動、共好」理念，期使課程發展以生命主體為起點，透過學習者核心素養的培養，讓潛能得以適性發展，進而運用所學、善盡責任，成為終身學習者（教育部，2014, p.1-2）。

十二年國民基本教育總綱推動之精神之一，即為透過持續強化中、小學課程之連貫性與統整性，實現素養導向的課程與教學方法，而透過此路徑使學生能適性發展，培育出具有終身學習、社會關心與國際視野的現代優良國民（教育部，2014）。其中，素養導向的課程、教學與評量為本次綱要改革的重點之一，而根據范信賢（2016）、與楊俊鴻（2015）所提出素養導向教學原則中，就指出教師在素養的教學應調整灌輸式的學科教學模式，可以多使用提問、討論、欣賞、展演、操作或體驗等活動，來啟發學生創造與實踐的機會，而學校教材也應跳脫知識內容的學習，強調以學習歷程與方法，使學生樂在學習。此外，知識學習以實踐導向為主，強調透過觀察現象、尋求關

係來解決問題，最終，促使學生在學校學習中，融入社會的參與，提供多元的情境以展現個人在社會中發展與生活的能力。整體而言，素養導向原則強調的是學生用以適應社會情境脈絡中，所習得的知識、能力、態度與策略。

因應素養導向教學與評量趨勢，由國家教育研究院主導之臺灣學生學習成就評量資料庫（TASA）評量內涵，也隨之面臨轉型需求，其原先內涵是建立在依據九年一貫課程綱要目標，以訂定之評量架構（臺灣學生學習成就評量資料庫網頁，2017），後續，隨著課程改革，乃必須轉換依據十二年國民教育之核心素養，進行架構建立，然而，在進行轉化前，對於「素養評量」的內涵、及其與過去九年一貫課程綱要所訂立的學科評量內涵的異與同，是有其釐清的必要性，而此也是導引本文回應 TASA 轉型內涵的脈絡背景。此外，由於核心素養涉及認知、情意、技能等層面，範疇過於廣泛，較難以於單章節內容中，完全釐清，因此，本章節仍先聚焦以認知層次評量為論述核心。

對於素養的意涵，若根據蔡清田（2011a, b）表示，素養係指為了成為一個健全的個人，人必須因應未來生活情境所需要的知識、能力與態度，其中，能力是不涉及態度、情意層面，僅界於為個人是否具有勝任某些任務能力、與潛力，再者，素養強調著教與價值的功能，是可為學習的成果，得經過後天努力以習得，而能力的養成主要透過父母遺傳、或後天努力。此外，依據不同領域，素養發展包括哲學、人類學、心理學、經濟學及社會學等面向，主要為人面對現代社會中所需要的知識、能力與態度，可用於回應現代社會的需求。若以此回應到教育部所制訂十二年國民基本教育總綱（教育部，2014）、國民中小學九年一貫課程綱要總綱（教育部，2012），其內涵即指出十二年國民基本教育與九年一貫課程綱要均以學生為中心，然而，九年一貫課程綱要重視學生基本能力的養成，然而，十二年國民基本教育則以培育學生的核心素養，成為一個現代公民，在面對生活挑戰時，能具備的知識、態度與能力為導向，其中，在受到社會的進步與變遷，學生的學習必須由能力轉型升級為素養之培養，而在素養導向的課程中，學生能具有連結不

同的領域、或科目的能力，並可以運用所學來解決特定生活情境的問題，對自己的實踐或行動進行反思。

核心素養的推動勢在必行，而其實際作為，就課程教學發展而言，國家教育研究院課程中心已研發各科的課程手冊，供現場老師教學使用，然而，評量工具的研發及其如何符合素養的精神，加諸考量 TASA 屬於大型評量，其施測時間、施測模式等，皆有其限制與規範，是有其必須折衷與協調處，據此，本文希冀藉由探究國際與國內十二年國教的核心素養內涵，再分析國際大型評量的素養形態，以於在 TASA 轉型中，對於素養導向的大型評量要素與方向提出論述，以為 TASA 轉型需求所參考。

二、國、內外對於素養內涵陳述

以下茲就各國際組織對於素養內涵進行說明，再導引自國內對於核心素養之論述，說明如下：

（一）UNESCO Institute for Education（2003）教育五柱

聯合國教育科學文化組織從 2003 年提出教育五柱的概念，認為現代國民的學習應該包含五個部分，分別是：1. 學習求知（learning to know），代表著個人應不斷的學習與融合廣博的普通知識、與專門學科知識；2. 學習做事（learning to do），代表應該具有工作所需的技能與情境中問題處理能力；3. 學會待人（learning to live together），代表尊重不同的文化，能與人相互了解，共同研議與管理人際關係的衝突；4. 學會發展（learning to be），代表重視個人的發展潛能，包括記憶力、推理能力、美感、體能與溝通力等；5. 學習改變（learning to change），代表透過學習以因應社會變遷的能力（UNESCO, 2003）。

