教育科學研究期刊　第五十六卷第一期
2011 年，56（1），33-65

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

任宗浩
國立臺灣師範大學
科學教育中心
助理研究員

譚克平
國立臺灣師範大學
科學教育研究所
副教授

張立民
澳洲墨爾本大學
評估研究中心
副教授

摘要

大型調查研究常採用多階段分層叢集抽樣，檢視臺灣學生歷年來參與 TIMSS 研究的結果顯示，其平均表現之標準誤常較他國稍大。應 TIMSS 專責抽樣之單位要求，本研究旨在推導設計二階段分層叢集抽樣時即可預估標準誤的公式，以利選擇較佳分層架構，達減少標準誤的目標，並進行三個分析檢查其有效性。分析一將 30 個參與 TIMSS 2007 的國家資料，用該公式與《TIMSS 技術手冊》建議之刀切重複取樣法，分別估算各國學生平均科學成績之標準誤，發現兩者之線性相關達 .98。分析二以 29 個連續參加 TIMSS 2003 和 2007 的國家資料，論述利用該公式與現有輔助變項預估將要進行調查的誤差之實用性。分析三探討當叢集的分層輔助變項爲連續量時，不同分層數與二階段分層叢集抽樣誤差間的關係，以預估臺灣學生在 TIMSS 2011 平均科學成績之標準誤。文後尚提出四階段的評估流程，供相關研究預估主要依變項平均值之標準誤時做參考。

關鍵字：大型評量、抽樣架構計畫、降低抽樣誤差、複雜抽樣設計、變異量估計

通訊作者：任宗浩，E-mail: tsunghau@ntnu.edu.tw
收稿日期：2010/09/30；修正日期：2011/02/08、2011/03/10；接受日期：2011/03/15。

壹、前言

近年來，隨著教育績效的觀念在國內外逐漸受到重視，我國一方面積極參與國際教育成就調查，包括由國際教育學習成就調查委員會（The International Association for the Evaluation of Education Achievement, IEA）所主持的「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）、「國際公民教育與素養調查計畫」（International Civic and Citizenship Education Study, ICCS）以及由經濟合作暨發展組織（Organisation for Economic Co-operation and Development, OECD）所主持的「學生能力國際評估計畫」（The Programme for International Student Assessment, PISA）等趨勢調查，希望能藉由國際比較來檢視我國教育實施之成效；另一方面，爲了能夠建立國內教育相關資料庫以作爲教育政策制定之依據，也委託研究單位進行國內教育相關調查，如中央研究院主持的「臺灣教育長期追蹤資料庫計畫」（Taiwan Education Panel Survey, TEPS）以及國家教育研究院所主持的「臺灣學生學習成就評量資料庫」（Taiwan Assessment of Student Achievement, TASA）等調查。由於所有的調查都有誤差，所以愈是要求精確的調查，可以經由如較多的測驗或問卷題目，或更多的樣本來使誤差變小。此外，任何一個大型調查在進行之前，應先針對主要的研究問題決定可容許誤差的大小，再評估所擁有的資源（經費、人力、時間、事前資訊），並設法控制調查誤差在可容許的範圍；倘若無法將調查誤差範圍縮小到理想範圍內，則應避免輕易進行調查，以免浪費時間與金錢。

以樣本估計值推論母群參數時所產生的誤差來源，一般而言可以分爲兩個部分：一部分是來自於由母群抽取代表性樣本時所產生的抽樣誤差（sampling error），另一部分來自於針對代表行爲實施測量以推估整體行爲的測量誤差（measurement error）。就個人行爲層級而言，測量誤差也可視爲對個人行爲抽樣所產生的誤差（Adams, 2005; Shavelson & Webb, 1991）。以樣本估計值推論母群參數的精確性，會和樣本大小以及抽樣方式有關，對於精確性的要求亦隨研究或調查的目的而定。在同樣的抽樣架構下，樣本愈大，對於母群參數的推估會愈精確，然而需要花費的人力、金錢與時間也愈多。以 TIMSS 調查爲例，雖然目的在調查參加國學生的數學和科學能力，而非排序其平均能力，但爲能協助各國評估其教育實施的成效，調查結果仍需適度呈現出各國學生平均能力的差異。於此，TIMSS 希望各國分數的 95%信賴區間範圍能夠小於全球參加學生分數標準差的十分之一，相當於要求各國的標準誤小於 0.05 個標準差，因此在隨機抽樣時樣本大小至少約爲 400 人（Joncas, 2008）。[1]由於考量實際施測成本與

[1] 在隨機抽樣的情況下，母群平均值的 95%信賴區間約為 $\pm 1.96 \times \sqrt{\frac{\text{標準差}}{N}}$ 。若要求其區間為 ±0.1 個標準差範圍內，則 N 至少約為 400。

可行性，TIMSS 和許多大型教育成就調查測驗（如 PISA 和 TASA）均非採用簡單隨機抽樣，而改採二階段分層叢集抽樣設計（two-stage stratified cluster sample design）：第一階段先針對調查學生母群所在之學校分布進行分層取樣，第二階段再針對學校內部的學生進行抽樣。不同的調查研究在第二階段抽樣時有些許不同：TIMSS（Joncas, 2008）和 TASA（http://tasa.naer.edu.tw）由抽樣各學校中隨機抽出 1 至 2 個班級作爲樣本，PISA 則是由抽樣各校中隨機抽出相同學生數進行調查（OECD, 2009）。然而，由於採取非隨機抽樣，對於所需樣本人數及抽樣誤差的事前估計，就變得複雜許多。已有統計學者（如 Cochran, 1963; Hansen, Hurwitz, & Madow, 1953; Kish, 1965）提出針對一階段分層抽樣以及叢集抽樣的誤差估計公式，但對於二階段分層叢集抽樣設計的誤差估計，似乎並無簡單的公式可用。本研究的主要目的在於導證二階段分層叢集抽樣設計的誤差估計公式，並利用 TIMSS 2007 的資料庫檢視該公式之有效性。

TIMSS 調查會利用跨屆調查的共同試題，將跨屆間的調查量尺分數予以等化（Foy, Galia, & Li, 2008），其量尺等化的方式以 TIMSS 1995 爲參照，將平均成績設爲 500，標準差設爲 100。根據 TIMSS 2007 調查結果，我國八年級學生的平均數學和科學成就分數爲 598 和 561，標準誤分別爲 4.5 和 3.7，因此數學和科學成就分數的 95%信賴區間範圍均小於十分之一個標準差而可接受。然而，如此的信賴區間範圍常使得幾個亞洲領先國家或地區（括我國、新加坡、韓國、日本、香港）彼此之間的差異無法被區分，因此 IEA 和負責進行 TIMSS 學校抽樣的加拿大統計局（Statistics Canada）希望經由抽樣誤差的減少，使得我國八年級生平均成就之標準誤能降到 3.0 量尺分數以下。而減少抽樣誤差的其中一種方法，可由事前評估以選擇較佳的分層架構來達成。爲了達到這個目標，本研究利用所推導出的公式，提出我國參加 TIMSS 2011 調查的學校分層架構，並針對此抽樣架構的抽樣誤差進行評估。

本文架構將包括以下幾個部分：首先，針對 TIMSS 2007 的抽樣架構及平均成就分數抽樣誤差之估計予以介紹。其次，研究者將導證用以估計二階段分層叢集抽樣誤差的公式。第三部分則包含三個分析：分析一，利用 30 個參與 TIMSS 2007 的國家（地區）的資料，檢驗該公式之有效性；分析二，經由該公式分析 29 個參加 TIMSS 2003 和 TIMSS 2007 兩屆調查的國家（地區），探討利用先前調查資訊以預估未來調查誤差之可行性；分析三，以我國即將進行的 TIMSS 2011 調查之抽樣架構爲實例，說明當分層輔助變項爲一連續變項時，第一階段針對叢集的分層數與抽樣誤差間之關係，並根據該公式預估八年級科學平均成就之抽樣誤差。最後，本研究針對二階段分層叢集抽樣之教育調查研究提出標準化的評估流程，使研究者在特定分層抽樣架構下，應用於不同的教育調查研究，據以估計主要調查變項母群平均值之誤差。

一、TIMSS 2007 抽樣介紹

TIMSS 2007 調查針對全國學校和校內班級進行二階段分層叢集抽樣，說明抽樣方式如後：

（一）學校抽樣階段

在學校抽樣階段，TIMSS 採用分層抽樣方法。先依各國所提供的分類法將全國國中或國小分爲若干層（strata），並依每一分類層人數占調查母群總人數之比例，估算出各分類層應抽樣的人數和班級數。之後，以各校每一層級學生人數占母群中該層級人數之比例，決定該校被抽取爲樣本的機率。此外，TIMSS 調查允許各參加國在學校抽樣階段，包含顯性分層（explicit stratification）和隱性分層（implicit stratification）兩種依據（Joncas, 2008）：

1.顯性分層

目的是想藉由調查以推論不同層（子群）的特性。例如某參加國欲瞭解不同學校形式（例如：公立學校和私立學校）的教育成果是否有差異，則相關的統計量必須要能精確推估學校形式的特性。因此，該參加國的抽樣架構，便應以涵蓋全國各種學校形式的母群爲基礎，以確保抽取出來的學校形式及學生樣本，能更準確地反映這些子群分布的比例。匈牙利（Hungary）便以地區發展程度爲顯性分層依據，將學校分爲城市學校或鄉村學校，臺灣及澳洲則單純地以行政或地理區域將學校分爲不同子群（Foy & Olson, 2009）。

2.隱性分層

目的是爲了減少學校層級的抽樣誤差，因此希望所抽取出來的樣本分布，能充分代表母群的實際分布情形。若學校層次的隱性分層變項與該校學生在 TIMSS 調查中的平均成就相關愈高，樣本的分布就愈能夠代表母群的分布。例如印尼（Indonesia）等參加國（Foy & Olson, 2009）先依據其他考試或成就調查結果將學校分爲高、中、低三個不同成就層，再根據各層占整體母群之比例，估算出每層所應抽取的樣本數，然後決定各層所應抽出的學校數量。此外，我國在 TIMSS 2007 的學校抽樣架構中，並未指定任何隱性分層變項，但由於 TIMSS 考慮到學校大小及該校被抽到的機率必須成正比（probability proportional to school size），亦即學校愈大被抽取爲樣本的機率也就愈高，這種抽樣會確保隸屬不同規模學校的母群學生人數分布能夠按比例被抽到（Joncas, 2008）；換言之，「學校大小」可以視爲內嵌於 TIMSS 學校分層架構的隱性分層變項。

因此，在 TIMSS 調查中屬同一分層的學校，亦即具有相同的的刀切變項（變項名稱 JKZONE）的學校，是指同時具有相同的顯性分層變項（變項名稱 ISATRATE）及隱性分層變項（變項名稱 ISATRATI），且學校大小接近者。

（二）校內班級抽樣階段

經過第一階段學校抽樣之後，第二階段的叢集抽樣（cluster sampling）係針對被抽到的學校進行班級抽樣。校內班級的抽樣以每班等機率方式隨機抽出受測班級，被抽到的班級，其整班學生均需接受調查施測。原則上每校抽取一班，但如果該校班級人數很少，可能會在同一學校內抽取兩班，或是合併其他類似背景學校的抽樣班級後，再將兩班學生合併後，視爲

