

論權重與似真值在大型成就評量資料分析的意義、重要性與影響

葛湘璋

致理技術學院國際貿易學系副教授

一、前言

近年來許多國家關心教育現象，政府機構陸續委託專業學術單位對學生、教師及學校單位進行大規模的國內或國際調查。如美國有全國教育進展評量（National Assessment of Educational Progress，簡稱 NAEP），德國有國家長期追蹤調查（National Educational Panel Study，簡稱 NEPS），台灣有台灣教育長期追蹤資料庫（Taiwan Educational Panel Survey，簡稱 NEPS）。著名的國際大型成就評量（large-scale assessment，簡稱 LSA），有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），全球學生閱讀能力進展研究（Progress in Reading Literacy Study，簡稱 PIRLS），學生基礎素養國際研究計畫（Programme for International Student Assessment，簡稱 PISA），國際公民教育與素養研究（International Civic and Citizenship Education Study，簡稱 ICCS）。而因為 LSA 資料庫包含豐富可貴的資訊，近年來使用 LSA 資料庫的研究越來越盛行。如 TIMSS 的相關論文自 1995 年至 1999 年的 340 篇增為 2005 年至 2009 年的 851 篇（Rutkowski, Gonzalez, Joncas, & von Davier, 2010）。而 TEPS 相關的期刊論文、研討會論文及碩博士論文共 155 篇（中央研究院調查研究專題中心，2012）。

LSA 的主要目的為了解學生的學習成就與家庭背景、學校環境、教師等影響因素的關係，並探討學生的學習特色、學習行為與教育相關議題，以做為教育政策擬定者改進教育參考。而國際 LSA 更進一步進行國際比較或國際教育的交流合作，提供跨國觀摩教育現象的機會，在拓廣教育視野、制定教育決策有其獨特之參考價值。

LSA 一般均採複合調查設計（complex survey design），在資料整理與建立模式比一般資料分析複雜。在 LSA 中，受測學生非隨機取樣，且學生只回答測驗題庫中的部分題目。這種形式的設計在資料分析時需使用抽樣權重（sampling weight），因此，LSA 資料的變異數計算須同時考慮差補變異數（imputation variance）與抽樣變異數（sampling variance）。一般教科書所列之變異數的計算公式僅適用於隨機取樣及學生回答了所有試題的情形，不適用 LSA 的資料（Olson, Martin, & Mullis, 2008; OECD, 2009）。再者，國際 LSA 資料庫在學生的成就以似真值（plausible values, 簡稱 PVs）來估計學生能力，如果直接將 PVs 取其平均或只使用使單個 PV 會低估能力分配的變異，所得的統計量的標準誤有偏差（Willms & Smith, 2005; NCES, 2001）。因此在分析 LSA 時，不

宜將 PVs 取其平均數或只取單個 PV 進行資料分析。由於 LSA 在學生抽樣、試題抽樣、及以 PVs 來估計學生能力的特性，必須採用合適的資料分析方法，才能獲得學生成就的不偏母體估計值。

如前所述，在使用 LSA 資料分析時，須將 PVs 與權重列入考慮，但因計算過程繁雜，須視所使用的軟體是否具使用權重與計算 PVs 合併的功能，因此大部分已發表的論文或研討會論文仍採不正確或簡便的統計方法求平均數、變異數或其他的分析（Carstens & Hastedt, 2010）。研究者常用的錯誤分析方式有僅使用一個 PV 或將 PVs 取平均當依變數，不採用權重或選擇錯誤的權重。本文從實務應用的觀點舉例說明不使用權重對資料的分析影響及可能獲得錯誤的結論，同時，也提出在 LSA 資料分析使用 PVs 的討論及舉例說明變異數的計算方式，以提供使用國際 LSA 資料庫的研究者在次級分析（secondary analysis）時參考。

二、權重

LSA 採複合調查設計，子樣本（subsample）有時未反應母體子群的真正比例。權重主要是考慮到被選取學校／教師／學生的機率並非完全相同，每個受測單位的回應需調整到能反映出學生在母體中被抽中的真正比例，這種調整即是抽樣權重（sampling weight）（Rutkowski et al., 2010），譬如在 LSA 調查設計時，抽樣考慮男女學生的比例需近似相等，然而如果母體中男女學生比例有差異，此時有使用權重及未使用權重的結果會有很大的差異。因此，資料分析需依研究目的，選擇適當的權重類型，以獲得正確結論。