UNESCO (1996) 亦提出在學校教育中，應使用不同的方式來增進學生閱讀、寫作、算術、科學與表達的能力，學習應著重在日常生活的情境以學習自然現象與社會互動。

(二) OECD (2005) 三類九項核心素養

經濟合作暨發展組織 (Organization for Economic Co-Operation and Development, OECD) 提出了核心素養的定義，其架構包括三大項，分別是 1. 運用工具互動的素養 (Use Tools interactively)，其內涵包括運用語言和符號，運用知識和資訊與運用科技互動；2. 異質團體中互動素養 (Interact in heterogeneous groups)，其內涵包括人際互動、團隊合作與處理衝突的能力；3. 自主行動素養 (Act autonomously)，其內涵包括依據情境脈絡行動的能力、個人規劃管理的能力與維護自身權益、興趣與需求的能力。

OECD (2005) 對於的素養的定義有以下幾大特點，第一，素養是在特定複雜的情境中，所展現的知識、技能與態度，例如，語言溝通的社交能力是一種素養，因為這必須利用個人的語言知識，社交技能與待人處世的態度；第二，素養是後天習得的，必非與生俱來，學習的環境包括家庭、社會、職業、經濟、政治、文化等多方面；第三，素養是跨領域且多功能，其中，在學校場域，能指稱跨學科，而在學校外，則能跨越不同的社會領域，而多功能係指素養能滿足個人、社會與職業的需求，幫助個人達成目標，解決不同情境的問題。同時，OECD 的國家同時啟動國際學生評比計畫 (Programme for International Student Assessment, PISA)，以提供理論基礎與評量框架。

(三) 美國 P21 四類素養 (2015)

美國 21 世紀學習夥伴 (Partnership for 21st Century Learning, P21) 組織，提出了四類素養 (P21, 2015)，其內涵包括有 1. 關鍵學科與議題 (Key

Subjects and 21st Century Themes），其學科內容包括英語、閱讀、外國語、藝術、數學、經濟、地理、歷史、公民，而議題包括全球意識、財經素養、公民素養、健康素養與環境素養；2. 學習創新技能，內容包括創造力與創新、問題解決、批判思考、溝通與合作等；3. 資訊、媒體及科技技能（Information, Media and Technology Skills），包括資訊素養、媒體素養與訊息溝通的科技素養；4. 生活及生涯技能，包含彈性及適應力、自我管理、社會及跨文化技能、生產力與績效、領導力與責任感。

在 P 21 的組織中，提出素養應超出讀、寫、算的技能，強調應將知識和技能應用在現代的生活情境中，核心學科的知識唯有透過與生活知能結合，才能產生意義。P21 亦提出關於評量的建議，包括應以多元的方式評量學習結果，在課堂中應兼顧行程性與總結性的評量，對於學生表現應給予即時回饋以提升學習成效。此外亦可考量使用檔案評量來評估學生的學習表現（P21, 2015）。

（四）歐盟核心素養

歐盟將核心素養定義為八大項，分別是 1. 母語溝通（communication in the mother tongue）、2. 外語溝通（communication in foreign language）、3. 數學素養和科學與科技基本素養（mathematical competence and basic competencies in science and technology）、4. 數位素養（digit competence）、5. 學習如何學習（learning to learn）、6. 社會與公民素養（social and civic competences）、7. 創發與創業精神素養（sense of initiative and entrepreneurship）、8. 文化意識與表達素養。其中，這八大素養都包含了批判思考、創造力、創發力、問題解決、風險評估、決定力及建設性管理力等情意態度（European Commission, 2006; 葉坤靈，2017）。

葉坤靈（2017）指出，歐盟的核心素養定義及其相關內涵仍流於抽象，如何能根據課程內容來彰顯核心素養的內涵，拆解原先的核心素養，成為可

評量、具有意義性、並可用於評估學生學習成果之評量規準，仍需再進行探究，而在現有的歐洲架構下，歐盟成員國已發展出可行的核心素養評量方法包括標準化測驗、實作評量、檔案評量、近年來更重視學生的數位學習歷程檔案的製作 (Pepper, 2011)。從歐盟的做法有兩項可以做為我國的參考，其一，核心素養需要拆解為較為細緻的是核心素養，以符應相關課程、教學與評量的活動。其二，評量的方法應需適當的獲取素養所涵蓋的知識、技能和態度情境脈絡資料，以作為形成性與總結性評量的基礎。