一個虛擬班級（pseudo-classroom）。

二、TIMSS 2007 調查的抽樣誤差估計

二階段分層叢集抽樣的精確性源自於兩方面所產生的效應：其一爲叢集抽樣，另一方面則來自於對叢集的分層方式，茲分別說明如後。

（一）叢集抽樣對抽樣誤差的影響

在叢集抽樣方面，因爲同學校或同班級的學生共用相同的學校資源，例如，有相同的老師、同樣的課程等等，所以同校或同班的學生會有較多相同的特質。又如，由於人們易因社經地位選擇居住區域，故當學生就讀於自家學區時，同一學區的學生便易來自相似的社經背景，其成就差異也可能比來自不同學區的學生來得小。以下研究者舉兩個有關叢集抽樣設計的特例加以說明：

1.假設所有的學生都是隨機地被分配到各個學校，此時，學校間同質性最高，而校內學生的同質性最低。在此情況下，先隨機選取 150 所學校，然後從每所學校裡抽出 30 名學生去推估母群平均值的抽樣誤差，等於從母群中簡單隨機抽樣 4,500 名學生的抽樣誤差。

2.假設任一所學校內的每位學生是完全同質的，此時校間的異質性最高。在此情況下，由於在任一校內所有學生的調查變項完全相同，因此只需抽出 1 位學生，就會知道同校其他所有學生的資訊。在這種狀況下，利用二階段分層叢集抽樣先抽取 150 所學校，再從中各抽出 30 名學生，相當於從 150 所學校內各抽 1 名學生；最後有效樣本數便等同於 150 人。此時，估計母群參數的抽樣誤差，等於隨機抽樣 150 個人的抽樣誤差。

實際上，沒有任何一個教育系統會符合這兩種極端的例子。一般而言，在相同樣本數下，簡單隨機抽樣的抽樣誤差比叢集抽樣小。因此，利用簡單隨機抽樣抽出 4,500 人，比起先抽出 150 所學校後，再從中各抽出 30 人的叢集抽樣方法，更能夠涵蓋母群的特質。

（二）叢集的分層對抽樣誤差的影響

決定二階段分層叢集抽樣精確性的因素，除了叢集抽樣產生的效果外，還有對叢集的分層方式。若第一階段針對對叢集的分層輔助變項（例如：TIMSS 調查中用來將學校分層的顯性和隱性變項）和主要調查變項有高相關，以確保不同水準的叢集都能夠按比例被抽到，如此叢集平均值的分布便較能表徵母群的實際分布。就可讓所抽出來的叢集平均值的分布比較接近母群的分布。

至此，我們可以理解分層抽樣所具備的兩難問題（dilemma）：如果沒有事先進行普查，如何得到所有學校的平均成就資料？然而，如果已經得到了平均成就的普查資料，那又何必再進行抽樣調查？的確，各個叢集（學校或班級）在調查變項上（如 TIMSS 測驗工具評量中的數學或科學成就）的平均值不可能事先取得，但實際作法上研究者可善用過去的調查或相

關資料庫的數據，以取得在叢集層次與調查變項相關的變項資料。以臺灣爲例，利用國中九年級生參加基本學力測驗的數學或自然科學平均成績，可爲 TIMSS 調查時對學校層級的分層依據。

（三）二階段分層叢集抽樣的誤差估計

二階段分層叢集抽樣的精確性來自兩個階段的效應：在叢集的分層階段，可以讓抽樣的結果比簡單隨機抽樣更具代表性；然而，在每一叢集層內的叢集抽樣結果，卻比簡單隨機抽樣來得較不具代表性。由於這兩個對立的效應，使得許多大型調查無法以簡單公式用以推估抽樣變異或統計值。但隨著資訊科技的發展，許多統計分析方法開始仰賴電腦快速運算來解決這類問題。其中，利用現有樣本進行重複取樣便可用來推估母群參數誤差，例如平衡重複取樣法（balanced repeated replications, BRR）（Plackett & Burman, 1946）、刀切重複取樣法（jackknife repeated replications, JRR）（Frankel, 1971），以及拔靴法（bootstrapping）（Efron, 1979）等等。除了重複取樣的技術外，有些軟體（如 SPSS）則採用泰勒線性近似法（Taylor linearization method）（Demnati & Rao, 2004），針對複雜的抽樣設計進行母群參數誤差的估計。根據 Lehtonen 和 Pahkinen（2004, p. 165）的研究結果，針對兩階段分層叢集抽樣調查的比例統計量（ratio estimators）之誤差估計，無論採用線性法、平衡重複取樣法、刀切重複取樣法以及拔靴法，所得到的誤差估計結果差異不大。

在 TIMSS 國際報告中，採用 Frankel（1971）針對二階段分層叢集抽樣所提之刀切重複取樣法的修訂方法來計算所有統計估計值的抽樣變異量（Foy et al., 2008）。根據 Frankel 的二階段抽樣刀切法，可設定重複抽樣加權值（replicate weights），在具有相同顯性分層變項的學校中，選取隱性分層變項與學校大小最接近的 2 所學校，將其定義爲同一個刀切抽樣區（jackknife sampling zone）。因此，150 所學校會配對成 75 組抽樣區。每一次重複抽樣，僅針對一個抽樣區隨機去除 1 所學校的資料，與之配對的另一所學校權重則爲二倍，而其他 74 個抽樣區的 148 所學校權重維持不變。之後，將原始樣本之統計值 $\hat{\theta}$（如：平均、相關或迴歸係數等）及重複抽樣 75 次所得之統計值 $\hat{\theta}_{(i)}$（i=1, 2, ..., 75）代入式(1)，以計算統計值 $\hat{\theta}$ 之抽樣變異量：

$$\sigma^2_{(\hat{\theta})} = \sum_{i=1}^{75} (\hat{\theta}_{(i)} - \hat{\theta})^2 \qquad (1)$$

一般而言，如果把 TIMSS 調查當作隨機抽樣，通常會低估統計量的標準誤，除非在學校層級的分層輔助變項與調查變項有高相關。以我國八年級學生在 TIMSS 2007 的平均數學成就估計爲例，如果直接將之視爲簡單隨機抽樣的結果，雖然對母群平均值（598 分）的估計不會產生偏誤，但是對於平均值標準誤的估計值（2.1 分），卻會低於考慮複雜抽樣架構的刀切重複取樣法所估計出來的結果（4.6 分）。而統計量標準誤的低估，會導致統計檢定不當推論的後果。

前述幾種估計複雜抽樣調查誤差的方法，均適用於調查完成後。然而，一個大型調查若無法在事前精準估計其誤差，待事後才發現誤差太大，已難收亡羊補牢之效。本研究所導證的公式，可供未來的研究者，藉由該公式與可利用的輔助資訊，於調查進行前及特定分層叢集抽樣架構下，預估主要調查變項的抽樣誤差。

三、二階段分層叢集抽樣調查的設計效果

此部分所導證的公式，可用以估計二階段分層叢集抽樣調查的抽樣誤差。首先，將設計效果（design effect）"D_{eff}"定義爲：在相同的樣本數下，特定抽樣架構相對於簡單隨機抽樣方式，其抽樣變異量的比值（Hansen et al., 1953; Kish, 1965）：

$$D_{eff} = Var_{SF}(\bar{x}) / Var_{SRS}(\bar{x}) \tag{2}$$

其中，$Var_{SF}(\bar{x})$和$Var_{SRS}(\bar{x})$分別代表特定抽樣架構與簡單隨機抽樣方式下，對母群平均值($\bar{x}$)的估計變異量。由式(2)可得到對 $\bar{x}$ 估計之標準誤 $SE_{SF}(\bar{x})$爲：

$$SE_{SF}(\bar{x}) \approx \sqrt{D_{eff}\frac{s^2}{n}} \tag{3}$$

其中 s^2 爲變異量的不偏估計，n 爲樣本數。Hansen 等（1953）推導一階段叢集抽樣設計（one-stage cluster sample design）的設計效果（$D_{eff\_cluster}$）爲：

$$D_{eff\_cluster} = [1+\rho(b-1)] \tag{4}$$

在式(4)中，b 爲叢集內樣本數，而 ρ 爲叢集內相關，並定義 ρ 爲：

$$\rho = 1 - \frac{s^2_{within\text{-}cluster}}{s^2}\frac{n}{n-1} \tag{5}$$

式(5)中，$s^2_{within\text{-}cluster}$ 爲叢集內樣本變異量的不偏估計。若各叢集樣本數不同，Donner 和 Klar（1994）建議可以利用加權後的平均叢集樣本大小 b' 取代式(4)中的 b。其數學表示如下：

$$b' = \sum_{i=1}^{m}\left(\frac{b_i}{\Sigma_{j=1}^{m} b_j}\right)b_i = \frac{\Sigma_{i=1}^{m} b_i^2}{\Sigma_{j=1}^{m} b_j} \tag{6}$$

式(6)中，m 是所抽取的的總叢集數量，而 b_i 爲第 i 個叢集內的樣本數。

對於二階段分層叢集抽樣而言，每一分層內的叢集抽樣就相當於一階段的叢集抽樣。因此，由式(3)和式(4)即可得到對於第 h 層內平均值的估計變異量爲：

$$\mathrm{Var}(\bar{x}_h) \approx [1+\rho_h(b_h^{'}-1)]\frac{s_h^2}{n_h} \tag{7}$$

其中下標 h 代表在第 h 個分層內，而 $\rho_h, b_h^{'}, s_h^2, n_h$ 則分別爲第 h 層內的叢集內相關、加權後的平均叢集內樣本數、層內變異量的不偏估計以及樣本大小。若 w_h 爲第 h 個分層之權重，由於 $\bar{x} = \Sigma_h \mathrm{w}_h \bar{x}_h$ ，根據簡單的變異數運算公式（Dorofeev & Grant, 2006, Eq. 3-2）以及式(7)：

$$\mathrm{Var}(\bar{x}) = \sum_h \mathrm{w}_h^2 \mathrm{Var}(\bar{x}_h) \approx \sum_h \mathrm{w}_h^2 [1+\rho_h(b_h^{'}-1)]\frac{s_h^2}{n_h} \tag{8}$$

如果以下三個假設可以成立，式(8)就可以被大幅簡化：

（一）假設一

跨層等抽樣比例（equal sample fraction across strata）。對一個均等機率的抽樣設計（Equal Probability of Selection Method, EPSEM）而言，每一個分層內所抽出的樣本數應該與母群在該層內的總人數呈正比。也就是：

$$f_1 = f_2 = f_3 = ... = f_h = \frac{n_h}{N_h} = \frac{n}{N} = f \tag{9}$$

其中， f_h 代表在第 h 個分層中，樣本數 n_h 與母群人數 N_h 的比例；n 和 N 爲總樣本數與母群人數。因爲每個人被抽到的機會相等，所以各層的權重（ w_h ）就會與該層的總人數成正比，因此

$$\mathrm{w}_h = \frac{N_h}{N} = \frac{n_h}{n} \tag{10}$$

（二）假設二

跨層等加權平均叢集大小（equal weighted average cluster size across strata），即滿足

$$b_1^{'} = b_2^{'} = ... = b_h^{'} = ... = b_H^{'} = b^{'} \tag{11}$$

（三）假設三

均等分層（equal stratum size），亦即各分層內所包含之母群人數均相同。這個假設配合前兩個假設，可以推論各層應抽出相同的叢集數量：

$$m_1 = m_2 = ... = m_h = ... = m_H = m \tag{12}$$

其中 m_h 爲第 h 分層抽出的叢集數。這個假設看似嚴苛，但先前提過在 TIMSS 調查的抽樣設計下，同一分層（具相同的 JKZONE 變項）僅抽出 2 所學校，每所學校又抽出約相同的人數（一個班級），所以各分層內抽出的樣本數大約相等；又因爲整個抽樣設計基本上遵守 EPSEM，各分層內樣本之總加權值也大約相等，所以各層內包含的母群人數應很接近。實際上，只要叢集的抽樣遵守叢集抽樣機率正比於叢集大小的原則（Probability Proportional to Size, PPS），我們都可以經過現有分層架構，加上叢集大小，定義出新的分層方式，將具有相同分層變項與大小最接近的兩個叢集視爲同一層。

