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：