(五) 臺灣十二年國教的核心素養

十二年國教所定義的核心素養為一個人為適應現在生活及未來挑戰，所應具備的知識、能力與態度。在十二年國教中，每一位接受十二年國民基本教育的學生，所應具備的基本且共同的素養，代表著各級各類學校的學生所應培養的最低共同要求 (潘文忠, 2014)。對於核心素養的內涵建構頗為完整，包括三個面向九大項目，簡述如下：

1. 溝通互動 (communicate interactively)

在此面向著重強調運用不同的工具，與外界的積極的互動，可能透過文字符號、資訊媒體或是美感表達來呈現 (教育部, 2014)。

2. 社會參與 (social participation)

在一個日益相互依賴的世界中，個人需要處理多元社會的多樣性，並且與人建立新的合作形式以及建立適宜的人際網絡以累積社會資本 (social capital)，個人亦需要發展與人及群體互動的能力，而這也是一種社會能力與跨文化能力。社會參與層面涵蓋公民責任與道德實踐、人際關係與團隊合作、國際理解與多元文化等具體內涵 (教育部, 2014)。

3. 自主行動 (act autonomously)

在此面向包括身心素質與自我精進、系統思考與解決問題、規劃執行與創新應變（教育部，2014）。

在各細項之定義如表 1。從細項的定義來看，大多數的面向是不容易評量的，過去以學科為導向、能力技能方面的評量，已經行之多年，然而，在十二年國教的核心素養中，有許多面向是態度、情意方面，屬於主觀、抽象的概念，要如何把這些概念具像化、客觀化，為評量的一大考驗。

在這樣的定義當中，可看出這些核心素養的概念多為指標性的、抽象性的定義，要能落實成為具體可評量的指標，仍需再經過轉化。舉例來說，在身心素質與自我精進方面，其內涵為具備身心健全的素質，透過自我的精進、探究人生的意義，規劃生涯的發展。試想要如何以一個客觀的評量工具來評估學生是否達到自我精進、能夠探究個人意義、規劃生涯發展的能力？以法國為例（Pepper, 2011; 葉坤靈，2017），透過各國對於核心素養的理解，將歐盟的自主與創發素養將以拆解，其拆解的過程中先定義具體評量項目為能自我評估並寫出自我的興趣和所獲得的能力，而後更具焦在學校環境中，學生在此素養的體現為使用學科工具、確認自己的優缺點、確認學校活動興趣、預期評量結果的衝擊、確認日常生活有益的活動等。

表 8-1 108 課程綱要三面九項內涵

面向	項 度	具 體 內 涵	評 量 的 重 點
自主行動	身心素質與自我精進	具備身心健全的素質，能透過自我的精進、探究人生的意義，規劃生涯的發展。	生涯規劃與人生意義。

面向	項 度	具 體 內 涵	評 量 的 重 點
自主 行動	系統思考與 解 決 問 題	具備問題解決、邏輯推理，思辨批判，能處理生活的問題，並自我反思、進行後設思考。	批判思考能力問題解決能力
	規劃執行與 創 新 應 變	具備規劃與執行的能力，能夠應用生活經驗，發揮創意來符應社會快速變遷，學生的適應力。	規劃執行的能力創新與應變的能力
溝通 互動	符號運用與 溝 通 表 達	能在日常生活及工作上運用語言、文字、數理與肢體藝術等方式來進行表達溝通與互動。	溝通能力
	科技資訊與 媒 體 素 養	善用科技與媒體應用的能力，培養倫理及媒體解讀的能力，且能分析思辨與批判科技、資訊媒體之間的關係。	科技運用的能力媒體識讀的素養能力
	藝術涵養與 美 感 素 養	具備有藝術感知、欣賞文化、透過生活美學、賞析、建構與分享美善的人事物。	具有美感素養的能力
社會 參與	道德實踐與 公 民 意 識	具備道德實踐的素養，能主動關心與參與社會的相關議題與活動，關懷自然生態與人類的永續發展，展現之善與行善的美德。	具有道德與公民的意識能力
	人際關係與 團 隊 合 作	能友善關懷他人，建立良好的互動關係，並能與人溝通協調、包容異己，發展團隊合作的能力。	能與人關懷互動具團隊合作的能力
	多元文化與 國 際 理 解	能尊重欣賞不同的文化，關心全球的脈動與情勢，並能順應社會的需求，發展國際理解、多元價值的胸懷。	培育多元文化的能力具有國際觀