真實的抽樣通常不符合上述三項假設，然而爲了符合真實的情況，將使得所推導的公式變得相當複雜，如此一來便失去了本研究希望能夠提供一個簡單的公式，讓研究者藉以提出適當的抽樣架構之本意；關於違反上述各項假設所可能導致的影響，將會是未來研究的努力方向。在許多時候，即使接受這些假設，仍然可以對抽樣誤差提供很好的近似估計，這一點本研究的分析一將以 30 個參與 TIMSS 2007 之國家（地區）的釋出資料作爲實例進行檢驗。若接受上述三個假設，式(8)可以改寫爲（導證過程請見附錄）：

$$\mathrm{Var}(\bar{x}) \approx \frac{\sigma^2_{\text{within-cluster}} + b'(\sigma^2_{\text{between-cluster}} - \sigma^2_{\text{auxiliary-variable}})}{n} \tag{13}$$

式(13)中，$\sigma^2_{\text{within-cluster}}$ 爲叢集內的變異量，$\sigma^2_{\text{between-cluster}}$ 爲叢集間的變異量，$\sigma^2_{\text{auxiliary-variable}}$ 爲被叢集層級的分層輔助變項所解釋的叢集間變異量；茲說明其意義如後。

我們利用圖 1 說明母群參數的變異量被叢集的分層輔助變項、叢集間以及叢集內解釋之比例。在一階段叢集抽樣的情況下，母群參數的變異量（σ^2_{total}）可以被分解成叢集間的變異量（$\sigma^2_{\text{between-cluster}}$）加上叢集內的變異量（$\sigma^2_{\text{within-cluster}}$）；如式(14)：

$$\sigma^2_{\text{total}} = \sigma^2_{\text{between-cluster}} + \sigma^2_{\text{within-cluster}} \tag{14}$$

在二階段分層叢集抽樣的設計下，叢集間變異量（也就是叢集平均值的變異量）的一部分（$\sigma^2_{\text{auxiliary-variable}}$）被叢集的分層輔助變項（例如以學校爲叢集單位時，學校分層的輔助變項可能爲學校大小、學校型態或學校所在區域的發展程度等等）所解釋，如式(15)：

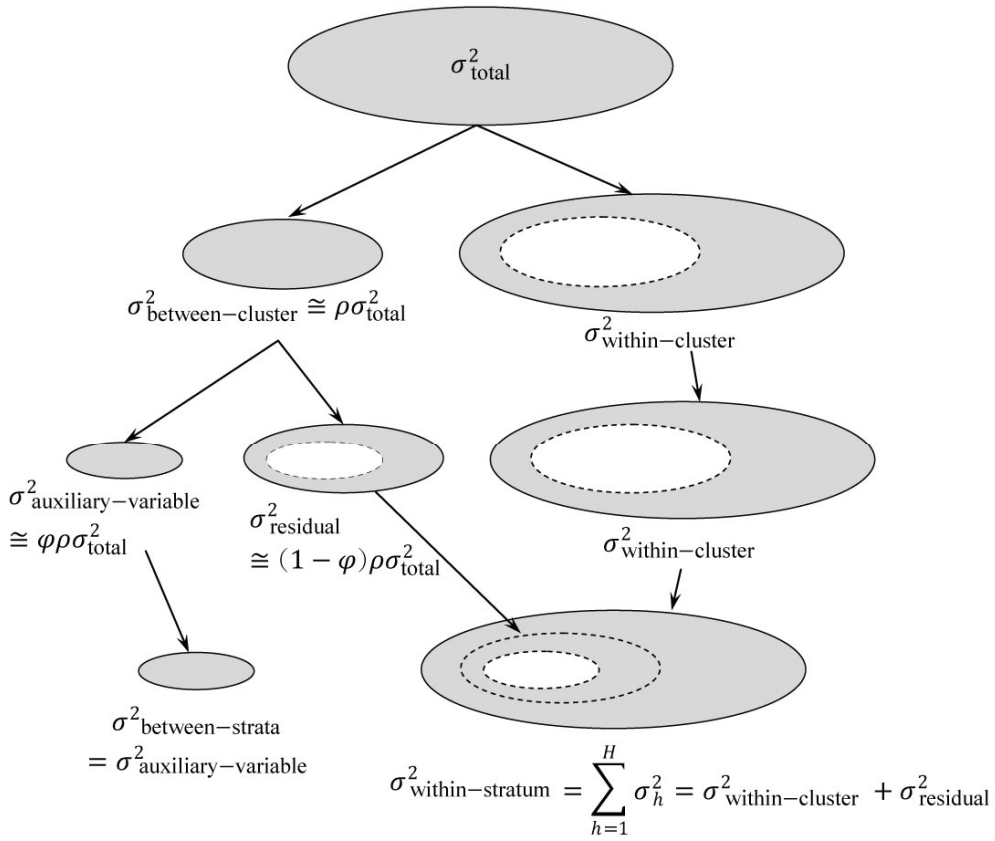

圖1　統計估計變異量的成分分析

$$\sigma^2_{\text{between-cluster}} = \sigma^2_{\text{auxiliary-variable}} + \sigma^2_{\text{residual}} \tag{15}$$

其中 $\sigma^2_{\text{residual}}$ 爲未被分層輔助變項所解釋的叢集間變異量。定義 φ 爲叢集的分層輔助變項解釋叢集間變異量的比例：

$$\varphi = \frac{\sigma^2_{\text{auxiliary-variable}}}{\sigma^2_{\text{between-cluster}}} \tag{16}$$

如果對叢集的分層將同一層內各叢集之分層輔助變項控制成完全相同（通常只發生在叢集的分層輔助變項爲類別變項時），按照圖 1 的概念，則所有層內的變異量總和應爲：

$$\sum_{h=1}^{H}\sigma^2_h = (1-\varphi)\sigma^2_{\text{between-cluster}} + \sigma^2_{\text{within-cluster}} \tag{17}$$

式(17)中，各個分層內樣本變異量之總和 $\Sigma_{h=1}^{H}\sigma^2_h$ 比起在式(14)中，所有樣本之總變異量 σ^2_{total} 少

了被分層變項解釋掉的部分 $\varphi\sigma^2_{\text{between-cluster}}$ 。同時，我們也可以化簡式(5)如下：

$$\begin{aligned}\rho &\approx 1-\frac{s^2_{\text{within-cluster}}}{s^2}\frac{n}{n-1}\\ &\approx \frac{\sigma^2_{\text{between-cluster}}}{\sigma^2_{\text{total}}}\end{aligned} \tag{18}$$

在簡化式(18)的過程中，我們將$(n-1)/n$ 和$(b'-1)/b'$當作 1，這個近似在叢集樣本很小的時候偏差會比較大。

進一步將式(16)和式(18)帶入式(13)式可以得到：

$$\begin{aligned}\text{Var}(\bar{x}) &\approx \frac{\sigma^2_{\text{within-cluster}}+b'(\sigma^2_{\text{between-cluster}}-\sigma^2_{\text{auxiliary-variable}})}{n}\\ &=\left[\frac{\sigma^2_{\text{within-cluster}}}{\sigma^2_{\text{total}}}+b'\left(\frac{\sigma^2_{\text{between-cluster}}}{\sigma^2_{\text{total}}}-\frac{\sigma^2_{\text{auxiliary-variable}}}{\sigma^2_{\text{total}}}\right)\right]\bullet\frac{\sigma^2_{\text{total}}}{n}\\ &\approx[(1-\rho)+b'(\rho-\rho\varphi)]\cdot \text{Var}_{\text{SRS}}(\bar{x})\\ &\approx\{[1+\rho(b'-1)]-\rho\varphi b'\}\times\frac{\sigma^2_{\text{total}}}{n-1}\end{aligned} \tag{19}$$

最後根據式(19)的結果，可以獲得經由二階段分層叢集抽樣進行母群平均值估計的標準誤爲：

$$\begin{aligned}SE_{\text{two-stage}} &\approx \sqrt{[1+\rho(b'-1)-\rho\varphi b']\times\frac{\sigma^2_{\text{total}}}{n-1}}\\ &=\sigma_{\text{total}}\sqrt{\frac{D_{\text{eff\_two-stage}}}{n-1}}\end{aligned} \tag{20}$$

其中

$$\text{D}_{\text{eff\_two-stage}} \approx 1+\rho(b'-1)-\rho\varphi b' \tag{21}$$

在式(20)中，σ^2_{total} 爲所有樣本的變異量，ρ 爲叢集內相關，b' 爲叢集平均大小，n 爲總樣本數，而 φ 則是叢集平均值的變異量被叢集的分層輔助變項所解釋的比例。考慮以下四種情況：

（一）當叢集的分層輔助變項解釋叢集平均值變異量的比例趨近於 0 時（ $\varphi=0$ ），就如同我國在 TIMSS 2007 的數據一樣（這一點稍後會在分析一的結果中進一步地討論），式(21)可以簡化爲 Hansen 等（1953）針對一階段叢集抽樣所推導出的設計效應（如式(4)）。換句話說，

在此種情況下，叢集的分層對於降低抽樣誤差並沒有任何的幫助。

（二）如果叢集內相關和叢集的分層輔助變項解釋叢集平均變異量的比例均爲 0（$\rho=\varphi=0$），式(20)的右方等於 1。在這個情況下，二階段分層叢集抽樣的母群平均值之抽樣誤差會等於以簡單隨機抽樣方式抽出相同樣本數所估計母群平均值的誤差。

（三）如果叢集內相關和叢集的分層輔助變項解釋叢集平均值變異量的比例均爲 1（$\rho=\varphi=1$），根據式(20)可以計算出對母群平均值的估計誤差爲 0。首先，叢集內相關等於 1，代表沒有叢集內的變異量。換句話說，個人層級的變異量可以完全被叢集間的變異量所解釋；其次，叢集間的變異量又完全可以被叢集的分層輔助變項所解釋；這表示利用這個分層抽樣的架構所抽出的樣本可以完全反映母群的分布，所以對母群平均數的估計誤差應爲 0。

（四）如果抽樣設計效果等於 1 且叢集內相關不爲 0，即：

$$1+\rho(b'-1)-\rho\varphi b'=1 \tag{22}$$

我們可以求出一個叢集平均與叢集的分層輔助變項之相關閾值（φ_s）：

$$\varphi_s=\frac{b'-1}{b'} \tag{23}$$

當叢集的分層輔助變項與叢集平均值間之相關大於φ_s時，設計效應有可能小於 1，此時對於母群平均值的估計誤差會比簡單隨機抽樣的情況來得更小。換句話說，雖然叢集抽樣的有效樣本數不如簡單隨機抽樣，但是透過在叢集層次有效的分層，所抽出來的樣本仍然有可能比簡單隨機抽樣的樣本更能代表母群的分布。

