第二節　測量模式

大型測驗的主要目的為評量群體的知識與技能，而評量牽涉到許多議題，首先進行測量前，應先編製所欲評量目標之試題，因應不同的評量目標需求，利用不同的試題類型進行施測，並確保試題的數量是夠多且可以涵蓋不同的難度範圍。施測時透過試題及題本的設計，受試者只施測部分的試題，以減輕受試者的負擔。

依據不同的試題類型，透過測量模式，提供不同的估計方法與模式的估計，將估計出來的試題與能力參數提供給受試者以及研究者。本節中，將介紹不同大型測驗中所使用的測量模式，針對模式估計的適合度進行探討，並羅列所使用估計試題與能力參數的測驗分析軟體。

以下將以 NAEP 1998 （Allen, Carlson, Johnson, & Mislevy, 1999）、TIMSS 2007（Foy, Galia, & Li, 2008）和 PISA 2003（OECD, 2005）的技術報告為主，針對這三大測驗所使用的試題類型、測量模式、模式適合度評估指標及測驗分析軟體作一整理說明。

壹、試題類型

目前在 NAEP 1998（Allen, Carlson, Johnson, & Mislevy, 1999）、TIMSS 2007（Foy, Galia, & Li, 2008）、PISA 2003（OECD, 2005）和 TASA 大型測驗中，測驗的題型大致上可以分為三大類：

一、選擇題（**multiple-choice items**）

在四種大型測驗中皆為四個選項的選擇題。

二、填充題

NAEP 的填充題（short constructed response items）可以分為答對答錯的二元計分，以及三點計分（0-2）兩種，PISA 的填充題可以分為封閉性填充題（closed-constructed response items）以及開放性填充題兩種，封閉性填充題是指固定單一答案，像是數學科中，需要學生填入的可能為一個數值，開放性的填充題則是學生的反應可以較廣泛的作答，而非單一的答案，TIMSS 的填充題（constructed-response items with just two response options）為二元計分的封閉性填充題。

三、開放性試題

NAEP 中的開放性試題可以分為三點計分（0-2）到六點計分（0-5）四種，PISA 中的開放性試題包括比較長的寫作、結論、摘要、批判等學生作答反應較廣泛之試題，TIMSS 中的開放性試題，大多為三點計分（0-2），各技術報告中的名詞如表 4-2-1 所示：

表 4-2-1 **NAEP、PISA、TIMSS、TASA 試題類型**

NAEP	PISA	TIMSS	TASA
選擇題（multiple-choice items）	選擇題（multiple-choice response）	選擇題（multiple-choice items）	選擇題
填充題（short constructed response items）	填充題（short answer）	填充題（constructed response items with just two response options）	無此題型
開放性試題（extended constructed response items）	開放性試題（open constructed response）	開放性試題（constructed response items）	開放性試題

貳、測量模式

不同大型測驗間針對不同測驗題型，使用不同的測量模式，常見的有二參數對數模式（two-parameter logistic model, 2PL）、三參數對數模式（three-parameter logistic model, 3PL）、一般化部分給分模式（generalized partial credit model, GPCM）以及多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM）。

一、二參數對數模式（**2PL**）

在 IRT 的 2PL 模式下，假設受試者 j 之能力為 θ_j，其作答試題 i 通過的機率如下（Birnbaum, 1968）：

$$P(X_{ij}=1|\theta_j,b_i,a_i)=\frac{1}{1+\exp[-a_i(\theta_j-b_i)]}$$

其中，X_{ij}為受試者j在試題i的作答反應，答對記為 1，答錯記為 0；a_i為試題i之試題鑑別度參數（item discrimination parameter），$-\infty < a_i < \infty$；b_i為試題i之試題難度參數，$-\infty < b_i < \infty$。

二、三參數對數模式（3PL）

在 IRT 的 3PL 模式下，假定測驗會發生猜題之現象，故假設受試者j之能力為θ_j，其作答試題i通過的機率如下（Birnbaum, 1968；Lord, 1980）：

$$P(X_{ij}=1|\theta_j,b_i,a_i,c_i)=c_i+\frac{(1-c_i)}{1+\exp[-a_i(\theta_j-b_i)]}$$