註：摘錄自教育部（2014）。十二年國民基本教育課程綱要總綱。2017年4月1日擷取自網址：
<http://12cur.naer.edu.tw/upload/files/96d4d3040b01f58da73f0a79755ce8c1.pdf>，p.4-6。

三、核心素養評量內涵

十二年核心素養推動之後，其導引議題在於素養要如何評量？過去的段考題目不是素養題嗎？抑或是評量試題是否一定要是非選擇題，才能考得出學生的素養能力？等，而由前述關於國內外核心素養評量的回顧可看出，素養的展現，重視生活情境的應用。要能全面的評估不同的核心素養，單純紙筆評量模式是有其難度，而大多數的面向，不是短期的外顯行為，而是長期的內隱行為。需仰賴教師，透過素養導向的教學，內化後將評量融入在課堂的教學之中，不會只有單次性的、短期的評量，而是多次性的、長期性的評量，不會是只有過去只重視學生是否習得知識和技能，同時也重視學生在學習中的態度與情意面向。教師可採用多元的評量方式，如可以使用表現本位評量（performance-based assessment），來評量學生共同合作以解決問題的能力；評量學生展示、實驗、團隊工作、訪談、角色扮演等能力。或採用卷宗評量（portfolio assessment）以評估學生的知識、技能與態度，運用在適當情境脈絡中的程度來進行評價，了解學生在某種學習項目上進步或改變的情形。或透過發展學生自我評量（self-assessment）的能力，由學生自己確認自己的學習結果，對於自己學到什麼樣的程度能進行自我的判斷。透過管理與控制自己的學習歷程，發展學生自我評量的能力，並提升「後設認知」（meta-cognitive）技能等（潘慧玲，2016）。

因此，在十二年國教的素養評量可彙整出以下幾個特點：

- （一）不僅評量學生的知識與技能，而且還有評量學生對於學習的態度。
- （二）不單只重視學習結果，也重視學習歷程。同時兼顧總結性與形成性的評量。
- （三）強調對於學生能整合所學並應用於生活情境的評量。

若以層次階段來畫分，素養導向的評量可分為不同的層次，如表 8-2 所

示，由張茵倩、楊俊鴻（2015）關於核心素養模式的分析可一窺究竟，其中，由最低層級的素養為學科素養，「學科素養」係由各個領域或科目內部的特性或重要內涵所發展出來的素養，例如：語文素養、數學素養或科學素養等。而「核心素養」則為較高的層級，係指一個人為適應現在生活及未來挑戰，所應具備的知識、能力與態度，例如：溝通表達、團隊合作、幸福感等。而核心素養的培養係奠基在學科素養的培養之基礎上。

表 8-2 不同層級之素養評量模式

模 式	說 明	例 證
模式一： 學科本位模式	關注單一學科素養的學習，與核心素養的關係較弱。不過，學科素養卻是培養核心素養的重要基礎。	例如：PISA2000-2012 數學素養、科學素養；國教院教學模組：指數率、熱傳播
模式二： 添加模式	在單一領域 / 科目或學科素養的學習過程中，開始關注到若干核心素養項目的培養。	例如：PISA2015 科學素養：合作式的問題解決
模式三： 折衷模式	單一領域 / 科目或學科素養與核心素養的學習並重，強調在兩類素養之間做有機的結合。	例如：公民意識與現代社會、系統思考與科學研究
模式四： 科際整合模式	以若干核心素養項目的培養作為重點，強調跨領域 / 科目的整合。	例如：PISA2012 問題解決、PISA2018 全球素養
模式五： 全人模式	強調透過學校教育（正式、非正式及潛在課程）、家庭教育及社會教育，全方位培養核心素養。	例如：PISA2018 青少年幸福感（草案）、學生學習歷程檔案評量

註：擷取自“十二年國民基本教育課程綱要之課程轉化與實踐：以臺南市保東國小為例”，張茵倩、楊俊鴻（2015，12月），載於國家教育研究院舉辦之「2015 邁向十二年國教新課綱的第一哩路」學術研討會論文集（頁 187-203），臺北市。

四、TASA 轉型內涵

在前述素養內涵論述、及其轉化為評量內涵說明背景下，臺灣學生學習成就評量資料庫（TASA）是有其轉型必要性，自目的、評量模式等，茲說明如下。隨著 107 課程綱要的實施，TASA 轉型之目的，大致分為以下幾點：