接下來，本研究將利用三個分析來檢驗式(20)有效性及實用性：分析一利用 30 個參加 TIMSS 2007 調查的國家（地區）之釋出資料，比較利用式(20)和刀切重複取樣法的誤差估計結果；分析二利用連續參加 TIMSS 2003 和 TIMSS 2007 的 29 個國家（地區）的資料，探討利用式(20)以及現有輔助資訊預估將要進行調查的誤差之實用性；分析三探討當叢集的分層輔助變項爲連續變項時，不同分層數與二階段分層叢集抽樣調查誤差間的關係，並藉以預估我國參加 TIMSS 2011 八年級生科學平均成就之誤差。

貳、分析一

分析一的主要目的爲針對二階段分層叢集抽樣設計，檢驗以式(20)估算母群平均值抽樣誤差之有效性。本研究利用 TIMSS 2007 資料庫中 30 個參加國（地區）的釋出資料（The International Association for the Evaluation of Education Achievement [IEA], 2009），計算式(20)中牽涉的各項變項，並比較利用式(20)所估計出來的抽樣誤差（$SE_{\text{two-stage}}$）與根據《TIMSS 2007

技術手冊》建議之二階段刀切重複取樣法所估計出的誤差（SE_{JRR}）（Martin et al., 2008）。

爲了估計 $SE_{two\text{-}stage}$，研究者必須知道式(20)中變項 n、 b'、ρ、σ_{total} 和 φ 等數值。其中 n 爲總樣本數；b' 爲平均叢集內樣本數。對於大部分的國家，除了班級人數很少的學校會抽兩個班組合成一個虛擬班級以外，每個學校只抽取一個班級，所以 b'大致可以視爲平均每個學校所抽取的樣本數；惟在 TIMSS 資料庫內認定之叢集並非完全以班級爲準，而是將同時具有相同的刀切區間（JKZONE）以及相同的刀切重複取樣變項（JKREP）之學生視爲同一叢集；ρ 爲叢集內相關，在這裡係指班級內學生科學成就分數之相關；σ_{total} 爲全國八年級生科學成就分數之標準差；最後，φ 爲學校分層的輔助變項（包含隱性變項和顯性變項）解釋叢集間學生科學平均成就變異量的比例。

利用 TIMSS 2007 資料庫中各國學生背景資料檔案（檔案名稱爲 bsg***m4.sav，***爲參加國代碼）中的學校代碼（變項名稱 IDSHOOL）、刀切區間變項、刀切重複取樣變項、學生總權重（變項名稱 TOTWGT）以及每位學生的 5 個似真值（plausible values）（變項名稱 BSSSCI01-BSSSCI05）可以計算得到 n、b'以及 σ_{total}。班級內相關 ρ 是利用 HLM 6.08 軟體求出叢集間與叢集內變異量後，依據式(18)計算獲得。φ 則是利用同一學校分層內的 2 所學校受測學生平均成績之相關求得，就像是古典測驗中兩個平行測驗觀察分數的相關相當於任一觀察分數與真分數相關的平方，我們很容易利用隨機誤差模型證明同一學校分層內任 2 校平均成績之相關，其期望值會等於於學校分層輔助變項解釋層內所有學校平均成績變異量之比例。須注意的是根據此法求出之 φ，在概念上應該介於 0 和 1 之間，但實際的數值是介於-∞到 1 之間；φ 爲負值的情況通常發生在真實的 φ 很接近 0 的時候，主要的原因來自於取樣的誤差（Krus & Helmstadter, 1993; Solomon, 2004），如表 1 中，韓國（Korea）、馬來西亞（Malaysia）和臺灣（Taiwan）的情況。在 TIMSS 資料庫中，所謂同一學校分層係指具有相同刀切區間變項而言，且 σ_{total}、ρ 和 φ 的計算均是經過將每個學生經由學生總權重加權後，分別利用學生科學成績的 5 組似真值求得 5 個統計量後取平均值而得。

表 1 針對參加 TIMSS 2007 的 30 個國家之八年級科學成就，呈現分別利用式(20)所計算出各國八年級平均科學成就分數之誤差（$SE_{two\text{-}stage}$）以及利用 TIMSS 2007 技術手冊中所建議之刀切重複取樣法所計算獲得之誤差（SE_{JRR}）；其中，各國（地區）SE_{JRR} 會比國際報告上的各國分數標準誤值稍小，因爲後者有加進測量工具的誤差。分析結果顯示，對於各國（地區）八年級生的科學平均成就來說，兩種方式所估計之誤差非常接近。除了對科威特（Kuwait）和黎巴嫩（Lebanon）2 個國家，兩種估計法所求出之誤差差異較大（分別爲 0.5 和 0.8 個量尺分數，約爲 19%和 16%）以外，其他國家的差異大都在 10%以內，有 20 個國家（地區）的差異介於 0.0 至 0.2 量尺分數之間。兩種方式所估計出來的誤差，其相關爲 0.98（如圖 2），這個結果顯示，雖然式(20)的推導基於三個較嚴苛的假設，但是用來作爲抽樣誤差的評估依據，仍具有非常好的近似結果。

表 1 針對 30 個 TIMSS 2007 參加國家（地區）之八年級生的科學平均成就，利用式(15)（$SE_{\text{two-stage}}$）以及二階段刀切重複取樣法（SE_{JRR}）所估計出的抽樣誤差結果比較

國家／地區	n	σ^2_{total}	b'	ρ	φ	$SE_{\text{two-stage}}$	SE_{JRR}
Amenia	4,689	10,230	37.7	0.30	0.03	5.1	5.5
Australia	4,069	6,452	25.7	0.48	0.42	3.5	3.5
Bahrain	4,230	7,402	32.5	0.21	0.80	1.9	1.7
Botswana	4,208	9,886	29.4	0.23	0.67	2.7	2.5
Cyprus	4,399	7,280	31.2	0.07	0.40	1.9	1.9
Egypt	6,582	9,877	52.6	0.25	0.24	4.0	3.6
England	4,025	7,293	31.4	0.47	0.14	4.9	4.4
Ghana	5,294	11,742	40.9	0.42	0.37	5.0	5.0
Hong Kong	3,470	6,557	29.5	0.54	0.22	4.9	4.8
Hungary	4,111	5,865	30.6	0.26	0.26	3.1	2.8
Indonesia	4,203	5,503	31.0	0.47	0.47	3.3	3.3
Iran	3,981	6,617	31.1	0.31	0.07	4.0	3.6
Israel	3,294	10,299	25.1	0.31	0.29	4.4	4.3
Italy	4,408	6,009	33.4	0.26	0.21	3.2	2.8
Japan	4,312	5,946	30.2	0.17	0.58	2.0	1.9
Jordan	5,251	9,549	40.2	0.28	0.19	4.2	4.0
Korea	4,240	5,755	30.0	0.07	-0.03	2.0	2.0
Kuwait	4,091	7,964	32.1	0.25	0.46	3.2	2.7
Lebanon	3,786	9,374	32.2	0.56	0.47	5.0	5.8
Lithuania	3,991	6,117	29.7	0.15	0.29	2.5	2.3
Malaysia	4,466	7,780	32.6	0.61	-0.14	6.4	6.0
Norway	4,627	5,369	35.3	0.07	0.06	2.0	2.1
Romania	4,198	7,727	32.9	0.30	0.32	3.7	3.7
Scotland	4,070	6,580	33.1	0.38	0.38	3.7	3.4
Slovenia	4,043	5,187	28.0	0.09	0.08	2.0	2.1
Sweden	5,215	6,090	36.2	0.12	0.06	2.4	2.5
Taiwan	4,046	7,971	27.9	0.22	-0.02	3.7	3.6
Ukraine	4,424	7,055	37.4	0.23	0.21	3.5	3.5
Tunisia	4,080	3,658	28.1	0.12	0.13	1.8	1.9
United States	7,377	6,769	55.8	0.33	0.53	2.9	2.9

註：n：總樣本數；σ^2_{total}：所有八年級生科學成就之變異量；b'：加權後的平均叢集大小；ρ：叢集內相關；φ：相同分層內（具有相同的 JKZONE 變項）叢集平均科學成就之相關

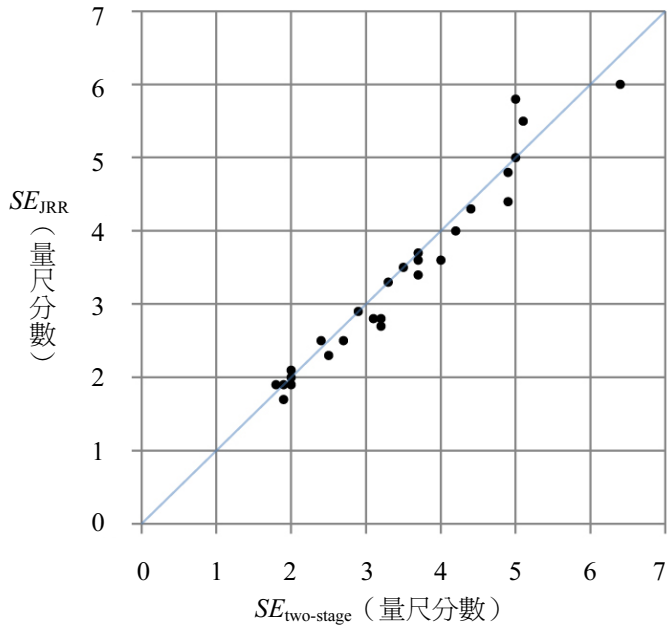

圖2　針對30個TIMSS 2007參加國（地區）八年級生科學成就平均之抽樣誤差，兩種估計法估計結果之分布相關

參、分析二

雖然分析一的結果顯示，藉由式(20)能夠相當準確地估計二階段叢集抽樣的調查誤差。但研究者若要利用此公式在調查進行之前先行評估調查的誤差，其中除了樣本大小（n）以及平均叢集大小（b'）可以在事前估計外，叢集內相關（ρ）、叢集的分層輔助變項解釋叢集平均值變異量的比例（φ）以及調查變項的標準差（σ_{total}）均無法在事前得知，幸而在某些特定的情況（例如 TIMSS、PIRLS 或 PISA 調查以學校或班級爲叢集調查某個年段學生的平均成就），透過其他調查所提供的數據通常相當穩定。

本研究根據 TIMSS 2003 和 TIMSS 2007 的資料（IEA, 2005, 2009），針對連續參加兩屆調查之 29 個國家或地區，利用統計軟體 HLM 6.08 計算出各國（地區）八年級生科學成就的班級內相關；且由式(20)可知班級內相關 ρ 對於誤差估計的影響小於 $\sqrt{\rho}$ 變化的比例，表 2 的結果顯示如果分別以 TIMSS 2003 調查的班內相關和 TIMSS 2007 調查的班內相關估計 TIMSS 2007 母群平均值的抽樣誤差，除了巴林（Bahrain）與瑞典（Sweden）有較大的差異外，其他 18 個國家（地區）對抽樣誤差的估計產生之差異小於 10%、9 個國家小於 20%。由此看出，將 TIMSS 2003 調查的班級內相關作爲 TIMSS 2007 調查時，班級內相關的近似值，對其中大部分國家（地區）已經可以提供很好的抽樣誤差估計。但兩屆調查畢竟間隔 4 年，如果能利用更近期的調查資訊，或許可以提供更好的近似估計。以臺灣中小學生學科成就的校內相

