TEPS-B 技術報告第 1 號

建構 TEPS-B Panel 1 SH 樣本

2009 年調查完訪成功樣本的機率權數

第一版

關秉寅 詹傑勝

民 國 一 0 四 年 五 月

目錄

一、前言

「台灣教育長期追蹤資料庫」(TEPS)的調查，是針對同一群受訪對象，在數個不同時間點進行追蹤調查。因此，追蹤樣本的代表性是研究分析時必須關切的重點。本技術報告以 2001 年及 2003 年接受 TEPS 調查之高中職五專受訪者（TEPS-B 稱之為 Panel 1 SH 樣本）為對象，分別以中央研究院調查研究專題中心學術調查研究資料庫釋出的 TEPS 及 TEPS-B 限制版及公共版資料，連結 TEPS 2001 年學生問卷、TEPS 2001 年家長問卷、TEPS 2003 年學生問卷，以及 TEPS-B 2009 年電話調查蒐集到的資料，建構研究分析時可使用的機率權數(probability weights)。此權數也稱之為抽樣權數 (sampling weights)。建立是因抽樣設計的某些原因，使得抽取的樣本無法完整的代表其所來自的母群體。透過樣本機率權數的建立，則分析時可以調整部份樣本之抽樣機率與其他樣本不同、部份樣本缺失，或是部份基本重要變項（如性別、年齡等）的分配與母體不一致等問題。

需要替 TEPS-B Panel 1 SH 樣本中成功為 2009 年調查追蹤樣本建構新的機率權數，則是因 2009 年的調查完訪人數 (N=10,546)，約佔 TEPS 樣本應接受正式電訪調查之 15,922 人的 66％左右。此追蹤調查得到的樣本，可能與未能追蹤到的樣本間有系統性的差異，而導致樣本選擇偏誤 (sample selection bias)。此選擇偏誤可能是因隨機性缺失（missing at random）造成的，也可能是具系統性的偏誤。但如假定未能追蹤到的樣本是一種隨機性缺失的話，則可透過原來 TEPS 樣本中一些重要基本變項來預測 TEPS-B 樣本成功追蹤完訪的機率，並以此來建

構調整原來 TEPS 提供的機率權數。本技術報告的目的為提供建構此新樣本機率權數的方式，提供研究者併檔 TEPS 及 TEPS-B Panel 1 SH 樣本為分析樣本時之參考。

新樣本機率權數建構主要是用 2001 年及 2003 年樣本一些基本人口、家庭背景，以及就讀學校特性等變項，求得 2009 年 TEPS-B 調查時成功完訪樣本的機率，後，再將此機率的倒數與 TEPS 第二波提供之樣本機率權數相乘而得。本技術報告也將比較使用原來的 TEPS 樣本及原來樣本機率權數求得的一些基本個人或家庭背景變項平均數或比例，與用 TEPS-B 2009 Panel 1 SH 樣本使用新建構之樣本機率權數得到之同樣變項的估計是否有明顯差異。如果使用新舊權數得到的估計，沒有明顯差異的話，則未來研究者分析 2009 年 TEPS-B Panel 1 SH 調查樣本時，應可參考此新樣本機率權數的建構方式。

二、TEPS 提供的機率權數

依據2011年12月1日修訂釋出之TEPS《資料使用手冊》[1]第7頁至第11頁有關母體與抽樣設計該節，說明TEPS的第一、二波的抽樣是考量了因果分析的目的、追蹤的流失率，以及多層次分析的需要等三項因素後，以台灣地區的城鄉、公私立學校，以及國中、高中／職、五專等學制等為多階段分層抽樣設計的依據後，以分層隨機抽樣方式進行抽樣。其抽樣步驟為：

1. 根據當年度教育部的學校資料，先區分國中、高中高職及五專等四種學程，分別抽出樣本學校。
2. 經取得這些樣本學校提供TEPS所需包括班級數量、班級特性、各班學生人數與完整學生名單後，先抽出班級，然後原則上再由樣本班級中隨機抽出15名學生為正取樣本，並將該班其他學生給予隨機編號的順序後，作為正取樣本無法接受調查時的遞補樣本。
3. TEPS的抽樣也考量一些特殊情況，而有非比例（non-proportional）抽樣的設計。例如，國中樣本的選取，即因原住民地區的學校中，原住民學生在各班級的分佈不均，而增加這類樣本學校被抽班級的比例。又如，當時五專學校紛紛改制升格，導致此類學校中五專班級數不如原先預期的多。在考量未來追蹤及分析的需要後，從五專學制抽出的班級，均改為全班調查。