（一）比較接受 99 與 108 課程綱要之世代學生成長之追蹤設計

因應 108 課程綱要推展，TASA 將進行長期追蹤設計。同時追蹤接受不同課程綱要之世代學生之成長表現，並評估其成長的幅度差異。有關縱貫性追蹤調查方面，其概念如圖 1 所示，第二年時，則是直接以第一年下半年所選取接受 99 課程綱要之 7 年級升至 8 年級學生，進行追蹤調查，第三年時，則是直接針對第二年所選取接受 108 課綱之 7 年級升至 8 年級學生，進行追蹤調查，以期探究國家課程轉換脈絡下，接受不同課程綱要之世代學生的成長差異。

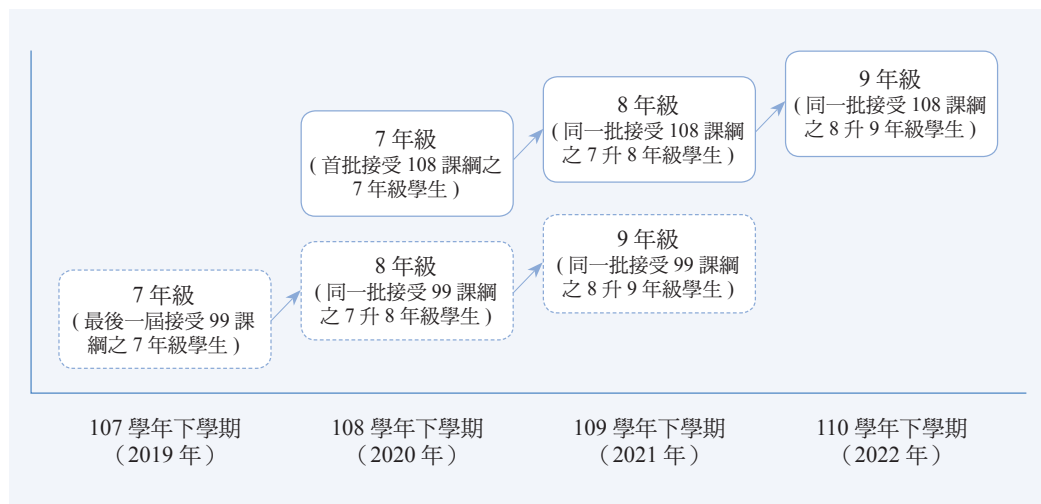


圖 8-1 TASA 轉型之長期追蹤設計示意圖

(二) 探討接受不同課程綱要世代學生對於關鍵素養或能力學習效果

108 課程綱要的內涵是一新概念，然而，其部分內容也是 99 課程綱要延伸發展而來，因此，TASA 轉型的規劃中，將針對新修的課綱內容加以強調，例如在新課綱中，修改數學的比率、幾何概念的調整等，英語文中增加邏輯思考、創造力等概念，這些都會是 TASA 未來轉型評估的重點。

(三) 探討背景與教學變項，以評估對學生素養表現影響。

透過長期追蹤設計，探討學生背景或教學變項（如學習策略、教學策略實施與否、學習態度），以探究不同變項對於對素養表現的關聯。

(四) 評估不同課程綱要的銜接教材

在推展 108 課程綱要之際，在九至十年級階段，教科書將推動銜接教材，使學生能順利接受新課程綱要，因此，TASA 轉型目的之一，也會針對銜接教材的部分進行評估。

(五) 接軌國際大型評量進行學生表現結果等化

在部分的施測年段，施測的樣本會與 PISA 和 TIMISS 重複，未來會考量與國際調查接軌和成果比較的可能性，然而，這部分尚需向國際評比組織提出申請，視國際評比組織回應再行規畫。整體而言，TASA 的主要目的為提供課綱的回饋，了解十二年國教不同課程綱要間轉換的差異，提供下輪課程綱要修訂的參考，其目的並非教室內的素養評量，提供給課堂老師作為學生的即時回饋，雖然教室內的素養評量是重要且關鍵的。另外，由於 TASA 的調查母群為全台灣的學生，並進行大規模的抽樣，考量施測成本與方式，很難將多元評量（如實作、檔案、觀察）等方式融入。僅能在學校所限定的施測時間內，進行短暫的、單次性的資料蒐集，然而，在此侷限下，TASA

