單向度試題反應理論之可能值方法於等化設計下之模擬實驗探究

許天維[1] 郭伯臣[2] *吳慧珉[3] 葉昶成[4]

摘要

許多國際大型測驗，多採用可能值方法來進行群體能力參數的估計。而可能值的資料型態，亦可讓資料分析者進行統計特性的描述。此外，一般大型測驗所評量的範圍都涵蓋了不同的認知向及難度，無法由單一受試者於短期間內全部完成，測驗題目都會進行不同的等化設計以減輕受試者負擔並達成測驗的目的。本研究係各以定錨不等組（non-equivalent groups with anchor test design, NEAT）及平衡不完全區塊（balanced incomplete block design, BIB）的垂直等化設計，並以可能值方法、納入背景變項的期望後驗法、期望後驗法及最大概度估計法等各種方法進行群體能力的平均數與標準差的估計，主要的目的在於探討可能值方法及其它估計法在群體參數估計的效果。本研究結果顯示在各種不同的等化設計下，群體能力平均數與標準差的估計，納入背景變項估計方法皆有較好的估計效果，特別是群體能力標準差的估計，可能值方法的估計結果遠優於各種估計方法。

關鍵字：大型測驗、參數估計、可能值方法、等化設計、試題反應理論

[1]許天維，國立臺中教育大學教育測驗統計研究所
[2]郭伯臣，國立臺中教育大學教育測驗統計研究所
[3]吳慧珉，國家教育研究院測驗及評量研究中心
[4]葉昶成，臺中市惠來國民小學
通訊作者：吳慧珉；Email: whm@mail.naer.edu.tw
地址：新北市三峽區三樹路 2 號 國家教育研究院測驗及評量研究中心

Exploring the Performance of Plausible Values Method Based On the Unidimensional Item Response Theory under Equating Designs -A Simulating Study

Tian-Wei, Sheu[1] Bor-Chen Kuo[2] *Huey-Min Wu[3] Chang-Chen Ye[4]

Abstract

The purpose of this paper is to explore the performance of plausible values method under BIB and NEAT designs based on simulated data. The major focus of large-scale assessments is always on the population statistics, such as means and standard deviations, and the plausible value method is usually used to estimate the population parameters. For large-scale assessments the spectrum of subject matter is usually wide, but the testing time is short. Therefore, in order to cover the proficiency domain sufficiently, multiple booklets are used. Balanced incomplete block design (BIB) and non-equivalent groups with anchor test design (NEAT) are two popular test equating methods for this condition. The experimental results show that the estimating method based on plausible values estimate better than that of other methods in equating designs, and as the test length increase, population parameters (means and standard deviations) are well estimated. In these experimental situations, the estimations of population parameters are not affected by sample size (16,128 and 10,920). Both linking designs, BIB and NEAT, can lead to more precision estimates by using plausible value method.

Keywords: equating design, item response theory, large-scale assessment, parameter estimation, plausible values

[1]Tian-Wei, Sheu, Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education
[2]Bor-Chen, Kuo, Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education
[3]Huey-Min, Wu, Research Center for Testing and Assessment, National Academy for Educational Research.
[4]Chang-Chen, Ye, Hui-Lai Elementary School, Taichung City
Corresponding Author: Huey-Min, Wu; Email: whm@mail.naer.edu.tw
Address: No.2, Sanshu Rd., Sanxia Dist., New Taipei City 237, Taiwan (R.O.C.)

壹、緒論

一、研究動機

近年來國際上陸續舉辦了數種大型測驗（large-scale assessments），諸如 Trends in International Mathematics and Science Study(TIMSS)、The Progress in International Reading Literacy Study (PIRLS)及 The Programme for International Student Assessment(PISA)等，另外如美國實施的 National Assessment of Educational Progress(NAEP)，這些大型測驗主要關注的焦點在於母群或母群中某些群體之能力表現，並長時間測量教育的發展情形等，而這些數據也將作為國家未來教育政策訂定之參考。這些大型測驗主要是採用可能值方法（plausible value method, PV）估計母群或次群體的參數，對於個體能力的描述則是以可能值（plausible value）的資料型態提供給次級資料的分析者，以進行群體能力表現時統計特性描述（Allen, Carlson, Johnson, & Mislevy, 1999; Foy, Galia, & Li, 2008; Graham, Christine, Alka, & Ebru, 2008; Martin, Mullis, & Kennedy, 2007; OECD, 2009）。

