TIMSS 2019 研究設計與資料分析（1）

蕭儒棠

國家教育研究院 測驗及評量研究中心

【轉載自：國際數學與科學教育成就趨勢調查 2019 國家報告第三章 (P.79-99)】

TIMSS 是每四年為一個調查週期的國際大型教育調查，在對學校、教師和學生產生最小干擾的原則下，對各國四、八年級學生的學習成就，提供有效且可靠的調查結果。其研究設計著重的是各國四年級學生（和八年級學生）在不同調查週期之間數學及科學學習成就的變化趨勢，而不在四年級與八年級學生的學習成就比較。TIMSS 2019 採用二階段隨機抽樣設計，第一階段由全國的學校中抽取學校樣本，第二階段則由從每個樣本學校中抽取一個或多個完整班級的學生。考量課程安排或教學活動皆以班級為單位，第二階段抽樣不以學生而以班級為單位，其目的在提供學生的學習與課程內容或教學經驗方面的訊息。此外，以班級為抽樣單位在執行上對學校、教師與學生的干擾也較少。本章第一節的重點在說明 TIMSS 2019 的抽樣設計與權重，內容包含國家抽樣計畫擬定、目標母群定義及規範，以及國家抽樣設計等。

TIMSS 採用嚴謹的學校和班級抽樣技術，透過抽樣學生的作答結果，估計該國整體學生母群的學習成就。TIMSS 的調查四、八年級學生的數學和科學學習成就，它的調查母群有兩個，一個是就讀四年級全體學生，另一個則是就讀八年級全體學生。參與 TIMSS 調查的國家可以選擇參加針對兩個或其中一個學生母群的調查。本章的第二節說明我國 TIMSS 2019 之母群抽樣分布，其中第一部份為我國參與 TIMSS 2019 之四、八年級 eTIMSS 抽樣，第二部分則為我國參與 TIMSS 2019 之四、八年級橋接測驗之抽樣。

TIMSS 2019 是調查形式由紙筆測驗（paperTIMSS）過渡至數位化測驗（eTIMSS）的調查週期，參與調查的國家可以選擇參與 paperTIMSS 或 eTIMSS。儘管這兩種作答方式的試題內容盡可能相似，但作答方式不同，二者之間不可避免地存在某些差異。TIMSS 是學習成就趨勢調查，無論是 TIMSS 2019 調查中兩種作答模式的結果連結，或是 TIMSS 2019 與過去的調查結果連結，都必須透過所謂的橋接資料（bridge data），以了解並控制作答模式對調查結果的影響。為了蒐集橋接資料，2019 年參與 eTIMSS 的國家中，受測學生在電腦或平板介面作答沿用自 TIMSS 2015 的趨勢試題，在相同調查期間，這些國家另外有一批學生以紙筆測驗形式完成相同的 TIMSS 2015 趨勢試題，透過這個方式比較可估計相同試

題在 paperTIMSS 和 eTIMSS 的差異。本章的第三節針對 TIMSS 2019 的橋接測驗與 eTIMSS，說明二者在試題分配、答對率以及平均量尺分數的差異。

第一節 抽樣設計與權重

一、 國家抽樣計畫擬定

參與 TIMSS 的國家需要定義其調查的目標母群，並依據 TIMSS 的抽樣設計，選取具全國代表性的學校和學生樣本。TIMSS & PIRLS 國際研究中心與加拿大統計局以及 IEA 位於德國漢堡的資料處理中心合作，為各國提供了一系列抽樣的操作手冊，確保各國完成符合國際標準的國家抽樣。而各國的國家抽樣設計，則由各參與國家的國家計畫協調人（National Research Coordinator，簡稱 NRC）和 TIMSS 的抽樣專家合作完成，並與加拿大統計局（Statistics Canada）共同確認國家抽樣計畫是否符合 TIMSS 標準。

加拿大統計局、IEA 資料處理中心與各國的 NRC 合作，確認各國的學校抽樣架構（從中抽取學校樣本的學校母群清單）是否完整；檢查被排除學生的類別定義是否合理、明確；確定樣本數量和分層計畫是否符合國際與國家目標等。每次 NRC 會議召開期間，加拿大統計局和 IEA 資料處理中心的抽樣人員，與 NRC 詳細討論各國的抽樣步驟與紀錄，同時也就各國的特有的問題，討論適合的解決方案。