僅能部分的評估 108 課程綱要的實施成效。而 TASA 轉型之素養題型研究方向，會朝向以下幾個方面：

1. 以生活情境化強調思考理解歷程為命題之目標導向

從十二年核心素養的精神來說，不管在三面九項的任何方面，都是需要有特定情境實踐，這種趨勢也和國外的核心素養框架相符合。因此在未來 TASA 的命題方面，將著重真實的情境與真實的問題，強調思考理解歷程及其應用在生活情境上的能力。過去的紙筆測驗，重視的多為知識和理解等認知層次，而素養導向的命題，則強調以思考、理解來解決日常生活情境或是學術的探究情境所會關注的問題。以下圖 8-2 英語文素養範例題為例，情境是學生常見生活廣告影片，廣告拍攝是描述一位受到家暴婦女尋求援助，然，因為種種因素無法實際說出求救本意，幸好碰到一位善聽接線生，理解該名婦女本意，但影片未顯示無法克服不良收訊影響，仍失去聯繫，而究其影片本意，顯示是廣告商為突顯電信公司品質所拍攝。據此，第一題為評量學生是否能直接找出訊息（locate explicitly stated information），指出文本內基本故事架構，例如人、事、時、地、物。

Q1 Who is the man receiving the call?

- ① a cook
- ② a mail man
- * ③ a 110 operator
- ④ a pizza delivery boy

第二題為評量學生是否能理解婦人打電話之本意，連結文本內訊息與自身背景知識，經詮釋與整合（interpret and integrate ideas and

information) 以理解文本細節關聯、及掌握訊息所欲顯示深層意涵 (implied or figurative meaning)。

Q2 Why does the woman call the man?

- * ① She called to ask for help.
- ② She called to order a pizza.
- ③ She called to report a lost child.
- ④ She called to ask a phone number.

第三題為評量學生是否能了解該影片是電信商廣告，其實際目的為增強該公司之正面形象，藉以吸引更多顧客，評量學生是否能反思評估訊息，連結背景知識以反思文本內容，以評估作者立場、觀點、事件發生可能性、及文本正確性。

Q3 What is the main purpose this video shot for?

- ① To tell people where is the pizza store.
- ② To tell people how to dial a proper call in need.
- ③ To tell people where to get more information about Bruce Telecom.
- * ④ To tell people what kind of company Bruce Telecom is.

以下圖 8-3 數學素養範例題為例，情境素材為學生日常生活中常見的斑馬線，斑馬線是生活中常見的幾何圖形，工程人員在繪製斑馬線除了利用工具外，過程中面臨的問題會應用數學素養去解決，才能畫出符合規範的斑馬線。這些試題的評量目標為學生是否能以數學的觀點，將生活常接觸的事物與數學相連結，屬於評量數學核心素養 A2 (系統

思考與解決問題)的試題。據此，第一題評量學生如何繪製出規範中傾斜 45 度的斜紋線，評量學生認知層次中之連結能力，亦即學生利用等腰直角三角形的性質，找出標記符號的對應關係才能解決此題。

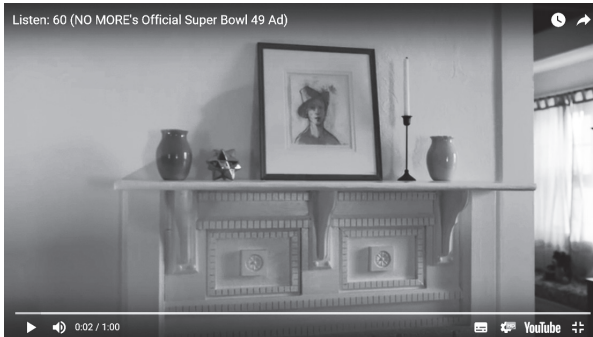
Q1 判斷阿明在畫斜紋線時，他須用哪一個線段當作第一個斜紋線的邊？

- (A) $\overline{a_2b_{20}}$
- (B) $\overline{a_2b_{21}}$
- (C) $\overline{a_2b_{22}}$ *
- (D) $\overline{a_2b_{23}}$

第二題評量學生是否能計算白色區域面積，在真實情境中相對應的問題是工程人員需評估完成一個工程需多少白色塗料。評量學生認知層次的分析並發展策略，學生須先分析情境條件，提出所發展出計算白色區域面積的策略，才有辦法解決此問題。

Q2 試求阿明在此次工程畫出的白色區域面積約為多少平方公分？

解答：分成右上三角形區域、左下三角形區域、中間平行四邊形區域，之後便可觀察到白色區域面積為全部矩形的一半，即約 8000 平方公分。



Phone dialing sound...

Man: This is 110. Where is the emergency?

Woman: 20 Zhongshan street.

Man: Okay, what's going on there?

Woman: I'd like to order a pizza for takeout.

Man: Ma'am, you are calling 110. This is a line for urgent and serious case.

Woman: Yes. Large with half chicken half tomato.

Man: Um, you know you call 110.

Woman: Do you know...how long it will be?

Man: Okay, ma'am. Is everything okay over there?

Woman: No.

