教育科學研究期刊　第五十七卷第一期
2012 年，57（1），29-50

缺失資料在因素分析上的處理方法之研究

王鴻龍
國立臺北大學
統計學系

楊孟麗
中央研究院
人文社會科學研究中心

陳俊如
國立臺北大學
統計學系

林定香
國立臺北大學
統計學系

摘要

因素分析常用來研究問卷及量表。當資料缺失過多或缺失機制為非完全隨機時，分析所得的共同因素個數或因素負荷常有偏差。本研究使用「台灣教育長期追蹤資料庫」，將其中的完整資料視為基準資料，並根據原有缺失結構，建構一至五倍缺失比率的資料集，以探討因素分析對缺失插補的敏感度。研究者比較了四種缺失處理法，包括：可用個體法、完整個體法、邏輯斯迴歸插補法與蒙第卡羅－馬可夫鏈（Monte Carlo Markov Chain, MCMC）插補法。結果顯示，缺失比率愈高時，所估計出來的變異數矩陣與基準資料的矩陣差異愈大。可用個體法在缺失比率較高時，萃取的共同因子的個數比基準資料多。在因素負荷上，可用個體法的誤差最嚴重，而完整個體法雖然和其他兩種插補法的誤差接近，不過會因缺失比率的增加與基準的誤差而隨之變大。研究者建議在缺失比率 20%～30%或以上時，使用邏輯斯迴歸插補法或是蒙第卡羅－馬可夫鏈插補法後再進行因素分析會有較小的誤差。

關鍵字：台灣教育長期追蹤資料庫、缺失資料、探索式因素分析、蒙第卡羅－馬可夫鏈（MCMC）插補法、邏輯斯迴歸插補法

通訊作者：林定香，E-mail: tinghlin@mail.ntpu.edu.tw
收稿日期：2011/04/26；修正日期：2011/09/15、2011/11/08；接受日期：2012/01/12。

壹、前言

問卷調查或教育測驗的研究，難免會有資料無法完整蒐集的情況，而這些不完整的資料，包括未繳卷、部分題目未填答或填答值為不知道、拒答，皆可被視為是不完整的作答反應。學者對不完整作答反應的定義也有不同的名詞，如缺失資料（missing data）、不完整資料（incomplete data）或無反應資料（non-response data）等。一般的套裝軟體多以刪除含有缺失資料的受訪者後，再以完整資料的部分進行分析。表面上看來似乎避免了分析不完整資料的問題，但被刪除的資料所隱藏的訊息則有被忽略的疑慮。

缺失資料研究領域發現，當缺失比率過高時，在許多統計分析法上，使用含有缺失資料的分析，相較於完全刪除缺失資料的分析，兩者所得的結果有明顯差異。因此，如果原始資料缺失過多，不但會降低統計上的檢力（power），使得標準誤變大，甚至資料的訊息也被扭曲或誤導。目前對於調查研究的缺失值處理方法，主要以插補進行事後的資料處理。

以量表作為測量方法的研究，多使用因素分析（factor analysis）來從量表調查的結果中，萃取幾個重要的共同因素（common factors）；由受試者在這些共同因素的因素分數（factor scores）來理解他在這些因素的表現。由於因素分析是透過相關係數矩陣（correlation matrix）或共變數矩陣（covariance matrix）來萃取共同因素，因此如果原始資料缺失過多，將使得相關係數矩陣或共變數矩陣的估計產生偏誤，連帶造成共同因素的萃取產生偏差，導致所得到的因素跟實際情況不同。

本研究將探討因素分析對遺失資料多寡的敏感度，以及不同的缺失處理方法對於修正因素分析結果的有效性。研究者將以不同的缺失處理方法，為不同缺失比率資料集做分析，並比較所估得的變異數矩陣之間的差異，以及在共同因素的個數或因素負荷量（factor loading）估計上的差別。研究者想探討的問題包括：當缺失資料大於何種比率時，上述變異數矩陣及因素負荷量的分析結果會產生顯著差異？當差異過大時，應該使用什麼樣的缺失資料處理方法才可以得到較穩定的結果？什麼樣的結果才是正確的？所謂分析結果的正確與否或好壞與否，是否有比較的基準？

研究者以「台灣教育長期追蹤資料庫」（Taiwan Education Panel Survey, TEPS）為探討的對象，以資料庫中的高中、高職、五專二年級的學生為首波調查對象，分別於 2001 年下半年（二年級上學期）及 2003 年上半年（三年級下學期）各做一波的資料蒐集。問卷蒐集內容包括各種學習相關資料及認知能力測驗。其中第二波的學生問卷資料，測量有關心理健康的題項共 7 題。這 7 題的缺失比率並不高，雖然其中 1 題缺失達到 6.7%，不過其他 6 題缺失都不到 1%。整體而言，TEPS 1 萬多人的樣本資料中，九成以上的人在這 7 題完全沒有缺失資料，適合用來探討本文的研究議題。研究者將從這 7 題的原始資料出發，以完整無缺失的部分作

為基準（baseline），並依據原始缺失的架構刪除不同比率的資料，用以探討這七個心理健康題項在不同的缺失比率下，若以不同的方法插補，對於因素分析結果的影響，並提出適當的缺失資料處理的建議。

貳、文獻探討

一、缺失資料的機制與處理方法

缺失資料的種類，分類方法大致有兩種：其一是以資料缺失的單位分類，第二種則是以資料缺失的機制分類（Elena, 2008）。簡述如下：

（一）以資料缺失的單位分類

可以分為兩類：觀察個體缺失（subject missing），亦即完全無法觀察到該個體的資料，像是受訪者或受測者拒答整份問卷；題目缺失（item missing）是指回收的測量資料中，有部分問項沒有回答，造成缺失。研究者面對個體缺失，多採用加權的方式（weighting）來彌補資料缺失可能造成的誤差；對於題目缺失的情況，則多採用插補法將缺失資料插補成完整資料，再進一步分析。

（二）以資料缺失機制分類

資料缺失機制可分為三類（Rubin, 1987）：

1. 完全隨機缺失（missing completely at random, MCAR）

缺失資料發生的機制，跟觀察到的資料及未觀察到的資料，都獨立無關，且是在研究者可控制之下。換言之，具有缺失資料的觀察個體可以視為母體的一組隨機樣本。由於缺失資料是隨機出現，因此 MCAR 是屬於可忽略的（ignorable）缺失機制。

2. 隨機缺失（missing at random, MAR）

缺失資料的發生與觀察到的資料有關，但是與未觀察到的資料之間是獨立無關的，亦即造成特定變數缺失的原因，只和其他已觀察到的變數有關。MAR 也是屬於可忽略的缺失機制。

3. 不隨機缺失（missing not at random, MNAR）

缺失資料發生的機制，與缺失資料本身的值有關，這違反了缺失資料是隨機出現的條件。例如，高所得者普遍傾向於拒絕回答收入問題，即為「不隨機缺失」，因為所得資料的缺失與否和所得的高低有關，故容易產生資料偏差的問題，因此，這是不可被忽略的（non-ignorable）缺失機制。

在缺失資料之統計分析方法的文獻中，有關缺失資料的因素分析的處理研究大致上分成插補法（imputation）及最大期望演算法（expectation maximization, EM）演算法兩大方向。至

於常用的缺失資料處理方法，大致有以下幾種（Allison, 2000; Enders, 2010; Schafer, 1997; Schafer & Graham, 2002）：