NRC 應確定與國際目標母群對應的年級；完成母群中所有目標年級學生就讀的學校清單，作為抽樣架構之用；根據 TIMSS 國際規範，確定全國母群的覆蓋率和排除率；與加拿大統計局合作完成國家抽樣計畫並確定合適的分層變項，確保分層變項存在並適用於所有學校；聯絡抽樣學校並確保學校參與意願；追蹤樣本學校和替代學校的參與情況；完成學校內的班級抽樣。由調查前的規劃乃至調查過程的追蹤，每個步驟都必須填寫一系列的表格，記錄每項任務的完成情況，並將記錄送交加拿大統計局。最後，TIMSS & PIRLS 國際研究中心，則根據加拿大統計局的建議，確認各國的國家抽樣計畫是否符合 TIMSS 的規範。

二、 目標母群定義

TIMSS 的目的在學生數學和科學學習成就的國際比較，其國際調查母群是所有在學的四年級和八年級學生。依據聯合國教科文組織（United Nations Educational, Scientific and Cultural Organization，簡稱 UNESCO）的國際教育標準分類（International Standard Classification of Education，簡稱 ISCED）（UNESCO, 2012），四年級和八年級對應的是國際教育標準分類中，第 1 級教育分類的第四年與第八年。此外，考量學生在認知領域的發展，TIMSS 也對受測學生加上平均年齡下限的限制。若對四（八）年級學生進行測驗時，

學生的平均年齡小於 9.5（13.5）歲，TIMSS 建議將調查的對象調整至更高的年級。TIMSS 針對四年級和八年級學生的國際調查目標母群定義如下：

- 四年級：就讀 ISCED 第 1 級起第 4 年的學生，調查時平均年齡至少為 9.5 歲。
- 八年級：就讀 ISCED 第 1 級起第 8 年的學生，調查時平均年齡至少為 13.5 歲。

TIMSS 調查的目的是描述四、八年級整體母群學生的學習成就，因此，各國的目標母群應涵蓋所有符合目標母群定義的學生。但在某些情況下，部分的學校或學生可能不得不被排除於各國的目標母群之外，例如，某些學校的課程不同於國家的課程標準，或是部分學生有學習障礙等。若調查時將這些學校或學生排除，將會降低整體目標母群的涵蓋率。TIMSS 訂有學校與學生排除調查的規範，作為各國 NRC 選取樣本時的依據。

（一）學生層級排除規範

- 有功能性障礙的學生：若學生的肢體障礙將影響學生的測驗表現，調查時可排除這類學生。
- 有智能障礙的學生：若學生經鑑定為智力、精神或情緒方面的障礙，調查時可排除這類學生。
- 非母語使用學生：若學生無法閱讀、口說測驗使用的語言，且測驗時無法克服語言障礙，調查時可排除這類學生。

（二）學校層級排除規範

- 學校位置偏遠無法到達訪問。
- 學校規模極小（例如，一個年級的學生數不足 5 人）。
- 學校的年級結構或課程不同於主流教育系統（例如，華僑中學、戲曲學校）。
- 學校招收對象為學生層級排除規範涵蓋的學生（例如，啟聰、啟明、啟智學校）。

（三）國家目標母群涵蓋率

依據 TIMSS 的要求，具有閱讀障礙或其它學習障礙的學生應盡可能納入調查，且不應單純因為學生在學科的表現不佳、一般的行為偏差而在調查時將學生排除。此外，為了使各國的國家樣本能準確代表國家目標母群，各國的 NRC 必須盡可能降低被排除學生的比例。TIMSS 要求：

- 因學生、學校層級排除規範被排除的學生人數不超過國家目標學生總數的 5%。
- 因就讀極小規模學校而被排除的學生人數不超過國家目標學生人數的 2%。

三、 目標母群規範

TIMSS 對抽樣的精確性、調查參與率以及抽樣工作都有其規範，目的在抽取最佳的國家樣本，並據以完成精確、無偏誤的可比較國際調查估計。

（一）抽樣精確性與樣本規模

TIMSS 的抽樣精確性要求，各國的學生樣本應使得全國平均成績的標準誤小於 0.35 個標準差，此標準差所對應的 95%信賴區間之平均分數的誤差在±7 分以內，而所對應的相鄰兩個調查週期的平均分數的誤差則在±10 分以內。