Man: Um... you aren't able to talk ...?

Woman: Right.

Man: Okay. Is someone in the ... with you? Just say yes or no.

Woman: ...Yes.

Man: Okay. I have a police ... a mile... I need you...on the... (lost connection due to interference)

Woman: Hello...

Man: Hello... are you still there? Hello

End of this video, then showed the following script and paused.

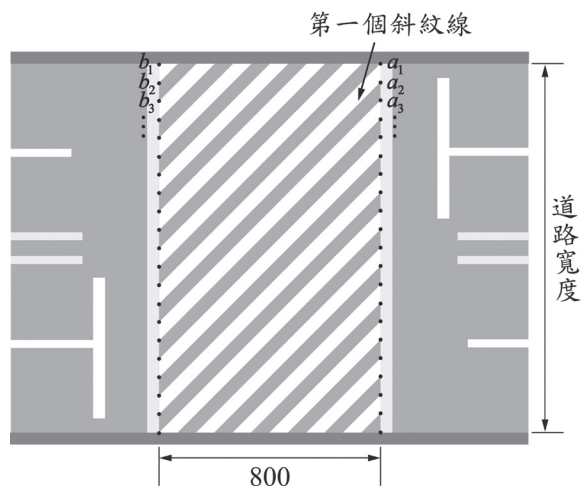
Always keep connected. This may change your life.

For more, please call 800-331-0500 or visit www.telecom.com.tw



圖 8-2 英語文素養範例題

行人穿越道，又稱斑馬線，是一種繪在馬路路面上的交通標線，一般行人穿越道用塗料在路口等特定地點繪製成相間的條紋。烏烏國對斑馬線的規定：它是由兩條橫跨道路的黃色平行線與內插白色斜紋線所組成，垂直路邊的兩平行線相距 800 公分，而斜紋線依行車方向自左上方向右下方傾斜 45° ，每個斜紋線的寬度與間隔均相等，其示意圖如下所示。



阿明被指派在一條寬度為 2000 公分的道路繪製斑馬線，為避免斜紋線不符合規定，他先在黃線標上記號，相鄰兩記號相距為 40 公分，在右方黃線標上的記號，由上至下依序命名為 a_1, a_2, a_3, \dots ，在左方黃線標上的記號，由上至下依序命名為 b_1, b_2, b_3, \dots 。最後，以 $\overline{a_1 a_2}$ 為第一個斜紋線的一邊，並將所有斜紋線畫完。

圖 8-3 數學素養範例題

2. 發展主軸以學科模式為主、折衷模式為輔之評量模式

真實世界中的問題是沒有學科領域的界線，在未來 TASA 的素養評量發展上以多元表徵、將閱讀理解的能力融合在不同學科，以及系統思考的能力，TASA 未來將發展具有跨學科甚至跨領域的素養題型。然而，這種跨學科的評量模式雖然在學理上具有重要性，但若以選擇題型來進行時，容易造成學生能力估計上的困難，試想若是數學試題和閱讀試題融合，在數學的試題需大量的閱讀知識才能解題，那在學生的能力估計上要如何置放閱讀和數學兩者的權重的？到底這樣的跨學科試題評量的是數學還是閱讀？還是融合兩者？在學生能力估計上的難題，仍需未來再做深入的探討。

3. 兼顧知識、能力與態度之評量

施測部分仍分為各學科素養測驗，過去 PISA 國際調查認為，透過生活化的測驗情境可以同時測得學生的能力加上態度，因為學生必須將所學內化成習慣，時時應用於日常生活中，才能在這類題目中有好的表現。在 TASA 的調查中，除了將生活情境應用在不同題型中，在態度與情意上的評量亦輔助問卷方式進行調查，以期能更多方了解影響學生學習成效的相關因素。

4. 以線上評量為主、紙筆評量為輔之施測模式

受限於數學非選題會有公式運算，學生不易在電腦上操作的狀況，在其他科目均會以線上施測的方式進行，除了減省成本之外，亦可較為快速、準確的進行資料蒐集，另外，線上施測也能有題型變化，如科學的探究性試題，可以透過多媒體的互動，讓學生實驗的操作，或是英語文的數位影音試題，數位閱讀試題等，這些多媒體的試題，能更廣泛的、多元的評估學生的素養能力。

5. 持續進行 108 課程綱要核心素養的研究，將指標轉化成可觀察的評量指標

在 108 課程綱要的推動下，雖然在課程的發展上已見雛形，但是在評量的模式上仍存在困難需要突破，許多的核心素養指標，都相當抽象難以評量，例如在情意與態度方面，要如何要具象化，以體現在紙筆評量中，還需更多測評專家和課程專家的共同努力。