（一）刪除缺失資料法（deletion）

刪除缺失資料法可以分為個體刪除（list-wise deletion）和配對變數刪除（pair-wise deletion）兩種。前者將出現任何缺失資料的觀察個體整筆排除，不納入分析的樣本，故又稱為完整個體法（complete case method, CC）。其缺點為是，可能造成樣本數的損失，而且只有在 MCAR 的假設成立時，這種方法所得到的參數估計值才不會有誤差的問題。配對變數刪除法，則是當缺失資料出現在所需要分析的變數時，樣本才會被刪除，比起前者，這種方法可減少樣本數的損失。不過這種方法會有不同成對變數可用樣本數不同的疑慮，如樣本相關係數矩陣。由於這個方法所使用的樣本，包含所有還可以用來分析的觀察個體，故又稱為可用個體法（available case method, AC）。

（二）最大概似估計法（maximum likelihood, ML）

最大概似估計法假設樣本所來自的母群為常態分配，具有未知的參數（平均值與變異數矩陣），ML 就是根據已觀察到的樣本，以使這個母群的機率（概似函數值）達到最高為原則來估計未知參數的方法。具體作法是先求出使概似函數（likelihood function）有最大值的參數聯立方程組，接著求出方程組的初始解，再持續修正參數值，直到達到最高概似值為止，而最後具有最高概似值的參數值即為該參數的最大概似估計值。最大概似估計法在處理缺失時並不需要刪除具有缺失資料的觀察個體，也不必事先處理缺失值，可以直接自具有缺失值的資料中直接估計出平均數向量與共變矩陣，並不屬於插補法。

（三）插補法（imputation）

主要分成只用一次缺失資料插補的單一插補法（single imputation），以及以一種插補方法，做多次插補的多重插補法（multiple imputation）。兩者的差別在於，單次插補後只產生一組插補後完整資料集，研究者將據此做後續的統計分析。而多重插補法則會產生多組（一般軟體多內建為三組）插補後的完整資料集，研究者再將各資料集分別做後續的統計分析，並將多組的分析結果合併以進行推論。而每一次的插補方法，則可以是傳統的熱卡插補法、平均數插補，或是迴歸插補、邏輯斯迴歸插補、蒙第卡羅－馬可夫鏈法等，分述如下：

1. 熱卡插補法（hot-deck imputation）

熱卡插補法依照輔助變項的不同條件，將完整觀察個體分類成為若干的「插補細格」（imputation cell）。依據其輔助變數，例如性別或是年齡層等，將出現缺失資料的觀察個體，從相對應的「插補細格」中找尋一個完整觀察的個體，以其最常被觀測所得的數值替代其缺失值。

2. 平均數插補法（mean imputation）

利用變數中的已觀察數值的平均值，取代未觀察的缺失值，其優點為簡單易用。然而，平均數插補法雖不會改變整體變數的平均值，但是插補後的變數，變異數較小，而且如果缺失值的比率很高時，插補後的資料會使整體的分配改變，形成高狹峰的分配。

3. 迴歸插補法（regression imputation）

此法假定，含缺失資料的連續型變數與其他已觀察到的變數之間有相關。缺失資料則以迴歸結果所得的預測值取代；此法與平均數插補法一樣，會使被插補變項的變異數降低，因此通常還會加上隨機誤差（random error），作為最後的插補值，以還原真實變異數。

4. 邏輯斯迴歸插補法（logistic regression imputation, LR）

此法一般用於離散型變數的缺失資料插補。此方法也假設缺失資料變數與其他已觀察到的變數之間有相關性。缺失資料則以邏輯斯迴歸結果所得的預測值取代。插補法中，迴歸插補法及邏輯斯迴歸插補法要求缺失資料結構必須符合單調缺失（monotone missing），其定義為含有缺失資料的所有變數經過排序後，有缺失資料的個體若排序前的變數有缺失格，排序後的所有變數也都會有缺失格。

5. 蒙第卡羅－馬可夫鏈（Monte Carlo Markov Chain, MCMC）

蒙第卡羅－馬可夫鏈是一個隨機變量序列，其中每個元素的分布只受前一個數值的影響。在蒙第卡羅－馬可夫鏈的模擬下，構建了一個具有穩定分配的馬可夫鏈，插補值則由這個馬可夫鏈中抽出。一般情況下，蒙第卡羅－馬可夫鏈有幾個步驟。首先，設起始值，利用這些起始值來估算先驗分布。第二步驟是插補步驟，從現有的先驗分布中重複隨機選擇一個值來取代缺失值，直到先驗分布達到穩定或前後兩個迭代的差異小於某一預設標準為止。第三步驟是後驗步驟，重新計算新的變異數矩陣。如果尚未得到最終估計值，新估出的變異矩陣可為下次迭代的插補步驟使用（Gilks, Richardson, & Spiegelhalter, 1995）。

（四）最大期望演算法（expectation maximization, EM）

此方法（Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997; Schafer, 1997）為兩階段的迭代程序，分別為 E-step（expectation）及 M-step（maximization）。E-step 在已觀察到的資料下可以利用各種方法來插補，由插補來達成估計及填補缺失值的目的；M-step 則是利用 E-step 插補後的資料，以最大概似法重新估算變異數矩陣及平均數向量，而更新過後的變異數矩陣，放回下一次迭代的 E-step 再做計算，產生新的插補資料給該次迭代的 M-step；重複 E-step 及 M-step 直到兩迭代間的共變數矩陣差異小於預先設定的標準為止。使用 EM 演算法時，需假定資料是常態分配且缺失機制為 MAR。

二、缺失資料對因素分析之影響的文獻

Bernaards 與 Sijtsma（2000）探討了十二種填補方法及兩種 EM 演算法在最大概似因素分析的影響。除了傳統的填補方法以外，Kamakura 與 Wedel（2000）提出以因素模式（factor model）填補資料的方法，並和一些傳統的填補方法做比較。Rubin 與 Thayer（1982）以及 Brown（1983）分別探討缺失資料因素分析的 EM 演算法。Liu 與 Rubin（1998）提出 EM 法的延伸演算法（expectation conditional maximization either, ECME）演算法，並使用該演算法比較完整資料與缺失資料在因素分析的最大概似估計上的差異。Schafer 與 Olsen（1998）則探討多變量缺失資料的多重填補方法。

結構方程模式分析方面，Enders 與 Bandalos（2001）以蒙第卡羅方法比較四種缺失資料處理方法在結構方程模式上的行為。Enders 與 Peugh（2004）討論 EM 演算法處理缺失資料在驗證式因素分析的效果。McArdle（1994）闡述了四種不同的缺失型態在結構方程模式分析方法上的研究方向。Allison（2003）除了呈現傳統處理方法對缺失資料在結構方程模式分析上的不良表現之外，認為多重填補方法與最大概似法對處理缺失資料，在結構方程模式分析上都有不錯的表現。不過在這些插補法或是演算法的討論上，多侷限在變異數矩陣的變數本身。而對於調查資料中的其他變項（例如受訪者基本資料等），是否或如何影響量表題項多未探討。且多數的研究多以模擬的方式呈現分析結果，並沒有真正的完整資料可供對照，無法針對不同方法造成的差異，在應用領域的詮釋做深入的探討。

在因素分析上，如果是由於某些機制使得某些人容易產生缺失值，刪除這些具缺失值的人的資料，會因而扭曲了變項間的相關性。如果因素分析的變項之缺失結構受到不同性別或年齡結構的影響，亦即缺失機制為隨機缺失，如不做適當的缺失處理，因素分析的結果就可能產生偏誤。本研究將針對這個部分做進一步的探討。