表 2　連續參加 TIMSS 2003 與 TIMSS 2007 兩屆調查之 29 個國家（地區）的班級內相關之比較

國家（地區）	ρ_{2003}	ρ_{2007}	ρ_{2003}/ρ_{2007}	$\sqrt{\rho_{2003}/\rho_{2007}}$
Armenia	0.21	0.30	0.70	0.84
Australia	0.39	0.48	0.81	0.90
Bahrain	0.10	0.21	0.48	0.69
Cyprus	0.09	0.07	1.29	1.13
Egypt	0.28	0.25	1.12	1.06
England	0.48	0.47	1.02	1.01
Ghana	0.32	0.42	0.76	0.87
Hong Kong SAR	0.48	0.54	0.89	0.94
Hungary	0.26	0.26	1.00	1.00
Indonesia	0.46	0.47	0.98	0.99
Iran	0.24	0.31	0.77	0.88
Israel	0.26	0.31	0.84	0.92
Italy	0.28	0.26	1.08	1.04
Japan	0.11	0.17	0.65	0.80
Jordan	0.21	0.28	0.75	0.87
Korea	0.07	0.07	1.00	1.00
Lebanon	0.44	0.56	0.79	0.89
Lithuania	0.16	0.15	1.07	1.03
Malaysia	0.50	0.61	0.82	0.91
Norway	0.08	0.07	1.14	1.07
Romania	0.40	0.30	1.33	1.15
Russian Federation	0.34	0.31	1.10	1.05
Scotland	0.45	0.38	1.18	1.09
Singapore	0.44	0.50	0.88	0.94
Slovenia	0.07	0.09	0.78	0.88
Sweden	0.22	0.12	1.83	1.35
Taiwan	0.25	0.22	1.14	1.07
Tunisia	0.14	0.12	1.17	1.08
United States	0.37	0.33	1.12	1.06

關係數爲例，利用 PIRLS 2006 的數據所估計臺灣小學四年級學生閱讀成就的班內相關爲 0.1，而 TIMSS 2007 的數據顯示小學四年級數學和科學成就的班內相關分別爲 .1 和 .08。因此，若利用 PIRLS 2006 臺灣小學四年級閱讀成就之班內相關作爲 TIMSS 2007 臺灣小學四年級之數

學或科學成就之班內相關，亦可提供很好的估計結果。

另一個影響估計誤差的重要變項是叢集的分層輔助變項解釋叢集平均值變異量的比例（φ）。這個數值會隨著分層輔助變項與調查變項的叢集平均值之間的相關愈高而愈大。如果在調查前能夠利用現有的資料庫，或其他的調查結果提供有利的資訊，作爲學校分層的輔助變項，就可以大幅減小調查結果的誤差。我們仔細檢視連續參加 TIMSS 2003 和 TIMSS 2007 的 29 個國家在兩屆調查的學校分層變項（Foy & Olson, 2009; Martin, 2005），發現僅有 6 個國家在兩屆調查中採用完全相同的學校分層變項；此外，臺灣在 TIMSS 2003 的學校抽樣架構中除了利用學校所在的地理區域外，比參加 TIMSS 2007 時多了一個班級性別的隱性分層變項。由於對於臺灣的學生而言，兩屆調查的結果均顯示無性別差異，所以是否加上性別作爲學校分層的隱性變項，差異應該不大。利用這 7 個國家在 TIMSS 2003 調查結果算出的班級內相關（ρ）與叢集的分層輔助變項解釋叢集平均值變異量的比例（φ）作爲各國在 TIMSS 2007 調查的對應變項的近似用以估計各國在 TIMSS 2007 的抽樣誤差，表 3 的結果顯示，除了亞美尼亞（Armenia）和馬來西亞（Malaysia）的估計結果與實際調查結果差距較大（分別低估 37%和 28%）；其他幾個國家（地區），兩者的差異從低估 2%至高估 19%。但作爲事前的預估，仍屬同一個數量級，並不算太差。

表 3　利用 TIMSS 2003 相同的抽樣架構所計算的班級內相關（ρ）與叢集的分層輔助變項解釋叢集平均值變異量的比例（φ）估計 TIMSS 2007 調查誤差與利用刀切重複取樣法所得到之誤差比較

國家（地區）	n	b'	ρ	φ	$SE_{(2003/2007)}$	SE_{JRR}	$SE_{(2003/2007)} / SE_{JRR}$
Armenia	4,689	37.65	0.21	-0.04	$0.034\sigma_{Armenia}$	$0.054\sigma_{Armenia}$	0.63
Cyprus	4,399	31.20	0.09	0.49	$0.023\sigma_{Cyprus}$	$0.023\sigma_{Cyprus}$	0.98
Indonesia	4,203	31.00	0.46	0.35	$0.047\sigma_{Indonesia}$	$0.044\sigma_{Indonesia}$	1.06
Israel	3,294	25.10	0.26	0.34	$0.034\sigma_{Israel}$	$0.042\sigma_{Israel}$	0.85
Italy	4,408	33.40	0.28	0.25	$0.043\sigma_{Italy}$	$0.036\sigma_{Italy}$	1.19
Malaysia	4,466	32.64	0.50	0.13	$0.049\sigma_{Malaysia}$	$0.068\sigma_{Malaysia}$	0.72
Taiwan	4,046	27.90	0.25	-0.04	$0.039\sigma_{Taiwan}$	$0.040\sigma_{Taiwan}$	0.97

最後一個決定各種估計方法所得誤差大小的最重要因素且無法在事前得知的重要數據就是樣本的變異量（σ^2_{total}），或是樣本的標準差（σ_{total}）。這一點倒無須過度擔心，在大部分的時候，尤其是非跨群體比較的調查，主要調查變項的標準化分數量尺都是以樣本的標準差（σ_{total}）作爲單位；換句話說，標準化之後的平均分數之標準誤相當於將式(20)中的 σ_{total} 用 1

來代替的情形。如果是在像 TIMSS 這類跨國比較的國際調查中，換算成統一的國際量尺時，就必須知道各國學生的標準差（σ_{total}）相當於多少國際量尺分數。表 4 列出同時參加 TIMSS 2003 與 TIMSS 2007 的各國（地區）八年級生科學平均成就之標準差，不難發現，除了亞美尼亞（Armenia）、香港（Hong Kong）、以色列（Israel）、馬來西亞（Malaysia）以及巴勒斯坦（Palestinian Nat'l Auth.）等 5 個國家（地區）的八年級生在兩屆調查時，科學成就標準差的差異比較大之外，其餘有 6 個國家（地區）的差異大約都在 15%，18 個國家（地區）的差異在 10%以內。因此，即使有需要在事前預估調查變項的標準差，利用過去類似的研究結果作爲估計的依據通常也可以作爲相當好的參考。

表 4 比較連續參加 TIMSS 2003 與 TIMSS 2007 趨勢調查國家（地區）在兩屆調查中八年級科學成就標準差的比較

國家（地區）	σ_{2003}	σ_{2007}	$\sigma_{2003}/\sigma_{2007}$
Armenia	81	101	0.80
Australia	75	80	0.94
Bahrain	74	86	0.86
Botswana	86	99	0.87
Bulgaria	93	103	0.90
Taiwan	79	89	0.89
Cyprus	79	85	0.93
Egypt	104	99	1.05
England	77	85	0.91
Ghana	120	108	1.11
Hong Kong, SAR	66	81	0.81
Hungary	76	77	0.99
Indonesia	79	74	1.07
Iran, Islamic Rep. of	73	81	0.90
Israel	85	101	0.84
Italy	78	78	1.00
Japan	71	77	0.92
Jordan	89	98	0.91
Korea, Rep. of	70	76	0.92
Lebanon	93	97	0.96
Lithuania	70	78	0.90
Malaysia	66	88	0.75
Morocco	69	79	0.87
Norway	70	73	0.96
Palestinian Nat'l Auth.	92	111	0.83

表 4　（續）比較連續參加 TIMSS 2003 與 TIMSS 2007 趨勢調查國家（地區）在兩屆調查中八年級科學成就標準差的比較

國家（地區）	σ_{2003}	σ_{2007}	$\sigma_{2003}/\sigma_{2007}$
Romania	91	88	1.03
Russian Federation	75	78	0.96
Saudi Arabia	72	78	0.92
Scotland	76	81	0.94
Serbia	84	85	0.99
Singapore	92	104	0.88
Slovenia	67	72	0.93
Sweden	74	78	0.95
Tunisia	60	60	1.00
United States	81	82	0.99

肆、分析三

在分析三中，我們進一步考慮當叢集的分層輔助變項爲連續變項的情形。此時，由式(20)可推論母群平均值的抽樣誤差下限爲：

$$SE_{\mathrm{LB}} \approx \sqrt{\{[1+\rho(b^{'}-1)]-\rho r^2 b^{'}\}\times\frac{\sigma^2}{n-1}} \tag{24}$$

其中 r 爲叢集的分層輔助變項與叢集平均值的相關，r^2 即爲該輔助變項解釋叢集平均值變異量的比例，相當於式(20)中的 φ。當在每一個針對叢集的分層內，所有叢集的分層輔助變項完全一樣時，式(24)中的等號才會成立。因此，若叢集的分層輔助變項爲連續變項時，除非叢集的分層數爲無限大，或是每個叢集自成一層，否則此一條件將無法滿足。本分析將探討當分層輔助變項爲連續變項時，分層數（H）與母群平均值估計誤差（SE）的關係。

當叢集的分層輔助變項（A）爲連續變項時，若假設所有叢集之分層輔助變項大小爲常態分布，則在每一層內各叢集輔助變項的變異量（$\sigma^2_{A_h}$）之總和（$\Sigma_{h=1}^{H}\sigma^2_{A_h}$）會和分層數（$H$）有關。一般而言，層內叢集輔助變項之變異量總和會隨著分層的數量增加而減少。Cochran（1963）證明在一階段分層隨機抽樣設計下，利用分層輔助變項進行最適配置（optimal allocation, OA）的分層抽樣時，分層輔助變項解釋總變異量比例約爲 $\left(1-\frac{1}{H^2}\right)r^2$。所謂最適配置的分層抽樣係指每一層內所抽取的樣本人數應正比於分層輔助變項在該層內的標準差；也就是說，分層輔助變項變異愈大的層內，應抽取愈多的樣本，此抽樣方式可以使得整體抽樣誤差達到最小值。

除了最適配置分層抽樣法之外，另一個常見的分層方法爲利用分層輔助變項的均等配置分層法（OA）；也就是在每一分層內抽取相同的樣本數，這種方法不考慮輔助變項的資訊，也因爲缺乏分層資訊，通常不會很精確。但我們採用的分層法是事先透過分層輔助變項，在調查之前將各叢集依其輔助變項（A）的大小排序，並列出各叢集所含標的母群的人數後，依叢集累加人數進行均等配置分層；依此法對所有叢集進行分層後，每一分層內所包含的母群人數皆相同，而後再從各層中抽取相同的樣本。爲了與一般的均等配置抽樣做區分，研究者稱這種方法爲均等分層下的等比例配置法（Proportional Allocation for Equal Stratification, PAES）。接下來的分析將暫時把抽樣單位設定爲叢集，比較利用輔助變項的 PAES 與 OA 兩種分層方法對母群平均值估計誤差的影響。