可能值方法可視為是階層式試題反應理論模式（hierarchical item response theory model），主要是將母群分佈以潛在迴歸模式表徵，在潛在迴歸模式中加入學生背景變項（background variables, BV），作為輔助變數（ancillary variables, AV）計算後驗分布，並從後驗分布中抽取可能值。可能值方法沒有先估計個體的能力再計算群體參數，而是使用學生的答題反應和背景變項資料直接估計母群參數，相較於集合個體能力估計值再進行母群參數估計，此方法所獲得之估計較為準確（Mislevy & Sheehan, 1989）。目前可能值方法的研究大多著重於估計方法的改善，如 Adams 等人透過提供背景變項的模式，提出 expectation-maximization（EM）方法估算試題和母群的參數，改善可能值估計方法，在能力估計中得到較小均方誤差（Adams, Wilson & Wu, 1997）。von Davier（2009）於馬可夫鏈蒙地卡羅法（Markov Chain Monte Carlo method, MCMC）中，加入 EM 的演算法，提升 MCMC 估計的速度以及估計效能。Wu（2005）探討在單組測驗設計下，不同估計方法，如期望後驗法（expected a posteriori , EAP）和 PV 對於母群之平均數與標準差之估計效能，研究結果顯示 PV 法對於母群標準差之估計精準優於其他方法。除此之外，少數學者探討納入的背景變項與能力值間的相關高低對可能值方法估計的影響，如 de la Torre（2009）的研究顯示所納入的輔助變項能解釋能力值變異的程度越大或能力值之間的相關越高，越能有效提升能力值的估計精準度。

上述研究主要是以單組測驗為主，而可能值方法主要是用於大型測驗中，大型測驗因題庫涵蓋不同認知程度及不同難度之試題，試題數量無法由單一受試者於短時間內完成，故多採用不同的等化設計進行；PISA 為採用 BIB（balanced incomplete block, BIB）等化設計（OECD, 2009）；NAEP 則在數學與科學使用 BIB 設計、閱讀與寫作方面則使用了 PBIB（partially balanced incomplete block, PBIB）設計（Andrew & Terry, 2001）；TIMSS2011 則是每個題本由四個試題區塊組合而成（每個題本均包含數學與科學各兩個試題區塊），而為了連結不同題本，每個試題區塊在題本中出現 2 次（Mullis, Martin, Ruddock, Sullivan & Preuschoff, 2009）；國內的「臺灣學生學習成就評量資料庫」（Taiwan Assessment of Student Achievement, TASA）也於不同年度不同科目，分別採用了 BIB、PBIB 以及定錨不等組設計（non-equivalent groups with anchor test design, NEAT）的等化設計（郭伯臣、曾建銘、吳慧珉，2012）。

關於 BIB 與 NEAT 設計之水平及垂直等化效果在個體能力估計的成效已有學者探究（郭伯臣，王暄博，2008），然而大型測驗皆使用可能值方法進行分析，且於題本設計上使用等化設計方法，但僅有少數的文章探討關於可能值方法應用於 BIB 設計上（von Davier, Gonzalez, & Mislevy, 2009），卻未探討不同的等化設計，對於可能值方法的參數估計之影響。除了水平等化設計外，大型測驗中亦需使用垂直等化設計進行跨年級的等化連結，以便能描繪出學生學習成長趨勢，進行學生學習成就長期性資料庫的建置（Glas & Geerling, 2009），然而垂直等化設計下可能值方法估計效果這方面的研究卻是付之闕如。本研究透過模擬研究，在不同的水平與垂直等化設計下，進行可能值方法及其它估計方法的估計效果探討。

二、研究目的

本研究以模擬研究的方式，探討在 NEAT 與 BIB 二種不同等化設計中，包含不同人數與不同題數的情況下，可能值等六種不同估計方法之估計成效。研究目的條列如下：

（一）水平等化設計下，不同的估計方法於不同的情境（不同受試人數與測驗題數）對於群體能力參數估計之成效探討。

（二）垂直等化設計下，不同的估計方法於不同的情境（不同受試人數與測驗題數）對於群體能力參數估計之成效探討。

（三）不同的估計方法於不同的等化情境（NEAT 與 BIB）對於群體能力參數估計之成效探討。

貳、文獻探討

本研究係以單向度單參數試題反應理論為基礎，使用模擬資料進行模擬實驗，以不同等化設計方式對進行模擬施測，並以可能值等數種估計法進行參數估計，探討不同估計方法對於群體參數的估計效果；因此，本節將針對單向度試題反應理論、參數估計方法以及測驗等化設計方法等相關研究進行分析整理。