參、研究方法

研究者將針對一般的因素分析模式、缺失資料對變異數矩陣的影響檢定，以及缺失資料處理對因素分析的影響等方面，說明本文的研究方法。

一、因素分析的模式

研究者以探索式因素分析（exploratory factor analysis）為主。假設隨機觀察向量 $\vec{X}=(X_1,\cdots,X_k)'$，期望值向量為 $E\vec{X}=\vec{\mu}=(\mu_1,\cdots,\mu_k)'$，共變數矩陣（covariance matrix）為 $Cov(\vec{X})=\mathbf{\Sigma}$，與共同因素 $F_1, F_2, ..., F_m$（不可觀察的隨機變數）及誤差項 $\varepsilon_1, \varepsilon_2, ..., \varepsilon_k$，其中共同因素與誤差項之間彼此互相獨立，而各誤差項之間也互相獨立。數學模式為：

$$
\begin{aligned}
X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\
X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\
&\vdots \\
X_k - \mu_k &= \ell_{k1}F_1 + \ell_{k2}F_2 + \cdots + \ell_{km}F_m + \varepsilon_k
\end{aligned}
$$

以矩陣型態呈現為 $\vec{X} - \vec{\mu} = \mathbf{L}\,\vec{F} + \vec{\varepsilon}$

其中 ℓ_{ij} 表示第 i 個變數，第 j 個共同因素的因素負荷量（factor loading），而以 $\mathbf{L} = (\ell_{ij})_{k\times m}$ 代表 $k \times m$ 維度的因素負荷矩陣。

假設 $\vec{F}$ 和 $\vec{\varepsilon}$ 互相獨立，且 $E(\vec{F}) = \vec{0}$ ， $\mathrm{Cov}(\vec{F}) = \mathbf{I}$ ， $E(\vec{\varepsilon}) = \vec{0}$ ， $\mathrm{Cov}(\vec{\varepsilon}) = \mathbf{\Psi}$ ，則變異數矩陣可表示成：

$$
\mathbf{\Sigma} = \mathrm{Cov}(\vec{X}) = E(\vec{X} - \vec{\mu})(\vec{X} - \vec{\mu})' = \mathbf{L}E(\vec{F}\vec{F}')\mathbf{L}' + E(\vec{\varepsilon}\vec{F}')\mathbf{L}' + \mathbf{L}E(\vec{F}\vec{\varepsilon}') + E(\vec{\varepsilon}\vec{\varepsilon}') = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}
$$

令 $\mathbf{T}$ 為正交矩陣， $\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$ ，則 $\vec{X} - \vec{\mu} = \mathbf{L}\vec{F} + \vec{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\vec{F} = \mathbf{L}^*\vec{F}^* + \vec{\varepsilon}$

其中 $\mathbf{L}^* = \mathbf{L}\mathbf{T}, F^* = \mathbf{T}'\vec{F}$ ，而原有的分配假設仍然相同，即 $E(\vec{F}^*) = \mathbf{T}'E(\vec{F}) = \vec{0}$ ， $\mathrm{Cov}(\vec{F}^*) = \mathrm{Cov}(\mathbf{T}'\vec{F}) = \mathbf{T}'\mathrm{Cov}(\vec{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \mathbf{I}$ 。研究者也可以推得 $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' + \mathbf{\Psi} = \mathbf{L}^*(\mathbf{L}^*)' + \mathbf{\Psi}$ 。亦即研究者可以透過正交矩陣將因素負荷矩陣轉置，仍然可以維持變異數矩陣與因素負荷的關係模式，不過不同的轉置可能產生不同的因素負荷。本研究有關因素分析實證分析結果僅有一個共同因子，所以沒有必要討論不同轉軸的方法。

由於因素分析是根據變異數矩陣或相關係數矩陣進行分析，因此，研究者將先比較不同的缺失比率及不同的缺失處理方法對矩陣的影響。而缺失比率及不同缺失處理方法對因素分析的影響之比較，則是以特徵值大於 1 的個數（共同因素個數）、特徵值、特徵值累積百分比以及負荷量等分析結果為主。各種缺失比率對於因素分析結果的影響，在特徵值與特徵值累積百分比的部分，將呈現與基準資料分析結果的差異，在負荷量方面，則除了呈現與基準資料的差異外，也計算平均差異平方和的平方根。

二、缺失資料對變異數矩陣的影響之檢定

在評估缺失資料對於變異數矩陣的影響上，研究者分兩方面探討：行列式值的比值及概似比統計量。將完整資料所計算出來的變異數矩陣視為已知，記為 $\mathbf{\Sigma}_0 = (\sigma_{ij})_{k\times k}$ 。使用缺失資料或是使用不同的填補方法所估計出來的變異數矩陣，記為 $\mathbf{S} = (s_{ij})_{k\times k}$ 。研究者先以兩矩陣的行列式值的比值， $\det(\mathbf{S}/\mathbf{\Sigma}_0)$ ，做初步的觀察。比值愈接近 1，兩矩陣愈相近。另外，研究者也採用 Anderson（2003）的「修正概似比統計量」（adjusted likelihood ratio statistics）來檢定母群變異數矩陣 $\mathbf{\Sigma}$（未知參數）和已知變異數矩陣 $\mathbf{\Sigma}_0$ 是否有顯著性差異。即檢定 H_0： $\mathbf{\Sigma} = \mathbf{\Sigma}_0$ ，而其修正概似比統計量為：

$$\lambda = (\frac{e}{n})^{\frac{1}{2}kn} \mid \mathbf{B\Sigma}_0^{-1} \mid^{\frac{n}{2}} e^{-\frac{1}{2}tr(\mathbf{B\Sigma}_0^{-1})}$$

其中 n 為樣本數減 1，$\mathbf{B} = n \times \mathbf{S}$。而 $-2 \log \lambda$ 服從近似卡方，自由度為 $\frac{1}{2}k(k+1)$，近似檢定之 p 值為 $P(\chi^2 > -2 \log \lambda)$。$\lambda$ 值愈小，表示兩個變異數矩陣差異愈大。p 值小於顯著水準 α 時，表示在顯著水準下，檢定結果具顯著性差異。研究者將比較不同缺失比率下，上述評估標準的差異。

三、缺失資料處理對因素分析的影響

為瞭解缺失資料對因素分析的影響，且探討不同缺失處理的效果，研究者將以 TEPS 第二波公共版的完整資料為基準資料集，先找出缺失機制的影響變數，根據原有的缺失結構，建立原有缺失比例一至五倍的缺失資料集。最後研究者將比較不同缺失處理方法，在不同缺失比例下，對因素分析結果的影響。

（一）不同缺失處理方法與分析結果之間的關係

一般套裝軟體多以可用個體法為內建方法，其方法在計算變異數矩陣時，每一對共變異數都盡可能的使用該配對所有的可用完整資料，所以變異數矩陣的各元素並不一定有相同的樣本數。除了此可用資料法外，研究者也將探討常見的完整個體法，變異數矩陣計算時，使用的資料為所有變項都無缺失值的觀察值。由於 TEPS 心理健康變項的缺失結構不具有單調缺失的特性，在插補方法上，本研究先以沒有缺失結構限制的蒙第卡羅－馬可夫鏈單鏈（single chain）單一插補法，簡記為 MCMC。另外，再以邏輯斯迴歸插補法，由缺失比率較小的變數開始逐一完成七個有缺失變數的插補，簡記為 LR，等兩種方法做插補，然後以插補後資料做因素分析。研究者將以上述四種缺失處理方法套用在五種不同比率的仿缺失資料上，比較各種缺失比率的仿缺失資料與基準資料，在變異數矩陣上的差異，並比較各種缺失比率的仿缺失資料，經過插補處理後，所獲得的因素分析結果，與基準資料的因素分析結果之間的差異，並觀察最適當的缺失處理方法與對基準資料的差異，是否在可接受的範圍。