其中，X_{ij}為受試者j在試題i的作答反應，答對記為 1，答錯記為 0；a_i為試題i之試題鑑別度參數，$-\infty < a_i < \infty$；b_i為試題i之試題難度參數，$-\infty < b_i < \infty$；c_i為試題i之試題猜測度參數（item guessing parameter），$0 \le c_i < 1$。

三、一般化部分給分模式（GPCM）

Muraki（1992）所提出，為各試題之間有不同的鑑別度參數。GPCM 模式假定一試題j具有m_j個等級類別（graded categories），越高的類別表示能力越高，而最高得分為m_j，GPCM 模式如下

$$P_{jk}(\theta)=\frac{\exp\left[\sum_{v=1}^{k}a_j(\theta-b_{jv})\right]}{\sum_{c=1}^{m_j}\left[\exp\sum_{v=1}^{c}a_j(\theta-b_{jv})\right]}=\frac{\exp\left[\sum_{v=1}^{k}a_j(\theta-b_j+d_v)\right]}{\sum_{c=1}^{m_j}\left[\exp\sum_{v=1}^{c}a_j(\theta-b_j+d_v)\right]}$$

其中

θ：表示受試者的潛在能力特質（$-\infty<\theta<\infty$）。

k：為受試者的回答所屬類別，從 1…m_j。

e：是底為 1.728 的指數。

m_j：為隨題目而變的變數，m_j則是第j題所有的類別數。

$P_{jk}(\theta)$：為潛在能力特質為θ的受試者在第j題得到第k類的機率（$0<P_{jk}(\theta)<1$）。

b_{jv}：$b_{jv}=b_j-d_v$。b_{jv}為第j題第 v 個的試題步驟難度參數（item step parameter）

或類別閾參數（category intersection parameter），隨著類別界線（category boundary）而變，相鄰在兩類別間，就有一個 b_{jv} 參數（$-\infty < b_{jv} < \infty$），即 b_{jk} 為 $P_{j,k-1}(\theta)$ 和 $P_{jk}(\theta)$ 的交點，同一試題內的試題步驟參數不需是有序的。b_j 為試題座標參數（item location parameter）、d_v 為閾參數（threshold parameter），d_k 為同一試題內的第 k 類和其他類別的相對難度（Andrich, 1982）。

a_j：試題 j 的斜率參數，同一試題在各類別選項有相同的斜率參數，但不同的試題有不同斜率。

四、多向度隨機係數多項洛基模式（MRCMLM）

MRCMLM是由Adams、Wilson與Wang（1997）等人所提出，MRCMLM為Rasch模式的衍生模式，是一個混合的co-efficients模型（mixed co-efficients model），試題參數是由未知的參數ξ所描述，而受試者的潛在變數 θ，是一個隨機變項，其模式定義如下：

$$P(X_{ik}=1;\mathbf{A},\mathbf{B},\xi \mid \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\xi)}{\sum_{k=1}^{K_i}\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\xi)}$$

其中，X_{ik}：受試者之做答反應組型

K_i：第 i 試題的計分類別數

$\boldsymbol{\theta}$：受試者的能力參數矩陣（多向度能力）

ξ：試題參數向量

$\mathbf{a}_{ik}$：第 i 題中第 k 個反應類別的設計向量（designing vector）

$\mathbf{b}_{ik}$：第 i 題在第 k 個反應類別上的計分向量（scoring vector）

A：整份測驗的設計矩陣（designing matrix）

B：整份測驗的計分矩陣（scoring matrix）

表 4-2-2 為整理各大型測驗中，不同的測驗題型所使用的模式。

表 4-2-2　大型測驗所使用之測量模式

題型	NAEP	TIMSS	PISA	TASA
選擇題	三參數對數模式（Birnbaum, 1968; Lord, 1980）	三參數對數模式（Birnbaum, 1968; Lord, 1980）	多向度隨機係數多項洛基模式，MRCML（Adams, Wilson & Wang, 1997）	三參數對數模式（Birnbaum, 1968; Lord, 1980）
填充題	二參數對數模式（Birnbaum, 1968）	二參數對數模式（Birnbaum, 1968）		無此題型
開放性試題	一般化部分給分模式(Muraki, 1992）.	一般化部分給分模式(Muraki, 1992）.		敘述統計分析