除前述抽樣設計TEPS的學生樣本需依其抽樣設計，給予不同的抽樣機率外，

[1] 可由 https://srda.sinica.edu.tw/group/scigview/2/8 下載。

TEPS《資料使用手冊》頁16也進一步說明，TEPS利用完訪樣本數來推估各分層的母體，計算出每個樣本的原始權數後，再考量母體與樣本在基本變項：「學校公私立別」、「城鄉類別」及「學程類別」上的次數分配，利用「多變數反覆加權法（raking）」的事後分層加權法，作第二階段的權數調整。TEPS建構學生樣本機率權數的方式，可說是一般大型調查資料庫的標準作法。

TEPS釋出的各波資料均提供該波樣本機率權數的變項。以第一波高中職五專學生樣本（TEPS-B Panel 1 SH）為例，此樣本機率權數變項名稱為 w1stwt1，此為標準化過後的學生樣本權數。第二波則提供了兩個權數，一為 w2stwt1，另一為 w2stwt2。此兩權數也均為標準化過後的學生樣本權數，但不同出在於後者是當分析時用到第二波IRT變項時才需要用的。

三、TEPS-B Panel 1 SH 樣本 2009 年調查完訪成功樣本之機率權數建構

TEPS-B 於 2009 年再次電話調查追蹤 TEPS 於 2001 年及 2003 年調查之高中職五專樣本 19,051 人。扣除電訪預試及實驗性電訪的樣本後，TEPS-B 正式電訪追蹤調查人數為 15,922 人，最後完訪樣本人數為 10,546 人，其中 10,084 人為電訪調查完訪者，462 人則是透過郵寄調查完訪者，全部完訪者占正式追蹤調查的 66.2%。這些成功追蹤到的樣本，自可能與未追蹤到的有系統性的差異。因此，我們分別以限制版及公共版的資料，以邏輯迴歸模型來估計 Panel 1 SH 樣本被追蹤成功的機率。[2]然後以此機率的倒數與 TEPS 第二波之 w2stwt2 相乘，以建構新的權數。使用被成功追蹤之機率倒數的意義，是使越不容易追蹤完訪者，能有更大的權重來代表與其有相同背景的人。此種作法的基本假定是納入邏輯迴歸模型的變項，可以解釋成功完訪者與無法完訪者之間的差異。當然此假定不一定正確，需要有其他的佐證來支持其合理性。以下即分別以研究者需要經過申請方可使用的 TEPS 及 TEPS-B 限制性資料，以及公共版的資料來瞭解前述建構新權數的作法是否合理。TEPS 限制性資料與公共版資料的不同在於前者釋出 100%受訪學生樣本，且有各學校的代碼等屬敏感性的資料，而後者則是由第一、二波受訪學生樣本中隨機抽取 70%（N=13,319），且學校層次資料只有公私立，以及學校所在之城鄉地區分層（分成鄉村、城鎮及都市等三類）。以下報告的分析均為使用

[2] Little, Roderick J. A. & Donald B. Rubin, 1987. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Stata 12.1。

以限制性資料估計 2009 年 TEPS-B 成功追蹤完訪機率時，納入邏輯迴歸模型的自變項包括：以 3-P IRT 模式估算的第一波學生綜合能力測驗分數、第二波綜合能力測驗分數，男性與否（此為 TEPS 樣本性別），是否與父母同住，族群虛擬變項組（包括台灣閩南、台灣客家、大陸各省市、原住民、其他），第一波父親教育程度虛擬變項組，第一波父親職業類型虛擬變項組，第一波母親教育程度虛擬變項組，第一波母親職業類型虛擬變項組，第一波家庭每月收入虛擬變項組，第二波學程類型虛擬變項組，第一波私立學校與否虛擬變項，第一波學校所在城鄉地區分層的虛擬變項，第二波學校代碼虛擬變項組（共 333 所高中職五專學校代碼，故有 332 個虛擬變項）。此迴歸模型分析的依變項為依據「接觸情形」建構之是否成功完訪（含電訪及郵寄）的虛擬變項，並以第二波學生樣本權數 w2stwt2 予以加權來執行分析。分析後，則進一步估計受訪成功機率，然後建立一個受訪成功機率倒數變項，並將此倒數變項乘上第二波學生樣本權數變項 w2stwt2，以建構一個新的樣本機率權數。此使用限制性資料建構的新權數變項，以下稱為 pw_os。以下是前述各步驟的 Stata 程式：

```
/* 以限制性資料分析是否成功完訪，依變項名稱為 contact1 */
xi: logit contact1 w2all3p w1all3p male coresid i.ethn i.w1faedu i.w1faocc i.w1moedu
i.w1moocc i.w1p515 i.w2pgrm i.w1priv i.w1urban i.w2sch_id [pw=w2stwt2]
/* 估計完訪成功的機率，機率變項名稱為 pscore1 */
predict pscore1
/* 建構 pscore1 的倒數 psw1，並建構新樣本機率權數 pw_os */
gen psw1=1/pscore1
gen pw_os=psw1*w2stwt2
```