（二）處理後資料與基準資料的因素分析結果的比較

針對缺失處理後的資料集與基準資料集的因素分析的結果比較，研究者將比較特徵值差異比率及因素負荷均方差平方根，其計算公式如下：

1. 特徵值差異比率：$\frac{|\lambda_t - \lambda_0|}{\lambda_0}$

其中 λ_t, λ_0 分別為缺失處理後的資料集 t 與基準資料集的因素分析的對應特徵值。差異比率愈小表示與基準資料的分析結果愈接近。

2.因素負荷均方差平方根：$\sqrt{\frac{\sum_{i=1}^{k}(\ell_{ij}-\ell_{i0})^2}{k}}$

其中 ℓ_{ij}, ℓ_{i0} 分別為缺失處理後的資料集 t 與基準資料集的因素分析的第 i 個因素負荷，$i=1,...,k$ 共有 k 個因素負荷（變數）。均方差平方根愈小表示與基準資料的分析結果愈接近。

肆、TEPS 資料的實證分析

本研究以 TEPS 第二波高中、高職、五專學生樣本資料為主，針對心理健康方面的七個題項，探討有無缺失、是否填補及在因素分析上的差異。首先探討缺失資料結構，並依據其缺失結構建構不同缺失比率的資料組，作為後續比較之用。

一、樣本與變項的概述

TEPS 第二波高中、高職、五專學生樣本之學生問卷資料，探討了有關學生該學期以來是否發生一些心理健康（mental health）上的症狀。其中心理健康題目為詢問受訪者「這學期以來曾有下列情形嗎？」:（一）不想和別人來往（變項名稱 W2S426）;（二）鬱卒（變項名稱 W2S427）;（三）想要大叫、摔東西、吵架或打人（變項名稱 W2S428）;（四）覺得搖晃、緊張或精神不能集中（變項名稱 W2S429）;（五）感到孤單（變項名稱 W2S430）;（六）睡不著、睡不好、很容易醒或做惡夢（變項名稱 W2S431）;（七）頭部緊緊的、身體感到發麻、針刺、虛弱或手腳發抖（變項名稱 W2S434）。以四點量表測量，1 代表「從來沒有」、4 代表「經常有」。因此，分數愈高表示心理衛生上的問題症狀愈多。本文以 0 代表「從來沒有」、1 代表「偶爾有」、2 代表「有時有」、3 代表「經常有」。

根據張苙雲（2009）的調查報告，高中、高職、五專學生第一波的資料，是以 2001 年臺灣及澎湖地區的高中、高職、五專二年級上學期的學生為抽樣架構，依據學校所在的都市化程度，學程種類（普通高中、高職、綜合學程、五專）及行政區域做分層後，先抽學校，再抽班級（每校平均四班），之後再從每個抽樣班抽約 15 位學生做調查。調查方式是由訪員到學校安排學生做集體調查。學生在填寫問卷之前，先做一份標準化測驗，以客觀評估學生的學習成就。第二波的資料則是在 2003 年，當這一批學生已經到了三年級下學期時蒐集的，同樣也有標準化測驗與調查問卷。第一波樣本接近 2 萬人，第二波樣本則由於學生輟學或轉學至非抽樣學校而流失近 1,000 人，但人數仍然近於 18,000 人。本研究所使用的資料則是該計畫所釋出的公共使用版。公共使用版本的資料是由計畫從整體資料中隨機抽取 70%的樣本，約 12,000 人。兩波的資料可經由相同的編號連結。

兩波都有關於心理健康的資料，但第一波有 14 題，第二波則有 7 題與第一波相同。第一波的資料，在心理健康方面的缺失值比率很低（0.2%～0.3%），第二波的資料裡，除了「不想

和別人來往」（變項名稱 W2S426）這題的缺失值比率特別高，達到 7%之外，其他各題的缺失值比率都不到 1%。整體而言，完整資料的部分至少有九成以上。為了能深入討論缺失資料對因素分析的影響，本研究僅就第二波資料加以分析。

連結後第二波共有 12,192 筆資料。其中 7 題全部都回答的完整資料有 11,191 筆，約占全部資料的 91.79%（11,191/12,192）。7 題心理健康題的缺失結構（missing pattern）如附表 1「T0」欄所示。

二、缺失結構的探討

首先探討缺失資料的結構，研究者希望能從受訪者的基本資料獲得額外訊息。研究者以基本資料為輔助變項，先個別檢驗心理健康的各題之缺失，是否隨著基本資料的不同而有所差異；若是有差異，則以這些基本資料作為輔助變項，該心理健康的缺失結構就符合隨機缺失的假設。接著，找出其缺失結構，並在與心理健康變項缺失有關的基本變項的交叉結構下，建構具相同心理健康缺失結構之各種缺失比例資料集。研究者從基本資料及七個心理健康變項都完整的資料中（樣本數 11,191，占原始資料的 91.79%，11,191/12,192），依照各缺失結構原始缺失比率，刪除具相同基本資料結構的心理健康問項資料，形成仿缺失資料。仿缺失資料的缺失比率，則分別設定為原始缺失比率的 1、2、3、4、5 倍。研究者將比較對照組（無缺失資料，共 11,191 筆，也稱為基準資料集）和這些不同缺失比率的仿缺失資料，在：共變數矩陣、特徵值（eignevalue）大於 1 的個數、特徵值及因素分析的因素負荷等的差異。

Yang 與 Tam（2004）表示心理健康題的缺失可能和受訪者的基本資料有關。研究者以第二波的全部資料（共有 12,721 筆）中，心理健康題缺失值較高的問項 W2S426 為例，初步分析顯示資料有缺失者與無缺失者的特質似有不同：

（一）兩性在這一題沒有回答的機率有顯著性差異（$p_m = .080, p_f = .065$），雙母體比率差異 z 檢定 p 值為 .001。顯示不同性別會影響缺失的機率。

（二）是否缺失與學程別有顯著相關，卡方獨立性檢定卡方值為 44.655，檢定 p 值小於 .001。顯示是否缺失在學程別上有顯著差異。

（三）是否缺失與城鄉別有顯著相關，卡方獨立性檢定卡方值為 11.05，檢定 p 值為 .004。顯示是否缺失在城鄉別上有顯著差異。

（四）公私立學生之間的缺失比率有顯著性差異（$p_0 = .063, p_1 = .084$），差異檢定 p 值小於 .001。顯示公立與私立學生的缺失比率有顯著差異。

（五）是否缺失與在家庭收入別（分為六個範圍選項）的差異達邊際顯著，卡方獨立性檢定卡方值為 9.959，檢定 p 值為 .076。雖然未達 .05 顯著水準，不過研究者並不排除不同的家庭收入選項可能會有不同的缺失比率。

（六）是否缺失與標準化測驗的平均分數有顯著差異（有遺漏值者，標準化測驗平均得分為-0.25，沒有遺漏值者，標準化測驗平均得分為 0.008），平均數差異檢定 t 值為 7.1，檢定

p 值為小於 .001。

從初步分析的結果可以看出來，完整資料與缺失資料在基本結構上有不同的表現，缺失資料所包含的訊息應該會在因素分析上產生不同的效應。

三、缺失資料對變異數矩陣的影響

為專注在心理健康資料缺失狀況對分析的影響，在 12,721 筆公共版第二波高中、高職、五專樣本中，研究者以學生問卷中性別（變項名稱 W2S445）、學程別（變項名稱 W2pgrm）、城鄉別（變項名稱 W2urban3）、公私立別（變項名稱 W2priv），以及家長問卷的家庭收入（變項名稱 W2S508），並連結學生的測驗成績別（由標準化成績（變項名稱 W2all3p）以等間距分成 7 組）六個變數作為基本資料。其中，這六個基本變數都無缺漏的樣本有 12,192 筆，稱為原始資料集。研究者以原始資料集探討缺失資料對因素分析的影響。原始資料集中，有 11,191 人在心理健康題完全沒有缺漏值，約占 91.79%（11,191/12,192），缺失結構如附表 1（T0 欄）。附表 1 的最左欄顯示有 26 組不同的缺失型態，第二欄到第八欄則呈現各組缺失型態在七個題項的缺失狀態，「X」表示無缺失，「・」表示有缺失。以原始資料集（T0 欄）為例，組 1（group 1）在 7 題都無缺失（在 12,192 人之中，11,191 的人都在這 7 題無缺失），是完全無缺失的樣本；其他 25 組則有不同的缺失模式。例如，組 2 僅在 W2S434 題有缺失，在 12,192 筆資料中，有 53 筆（約占 0.43%）具有這樣的缺失模式，其他以此類推。研究者將「T0」的組 1 之 11,191 筆資料視為基準資料集，以之算出變異數矩陣（7×7 矩陣，記為$\boldsymbol{\Sigma}_0$）稱為基準變異數矩陣：用來跟其他仿缺失資料集（T1～T5）的組 1 資料所得的變異數矩陣相比較。