五、綜合討論與建議

目前國際大型測驗在選擇題題型部分 NAEP、TIMSS 皆使用三參數對數模式，PISA 則使用多向度隨機係數多項洛基模式，目前 TASA 為使用三參數對數模式，未來建議繼續使用此模式進行選擇題的分析，題組試題的部分，建議使用題組模式分析；開放性試題部分 NAEP、TIMSS 皆使用一般化部分給分模式，PISA 則使用多向度隨機係數多項洛基模式，目前 TASA 針對開放性試題僅進行敘述統計分析，未來建議使用一般化部分給分模式進行試題分析。

參、模式適合度評估方法

在模式適合度方面，NAEP、TIMSS 中的模式適合度是使用圖形化判斷方法，PISA 是以標準殘差（standardised residual）為基礎，建立非權重的適合度統計量（unweighted fit statistic）和權重適合度統計量（weighted fit statistic）。詳述如下:

一、圖形化判斷方法

在NAEP、TIMSS中使用試題適合統計量，因為沒有一個真正的 χ^2 分佈，測量試題適合統計量為比較真實與理論上的試題反應函數，看試題對於模式而言是否較不合適，像是多點計分試題中某個類別的得分較低，或者某一題的反應與理論上不符合。對於IRT模式適合度的方法，為比較同一量尺上的觀察值以及理論上的試題反應函數所產生的曲線，其中理論上的曲線（theoretical curves）是根據試題參數的估計值所畫出來的，而觀察值則是依據有施測該試題的學生所產生的

後驗分佈而得。對二元計分試題而言，能力值為$\boldsymbol{\theta}$的學生答對該試題的後驗分佈加上能力值為$\boldsymbol{\theta}$的學生遺漏該試題的後驗分佈，此方法相似於答對該題的學生加上遺漏該題的學生。在每個能力值上學生施測該試題（receiving the item）的後驗分佈機率值的總和，相似於在各能力點上施測該試題的學生數。最後的試題反應曲線的值（plotted values）為各能力值的個別後驗機率值的加總，利用答對該題的後驗機率值加上遺漏該題的後驗機率值除以施測該題的後驗機率值，在估計完試題參數後，通常是透過估算試題的適配程度代表IRT模式的適合度。利用圖形的分析做為模式適合度的評估準則，比較實際值與理論值的曲線來做為判斷式，兩者的圖形越接近重疊，則適配的情形越好。

二、適合度統計量

PISA中使用ConQuest 軟體所提供的適合度檢定方法，ConQuest 軟體針對每個需要估計的參數，提供一個適合度的檢定，此檢定是Wu（1997）以Wright 及Masters（1982）所提出的論點為基礎所發展出來的，Wu（1997）將它延伸到兩個面向。第一，將它他應用到更廣泛的模式中，提供參數的適合度檢定，而非原始的試題的適合度檢定。第二，Wright 和 Masters(1982)所提出之適合度統計方法適用於非條件式的最大概似估計法中（unconditional maximum likelihood estimates），而Wu（1997）將其延伸至可用至邊際最大概似估計法中（marginal maximum likelihood estimates）。

令A_p為設計矩陣A中的第p行，Wu（1997）的適合度統計是以標準殘差為基礎的。

$$z_{np}(\theta_n)=\frac{A'_p x_n - E_{np}}{\sqrt{V_{np}}}$$

其中，$A'_p x_n$為受試者n在參數p上的充份統計量，E_{np} 和 V_{np}分別為$A'_p x_n$的條件期望值與變異數，建立一個非權重的適合度檢定，殘差的平方為個別後驗機率分佈積分的平均。

$$Fit_{out,p}=\int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_N}\left[\frac{1}{N}\sum_{n=1}^{N}\hat{z}_{np}^2(\theta_n)\right]\prod_{n=1}^{N}h_\theta(\theta_n;Y_n,\hat{\xi},\hat{\beta},\hat{\sigma}^2\mid x_n)d\theta_N d\theta_{N-1}\cdots d\theta_1$$