在均等分層的等比例配置（PAES）叢集抽樣設計下，爲了估計叢集之分層輔助變項在各分層內的變異量，我們假設所有叢集的分層輔助變項（A）成常態分布。首先，因爲每個叢集被抽到的機率與叢集大小成正比，各層內所有叢集被抽到的機率累加後應該正比於該層內叢集所包含調查母群人數的總和；因此在均等分層的抽樣架構下：

$$\int_{-\infty}^{a_1} P(A)\cdot dA = \int_{a_1}^{a_2} P(A)\cdot dA = \ldots = \int_{a_{h-1}}^{a_h} P(A)\cdot dA = \ldots = \int_{a_{H-1}}^{\infty} P(A)\cdot dA$$

$$= \frac{1}{H} \tag{25}$$

式(25)中的 $P(A)$爲常態分布，爲了簡化計算，我們將 A 的平均值設爲原點（如圖 3），這個假設並不會影響對於叢集的分層輔助變項 A 之變異量估計：

$$P(A) = \frac{A}{\sqrt{2\pi\sigma_A^2}} \cdot e^{-\frac{A^2}{2\sigma_A^2}} \tag{26}$$

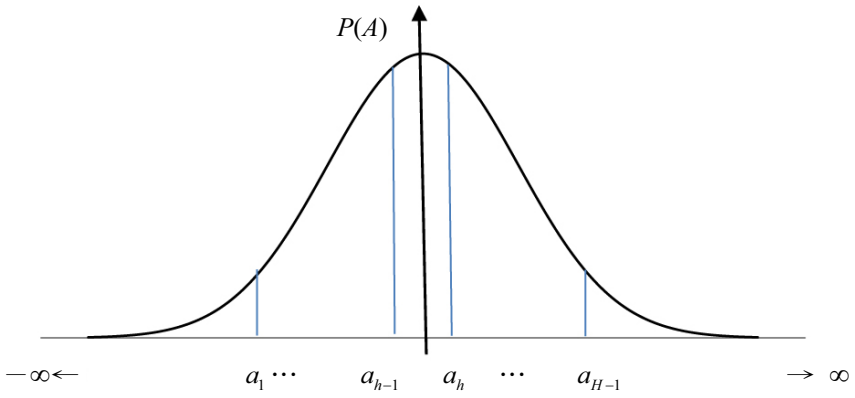

圖3　所有叢集之分層輔助變項（A）大小之機率分布

其中，σ_A^2 爲所有輔助變項 A 之變異量，我們根據式(25)和式(26)求出當叢集的分層數 H 分別爲 2、3、4、…、10 時，根據輔助變項 A 進行等人數分層的分層區間（$-\infty, a_1$）、（a_1, a_2）、…、（a_{h-1}, a_h）、…（a_{H-1}, ∞）。接下來，根據式(27)、(28)以及附錄中的積分公式（A-5 和 A-6）估計出每個分層內各叢集輔助變項之平均值 $E(A)$以及變異量 $E(\sigma_A^2)$。其中 $E(A \mid a_{h-1} \le A < a_h)$ 爲第 h 分層內各叢集的分層輔助變項（A）之平均值（$\overline{A}_h$）：

$$\overline{A}_h = E(A \mid a_{h-1} \le A < a_h) = \frac{\int_{a_{h-1}}^{a_h} P(A) \cdot A \cdot dA}{\int_{a_{h-1}}^{a_h} P(A) \cdot dA} \tag{27}$$

而第 h 分層內，$E(\sigma_A^2 \mid a_{h-1} \le A < a_h)$ 即爲分層輔助變項（A）之變異量爲（$\sigma_{A_h}^2$）：

$$\sigma_{A_h}^2 = E(\sigma_A^2 \mid a_{h-1} \le A < a_h) = \frac{\int_{a_{h-1}}^{a_h} P(A) \cdot [A - E(A)]^2 \cdot dA}{\int_{a_{h-1}}^{a_h} P(A) \cdot dA} \tag{28}$$

表 5 列出利用數值分析方法對於採用均等分層的等比例配置（PAES）進行分層叢集抽樣時，不同分層數的輔助變項分層區間，及對於分層輔助變項 A 而言，各分層內叢集變異量之總和（$\Sigma_{h=1}^{H}\sigma_{A_h}^2$）。

根據表 5 的數值分析結果，當分層數愈多的時候，各分層內叢集輔助變項變異量之總和愈小，亦即叢集輔助變項的變異量被分層解釋的部分愈大。如果分層數趨近無限大，使得各層內叢集輔助變項之變異趨近 0，式(24)的等號就會成立，此時母群平均值估計誤差最小。而當分層數爲有限數值時，各分層內輔助變項無法完全被控制成完全相等，此時分層解釋輔助變項之比例 $R^2(H)$ 爲：

$$R^2(H) = 1 - \frac{\Sigma_{h=1}^{H}\sigma_{A_h}^2}{\sigma_A^2} \tag{29}$$

其中，$R^2(H)$ 爲叢集的分層數（H）之函數。又輔助變項解釋主要調查變項的叢集間變異量之比例爲 r^2，所以式(20)中的 φ 必須用 $r^2R^2(H)$ 取代，改寫式(20)爲：

$$SE_{\text{two-stage}} \approx \sqrt{[1 + \rho(b' - 1) - \rho r^2 R^2(H) b'] \times \frac{\sigma_{\text{total}}^2}{n-1}} \tag{30}$$

表 5　利用數值分析法，估計對叢集採 PAES 抽樣時，不同分層數對應之輔助變項分層區間以及各區間輔助變項變異量之總和

分層數（H）	輔助變項分層區間（單位：σ_A）	變異量總和（$\Sigma_{h=1}^{H}\sigma_{A_h}^2$）
1	(-∞, ∞)	$1.000\,\sigma_A^2$
2	(-∞, 0), (0, ∞)	$0.363\,\sigma_A^2$
3	(-∞, -0.431), (-0.431, 0.431), (0.431, ∞)	$0.207\,\sigma_A^2$
4	(-∞, -0.674), (-0.674, 0), (0, 0.674), (0.674, ∞)	$0.139\,\sigma_A^2$
5	(-∞, -0.842), (-0.842, -0.253), (-0.253, 0.253), (0.253, 0.842), (0.842, ∞)	$0.103\,\sigma_A^2$
6	(-∞, -0.967), (-0.967, -0.431), (-0.431, 0), (0, 0.431), (0.431, 0.967), (0.967, ∞)	$0.081\,\sigma_A^2$
7	(-∞, -1.068), (-1.068, -0.566), (-0.566, -0.180), (-0.180, 0.180), (0.180, 0.566), (0.566, 1.068), (1.068, ∞)	$0.066\,\sigma_A^2$
8	(-∞, -1.150), (-1.150, -0.674), (-0.674, -0.319), (-0.319, 0), (0, 0.319), (0.319, 0.674), (0.674, 1.150), (1.150, ∞)	$0.055\,\sigma_A^2$
9	(-∞, -1.221), (-1.221, -0.765), (-0.765, -0.431), (-0.431, -0.140), (-0.140, 0.140), (0.140, 0.431), (0.431, 0.765), (0.765, 1.221), (1.221, ∞)	$0.047\,\sigma_A^2$
10	(-∞, -1.282), (-1.282, -0.842), (-0.842, -0.524), (-0.524, -0.253), (-0.253, 0), (0, 0.253), (0.253, 0.524), (0.524, 0.842), (0.842, 1.282), (1.282, ∞)	$0.041\,\sigma_A^2$

如前所述，當採最適配置（OA）時，$R^2(H)$ 即爲 $\left(1-\frac{1}{H^2}\right)$；而採均等分層的等比例配置（PAES）時，雖沒有一簡單公式可循，但根據表 5 的數值分析結果，可以比較當利用輔助變項對所有叢集進行分層時，不同分層數的兩種配置抽樣法解釋叢集平均值變異量之比例。表 6 和圖 4 顯示比較結果，當分層數愈大時，兩種抽樣方式的誤差愈接近。雖然採最適配置的方式對叢集進行分層能夠讓相同樣本下的抽樣誤差最小，但是事前的計算較 PAES 來的複雜許多，且由於採 PAES 法抽樣時，盡可能讓每一位樣本的代表性差異最小，加權值的計算也較容易。

最後，研究者利用分析的結果對於我國參與 TIMSS 2011 調查提出一個學校層級的抽樣架構：由於 TIMSS 參加國必須在正式調查開始的 2 年以前，提供加拿大統計局該國學校層級的分層抽樣架構。本研究根據我國參加 TIMSS 2007 的數據以及不違反保密原則的情況下，透過國立臺灣師範大學基本學力測驗中心的協助，估計出我國參加 TIMSS 2007 調查之各個國中八年級平均科學成就與該校九年級生同年參加基本學力測驗數學科平均成績之相關爲 0.71，且同一學校每連續 2 年基測數學平均成績相關爲 0.97；所以若由正式調查 2 年前（2009 年）各

表 6　在不同分層數的情況下，採均等分層下的等比例配置（PAES）叢集抽樣與最適配置（OA）叢集抽樣對叢集間變異量解釋之比例

分層數 (H)	PAES抽樣 $R^2(H)$	OA抽樣 $\left(1-\frac{1}{H^2}\right)$
1	.00	.00
2	.64	.75
3	.79	.89
4	.86	.94
5	.90	.96
6	.92	.97
7	.93	.98
8	.95	.98
9	.95	.99
10	.96	.99

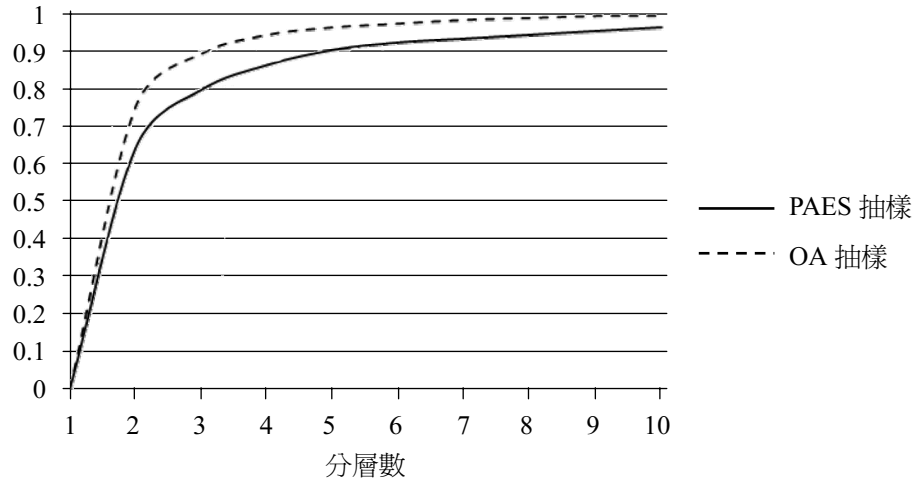

圖4　在不同分層數的情況下，採均等分層下的等比例配置（PAES）叢集抽樣與最適配置（OA）叢集抽樣對叢集間變異量解釋之比例

校之九年級生參加基本學力測驗之平均數學成績作爲 TIMSS 2011 正式調查時，[2]學校分層之輔助變項，估計該輔助變項與我國參加 TIMSS 2011 調查之各校八年級生平均科學成就相關之