仿缺失資料集的建構方面，研究者以原始資料集中，六個基本資料變數的所有交叉組合中的缺失結構為準。在六個基本資料變數的各個交叉組合下，採系統隨機的方法，從基準資料集中，依據相同的缺失結構，以特定比率刪除對應的心理健康變數格，以建構仿缺失資料集。仿缺失資料集的比率則是以原始缺失比率（8.21%）的 1、2、3、4 及 5 倍等五種。由於六個基本資料變數的各個交叉組合資料的樣本數並不均等，有些組合並沒有足夠的樣本讓研究者依據缺失結構刪除，最後的缺失比率可能比原先設定的比率稍低。五組仿缺失資料集，真正的缺失比率分別為 8.82%（＝1－10,204/11,191）（附表 1 的 T1 之組 1）、17.51%（＝1－9,232/11,191）（附表 1 的 T2 之組 1）、25.91%（＝1－8,291/11,191）（附表 1 的 T3 組 1）、34%（＝1－7,386/11,191）（附表 1 的 T4 組 1）及 41.48%（＝1－6,549/11,191）（附表 1 的 T5 組 1）。接著，研究者分別以五組仿缺失資料的完整資料部分（組 1）為樣本，分別將他們與基準資料集的組 1 做比較，比較的內容則包括行列式值的比值、概似比檢定統計量，以及其檢定 *p* 值。結果呈現如表 1 所示。

表 1

不同缺失比率對變異數矩陣估計的評估比較

變異數矩陣之評估值	缺失比率倍數				
	T1	T2	T3	T4	T5
缺失比率	8.820%	17.510%	25.910%	34.000%	41.480%
行列式值比	0.974	0.995	0.961	0.941	0.964
修正的概似比統計量	3.477	6.544	16.194	21.428	27.180
（檢定 *p* 值）	（1.000）	（0.999）	（0.962）	（0.807）	（0.508）

從表 1 所示，行列式值比都小於 1，顯示仿缺失資料集的變異數，都比基準資料集的變異數矩陣小，但由於行列式值是一種綜合性統計量，不容易看出共變異數的些微變化，研究者主要的參考依據還是「修正的概似比統計量」，以觀察兩變異數矩陣是否相同。由表 1 可以清楚觀察到，雖然檢定（*p* 值）沒有達到顯著水準，然而，隨著缺失比率的增加，統計量也跟著遞增，檢定 *p* 值也隨著遞減的趨勢。

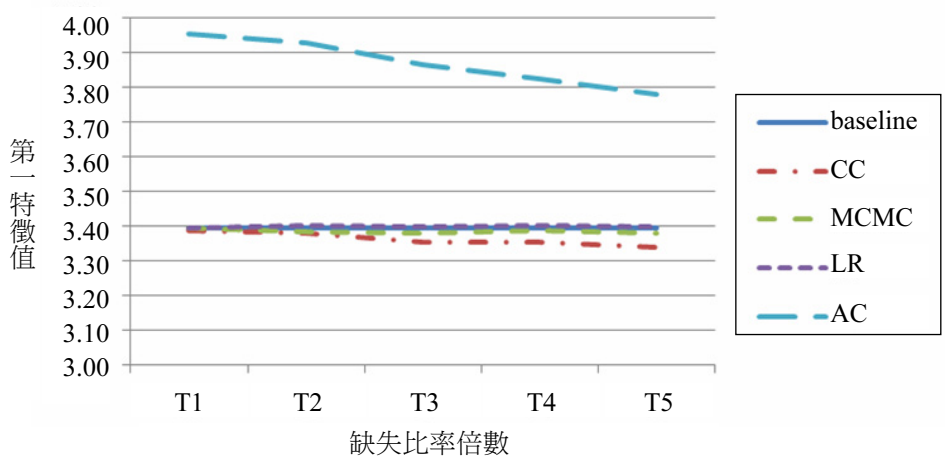
四、缺失處理方法對因素分析結果的影響

研究者接著探討，以四種缺失處理方法分析不同比率的仿缺失資料集時，如何影響分析結果。研究者以基準資料集（樣本數 11,191）的結果為標準。以上一節所述的方式建構五個不同缺失比率的仿缺失資料集。

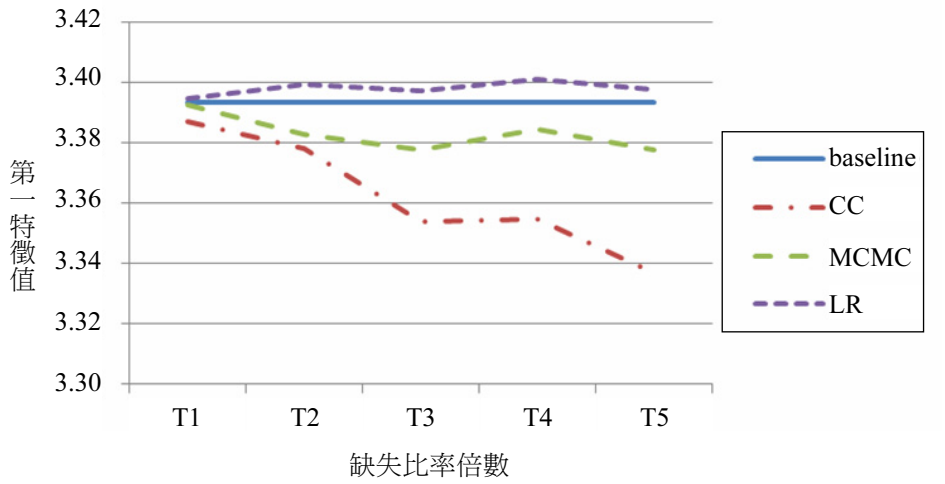
針對組仿缺失資料，研究者除了以常用的可用個體法、完整個體法分析資料外，也分別以不同的方式插補，包括：以邏輯斯迴歸插補法、蒙第卡羅－馬可夫鏈之單鏈單次插補並將結果做四捨五入，先做資料插補，再做後續分析。四種缺失處理法都將與基準組的分析結果做比較。

（一）特徵值的比較

研究者先輸入相關係數矩陣，使用主成分法萃取因素，以特徵值達到 1 或以上作為判斷準則，決定抽取因素個數。基準資料集的因素分析結果顯示共同因子（特徵值大於 1）的個數只有一個，數值是 3.393，變異數解釋量百分比是 48.47%。可用個體法缺失處理法在五組資料的因素分析第一特徵值都嚴重高估（如圖 1(a)），且兩個缺失比率達 30%以上的資料集（T4 與 T5）都有兩個大於 1 的特徵值，萃取出兩個共同因子。以完整個體法、蒙第卡羅－馬可夫鏈、邏輯斯迴歸插補法等三種方法處理後，五個資料集的都只得到一個大於 1 的特徵值與基準資料相同；但三種方法所得到的特徵值不甚相同：使用邏輯斯迴歸插補法，結果都稍微高估特徵值，蒙第卡羅－馬可夫鏈則都稍微低估，不過兩者都與基準組相近。完整個體法也低估特徵值，不過低估的情形隨著缺失比率增加而更嚴重，其中 T3～T5 的低估值有陡降的現象