以公共版資料建構新樣本機率權數的程式大同小異。不同之處，因公共版未提供第二波學校代碼之變項，故其邏輯迴歸模型無此組虛擬變項：

```
xi: logit contact2 w2all3p w1all3p male coresid i.ethn i.w1faedu i.w1faocc i.w1moedu
   i.w1moocc i.w1p515 i.w2pgrm i.w1priv i.w1urban [pw=w2stwt2]
predict pscore2
gen psw2=1/pscore2
gen pw_pr=psw2*w2stwt2
```

以上的程式，因資料不同，故依變項、估計之受訪成功機率等變項，均另外命名，最後估計出的樣本權數變項為 pw_pr。

檢證追蹤調查成功的樣本是否有選擇偏誤，最基本的方式是觀察邏輯迴歸模型中，是否有任何自變項的迴歸係數達顯著水準。如果邏輯迴歸模型中有不少自變項達顯著的話，則表示追蹤調查成功與否容易受到不少樣本背景因素的影響。反之，則表示無法成功追蹤者，比較接近隨機缺失的情況。如以 $p<.05$ 來判定自

變項是否達顯著水準的話，使用限制版資料的邏輯迴歸模型中，[3]自變項達顯著水準者包括男性、母親的教育程度為一般大學、母親教育程度為研究所、母親職業是特定專業人員（律師、法官、醫師、工程師或會計師）、以及五所學校代碼的虛擬變項。其中樣本為男性者、母親職業是特定專業人員，以及就讀某一所學校者，會顯著地降低成功完訪機率。至於母親的教育程度為一般大學或研究所，以及就讀其他四所學校的六個達顯著水準的自變項，則會增加成功完訪的機率。其他自變項則都未達統計顯著水準。

公共版資料的分析結果顯示，自變項達顯著水準的，則只有父親職業類別為生產設備操作及體力工，母親職業為一般事務及業務人員，以及其他類職業者。其中父親職業為生產設備操作及體力工者，會顯著地增加完訪機率，母親的兩項職業類別，則會減少完訪機會，其他自變項則都未達統計顯著水準。不論是限制性資料或公共版資料的分析結果均顯示，追蹤調查成功的樣本，如有選擇偏誤的話，是只受到少數變項的影響，且整體言，偏誤並不大。

另一個簡單的檢證追蹤調查成功樣本是否有偏誤的方式，則是看新的權數與 TEPS 提供之學生樣本權數的相關程度為何。如果相關程度越高，則表示前述納入邏輯迴歸模型可以相當程度的解釋成功完訪與否。表一即呈現兩個新建構變項與 TEPS 第二波學生樣本權數 w2stwt2 間相關係數，以及納入分析的樣本數。表一顯示新舊機率權數的相關極高，而且即使公共版資料是用比較簡略代表學校類

[3]因篇幅限制，此處僅報告達顯著水準之自變項。

型及城鄉分層的變項，其相關近乎完美。整體言之，以 2009 年 Panel 1 SH 追蹤成功樣本新建構的權數，只是微調了原來 TEPS 的學生樣本機率權數。

表一、TEPS 第二波學生樣本權數與兩個新建構樣本機率權數之相關

	pw_pr	pw_os
w2stwt2	0.997	0.988
N	9,688	10,268

四、以新舊權數加權估計平均數或比例的比較

依據 TEPS《資料使用手冊》頁 16 至頁 17 的說明，TEPS 成功樣本經加權處理後，在性別、公私立別、學程類型，以及城鄉類別等分佈上與母體分佈無差異。因此，TEPS-B 新建構的樣本機率權數是否合理的另一檢證方式，是以新建權數加權估計這些基本變項的平均數或比例，並檢視這些估計是否與使用 TEPS 原來樣本及權數加權所得的估計一致。