[2] 依據分析結果，各參與 TIMSS 2007 調查之國中，不論是 TIMSS 調查的八年級生數學或科學平均成就，和該校同年度九年級生參加基測之數學平均成績之相關，均比與基測科學成績的相關來得高。

下限約爲 0.67（$=0.97^2\times0.71$）。由於推導式(30)時，我們假設叢集之分層輔助變項平均值成常態分布，所以針對我國所有國中於 2009 年之第一次基測數學平均成績進行常態分布之 Kolmogorov-Smirnov 檢驗，結果顯示該假設可以被接受（$p=$ .09）。因此，我國 TIMSS 調查中心委請基本學力測驗中心協助將所有包含八年級生之中學，按照各校在 2009 年第一次基本學力測驗的數學平均成績，依據均等分層的方式分成八層，其中每一層所包含的學生人數大致相同。依此架構作爲我國參加 TIMSS 2011 八年級群之學校分層抽樣架構，如果我國八年級生在 TIMSS 2011 與 TIMSS 2007 的表現情況大致相同時（$n=$4,046, $\rho=$0.22, $\sigma^2_{\text{total}}=$7,971, $b'=$27.9），且當叢集的分層輔助變項與叢集平均值相關（r）爲 .67 且分層數（H）爲 8 的情況下，由式(30)可預估我國參加 TIMSS 2011 八年級生科學平均成績之抽樣誤差約爲 2.9 量尺分數。

伍、結論

近年來，我國愈來愈重視國際與國內的大型教育調查研究，且這類調查大多採二階段分層叢集抽樣設計。許多研究者在進行二階段分層叢集抽樣調查之前，並不清楚抽樣架構、樣本數與抽樣誤差間的關係。本研究導證一個簡易且方便計算的公式，提供研究者在進行二階段分層叢集抽樣調查之前，能夠藉由相關的輔助訊息，估算出主要調查變項的抽樣誤差。分析一和分析二以 TIMSS 參加國（地區）的數據爲實例，檢驗該公式之有效性和實用性。此外，爲了能夠讓我國在參加 TIMSS 2011 調查時，八年級生科學平均成就的抽樣誤差減少至 3.0 量尺分數以下，利用各國中在基本學力測驗的平均表現作爲該調查中，學校分層的輔助變項；透過分析三，除了推論在二階段分層叢集抽樣設計下，叢集的分層數與抽樣誤差間之關係，並據以預估我國在 TIMSS 2011 調查中，八年級生科學平均成就之抽樣誤差約爲 2.9 量尺分數。

最後，本研究嘗試提出一個四個階段的標準化評估流程，協助研究者在進行二階段分層叢集的抽樣調查前，對主要調查變項母群平均值之抽樣誤差進行評估。分別說明各階段如下：

（一）階段一

根據過去的調查和文獻，訂定合理且可以回答研究問題的誤差範圍。

舉例來說，如果想在臺灣進行一項教育調查，探討各縣市中小學生成就的差異，首先必須根據過去的研究或是相關的資料庫研判各縣市中小學生學科成就可能的差異大小。若依據過去的研究調查結果，如澳洲的 NAPLAN（National Assessment Program － Literacy and Numeracy）（Australian Curriculum, Assessment and Reporting Authority [ACARA], 2010）和芝加哥的 CARP（Chicago Annenberg Research Project）（Newmann, Smith, Allensworth, & Bryk, 2001），國小學生基本能力的發展（如數學或語言能力），每一年的進步大約爲 0.5-0.7 個邏輯

斯（logits），這也大約相當於 0.5-0.7 個學生能力的標準差；到了中學，每一年的進步甚至可能小於 0.2-0.3 個標準差以下。我國中小學因採標準課綱，所以各縣市同一年級學生平均學科能力的差異應該小於 0.1 個標準差。因此，若想藉由抽樣調查分辨各縣市的差異，則調查的誤差至少應小於 0.05 個標準差，否則調查結果必然無法分辨其間差異。又以我國將參加 TIMSS 2011 調查之八年級生科學平均成就的抽樣誤差範圍定爲 3.0 量尺分數爲例：由於我國參加 TIMSS 2007 調查時，八年級生的科學成就的標準差約爲 89 量尺分數，抽樣誤差小於 3.0 量尺分數的要求，相當於要求抽樣誤差小於 0.034 個標準差。接下來的階段均將依此目標作爲實例，示範對我國參加 TIMSS 2011 之抽樣架構的評估流程。

（二）階段二

根據過去的研究或相關資料庫的數據，計算出叢集內相關，藉以估算抽樣設計效應以及所需最大樣本數的近似值。

爲了達到預定的精確性，採不同的抽樣方式需要不同的樣本數。以母群平均值的標準誤爲 0.034 個標準差爲例，若採簡單隨機抽樣，大約需抽取 865 個樣本。若是以學校或校內班級爲單位進行叢集抽樣，則調查的精確度與調查變項的校內或班級內相關有關。當採二階段分層叢集抽樣設計時，若僅有叢集內相關的資訊，而缺乏分層輔助變項與叢集平均值的相關時，可以先假設後者爲 0，此時等同於採一階段的叢集抽樣。例如，可利用 TIMSS 2007 調查結果得到我國八年級生在該年度科學成就之班級內相關（0.22）以及平均班級學生人數（約 28 人），以此作爲我國參加 TIMSS 2011 時，八年級生科學成就班內相關以及平均班級人數之近似值。以抽樣誤差爲 0.034 個標準差的精確性要求而言，若採學校或班級爲叢集單位進行一階段叢集抽樣調查，且叢集內樣本數爲 28 人時，可利用 Hansen 等（1953）推導的公式（式(4)）估計抽樣設計效應爲 6.94，亦即所需樣本數約爲採隨機抽樣人數的 6.94 倍，也就是需要抽取約 6,000 名樣本，這個數量超過 TIMSS 調查預計從 150 所學校中各抽出約 28 名學生（共約 4,200 名）的樣本數。換句話說，若無法藉由對叢集的分層以減少抽樣誤差，且按 TIMSS 調查預計抽取約 4,200 個學生的設計，其抽樣誤差應無法滿足小於 3.0 量尺分數（0.034 個標準差）的要求。

（三）階段三

根據過去的研究或相關資料庫的數據計算出叢集的分層輔助變項與叢集平均值的相關，藉以估算出抽樣設計效應以及所需樣本數的近似下限。

如果進行叢集抽樣之前能夠先將所有的叢集根據與主要調查變項相關的輔助變項對所有的叢集進行分層，可以在與一階段叢集抽樣具相同精確性的要求下，令所需的樣本大幅減少。若事前可以估計出叢集的分層輔助變項與叢集平均值的相關，藉由本研究所推導的式(21)加上叢集內相關的近似值，研究者得以估計抽樣設計效應的下限。延續先前的例子，假若已知學

校分層輔助變項與學校平均值的相關爲 0.67，且班級內相關爲 0.22，班內樣本大小約爲 28，則由式(21)可求得抽樣設計效應約爲 4.2。也就是說，欲達到抽樣誤差爲 0.034 個標準差的精確性要求，所需要的樣本數約爲 3,600，這大約是一階段叢集抽樣所需樣本數的五分之三。

（四）階段四

根據所擁有的人力、時間和經費等資源，判斷採二階段分層叢集抽樣是否可以滿足對於誤差範圍的要求？若所需樣本數的下限能夠被接受，進一步估算叢集的最小分層數量以確保叢集的分層作業能夠達到對抽樣誤差的要求。

由前兩個步驟的評估結果可以得知，在誤差小於 0.034 個標準差的要求下，利用與叢集平均值相關爲 0.67 之分層輔助變項，對於叢集進行分層後，再在各層內進行叢集抽樣，若叢集樣本大小爲 28，則最少需樣本數爲 3,600，此時各層內所有叢集之分層輔助變項必須相同；但若叢集的分層輔助變項與叢集平均值零相關時，等同於進行一階段的叢集抽樣，則需要 6,000 樣本才能達到相同精確度的要求。

由於樣本數會直接影響調查所需要的人力、經費與時間，例如實測、閱卷、資料輸入等步驟，經常是耗費資源的重要原因。因此研究者在擬訂計畫時，應考慮每個不同的步驟所需要花費的經費與時間，決定可以負擔的樣本數。若這個數目大於前一階段所估算出利用輔助變項對叢集進行分層後，再進行叢集抽樣的最少樣本數，就表示只要分層數夠多，應可達到研究對於誤差範圍的要求。對於最少分層數的要求則需要考慮預計的總樣本數（n）、叢集內樣本數（b'）、叢集內相關（ρ）、分層輔助變項與叢集平均值的相關（r），以及可容許的最大抽樣誤差等資訊（$SE_{\text{two-stage}}$），帶入式(30)算出在特定分層數下，叢集間變異量被分層解釋的比例（$R^2(H)$）之最小值，再利用表 6 找出對應的分層數（H），即可決定出最少分層數。

以我國參加 TIMSS 2011 八年級調查的抽樣架構爲例：預估學校分層輔助變項與主要調查變項學校平均值的相關（r）爲 0.67，主要調查變項之校內相關（ρ）爲 0.22，校內樣本大小（b'）爲 28。由前一步驟可以得知，在可容許抽樣誤差爲 0.034 個標準差（約相當於 3.0 量尺分數）的要求下，估算出所需最小樣本數爲約爲 3,600。根據我國參加 TIMSS 2011 的預算與抽樣方式，預估樣本數（n）約爲 4,200，大於所需最小樣本數，表示有機會達到對於抽樣誤差小於 0.034 個標準差的要求。接下來可由式(30)和表 6 推論出採均等分層的等比例配置法（PAES）進行抽樣時，能滿足抽樣誤差要求的最小的分層數。首先，將上述數據帶入式(30)，可求得當 $R^2(H)$等於 0.75 時，抽樣誤差剛好爲 0.034 個標準差；再比對表 6 的數據，可知最小分層數爲 3，才能令 $R^2(H)$的數值大於 0.75。由此這個結果可以推論，當調查參數與近似參數大約相同時，至少需將學校依據其基測平均成績以及均等分層的方式分成三層，始能達到抽樣誤差小於 3.0 量尺分數的要求。但由於實際抽樣時的參數值可能與預估值不同，且分層數愈多，抽樣誤差愈小，所以建議條件許可的話，可以增加抽樣架構中叢集的分層數。這也是我國在 TIMSS 2011 的學校抽樣架構中，將八年級群學校分層數定爲八層的主要原因。

陸、研究限制

本研究期能藉由所推導的公式，幫助研究者在事前對於調查抽樣的精確性進行有效的評估。一個調查研究誤差的主要來源，如本研究在一開始所描述，除了抽樣誤差所造成的影響，另一個部分爲測量誤差。本研究的討論範圍僅限於抽樣誤差的估計，在 TIMSS 和 PISA 等大型調查的研究中，對母群的平均值而言，抽樣誤差遠比測量誤差來得大；以 TIMSS 2007 爲例，大部分國家（地區）成就平均值的測量誤差對標準誤的校正爲 0.1-0.2 個量尺分數，約爲抽樣誤差的數十分之一（IEA, 2009）。此外，本研究所推導的公式乃立基於三個較爲嚴格的假設上，當這些假設違犯程度不同時，會對所推導公式的偏誤有多大的影響，也將是研究者下一步的研究方向。

誌謝

本研究承蒙行政院國家科學委員會科學教育發展處專案計畫經費補助（計畫編號：NSC97-2511-S-003-045-MY5、NSC97-2522-S-003-001）以及國立臺灣師範大學基本學力測驗中心協助提供相關數據；此外，審查者所提供的專業建議也提升本研究的可讀性，特此感謝。

參考文獻

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162-172.