茲舉例說明權重對估計值的影響如下。假設一個班級有 24 個學生，16 位男生，8 個女生。如果要從此班隨機抽取六位學生進行調查以研究學生每週上網時間。依照母體比例，應抽 4 位男生，2 位女生。然而如果抽樣要男女生人數相等（亦即要抽取三位男生，三位女生），同時也要考慮到母體男女生真正比例的情形（母體女生比例小於男生），此時被抽取的學生的比例必須能反應真正的母群體男女生比例，因而在估計時必須要調整，此即所謂的抽樣權重。有採取及未採取抽樣權重的估計值會因子樣本（subsample）的比例不同而不同。以上例而言，女生的權重為 2.667，男生的權重為 5.333。假設此三位女生每週各花 2、3.5、1.5 小時上網，此三位男生每週各花 0.5、1、1.2 小時上網，如果沒有採取權重，則此六位學生的平均數為 $(2+3.5+1.5+0.5+1+1.2)/6=1.617$ 小時，而如果採取抽樣權重，則平均數為 $(2.667*(2+3.5+1.5)+5.333*(0.5+1+1.2))/24=1.378$ ，比未加權平均數少了 14.8%。另舉一例，如果此班有 24 位學生，10 位男生，14 位女生，要從中抽取六位學生，若依母體比例應抽取 2.5 位男生 3.5 位女生。若要抽取相等人數的男女生，即男生 3 位，女生 3 位，此時男女生權重各為 3.333 及 4.667。此時採抽樣權重所得平均數為 $(4.667*(2+3.5+1.5)+3.333*(0.5+1+1.2))/24=1.736$ ，比未加權的平均數高 7.39%。表 1 列出此兩例的比較。由此表可看出，使用權重與未使用權重，子樣本對估計值的影響。未使用權重可能高估或低估母體估計值，使用權重時，樣本數較小的子樣本會比未使用權重時對估計值的影響較大。

表 1 加權與未加權平均數例題比較

母體數		應抽樣本數		權重		上網時數						平均數		
男	女	男	女	男	女	男 1	男 2	男 3	女 1	女 2	女 3	未加權	加權	增減比例
16	8	4	2	5.333	2.667	0.5	1	1.2	2	3.5	1.5	1.617	1.378	-14.8%
10	14	2.5	3.5	3.333	4.667	0.5	1	1.2	2	3.5	1.5	1.617	1.736	7.4%

以 TIMSS 為例，TIMSS 使用二階段抽樣（Martin, Mullis, & Kennedy, 2007; Olson et al., 2008; Joncas, 2008; OECD, 2009），即先抽學校再抽班級，各年級抽樣人數有 150 所學校及 5000 位學生。各抽樣單位（學生 / 班級 / 學校）並非在相同機率下被抽，抽樣單位需適當加權，以代表母群體中子群的真正比例。TIMSS 採多種權重，分別為學生總數權重（Total Student Weight）、學生樣本權重（Student House Weight）、學生等數權重（Student Senate Weight）、整體及學科教師權重（Overall and Subject Teacher Weight）及學校權重（School Weight）。Rutkowski et al.（2010）、Foy 與 Olson（2008）、Williams, Ferraro, Roey, Brenwald, Kastberg, Jocelyn, Smith 與 Stearns（2009）在 TIMSS 資料分析時，對權重的選擇提出建議以作為研究者在使用 TIMSS 資料分析時參考。

茲將 TIMSS 權重的使用原則綜合歸納如下，表 2 為 TIMSS 的權重名稱、適用時機及分析層次之綜合整理，其他國際 LSA 資料庫的使用者在資料分析前亦須參考官方技術報告及相關文獻對權重類型與使用方式的說明。

（一）對樣本數敏感的分析（如卡方檢定）採用正規化權重（normalized weights），在 TIMSS 稱為 House weight。House Weight 是 Total Student

Weight 的線性轉換，總合等於樣本數。

（二）在做跨國分析時，合併多國資料後，必須注意到母體人數會影響到結果。例如美國八年級生有三百多萬人，賽浦勒斯八年級生有 9,237，馬爾他的八年級生只有 4,943 人（Olson et al., 2008）。在做國際數學成就的比較分析時，如果採用 Total Student Weight，結果會受到美國的大母體的影響。TIMSS 提供 Student Senate Weight 作為跨國比較。Student Senate Weight 是 Total Student Weight 的轉換，每個國家的權重學生樣本數（weighted student sample size）為 500，因此在跨國比較時，不論母群體的大小，每個國家的貢獻是相等的。在作跨國比較分析時，應採用 Student Senate Weight，以避免母體數差異太大造成研究結果偏誤。