(a)AC、CC、MCMC、LR與baseline的比較



(b)排除AC後，CC、MCMC、LR與baseline的比較

圖1.　缺失資料處理後第一特徵值

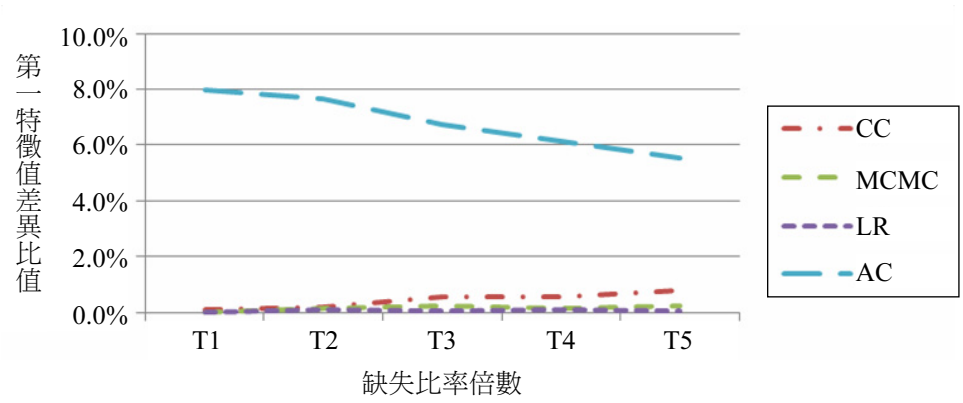
（如圖 1(b)）。為了方便對照，研究者在圖 1 中 T1～T5 都標示了的基準資料集的特徵值，$y=3.393$ 的直線。

研究者以缺失處理後的資料集做因素分析並獲得第一特徵值，再求其與基準組的第一特徵值之差異，並觀察此差異相對於基準組第一特徵值的比率為何。比值愈小表示差異愈小，故處理方法愈值得推薦。表 2 顯示不同處理方法與基準的差異比率的分析結果，其中可用個體法的差異比率最大，五種缺失比率資料組的分析結果都與基準組有 5.55%～7.99%的差距，如圖 2(a)。其他三種缺失處理法的比值都小於 1%，其中以邏輯斯迴歸插補法最好，差距不到 0.1%。蒙第卡羅－馬可夫鏈插補法與基準的差距約為 0.011%～0.22%之間，完整個體法與基準

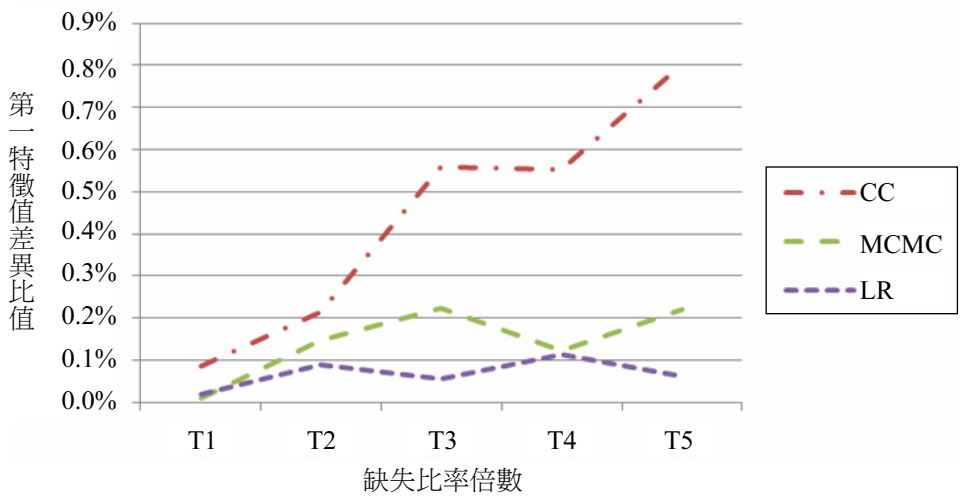
表 2

缺失資料處理後第一特徵值與基準的差異比率

插補方法	缺失比率倍數				
	T1	T2	T3	T4	T5
AC	7.995%	7.642%	6.711%	6.135%	5.522%
CC	0.087%	0.214%	0.560%	0.552%	0.804%
MCMC	0.011%	0.147%	0.225%	0.124%	0.220%
LR	0.020%	0.089%	0.054%	0.112%	0.063%



(a)AC、CC、MCMC、LR與baseline的比較



(b)排除 AC 後，CC、MCMC、LR 與 baseline 的比較

圖2. 缺失資料處理後第一特徵值與基準的差異比率

的差距約 0.087%～0.80%之間，如圖 2(b)。由上述的分析結果可知，一般常用的可用個體法與完整個體法等缺失處理方法，可用個體法的誤差遠大於完整個體法。完整個體法雖然有相對小的誤查，不過卻有缺失比率愈高處理後的分析結果誤差愈大的情形，其中 T3～T5 的差異比例有陡升 2 倍多的現象（如圖 2(b)）。使用適當的插補方法，如邏輯斯迴歸插補法或蒙第卡羅－馬可夫鏈則有較小且穩定的誤差。詳細結果如表 2 及附表 2 所示。

（二）因素負荷的比較

基準資料所獲得的一個共同因素，各題項的因素負荷係數都是正值，依照變數順序 W2S426～W2S431 及 W2S434 分別為 0.63944、0.79633、0.73170、0.73118、0.77130、0.64109、0.52365（如附表 3(a)）。除了可用個體法之外，五組不同缺失比率資料在其他三種缺失處理後，各題項的因素負荷量係數的正負號與標準資料的結果一致，不過因素負荷量平均誤差平方和平方根的表現則有不同。

使用可用個體法的結果最差，其五種缺失比率資料集的均方差平方根都在 0.19～0.22 之間。完整個體法的均方差平方根，則介於 0.0013～0.0085 之間，且隨著缺失比率的上升而增加。蒙第卡羅－馬可夫鏈的均方差平方根在五種缺失比率資料集的分析結果呈穩定狀態，且介於 0.0013～0.0052 之間。邏輯斯迴歸插補法在五種缺失比率資料集的分析結果也呈穩定狀態，且均方差平方根最小，介於 0.00063～0.004222 之間。詳細均方平方和平方根如表 3 與圖 3 所示。大致來說，常用的缺失處理方法：可用個體法與完整個體法，可用個體法的均方差平方根遠大於完整個體法，不過兩者都有缺失比率愈高處理後的分析結果的均方差平方根愈大的現象，其中完整個體法在 T3～T5 的均方差平方根有上升至近 2 倍以上的現象（如表 3、圖 3(b)）。而使用適當的插補方法，如邏輯斯迴歸插補法或蒙第卡羅－馬可夫鏈則可得穩定且相對小的均方差平方根。

伍、結論與建議

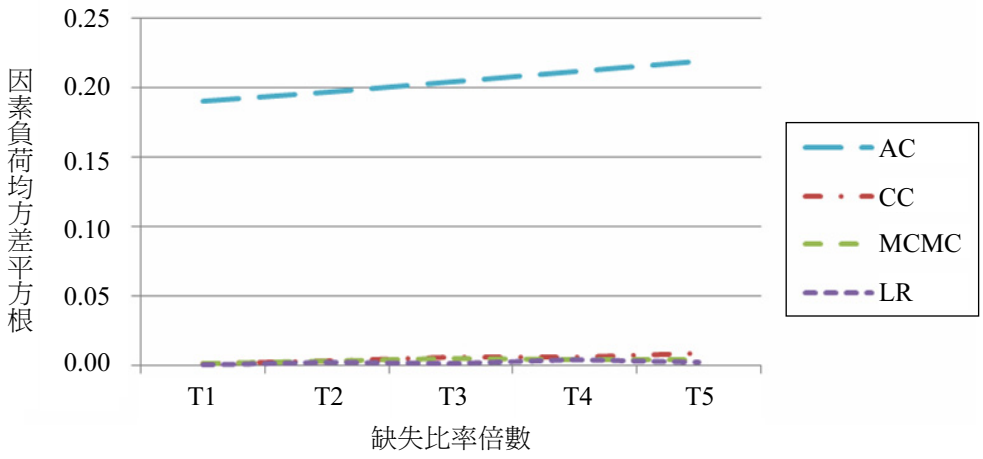
缺失資料對分析結果的影響顯示，缺失比率愈大，缺失資料所估計的變異數矩陣與基準資料集所算出的變異數矩陣在「修正的概似比統計量」的差異愈大。直接使用缺失資料透過變異數矩陣或是相關係數矩陣的後續分析，例如因素分析，將有顯著性的偏誤。