一、單向度試題反應理論

本研究係使用單參數對數模式進行探討，單參數對數模式又稱為 Rasch 模式，假設受試者 j 之能力為 θ_j，其答對試題 i 的機率為 $P_i(\theta_j)$，其數學公式如下（Rasch, 1960）：

$$P_i(\theta_j) \equiv P(X_{ij}=1 \mid \theta_j, b_i) = \frac{1}{1+\exp[-(\theta_j - b_i)]} \quad i=1,2,3,...n \quad j=1,2,3,...N \quad \text{（公式 2-1）}$$

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，

b_i 為試題 i 之試題難度參數（item difficulty parameter）；

n 為試題長度；N 為受試者人數。

二、參數估計方法

本研究乃以傳統的點估計方法、加入輔助變數之期望後驗估計法與可能值方法等進行探討，在傳統點估計方法中主要使用最大概似估計法、加權概似估計法（weighted likelihood estimation, WLE）與期望後驗估計法，此三種估計法之詳細過程有興趣的讀者請參閱 Baker & Kim（2004）與 Warm（1989）。以下介紹加入輔助變數之期望後驗估計法與可能值方法。

（一）加入輔助變項之期望後驗估計法（expected a-posteriori with ancillary variables, EAP_AV）

期望後驗估計法（expected a-posteriori, EAP）的估計過程是依據貝氏理論，如公式(2-2)，其中公式(2-2)的分母可以表示如公式(2-3)(Baker & Kim, 2004)。

$$g(\theta_j \mid u_j, \xi) = \frac{P(u_j \mid \theta_j, \xi) g(\theta)}{P(u_j)} \quad \text{(公式 2-2)}$$

$$P(u_j) = \int_{\theta} P(u_j \mid \theta) g(\theta) d\theta \quad \text{(公式 2-3)}$$

其中， θ_j 受試者 j 的能力，ξ 是試題參數。在局部獨立的假設下，受試者 j 的作答反應向量為 $\mathbf{u}_j = (X_{1j}, ..., X_{nj})$，其答對機率如公式（2-4）

$$P(\mathbf{u}_j \mid \theta_j, \xi) = \prod_{i-1}^{n} P_i(\theta_j)^{X_{ij}} Q_i(\theta_j)^{1-X_{ij}} \quad \text{（公式 2-4）}$$

其中 X_{ij} 指受試者 j 在第 i 題的作答反應；$P_i(\theta_j)$ 指受試者 j 在第 i 題的答對機率；$Q_i(\theta_j)$ 指受試者 j 在第 i 題的答錯機率，$Q_i(\theta_j) = 1 - P_i(\theta_j)$。

加入輔助變項之期望後驗估計法（EAP_AV）主要是以 EAP 法為基礎如公式（2-2）與（2-3），然後加入輔助變項進行後驗分布估計。在 EAP 中，因假設抽樣的學生是來自於一個常態分布的母體，其平均數為 μ，變異數為 σ^2，故公式（2-2）之 g(θ)如下：

$$g(\theta) \equiv f_{\theta}(\theta; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(\theta - \mu)^2}{2\sigma^2} \right] \quad \text{（公式 2-5）}$$

Adams（1997）等人將平均數 μ 以迴歸模式 $Y_j^T \beta$ 取代，假設有 d 輔助變項，則 Y_j 是一個 d 的向量，對於受試者 j，Y_j 是固定且已知的條件變數（如性別或社經地位等輔助變數），β 是一個相對應的迴歸係數向量，則學生 j 的母群模式可表示為 $\theta_j = Y_j^T \beta + E_j$。其中，假設 $E_j \overset{iid}{\sim} N(0, \sigma^2)$，此種方法本研究稱為加入輔助變數之期望後驗估計法（EAP_AV）