對於多數國家，抽樣學校樣本數 150 所且學生樣本 4000 位，可達到 TIMSS 要求的精確度。例如，若該國的平均班級人數為 27 人，假設所有樣本學校與學生皆參與調查，則 150 所學校將有 4,050 位學生參與調查。一般而言，每個樣本學校抽取一個班級便可達到足夠的精確度，但也有某些國家選擇在每個學校選取更多的班級，藉此提高學生樣本的規模，或更精準地估計學校層級的效應。

一般而言 150 所學校樣本即足夠，若有下列情況，則需要選取更多學校樣本：

- 該國的班級規模普遍較小，選取 150 校，每校選取二個班級以上仍無法達到設定的學生樣本規模。
- TIMSS 過去調查的數據顯示，需要較大的學校樣本規模才能達到設定的抽樣精確度。
- 學校依據學生的表現分班（八年級比四年級更常見）。這種狀況將增加班級間學生成績的差異，並影響抽樣的精確度。因應的方法除了選取更多的樣本學校，也建議每所學校盡可能抽取二個班級的學生參與調查。
- 預期將有高的未作答比例，進而導致樣本流失和樣本規模降低。請注意，儘管較大的學校樣本有助於維持樣本規模，但它無法彌補未作答偏差（non-response bias）。

若參與國準備由 TIMSS 過渡到 eTIMSS，則該國至少必須額外增加 1,500 名學生參與橋接測驗，蒐集過渡到 eTIMSS 所需的數據。橋接樣本的選取有三種方式，方式一，從樣本學校的子集中另外再選取一個班級；方式二，另外再選取不同的學校樣本，或同時運用方式一、二。

儘管 TIMSS 試測進行的時間是實測的前一年，但試測和實測的學校樣本是在相同時間，由相同學校母群中抽取得。試測的樣本規模要求每個試測的成就評量題本需要 200 名學生作答，試測的題本越多，需要的學生數就越多。TIMSS 2019 的 paperTIMSS 的試測中，每個年級有 5 種題本，因此，每個年級需要 1,000 名學生作為試測的樣本。eTIMSS 的試測

則有 5 個試題區塊組合，同樣每個年級需要 1,000 名學生。此外，eTIMSS 還有 3 個 PSI 的試題區塊組合，這三個區塊組合需要再增加 300 名學生。臺灣於 TIMSS 2019 參與 eTIMSS 和 PSI 試題的試測，共需要 1,300 名學生樣本。

（二）學校與學生調查參與率

TIMSS 的目標是樣本學校、班級和學生能 100% 參與，但拒絕參與或缺席仍不可避免。為降低未作答偏差，TIMSS 要求各國的國家樣本必須滿足下列任一條件：

- 最初的學校樣本中，學校參與率達 85%，且
- 最初的學校樣本和替代學校樣本中，班級參與率達 95%（學生參與率低於 50% 的班級視為未參與），且
- 學校樣本和替代學校樣本中，學生參與率達 85% 或
- 根據最初的學校樣本（儘管班級和學生的參與率可能包括替代學校），學校、班級和學生的最低綜合參與率達 75%。

四、 國家抽樣設計

各國的國家抽樣設計由該國的 NRC 負責制定和執行，但加拿大統計局和 IEA 資料處理中心抽樣小組仍與 NRC 密切合作，確保抽樣計畫能納入各國的需求，且符合 TIMSS & PIRLS 國際研究中心設定的標準。TIMSS 的國家抽樣設計採二階段分層叢集抽樣（stratified two-stage cluster sample design），第一階段為學校抽樣，第二階段為學校內部的班級抽樣：

（一）學校抽樣

第一階段的學校抽樣，TIMSS 採用分層抽樣的方式，先依某些重要的人口統計學變項分層、分類或分群（strata），接著以 PPS（probability proportional to their size）的原則，根據調查母群於各分層中的人數，估算各分層的抽樣人數和班級數，最後再以各分層中抽樣人數佔母群的機率，隨機選出樣本學校。透過 PPS 的抽樣方式，學生數較多、規模較大的學校被抽取的機率相對較高，而規模相對較小學校，被抽取的機率則相對較低。

抽樣的分層指的是目標母群中，由某些學校的共同特徵（例如，地理區域、學校類型）形成的群組或階層。TIMSS 使用的學校分層變項包含國家／地區（例如，州或省）、學校類型或經費來源（例如，公立或私立）、教學語言、都市化程度（例如，城市或農村地區）、社會經濟指標，或學校在全國性考試的表現。學校分層可分為顯性分層（explicit stratification）和隱性分層（implicit stratification）二種。