在因素分析的結果方面，除了可用個體法在仿缺失比率較高時，共同因素的個數會跟基準資料集的結果不同外，完整個體法、邏輯斯迴歸插補法、蒙第卡羅－馬可夫鏈等三種缺失處理方法所得的共同因素的個數大致不受到缺失比例的影響。使用可用個體法的分析結果，在特徵值上的誤差也很大，不建議研究者使用。採用完整個體法分析所獲得的特徵值與基準資料的差異雖然相對小，但其誤差會隨著缺失比率的增加而變大，在缺失比例達 20%～30%或以上時，誤差有陡升的現象。而使用適當的插補方法，例如蒙第卡羅－馬可夫鏈法以及邏

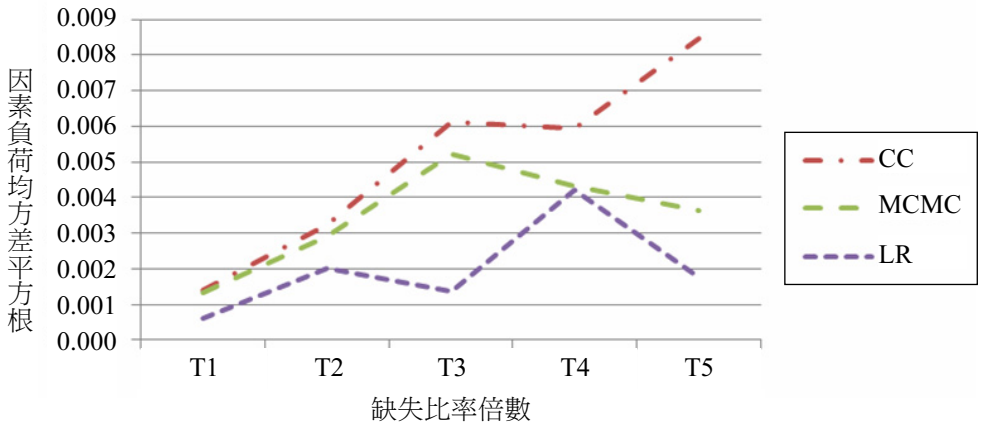
表 3

缺失資料處理後與基準組因素負荷均方差平方根

插補方法	缺失比率倍數				
	T1	T2	T3	T4	T5
AC	0.1897	0.1968	0.2035	0.2109	0.2187
CC	0.0014	0.0032	0.0061	0.0059	0.0084
MCMC	0.0013	0.0029	0.0052	0.0043	0.0036
LR	0.0006	0.0020	0.0013	0.0042	0.0017



(a)AC、CC、MCMC、LR 的比較



(b)排除 AC 後，CC、MCMC、LR 的比較

圖3. 不同缺失資料處理後與基準組因素負荷均方差平方根

輯斯迴歸插補法則有相對小的特徵值誤差，且在較高缺失比率時誤差也都呈穩定狀態。各題項的因素負荷量之估計方面，可用個體法的誤差最為嚴重，其均方差平方根是其他缺失處理方法的數十倍以上。完整資料法雖然有相對小的均方差平方根，不過會有隨著缺失比率愈高，均方差平方根也愈大的現象，在缺失比例達 20%～30%或以上時，均方差平方根有變大二倍以上的現象；若使用適當的插補方法，例如蒙第卡羅－馬可夫鏈或邏輯斯迴歸插補法則有相對小且穩定的均方差平方根，不會隨著缺失比率改變而有太大的變動。

本研究發現，心理健康的缺失與否和性別（W2S445）、學程別（W2pgrm）、城鄉別（W2urban3）、公私立別（W2priv），以及家長問卷的家庭收入（W2S508）等基本問項有關，在缺失機制符合隨機缺失的假設下，以這些基本問項作為資料是否缺失的影響變數，透過蒙第卡羅－馬可夫鏈單鏈單次插補法或是邏輯斯迴歸逐一插補法，做缺失處理後再進行因素分析都有降低誤差的效果，即使缺失比率較大，誤差也不致過大而且呈現穩定狀態。如果不確定缺失機制而進行插補，則完整個體法也有不錯的效果，不過缺失比率較大時，這個方法的誤差有增加的趨勢。

本研究僅針對 TEPS 第二波高中、高職、五專學生樣本的資料中有關心理健康問項作探討，或許可以作為其他類似調查資料再做因素分析時的參考。不過仍需做更深入的研究，才能針對一般的因素分析缺失資料插補方法做更明確的建議。

參考文獻

一、中文文獻

張芷雲（2009）。**台灣教育長期追蹤資料庫—學生問卷調查結果報告**。中央研究院調查報告。臺北市：中央研究院。

【Chang, L.-Y. (2009). *Taiwan Education Panel Survey: Student questionnaire report*. Project report from Academic Sinica. Taipei, Taiwan: Academic Sinica.】

二、外文文獻

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28*(3), 301-309.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: John Wiley & Sons.

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item non-response in questionnaire data is non-ignorable. *Multivariate Behavioral Research, 35*(3), 321-364.

Brown, C. H. (1983). Asymptotic comparison of missing data process for estimating factor loadings. *Psychometrika, 48*(2), 269-291.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Elena, E. D. (2008). An overview of prevention and correction methods for non-response in surveys. *Analele Stiintifice ale Universitatii "Alexandru Ioan Cuza" din Iasi - Stiinte Economice, 55*, 371-380.

Enders, C. K. (2010). *Applied missing data analysis* (1st ed.). New York, NY: Guildford Press.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(3), 430-457.

Enders, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: Choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(1), 1-19.

Gilks, W., Richardson, S., & Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in practice*.

London, UK: Chapman and Hall.

Kamakura, W. A., & Wedel, M. (2000). Factor analysis, missing data, discrete variables, data fusion, item non-response. *Journal of Marketing Research, 37*(4), 490-498.

Liu, C., & Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statistica Sinica, 8*, 729-747.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29*(4), 409-454.

McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, NY: John Wiley & Sons.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika, 47*(1), 67-76.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York, NY: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*(4), 545-571.

Yang, M.-L., & Tam, T. (2004, May). *Mental health inequality in the adolescent society: Family background and the paradox of academic success in Taiwan*. Paper presented at the conference on Social Stratification, Mobility, and Exclusion, the Research Committee on Social Stratification and Mobility (RC28) of the International Sociological Association, Neuchatel, Switzerland.