（三）如果要將教師變數當為學生特質的分析時，適合使用 Overall Teacher Weight。如果是做數學成就分析時，則使用數學教師權重（Mathematics Teacher Weight），如果是做科學成就分析時，則使用科學教師權重（Science Teacher Weight）。但研究者須注意到，TIMSS 的教師並非隨機取樣，而是被抽選到的學生的教師回答教師問卷，因此在分析教師變數時，須採用教師 - 學生資料進行分析。

（四）如果研究興趣是在學校層次的特質，則使用學校層級的權重。

多層次的統計軟體如 HLM (Raudenbush, Bryk, Cheong, & Congdon, 2004) 及 Mplus (Muthen & Muthen, 1998-2010) 能讓使用者指定各層次的權重並將似真值列入模式參數計算的選項，而 TIMSS 計畫的主持單位國際教育成就評比學會 (International Association for the Evaluation of Educational Achievement，

簡稱 IEA) 發展出 TIMSS 資料的 IEA 國際資料庫分析軟體 (IEA International Database Analyzer Software，簡稱 IEA IDB Analyzer) 也提供了 TIMSS 研究者資料分析使用權重與 PVs 的選擇。在資料分析時，研究者需參考軟體使用手冊，官方的使用手冊及技術報告，配合研究目的與問題，選擇各層次合宜的權重。

表 2 TIMSS 的權重名稱、適用時機及分析層次

權重	變數名稱	說明	適用時機	分析層次
Total Student Weight	TOTWGT	加總等於國家受測年級母群體	單一國家研究	學生
Student Senate Weight	SENWGT	每個國家權重學生樣本數 合併多國資料，跨國研究 加總為 500		學生
Student House Weight	HOUWGT	加總為每個國家的學生樣本數 樣本數敏感的顯著性檢定		學生
School-level Weight	SCHWGT	學校層級的權重	學校	學校
Overall Teacher Weight	TCHWGT	整體教師的權重	教師及學生	學生 / 教師
Mathematics Teacher Weight	MATWGT	數學教師權重	數學教師及學生	學生 / 教師
Science Teacher Weight	SCIWGT	科學教師權重	科學教師及學生	學生 / 教師

三、似真值

在測驗施測時，測量誤差主要來自於測驗內容題項、學生的身心狀況及測驗情境 (OECD, 2005)。而在大型成就評量時，通常會從題庫中抽取題目編製成不同版本的題本，每個版本的題本雖然題目不同，但所要測量的特質與內涵是相同。由於題庫測驗題數相當多，每個學生只回答其中部分題目，學生的能力無法精確得知，使用不同的題本，有些學生的能力可得到良好估計，有些則否。因此，一般國際 LSA 以 PVs 來估計學生成就而不以一個測驗總分來代表學生的測

驗結果。以 TIMSS 而言，TIMSS 建立試題庫，採用複合輪換題本設計 (complex rotated booklet design) 以確保每題試題都能充分使用及題本試題數足以測量出學生的能力，並以五個 PVs 來估算學生數學成就分數 (Foy & Olson, 2008)。同時 TIMSS 將測驗時間限制於 90 分鐘，在複合輪換題本設計及時間限制下，傳統估計學生能力方式所得的母體參數變異數是有偏誤的 (Mislevy, Beaton, Kaplan, & Sheehan, 1992)。亦即從一個測驗題庫隨機抽取部分測驗題來估計學生的能

力，有其不確定性（uncertainty）存在，估計的準確度會受影響。在成就測驗中，研究者通常希望能得到一個似乎學生完成整個測驗題庫的分數，這與插補遺漏（imputing missing）的觀念相似，因此，就有一個 PV 分數（假設學生完成了所有測驗題目）及其測量誤差（Willms & Smith, 2005）。PVs 所代表的意義可說是學生能力的各種可能值的範圍及其機率，而非學生個人成績，學生個人的 PVs 則從此機率分配中隨機產生（von Davier, Gonzalez, & Mislevy, 2009）。