針對權重的適合度檢定，殘差平方的權重平均可以表示如下式所示：

$$Fit_{in,p} = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_N}\left[\frac{\sum_{n=1}^{N}\hat{z}_{np}^2(\theta_n)V_{np}(\theta_n)}{\sum_{n=1}^{N}V_{np}(\theta_n)}\right]\prod_{n=1}^{N}h_\theta(\theta_n;Y_n,\hat{\xi},\hat{\beta},\hat{\sigma}^2\mid x_n)d\theta_N d\theta_{N-1}\cdots d\theta_1$$

在ConQuest中，蒙地卡羅方法用來逼近上述方程式的積分，Wu（1997）表示上述方程式近似於卡方分配，利用Wilson-Hilferty transformations轉換方法將統計量轉換成近似於常態。

$$t_{out,p} = \frac{(Fit_{out,p}^{\frac{1}{3}} - 1 + \frac{2}{(9rN)})}{(\frac{2}{9rN})^{\frac{1}{2}}}$$

和

$$t_{in,p} = \left[Fit_{out,p}^{\frac{1}{3}} - 1\right]\times\frac{3}{\sqrt{Var(Fit_{in,p})}} + \frac{\sqrt{Var(Fit_{in,p})}}{3}$$

其中，r 為蒙地卡羅法的抽取次數。

$$Var(Fit_{in,p}) = \left[\frac{1}{\sum_n V_{np}}\right]^2\times\frac{3}{\sqrt{Var(Fit_{in,p})}} + \frac{\sqrt{Var(Fit_{in,p})}}{3}$$

詳細的推導過程請詳閱 Wu（1997）。表 4-2-3 是各大型測驗中模式適合度評估之方法。

表 4-2-3　模式適合度評估方法

NAEP	TIMSS	PISA	TASA
圖形化判斷	圖形化判斷	Unweighted fit statistic and Weighted fit statistic	無

三、綜合討論與建議

模式適合度評估方法目前 NAEP、TIMSS 皆使用圖形化判斷方式進行評估，PISA 則使用 ConQuest 軟體所提供的適合度檢定方法，Unweighted fit statistic and Weighted fit statistic，但由於此方法僅適用於 Rasch 模式下，因此不適合目前為使用三參數對數模式的 TASA，TASA 目前尚未使用任何方法進行模式適合度評估，未來建議參考 NAEP、TIMSS 使用圖形化判斷方式進行模式適合度評估。

肆、測驗分析軟體

不同的大型測驗使用不同的分析軟體進行參數之估計，TIMSS 中分別使用 BILOG 進行二參數對數模式、三參數對數模式試題進行分析，使用 PARSCALE 進行一般化部分給分模式試題進行分析；NAEP 中使用結合 BILOG 和 PARSCALE 的 NAEP BILOG/PARSCALE 軟體進行二參數對數模式、三參數對數模式以及一般化部分給分模式試題分析；PISA 中使用的模式為多向度隨機係數多項洛基模式，因此其參數估計軟體為使用適合多向度隨機係數多項洛基模式的 ConQuest 進行分析，各大型測驗所使用的估計軟體整理如表 4-2-4 所示。

表 4-2-4　大型測驗使用之測驗分析軟體

NAEP	TIMSS	PISA	TASA
BILOG（Mislevy & Bock's, 1982） PARSCALE （Muraki & Bock's, 1991）	BILOG（Mislevy & Bock's, 1982） PARSCALE （Muraki & Bock's, 1991）	ConQuest （Wu, Adams, & Wilson, 1998）	BILOG（Mislevy & Bock's, 1982） SCORIGHT （Wang, Bradlow & Wainer, 2004）

一、綜合討論與建議

目前大型測驗中 NAEP、TIMSS 依題型選擇題、填充題以及開放性試題使用分析模式分別為三參數對數模式、二參數對數模式和一般化部分給分模式，因此分別使用 BILOG、PARSCALE 軟體進行分析，PISA 則是使用多向度隨機係數多項洛基模式進行三種題型的試題分析，因此選用 ConQuest，目前 TASA 針對選擇題部分使用 BILOG 進行分析，未來建議 TASA 繼續使用 BILOG 進行選擇題分析，並增加 PARSCALE 進行開放性試題分析。