附錄

附表 1

原始資料、仿缺失比率一、二、三、四、五倍資料集的缺失結構

Missing pattern								T0		T1		T2		T3		T4		T5	
Group	S426	S427	S428	S429	S430	S431	S434	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent	Freq	Percent
1	×	×	×	×	×	×	×	11,191	91.79	10,204	91.18	9,232	82.49	8,291	74.09	7,386	66	6,549	58.52
2	×	×	×	×	×	×	.	53	0.43	53	0.47	106	0.95	156	1.39	204	1.82	249	2.23
3	×	×	×	×	×	.	×	19	0.16	18	0.16	37	0.33	53	0.47	70	0.63	87	0.78
4	×	×	×	×	×	.	.	6	0.05	6	0.05	12	0.11	18	0.16	24	0.21	30	0.27
5	×	×	×	×	.	×	×	10	0.08	10	0.09	20	0.18	30	0.27	40	0.36	49	0.44
6	×	×	×	×	.	×	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
7	×	×	×	×	.	.	×	2	0.02	2	0.02	4	0.04	6	0.05	7	0.06	8	0.07
8	×	×	×	.	×	×	×	10	0.08	10	0.09	20	0.18	30	0.27	40	0.36	50	0.45
9	×	×	×	.	×	×	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
10	×	×	.	×	×	×	×	9	0.07	9	0.08	18	0.16	27	0.24	35	0.31	43	0.38
11	×	.	×	×	×	×	×	3	0.02	3	0.03	6	0.05	9	0.08	12	0.11	15	0.13
12	×	.	×	×	×	×	.	1	0.01	1	0.01	2	0.02	3	0.03	3	0.03	3	0.03
13	×	.	×	×	×	.	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
14	×	.	.	×	×	×	×	2	0.02	2	0.02	4	0.04	6	0.05	8	0.07	9	0.08
15	×	.	.	.	.	.	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
16	.	×	×	×	×	×	×	838	6.87	827	7.39	1,640	14.65	2,432	21.73	3,193	28.53	3,896	34.81
17	.	×	×	×	×	×	.	3	0.02	3	0.03	6	0.05	9	0.08	12	0.11	15	0.13
18	.	×	×	×	×	.	×	5	0.04	5	0.04	9	0.08	13	0.12	17	0.15	20	0.18
19	.	×	×	×	×	.	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
20	.	×	×	×	.	×	×	2	0.02	2	0.02	4	0.04	6	0.05	8	0.07	10	0.09
21	.	×	×	×	.	.	×	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
22	.	×	×	.	×	×	×	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
23	.	×	.	×	×	×	×	2	0.02	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
24	.	.	×	×	×	×	×	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	4	0.04
25	.	.	×	.	×	.	.	1	0.01	1	0.01	2	0.02	3	0.03	4	0.04	5	0.04
26	.	.	.	.	.	.	.	27	0.22	26	0.23	51	0.46	72	0.64	92	0.82	109	0.97
Total								12,192	100.0	11,191	100.0	11,191	100.0	11,191	100.0	11,191	100.0	11,191	100.0

註：T0：原始資料集中的缺失架構，共二十六種缺失模式，每種模式的缺失率不同。T1 到 T5 是從基準資料集（樣本數 11,191 筆）中，依據 T0 的架構刪除心理健康題項中的某些觀察值而成。T1 的缺失架構與 T0 同，T2 在心理健康各題項的缺失比率，大致為 T0 的兩倍，T3 為 T0 的三倍，以此類推。

（續）

附錄（續）

附表 2

原始資料、仿缺失比率一、二、三倍資料集，因素分析的第一特徵值（基本資料集 3.393147）

插補方法	缺失比率倍數				
	T1	T2	T3	T4	T5
AC	3.952788	3.928112	3.862948	3.822590	3.779685
CC	3.387029	3.378176	3.353925	3.354489	3.336885
MCMC	3.392351	3.382856	3.377404	3.384456	3.377741
LR	3.394562	3.399376	3.396941	3.401005	3.397579

附表 3(a)

仿缺失比率一、二、三、四、五倍資料集，在 CC、AC 缺失處理後，因素分析的因素負荷與誤差均方根

因素一	基準資料集	CC					AC				
		T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
W2S426	0.639440	0.639710	0.638060	0.637970	0.639850	0.641670	0.211250	0.187110	0.159420	0.133580	0.106580
W2S427	0.796330	0.796850	0.799290	0.797350	0.797280	0.793600	0.893090	0.893230	0.889620	0.889010	0.889250
W2S428	0.731700	0.729010	0.730630	0.728810	0.725140	0.726040	0.848540	0.847680	0.841970	0.840250	0.836880
W2S429	0.731180	0.730360	0.728360	0.726540	0.727540	0.723640	0.833330	0.832560	0.825680	0.820930	0.815130
W2S430	0.771300	0.771050	0.770990	0.768870	0.768700	0.769320	0.826700	0.824630	0.817490	0.813310	0.809360
W2S431	0.641090	0.638890	0.634450	0.628040	0.629210	0.625760	0.764940	0.762810	0.758790	0.754210	0.749000
W2S434	0.523650	0.524620	0.520490	0.516260	0.517180	0.511000	0.653850	0.649330	0.641400	0.635720	0.629840
RMSE		0.001419	0.003250	0.006140	0.005941	0.008454	0.189795	0.196824	0.203562	0.210931	0.218780

附表 3(b)

仿缺失比率一、二、三、四、五倍資料集，在 MCMC、LR 缺失處理後，因素分析的因素負荷與誤差均方根

因素一	基準資料集	MCMC					LR				
		T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
W2S426	0.639440	0.637000	0.632490	0.625840	0.628920	0.630410	0.639900	0.644420	0.642020	0.650120	0.642420
W2S427	0.796330	0.795920	0.795870	0.796290	0.794730	0.796740	0.795670	0.797560	0.798040	0.795840	0.796480
W2S428	0.731700	0.731410	0.730910	0.732180	0.732510	0.729960	0.731140	0.731010	0.730800	0.732260	0.731650
W2S429	0.731180	0.731910	0.731560	0.733270	0.733460	0.732040	0.731460	0.731070	0.730360	0.729540	0.730710
W2S430	0.771300	0.770700	0.770620	0.771490	0.772050	0.771870	0.771400	0.770990	0.772410	0.771460	0.774210
W2S431	0.641090	0.641900	0.639600	0.639810	0.639860	0.638970	0.642190	0.640090	0.640300	0.638380	0.639380
W2S434	0.523650	0.525770	0.526360	0.522810	0.526610	0.522250	0.524350	0.524440	0.523240	0.523320	0.522530
RMSE		0.001323	0.002911	0.005236	0.004308	0.003630	0.000627	0.002019	0.001367	0.004222	0.001764

Journal of Research in Education Sciences
2012, 57(1), 29-50

Missing Data Techniques for Factor Analysis

Hong-Long Wang
Department of Statistics,
National Taipei University

Meng-Li Yang
Research Center for Humanities and Social Sciences,
Academia Sinica

Chun-Ju Chen
Department of Statistics,
National Taipei University

Ting-Hsiang Lin
Department of Statistics,
National Taipei University

Abstract

Factor analysis is frequently employed to analyze scales and questionnaires. However, when the proportion of missing data is high or the missing data are not random, the number of factors extracted can be biased. We used the Taiwan Education Panel Survey (TEPS) and constructed 5 data sets with different missing proportions to assess the effects of missingness on factor analysis imputation. Complete observed data were used as a baseline for comparison. We compared the 4 treatments: available case method (AC), the complete case method (CC), MCMC single imputation (MCMC), and step-wise logistic regression single imputation (LR). The results show that the higher the missing proportion, the greater the discrepancy between the covariance matrix of the constructed data set and that of the baseline. For the AC method, the higher the proportion of missing data, the more the number of extracted factors exceeds that of the baseline. The AC method possessed the largest bias in factor loadings. The bias in factor loading of the CC method increased as the missing portion also increased. Thus, we recommend not applying the list-wise deletion method for factor analysis when the missing proportion is 20% or more.

Keywords: TEPS, missing data, exploratory factor analysis, MCMC imputation, logistic regression imputation

Corresponding Author: Ting-Hsiang Lin, E-mail: tinghlin@mail.ntpu.edu.tw