參考文獻

- 范信賢 (2016)。核心素養與十二年國民基本教育課程綱要：導讀《國民核心素養：十二年國教課程改革的 DNA》。國家教育研究院教育脈動電子期刊，5，1-7。
- 張茵倩、楊俊鴻 (2015，12 月)。十二年國民基本教育課程綱要之課程轉化與實踐：以臺南市保東國小為例。載於國家教育研究院舉辦之「2015 邁向十二年國教新課綱的第一哩路」學術研討會論文集 (頁 187-203)，臺北市。
- 教育部 (2012)。國民中小學九年一貫課程綱要總綱。2017 年 4 月 1 日擷取自網址：http://teach.eje.edu.tw/9CC2/9cc_97.php
- 教育部 (2014)。十二年國民基本教育課程綱要總綱。2017 年 4 月 1 日擷取自網址：<http://12cur.naer.edu.tw/upload/files/96d4d3040b01f58da73f0a79755ce8c1.pdf>
- 楊俊鴻 (2015)。核心素養在教學現場的展現。載於洪詠善、范信賢主編，同行：走進十二年國民基本教育課程綱要總綱 (16-17 頁)。新北市：國家教育研究院。
- 葉坤靈 (2017)。由歐盟核心素養的評量審查我國中小學核心素養評量之相關議題。臺灣教育評論月刊，6 (3)，7-14。
- 臺灣學生學習成就評量資料庫網頁 (2017)。臺灣學生學習成就評量資料庫建置計畫。2017 年 4 月 9 日擷取至網頁：<http://www.naer.edu.tw/files/11-1000-1408-1.php?Lang=zh-tw>
- 潘文忠 (主編) (2014)。十二年國民基本教育課程發展建議書。新北市：國家教育研究院。

- 潘慧玲（主編）（2016）。十二年國民基本教育普通高中課程規劃及行政準備手冊。新北市：國家教育研究院。
- 蔡清田（2011a）。課程改革中的「素養」（competence）與「知能」（literacy）之差異。教育研究月刊，203, 84-96。
- 蔡清田（2011b）。素養：課程改革的DNA。台北市：高等教育。
- European Commission.(2006). *Key Competences for Lifelong Learning: European Reference Framework*. Retrieved from <http://www.erasmusplus.org.uk/file/272/download>.
- OECD(2005). *Definition and Selection of key competencies: Executive summary*. Retrieve 9 April, 2017 from <https://www.oecd.org/pisa/35070367.pdf>
- Partnership for 21st Century Learning[P21](2015). *P21 Framework Definitions*. Retrieved 9 April, 2017 from <http://www.p21.org/our-work/p21-framework>
- Pepper, D.(2011). Assessing key competences across the curriculum and Europe. *European Journal of Education*, 46 (3), 335-353.
- UNESCO (1996). *Learning: The treasure within*. Retrieved from <http://unesdoc.unesco.org/images/0010/001095/109590eo.pdf>
- UNESCO (2003). *Nurturing the treasure: Vision and strategy 2002-2007*. Hamburg, Germany: Author.

大型教育調查研究實務：以 TASA 為例 / 蕭儒棠等著 . -- 初版 .
-- 新北市：國家教育研究院，民 106.12 面；公分
ISBN 978-986-05-5120-4（平裝）
1. 教育測驗 2. 資料庫
521.3029 106025427

書名：大型教育調查研究實務：以 TASA 為例
著者：蕭儒棠、曾建銘、謝佩蓉、黃馨瑩、吳慧珉、陳冠銘、蔡明學、
謝名娟、謝進昌
出版機關：國家教育研究院
地址：新北市三峽區三樹路 2 號
網址：<http://www.naer.edu.tw>
電話：(02) 7740-7890
出版年月：民國 106 年 12 月
版次：初版
其他類型版本說明：本書另有電子版本，網址為：<http://teric.naer.edu.tw>；
附錄網址為：<http://trac.naer.edu.tw/106TASA>
定價：新臺幣 550 元
展售：政府出版品展售中心
· 五南文化廣場：臺中市中山路 6 號
電話：04-22260330；傳真：04-22258234
網址：<http://www.wunan.com.tw/>
· 國家書店松江門市：臺北市松江路 209 號 1 樓
電話：02-25180207；傳真：02-25180778
網址：<http://www.govbooks.com.tw/>
GPN：1010602797
ISBN：978-986-05-5120-4
美術設計：尚暉文化事業有限公司
地址：新北市板橋區板新路 103 號 4 樓之 1
電話：(02)2958-6010