在 TIMSS 中，學生在有限的時間裡回答了每個領域的部分試題，學生的 PVs 隨機從所估計的能力分配所得，而這能力分配主要從有相似的試題回答方式（item response pattern）及相似背景資料的受試者所估計而得。在使用 PVs 時，同時也必須計算由於測驗缺乏精確性所造成的不確定性，如果測驗是精確的，那麼測驗誤差為零，而每個 PV 分析所得的統計數值應相等。然而，完全精確無誤的測驗幾乎是不可能，因此會有測量變異，通常稱為差補變異數（imputation variance）。LSA 研究中，和 PVs 有關的資料分析須同時考慮到差補變異與抽樣變異，由於計算繁雜，一般研究者在使用 PVs 分析資料時，通常會採取兩種簡便的方式，一種為只使用一個 PV 作分析，因為和力量測有關的不確定性被忽略（如抽樣誤差），即標準誤的估計值沒有考慮抽樣變異，因此這種分析方式所得的統計量的標準誤會被低估。第二種方式是取 PVs 的平均值作分析，這種採用平均 PVs 的方式比只採用一個 PV 的分析方式更嚴重的低估了標準誤（von Davier et al.,

2009），這兩種分析方式均會使推論不正確。

因此，在和測驗有關的資料分析中，若使用到 PVs，必須對每個變數依公式估計其參數的變異。例如，在迴歸分析時，因 TIMSS 有五個 PVs，因此必須對每個 PV 作個別分析，重複五次，然後再依公式求其合併後之係數與標準誤（Willms & Smith, 2005; Foy & Olson, 2008）。即若為母體統計數值，是第 i 個（ $i=1, \dots, k$ ）PV 的分析所得的統計數值，則

$$\theta = \frac{1}{k} \sum_{i=1}^k \theta_i, i=1, 2, 3, \dots, k, \quad (1)$$

設 B_k 為差補變異數

$$B_k = \frac{1}{k-1} \sum_{i=1}^k (\theta_i - \theta)^2, \quad (2)$$

抽樣變異數（sampling variance）則為

$$U = \frac{1}{k} \sum_{i=1}^k u_i, \quad (3)$$

u_i 為第 i 個 PV 分析的誤差變異數（error variance），再將抽樣變異數與插補變異數相加，得合併後得係數之變異數，即

$$V = U + (1 + \frac{1}{k})B_k, \quad (4)$$

例如，TIMSS 有五個 PVs，因此

$$V = U + 1.2B_k \quad (5)$$

茲舉例說明如下：假設在以 TIMSS 資料作迴歸分析，以 PV 為依變數，截距、變數一、變數二之迴歸係數分別為 $g00$ 、 $g10$ 、 $g20$ 。對每個 PV 作個別分析，重複作五次。再依公式計算每個係數的 θ 、 B_k 、 U 、 V ，再以合併後的係數及標準誤計算最後的 t_i 值為 $\frac{\theta_i}{\sqrt{V_i}}$ 。表 3 為五次迴歸分析之迴歸係數及標準誤、及 θ 、 B_k 、 U 、 V 、 t 值的應用實例。

表 3 五次迴歸分析之迴歸係數及標準誤、及 θ 、 B_k 、 U 、 V 、 t 值

	g00	SE(g00)	g10	SE(g10)	g20	SE(g20)
pv1	531.12	6.88	5.44	2.73	18.30	1.57
pv2	535.47	7.11	2.47	2.79	18.09	1.56
pv3	530.21	7.04	5.12	3.03	19.06	1.60
pv4	533.16	7.07	5.44	2.84	18.44	1.63
pv5	533.28	7.35	3.90	2.94	17.95	1.65
θ	532.65	7.09	4.47	2.87	18.37	1.60
B_k		50.30		8.24		2.58
U		4.23		1.66		0.19
V		55.37		10.24		2.80
t	71.58		1.40		10.98	

四、結論與討論

LSA 計劃的專責單位在完成調查後釋出豐富的資料供教育專家學者及研究者進行二次分析。然而由於 LSA 的調查設計不同於一般隨機取樣調查，在學生抽樣、試題抽樣及學生能力估算的複雜性，使用一般統計公式及分析技巧，所得的結果通常是有偏誤。本研究舉例說明使用 / 不使用權重在估計值的差異及整合 PVs 標準誤的計算。同時，也指出在 LSA 調查設計，由於學生抽樣及試題抽樣的複雜性，學生在有限的測驗時間及試題限制下，如只選用一個 PV，因為沒有考慮到抽樣的誤差，因此標準誤將被低估且能力分配測量（measurement of proficiency distribution）的不確定性將被忽略。將 PVs 取平均比只用單個 PV 更低估了標準誤，因此，在有 PVs 的 LSA 資料庫中，最不宜將 PVs 取平均做二次分析。單個 PV 及 PVs 的平均適用於每個學生均回答了所有試題的情形，不適用於如 TIMSS 採用複合輪換題本設計，亦即 LSA 資料分析時，PVs 的處理須同時考慮差補變異與抽樣變異，以獲得不偏的母體估計值。使用綜合 PV（每個 PV 作個別分析，然

