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