以下以不同的追蹤樣本及新舊權數來估計並比較：第一波綜合能力測驗平均數、樣本男性比例、父母是為結婚狀態比例、就讀私立學校比例、高中職專科等學程類型比例，以及城鄉類別比例。

1. 使用 TEPS 限制版樣本的估計

表二顯示的是以 TEPS 第一、二波合併之高中職五專限制版樣本(N=17,860)，以及使用 w2stwt2 加權估計前述各變項平均數及比例的結果。

表二、TEPS 限制版樣本及 TEPS 機率權數加權後之估計（N=17,860）

	平均數 / 比例	標準誤	95% 信賴區間	
第一波綜合能力測驗	1.6256	0.0165	1.5932	1.6580
男性	0.4894	0.0069	0.4758	0.5029
父母為結婚狀態	0.8860	0.0044	0.8774	0.8946
就讀私立學校	0.4457	0.0069	0.4322	0.4593
學程類型				
普通	0.3595	0.0058	0.3481	0.3708
綜合	0.1080	0.0038	0.1005	0.1155
職業	0.4367	0.0074	0.4222	0.4512
專科	0.0959	0.0023	0.0914	0.1003
城鄉類別				
鄉村	0.0469	0.0027	0.0416	0.0522
城鎮	0.4290	0.0069	0.4155	0.4425
都市	0.5241	0.0069	0.5105	0.5377

以下表三則顯示以合併 TEPS 第一、二波高中職五專限制版樣本及 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本（N=10,263），並以新建構之 pw_os 加權估計的結果。

表三、合併 TEPS 限制版及 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本及新建構 pw_os 之加權估計（N=10,263）

	平均數 / 比例	標準誤	95% 信賴區間	
第一波綜合能力測驗	1.6174	0.0217	1.5747	1.6600
男性	0.4890	0.0091	0.4711	0.5069
父母為結婚狀態	0.8906	0.0055	0.8798	0.9015
就讀私立學校	0.4467	0.0091	0.4289	0.4645
學程類型				
普通	0.3620	0.0077	0.3469	0.3772
綜合	0.1084	0.0051	0.0984	0.1183
職業	0.4329	0.0097	0.4138	0.4520
專科	0.0967	0.0030	0.0908	0.1026
城鄉類別				
鄉村	0.0478	0.0033	0.0413	0.0544
城鎮	0.4247	0.0090	0.4070	0.4424
都市	0.5275	0.0091	0.5096	0.5453

比較表二及表三呈現的結果後可看出，不論是用哪個樣本或是樣本機率權數加權來估計，其數值都相當接近。以第一波綜合能力測驗而言，以 TEPS 限制版樣本及原來樣本機率權數加權估出的平均數是 1.6256，95%信賴區間則是在 1.5932 與 1.6580 之間。TEPS 限制版樣本與 TEPS-B 2009 年 Panel 1 SH 成功追蹤樣本合併後，並使用 pw_os 加權估計得到的平均數為 1.6174，與用原來樣本及權數估計得到平均數的差距小於 0.01，其 95%信賴區間介於 1.5747 與 1.6600 之間，也是與原來 TEPS 樣本及權數估計得到的信賴區間大致重疊。

以不同追蹤樣本及權數估計其他變項所得到的比例與 95%信賴區間，也都相當接近。以兩種估計得到之比例差距的絕對值言，差距最小的是男性樣本的比例，TEPS 樣本估得比例略高約.0004；差距最大的則是父母為結婚狀態的比例，TEPS 樣本估計的比例略低約.0046。

2. 使用 TEPS 公共版樣本的估計

先前表一已顯示以 TEPS 公共版樣本及 TEPS-B 2009 年 Panel 1 SH 樣本合併後估計出的樣本機率權數與原來 TEPS 第二波學生樣本權數 w2stwt2 的相關高達 0.997。因此，我們可預期以 TEPS 公共版樣本及原來權數估計前述各變項的平均數或比例，以及其 95%信賴區間，也應與使用合併 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本後得到的估計相當接近。從表四及表五呈現的結果來看，也的確是如此。表四呈現使用 TEPS 第一及第二波公共版合併樣本（N=12,443），並用 w2stwt2 樣本權數加權估計各變項平均數或比例的結果。