TIMSS 使用顯性分層的主要原因是，抽樣時考量某些與比例無關的變項，目的是為了確保研究變項的教育成果時，該變項涵蓋的學校有足夠的樣本數。例如，為了估計地理區

域的影響時，可以使用按區域劃分的顯性分層，確保每個區域都能選取一定比例的樣本學校。此外，參與國家的學生必須參與國家舉辦的全國性成就測驗時，若以各校學生在成就測驗的平均成績，作為隱性分層的劃分依據，可減少在學校層級的抽樣誤差，提高 TIMSS 估計學生學習成就的精確性。

原則上所有依抽樣設計選取的學校都應參加調查，但並非所有國家都能達到 100% 的調查參與率。為了避免樣本流失，可預先為每所樣本學校設定二所替代學校，作為樣本學校無法參與調查時的備案，避免樣本流失影響代表性。依據學校抽樣架構製作的學校列表，在緊鄰樣本學校的前後各選擇一所學校作為替代學校。替代學校應與原始學校屬於相同的顯性分層，若原始學校是隱性分層的第一或最後一所學校，則替代學校可能來自不同的隱性分層。

儘管使用替代學校並無法完全消除學校不參與調查產生的誤差，但採用隱性分層並依據學校規模排序製作學校抽樣架構，作為替代學校的選取參考，有助於增加替代學校與不參與調查學校間的相似性。TIMSS 設置替代學校的目的，既可維持所需的樣本規模，也能確保有足夠的樣本量用於分析母群中某些群體的差異。

參與 TIMSS 調查的國家若選擇同時參加四、八年級的調查，學校抽樣時可依據該國的學校管理特性，彈性選擇四、八年級的調查學校是同一所或不同學校。

（二）班級抽樣

完成學校抽樣後，再由符合調查目標的班級清單中，以 WinW3S 程式，系統性隨機抽樣的方式選取一個或多個完整的班級。選取校內的班級時，每個班級以等機率方式隨機抽樣，原則上 1 校抽 1 個班級，若學校內普遍班級人數較少，則可抽取 2 個班級，或將校內某些班級合併為一個虛擬班級（pseudo-class）後再做校內的班級抽樣。

對於需要橋接樣本的 eTIMSS 國家，則需要額外的抽樣步驟。橋接樣本必須盡可能與 eTIMSS 主測的樣本相近，參與的學生則作答趨勢試題區塊的紙本試題。此外，由於兩者的試題內容相近，同一個班級的學生並無法同時接受 eTIMSS 與橋接測驗。因此，應由 eTIMSS 樣本的學校或另一所學校中，另選一個班級作為橋接樣本。eTIMSS 與橋接測驗的樣本選取同樣使用 IEA 資料處理中心和加拿大統計局開發的校內抽樣程式（WinW3S）。

（三）抽樣權重

樣本的權重表示其母群代表性，權重的大小與樣本在母群中被選取的機率成反比，樣本抽取機率越小者其權重越大，其母群代表性越大；樣本抽取機率越大者其權重越小，其母群代表性也越小。若樣本的選取由簡單隨機抽樣所取得，則所有的樣本皆具有相同的抽樣權重（sampling weights），然而，TIMSS 採取的二階段分層叢集抽樣屬於複雜抽樣（complex

sampling method），此時樣本的抽樣權重並不完全相同。

TIMSS 的學生抽樣權重是學校、班級、學生三種權重的組合，這三種權重各自反映學校、班級和學生被選取的機率和抽樣的結果。在每個層級上，權重的計算方式都是，該層級被抽取機率倒數，再針對未作答情況作調整。每個學生的整體抽樣權重則是學校、班級（校內）和學生（班內）這三種權重的乘積。

在 TIMSS 中，一個國家的每個目標母群會搭配一套抽樣權重，同時參與四、八年級的國家則會有二套抽樣權重。然而 TIMSS 2019 的題本設計中導入數位化的 PSI 試題，因此，若學生只作答一般數位化試題，其權重標記為 TIMSS weights，對於 eTIMSS 和 paperTIMSS 國家而言，其抽樣權重皆為 TIMSS weights；若學生作答的試題包含 PSI，其權重標記為 TIMSS+PSI weights。