後再依公式求其合併後之係數與標準誤）比使用 PV 取平均或僅使用單個 PV 有較大的標準誤，小標準誤可能導致不同的結果。在 Carstens 與 Hastedt（2010）的研究中以 TIMSS 資料分析也發現，使用單個 PV 或使用 5 個 PVs 的平均來做國家排比時，排比順序會改變，使用不適當的分析方法會使原來未具顯著差異變為顯著不同。研究者若將 PVs 當依變數，可考慮資料分析時以單個 PV 做初步的探索性分析，最後仍應以對每個 PV 作個別分析，然後以合併後之係數與標準誤作最後結論。

LSA 提供了豐富的教育資訊，一個 LSA 資料庫的產生，通常是教育相關領域的政府單位、專家學者、學術單位，實務工作者及研究團隊投注多年心力、時間與資源共同努力的結果，譬如 TEPS 從 2001 年開始至今已執行近 12 年，TIMSS 從 1995 年執行至今近 18 年，TIMSS 資料庫長久以來，提供國內 / 國際教育專家學者研究國內 / 國際學生數學及科學成就及相關教育議題的資訊，是數學及科學教育研究的可貴來源，TIMSS

2011 資料亦於今年（2013）二月釋出，這樣的資料庫是研究者探討教育相關議題的次級分析的可貴來源。然而 LSA 在試題設計、抽樣設計、能力分數的估算不同於一般橫斷資料，因此不宜以一般的統計技巧做資料分析。研究者在使用 LSA 資料須了解資料庫變數的特性外，更應參考官方使用手冊如抽樣、權重、學生

能力估算及資料分析應注意事項等議題，配合研究目的，選用正確的分析方式，才能獲得正確的母體估計值與做出正確的結論。本研究提出次級分析時須注意的議題，期能使 LSA 研究者瞭解權重及 PVs 在 LSA 資料分析的重要及錯誤的分析方式可能導致錯誤的結論，提供研究者在進行次級資料分析時參考。

參考文獻

- 中央研究院調查研究專題中心（2012）。台灣教育長期追蹤資料庫資料使用工作坊。台灣教育長期追蹤資料庫資料使用工作坊會議資料，中央研究院調查研究專題中心。
- Carstens, R. & Hastedt, D. (2010). *The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data*. Retrieved August 10, 2011, from http://www.iea.nl/fileadmin/user_upload/IRC/IRC_2010/Papers/IRC2010_Carstens_Hastedt.pdf
- Foy, P., & Olson, J. F. (Eds.) (2008). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Joncas, J. (2008). TIMSS 2007 Sampling weights and participation rates. Chapter 9 in J. F. Olson, M. O. Martin, and I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp.153-192). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Martin, M., Mullis, I., & Kennedy, A. (Eds.). (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Education Measurement*, 29, 133-161.
- Muthen, L., & Muthen, B. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthen & Muthen.
- NCES(2001). User's guide for the Third International Mathematics and Science Study(TIMSS). NCES 2001-065 by B. Chaney, L. Jocelyn, D. Levine, T. Mule, L. Rizzo, S. Roey, K. Rust. & T. Williams Washington, DC: Government Printing Office.
- Olson, J., Martin, M., & Mullis, I. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Organisation for Economic Co-operation and Development. (2005). *PISA 2003 data analysis manual*. Retrieved May 8, 2010 from www.pisa.oecd.org/dataoecd/53/22/35014883.pdf
- Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris: Author.
- Raudenbush, S., Bryk, A., Cheong, Y., & Congdon, R. (2004). HLM 6 [Manual]. Lincolnwood, IL: Scientific Software International.
- Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39, 142-151. Retrieved August 10, 2010, from <http://edr.sagepub.com/content/39/2/142>.
- von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them?

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2, 9–36.

- Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., Smith, C., & Stearns, P. (2009). *TIMSS 2007 U.S. technical report and user guide (NCES 2009–012)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
- Willms, J., & Smith, T.(2005) . *A manual for conducting analysis with data from TIMSS and PISA*. from www.unb.ca/crisp/pdf/manual_TIMSS_PISA_2005-0503.pdf