表四、TEPS 公共版樣本及 TEPS 機率權數加權後之估計（N=12,443）

	平均數 / 比例	標準誤	95% 信賴區間	
第一波綜合能力測驗	1.7198	0.0198	1.6810	1.7587
男性	0.4823	0.0083	0.4659	0.4986
父母為結婚狀態	0.8815	0.0052	0.8713	0.8918
就讀私立學校	0.3727	0.0080	0.3570	0.3884
學程類型				
普通	0.3584	0.0070	0.3448	0.3721
綜合	0.1040	0.0046	0.0951	0.1130
職業	0.4443	0.0088	0.4270	0.4616
專科	0.0932	0.0027	0.0880	0.0985
城鄉類別				
鄉村	0.0329	0.0026	0.0279	0.0380
城鎮	0.4223	0.0083	0.4061	0.4386
都市	0.5447	0.0083	0.5284	0.5610

表五則呈現使用合併 TEPS 第一、二波公共版及 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本(N=9,690)，並以新建構之 pw_pr 機率權數加權後估計各變項的結果。

表五、合併 TEPS 公共版及 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本及新建構 pw_pr 之加權估計（N=9,690）

	平均數 / 比例	標準誤	95% 信賴區間	
第一波綜合能力測驗	1.7197	0.0221	1.6764	1.7629
男性	0.4824	0.0094	0.4640	0.5008
父母為結婚狀態	0.8809	0.0060	0.8692	0.8926
就讀私立學校	0.3730	0.0090	0.3554	0.3907
學程類型				
普通	0.3590	0.0079	0.3435	0.3745
綜合	0.1045	0.0053	0.0941	0.1149
職業	0.4430	0.0100	0.4235	0.4625
專科	0.0936	0.0030	0.0876	0.0995
城鄉類別				
鄉村	0.0329	0.0028	0.0273	0.0385
城鎮	0.4220	0.0093	0.4038	0.4403
都市	0.5451	0.0094	0.5267	0.5634

因為新舊樣本機率權數高度相關的原因，表四及表五估計出的平均數或比例的差距，比用限制性樣本所估計得到的更小。以第一波綜合能力測驗的平均數言，用原來 TEPS 公共版樣本估計的平均數（1.7198）與用其合併 TEPS-B 2009 年 Panel 1 SH 成功完訪樣本估計得到平均數（1.7197）幾乎是完全一樣。各變項比例差距最大的是學程類型中屬職業學程者，以原來 TEPS 樣本估計的比例，約比合併 TEPS-B 2009 年 Panel 1 SH 樣本後估計的比例大 0.0013。

五、結論

2009 年 TEPS-B 透過電訪調查追蹤於 2001 年及 2003 年接受 TEPS 調查之高中職五專樣本(即 TEPS-B Panel 1 SH 樣本)。扣除預試樣本後，追蹤完訪的 10,546 人，佔應接受正式電訪調查之 15,922 人的 66％左右。如以原來全部 TEPS 樣本人數 19,051 人來計，則只佔 55%左右。由於 2009 年追蹤調查 TEPS-B Panel 1 SH 樣本時，已距離前次調查達六年之久，且其間並無更新樣本的聯絡方式，因此無法順利追蹤到一定比例的原來樣本是可預期的。但無法追蹤到的樣本是否與可追蹤到者之間有重要特性上的差異，亦即可追蹤到的樣本是否仍具代表性，自是長期追蹤調查計畫必須關心的重點之一。

從新建構樣本機率與原有機率的高度相關，以及以新舊樣本機率估計 TEPS 限制版或公共版樣本一些基本變項的平均數或比例只有極小差異等分析，可知 TEPS-B 2009 年成功追蹤完訪之 Panel 1 SH 樣本與原來的 TEPS 樣本應無太大差異。如有差異的話，分析時也可考慮以本技術報告建議的方式建構新的樣本機率來做調整。由於 TEPS 原來的多階段分層抽樣設計即有非同等比例抽樣（non-proportional sampling）及集聚式抽樣（clustering sampling）的特性，因此 TEPS《資料使用手冊》頁 71 提醒使用者做描述統計時，應該要使用樣本機率權數來加權處理，才能得到不偏的估計值。至於，研究者以統計模型進行分析時，是否要用權數調整，則是一頗複雜的議題。因此，不論是使用 TEPS 或 TEPS-B 資料分析時，建議讀者應仔細閱讀 TEPS《資料使用手冊》對加權處理此一議題的說

明，也要參閱其他與此議題相關的文獻。

※建議引用本文之參考文獻格式：

關秉寅、詹傑勝（2015）。**TEPS-B 技術報告第一號：建構 TEPS-B Panel 1 SH 樣本 2009 年調查完訪成功樣本的機率權數**。取自：http://tepsb.nccu.edu.tw/download/TEPS-B_technical_report_no1.pdf