Australian Curriculum, Assessment and Reporting Authority. (2010). *NAPLAN summary report: Achievement in reading, writing, language conventions and numeracy*. Sydney: Author.

Cochran, W. G. (1963). *Sampling techniques*. New York: Wiley.

Demnati, A., & Rao, J. N. K. (2004). Linearization variance estimators for survey data. *Survey Methodology, 30*(1), 17-26.

Donner, A., & Klar, N. (1994). Methods for comparing event Rates in intervention studies when the unit of allocation is a cluster. *American Journal of Epidemiology, 140*(3), 279-289.

Dorofeev, S., & Grant, P. (2006). *Statistics for real life sample surveys: Non-simple-random samples and weighted data*. New York: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*(1), 1-26.

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 225-279). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Foy, P., & Olson, J. F. (Eds.). (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Frankel, M. R. (1971). *Inference from survey samples*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory*. New York: Wiley.

Joncas, M. (2008). TIMSS 2007 sample design. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 77-92). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kish, L. (1965). *Survey sampling*. London: John Wiley & Sons.

Krus, D. J., & Helmstadter, G. C. (1993). The problem of negative reliabilities. *Educational and Psychological Measurement, 53*(3), 643-650.

Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys, statistics in practice* (2nd ed.). New York: John Wiley & Sons.

Martin, M. O. (Ed.). (2005). *TIMSS 2003 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., Olson, J. F., Erberber, E., Preuschoff, C. et al. (2008). *TIMSS 2007 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Newmann, F. M., Smith, B., Allensworth, E., & Bryk, A. S. (2001). *School instructional program coherence: Benefits and challenges*. Chicago: Consortium on Chicago School Research.

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris: Author.

Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika, 33*(4), 305-325.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Solomon, D. J. (2004). The rating reliability caculator. *BMC Medical Research Methodology, 4*(1), 1-3.

The International Association for the Evaluation of Education Achievement. (2005). *TIMSS 2003 international database*. Retrieved March 31, 2009, from http://timss.bc.edu/timss2003i/userguide.html

The International Association for the Evaluation of Education Achievement. (2009). *TIMSS 2007 international database*. Retrieved March 31, 2009, from http://timss.bc.edu/timss2007/idb_ug.html

附錄

一、式(13)之推導過程

由式(8)、式(10)以及式(11)可得：

$$\begin{aligned}\mathrm{Var}(\bar{x}) &\approx \frac{\Sigma_h n_h^2[1+\rho_h(b_h^{'}-1)]\frac{s_h^2}{n_h}}{n^2}\\ &\approx \frac{\Sigma_h n_h[1+\rho_h(b^{'}-1)]\sigma_h^2}{n^2}\\ &\approx \frac{\Sigma_h n_h(\sigma^2_{\text{within-cluster}_h}+b^{'}\sigma^2_{\text{between-cluster}_h})}{n^2}\end{aligned} \tag{A-1}$$

上式最後一個近似等號的成立用到了式(5)對於叢集內相關的定義：

$$\begin{aligned}\rho_h &= 1-\frac{s^2_{\text{within-cluster}_h}}{s_h^2}\frac{n_h}{n_h-1}\\ &\approx \frac{\sigma^2_{\text{between-cluster}_h}}{\sigma^2_{\text{between-cluster}_h}+\sigma^2_{\text{within-cluster}_h}}\end{aligned} \tag{A-2}$$

以及在第 h 分層內，母群參數的變異量（$\sigma^2_{\text{total}_h}$）可以被分解成叢集間的變異量（$\sigma^2_{\text{between-cluster}_h}$）加上叢集內的變異量（$\sigma^2_{\text{within-cluster}_h}$）之關係式：

$$\sigma^2_{\text{total}_h}=\sigma^2_{\text{between-cluster}_h}+\sigma^2_{\text{within-cluster}_h} \tag{A-3}$$

根據變異量的定義，式(A-1)可改寫爲：

$$\begin{aligned}\mathrm{Var}(\bar{x}) &= \frac{\Sigma_{h=1}^{H} n_h\left[\frac{\Sigma_{j=1}^{m_h}\Sigma_{i=1}^{b'}(x_{ijh}-\bar{x}_{jh})^2}{n_h}+b^{'}\frac{\Sigma_{j=1}^{m_h}(\bar{x}_{jh}-\bar{x}_h)^2}{m}\right]}{n^2}\\ &= \frac{\Sigma_{h=1}^{H}\left[\Sigma_{j=1}^{m_h}\Sigma_{i=1}^{b'}(x_{ijh}-\bar{x}_{jh})^2+b'^{2}\Sigma_{j=1}^{m_h}(\bar{x}_{jh}-\bar{x}_h)^2\right]}{n^2}\\ &= \frac{\frac{\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}\Sigma_{i=1}^{b'}(x_{ijh}-\bar{x}_{jh})^2}{n}+\frac{b'^{2}\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}(\bar{x}_{jh}-\bar{x}_h)^2}{Hmb^{'}}}{n}\end{aligned}$$

$$= \frac{\frac{\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}\Sigma_{i=1}^{b'}(x_{ijh}-\bar{x}_{jh})^2}{n} + \frac{b'\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}(\bar{x}_{jh}-\bar{x}_h)^2}{Hm}}{n}$$

$$= \frac{\frac{\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}\Sigma_{i=1}^{b'}(x_{ijh}-\bar{x}_{jh})^2}{n} + \frac{b'\Sigma_{h=1}^{H}\Sigma_{j=1}^{m_h}\left[(\bar{x}_{jh}-\bar{x})^2-(\bar{x}_h-\bar{x})^2\right]}{Hm}}{n}$$

$$= \frac{\sigma^2_{\text{within-cluster}} + b'(\sigma^2_{\text{between-cluster}} - \sigma^2_{\text{auxiliary-variable}})}{n} \tag{A-4}$$

其中下標 h, j, i 分別對應於第 h 個分層、第 j 個叢集以及第 i 個樣本，且 $i = 1, \ldots, b'$，$j = 1, \ldots, m_h$，$h = 1, \ldots, H$，以及 $\Sigma_h m_h = Hm$ 。

二、式(27)與式(28)之積分公式

$$\bar{A}_h = E(A \mid a_{h-1} \le A < a_h) = \frac{\int_{a_{h-1}}^{a_h} P(A)\cdot A\cdot dA}{\int_{a_{h-1}}^{a_h} P(A)\cdot dA}$$

$$= \frac{\int_{a_{h-1}}^{a_h} \frac{A}{\sqrt{2\pi\sigma_A^2}}\cdot e^{-\frac{A^2}{2\sigma_A^2}}\cdot dA}{\int_{a_{h-1}}^{a_h} P(A)\cdot dA}$$

$$= \frac{\frac{\sigma_A}{\sqrt{2\pi}}\left(e^{-\frac{a_{h-1}^2}{2\sigma_A^2}} - e^{-\frac{a_h^2}{2\sigma_A^2}}\right)}{\int_{a_{h-1}}^{a_h} P(A)\cdot A\cdot dA}$$

$$= \frac{H\sigma_A}{\sqrt{2\pi}}\left(e^{-\frac{a_{h-1}^2}{2\sigma_A^2}} - e^{-\frac{a_h^2}{2\sigma_A^2}}\right) \tag{A-5}$$

$$\sigma^2_{A_h} = E(\sigma_A^2 \mid a_{h-1} \le A < a_h)$$

$$= \frac{\int_{a_{h-1}}^{a_h} P(A)\cdot[A - E(A)]^2\cdot dA}{\int_{a_{h-1}}^{a_h} P(A)\cdot dA}$$

$$= \frac{\int_{a_{h-1}}^{a_h} \frac{e^{-\frac{A^2}{2\sigma_A^2}}}{\sqrt{2\pi\sigma_A^2}} \cdot \left[A - \frac{H\sigma_A}{\sqrt{2\pi}} \left(e^{-\frac{a_{h-1}^2}{2\sigma_A^2}} - e^{-\frac{a_h^2}{2\sigma_A^2}} \right) \right]^2 \cdot dA}{\int_{a_{h-1}}^{a_h} P(A) \cdot dA}$$

$$= \frac{H\sigma_A}{4\sqrt{2\pi}} \cdot e^{-\frac{2a_{h-1}^2 + 2a_h^2 + A^2}{2\sigma_A^2}}$$

$$\cdot \left[4\sqrt{2} H\sigma_A \cdot e^{\frac{a_{h-1}^2 + a_h^2}{2\sigma_A^2}} \cdot \left(e^{\frac{a_h^2}{2\sigma_A^2}} - e^{\frac{a_{h-1}^2}{2\sigma_A^2}} \right) - 4\sqrt{\pi} \cdot A \cdot e^{\frac{a_{h-1}^2 + a_h^2}{\sigma_A^2}} \right.$$

$$+ \sqrt{2}\sigma_A \cdot e^{\frac{A^2}{2\sigma_A^2}} \cdot erf\left(\frac{A}{\sqrt{2}\sigma_A} \right)$$

$$\left. \cdot \left(H^2 \cdot e^{\frac{a_{h-1}^2}{\sigma_A^2}} + H^2 \cdot e^{\frac{a_h^2}{\sigma_A^2}} - 2H^2 \cdot e^{\frac{a_{h-1}^2 + a_h^2}{2\sigma_A^2}} + 2\pi \cdot e^{\frac{a_{h-1}^2 + a_h^2}{\sigma_A^2}} \right) \right] \Bigg|_{A=a_{h-1}}^{A=a_h} \qquad \text{(A-6)}$$

Journal of Research in Education Sciences
2011, 56(1), 33-65

An Estimation of the Design Effect for the Two-Stage Stratified Cluster Sampling Design

Tsung-Hau Jen
Science Education Center, National Taiwan Normal University
Assistant Research Fellow

Hak-Ping Tam
Graduate Institute of Science Education, National Taiwan Normal University
Associate Professor

Margaret Wu
Assessment Research Centre, University of Melbourne
Associate Professor

Abstract

Most large-scale educational surveys utilize a multi-stage stratified cluster sampling design. Past findings revealed that the standard errors of Taiwan students' mean performances were slightly larger than other countries'. In response to the request by the institute in charge of TIMSS sampling, this study was launched to derive a formula that could estimate the standard error of population mean prior to conducting a two-stage stratified cluster sampling design. This formula could then be used to select an optimal stratification framework that could reduce the size of standard error to an acceptable level. Its validity was investigated in three subsequent studies. In the first study, standard errors for 30 TIMSS 2007 participating countries were estimated according to the newly derived formula as well as by the jackknife replication. The correlation between the two sets of standard errors amounted to 0.98. The second study investigated the practicality of using the new formula in addition to auxiliary variables for predicting standard errors on the data of 29 countries that participated in both TIMSS 2003 and 2007. The third study explored the relationship between the number of stratum and the standard errors under a two-stage stratified cluster sampling design when the auxiliary variables for stratification were continuous. This paper closed by suggesting a four-step procedure to facilitate researchers in estimating standard errors of means during the planning stage of sampling design.

Keywords: large-scale assessment, planning sampling framework, sampling error reduction, complex survey design, variance estimation

Corresponding Author: Tsung-Hau Jen, E-mail: tsunghau@ntnu.edu.tw