由於 TIMSS 調查時，並非每位受調查的學生被抽取的機率均相等，為了能正確地推估母群參數，必須對樣本資訊進行正確的加權，在 TIMSS 資料庫中，學生權重變項名稱為 TOTWGT，分析學生層級資料時，使用此權重才能正確推估母群參數。除了使用 TOTWGT 可以用在對於各參與國家內部學生母群參數的推估外。TIMSS 資料庫也提供一些不同的權重，應用於某些特定的資料分析。表 3-1 為 TIMSS 資料庫中，主要的權重變項以及這些權重變項的使用時機（Joncas & Foy, 2012；任宗浩、陳冠銘，2018）。

表 3-1　TIMSS 2019 權重變項、適用時機與研究層次之關係

變項名	全名	適用時機	研究層次
TOTWGT	Total Student Weight	各國研究	學生
SENWGT	Student Senate Weight	跨國比較	學生
HOUWGT	Student House Weight	顯著性檢定	學生
TCHWGT	Overall Teacher Weight	教師與學生	學生
SCHWGT	School-level Weight	學校分析	學校

（四）標準誤估計

考量國際調查的執行成本與現實，國際大型教育調查並不可能以普查的形式進行，簡單隨機抽樣在執行上也有許多困難。為了考量執行實務上的困難，又能獲得精確的學生能力估計，TIMSS 2019 運用複雜的抽樣技術，由各國的四年級和八年級學生母群取得樣本。此外，TIMSS 數學和科學評量架構涵蓋的試題數多，為了讓評量架構涵蓋的試題在調查中能完整呈現，每位學生在受測時間內僅完成部分試題。TIMSS 在受測學生抽樣及評量工具的設計，雖然有效地降低了調查執行的複雜度，也將學生的作答負擔降至最低，其代價則是估計得到的統計量存在某些不確定性。

為了在調查結果的精確性與調查執行的可行性之間取得平衡，調查的設計必須由樣本

的統計量估計母群統計量的精確性，藉此量化相關的不確定性。TIMSS 2019 國際報告中的每個統計量均附有其標準誤，若統計量表示的是兩個估計值的比較，標準誤則可用於計算信賴區間或顯著性檢驗。TIMSS 學生能力估計的標準誤計算有兩個部份，第一個部份稱為抽樣變異量（sampling variance）估計，其不確定性來自樣本統計量推估母群統計量的過程；第二個部份稱為差補變異量（imputation variance）估計，產生此不確定性是因為 TIMSS 依據樣本學生在部份試題的作答表現以及其它與學生能力相關的資訊，估計得到母群學生的能力值。

1. 抽樣變異量估計

TIMSS 由全國學生母群中抽取部分學生做為樣本，藉此估計全國學生母群的能力，抽樣時先由全國抽取樣本學校，再由樣本學校中抽取一或二個班級做為學生樣本。由於以相同的抽樣方式對母群抽樣時，每次抽取的樣本只是眾多樣本中的一組樣本，因此，每一組樣本總有其母群代表性的不確定性，而來自抽樣的不確定性可透過抽樣誤差變異量估計得到。這種多階層叢集抽樣設計，並不適合使用簡單隨機抽樣的估計方法抽樣變異量，常用的方法為重複抽樣法（resampling schemes）。TIMSS 採二階段分層叢集抽樣，估計抽樣變異量時使用的則是刀切重複抽樣法（Jackknife Repeated Replication，簡稱 JRR）。TIMSS 第一階段抽取學校時，同時考量了學校的規模、分布的地域，或學校的類型，適合以刀切重複抽樣法估計抽樣變異量。

對於多數國家，抽樣學校樣本數 150 所且學生樣本 4000 位，可達到 TIMSS 要求的精確度。這 150 所學校依據學校分層變項、學校規模等相近的特質配對成 75 個刀切抽樣區，每個樣區包含 2 所學校，接著再進行重複抽樣（任宗浩、譚克平和張立民，2011）。其方法是隨機地移除一個刀切抽樣區中一所學校的資料，再以另一所學校的資料代替被移除的學校。實務上的做法是運用重複抽樣加權值（replicate weights）的概念，將被移除學校的加權值設為 0，而另一所學校的加權值設為 2，其餘學校的加權值則為 1，接著再以同樣的程序重複操作剩餘的 74 個刀切抽樣區。研究者可利用原始樣本之統計量$\hat{t}$，及產生出的許多重複樣本計算欲得之統計量$\hat{t}_{(j)}$。再利用重複樣本所算出的值與原始樣本之差異得出抽樣變異量（Foy & LaRoche, 2016；任宗浩等人，2011）：

$$\sigma^2_{(\hat{t})} = \sum_{j=1}^{75} (\hat{t}_{(j)} - \hat{t})^2$$

【待續】