（二）可能值方法（PV）

試題反應模式中，無法直接觀察到個體的能力值，亦即表示個體能力的測量含有不確定性；而在計算群體統計量和相關連的標準誤時，應考量這些不確定性（Allen, Donoghue, & Schoeps, 2001; Mullis, Martin, & Foy, 2008; OECD, 2009）。PV 法是從後驗分布中隨機抽取學生的可能值，則能考量到上述提及的不確定性；而且 PV 法並沒有先估計個體的能力再計算群體參數，而是直接估計母群的參數，這樣可以使參數的估計更為精準（Mislevy & Sheehan, 1989）。

PV 法是以潛在迴歸模式，加入學生答題反應和相關條件變項後計算每一位學生的後驗機率分布，並從後驗分布中隨機抽取學生可能的能力值，以利於次級資料分析者使用。當模式被正確界定時，可能值可以提供群體參數的一致性估計，但並非個體能力的不偏估計，使用可能值的平均並不能代表個別學生的能力（Mislevy, Beaton, Kaplan, & Sheehan, 1992）。

試題反應理論是以能力值 θ 為條件而產生試題反應的過程，此模式需要界定能力值 θ 的密度函數 $f_{\theta}(\theta; \alpha)$。令 α 為 θ 分布的參數集。當定義單向度邊際試題

反應模式（uni-dimensional marginal item response models），常假設抽樣的受試者是來自於一個常態分布的母體，其平均數為 μ，變異數為σ^2。也就是公式 2-5，或者同義的式子：

$$\theta = \mu + E \quad \text{（公式 2-6）}$$

其中，$E \sim N(0,\sigma^2)$。

Adams（1997）等人使用迴歸模式$Y_j^T\beta$取代平均數 μ，則受試者 j 的母群模式可表示為

$$\theta_j = Y_j^T\beta + E_j \quad \text{（公式 2-7）}$$

其中，假設$E_j \overset{iid}{\sim} N(0,\sigma^2)$。$E_j$的分布應該會和$\theta_j$相同，只是將其轉換為平均數為 0，利用迴歸模式$Y_j^T\beta$取代平均數 μ，其中$Y_j$為 u 的向量，β 為迴歸係數，則母體的模式可以被替換為如下：

$$f_j(\theta_j;Y_j,\beta,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp[-\frac{1}{2\sigma^2}(\theta_n - Y_j^T\beta)'(\theta_j - Y_j^T\beta)] \quad \text{（公式 2-8）}$$

公式（2-8）為一常態分配，平均數為$Y_j^T\beta$，及變異數為σ^2，若使用公式（2-8）估算母體分配，則需要估算的參數為 β，σ^2和 ξ（試題參數），其邊際後驗機率可以被表示如公式（2-9），其中，X_j為受試者 j 的作答反應

$$h_\theta(\theta_j;Y_j,\xi,\beta,\sigma^2 \mid X_j) = \frac{f_j(X_j;\xi \mid \theta_j) f_\theta(\theta_j;Y_j,\beta,\sigma^2)}{f_X(X_j;Y_j,\xi,\beta,\sigma^2)} \quad \text{（公式 2-9）}$$

可能值方法是從公式 2-9 的分佈中，抽取五個可能值（Foy, Galia, & Li, 2008）

一般而言，在大型測驗資料的分析上常遇到兩種錯誤的資料分析方式，一種是以傳統的點估計方法進行群體參數之估計，一種是錯誤的可能值使用方式，分別說明如下：

1.傳統的點估計方法進行群體參數估計

目前國內許多大型測驗相關研究，大多是直接計算個別受試者能力值的平均值與變異數，並未使用 PV 法進行分析，例如洪碧霞、林素微、林娟如（2006）是將 TASA 數學科施測資料採用上述方式進行分析研究。有學者研究指出，利用傳統點估計方式去推論群體參數容易造成偏誤（Mislevey, 1991; Mislevy, et al., 1992; OECD, 2005; Lee, Grigg, & Dion, 2007）。此外，von Davier, et al.（2009）比較傳統的點估計方法（最大概似估計法、期望後驗估計法）以及 PV 法間的估計效果，顯示 PV 法對於群體參數有較好的估計準確度。Wu（2005）比較點估計方法與 PV 法，估計母群平均數時兩種方法並沒有明顯的差異，但在變異數估計中，PV 法得到較好的結果。

2.錯誤的可能值使用方式導致估計偏誤

PV 法正確的使用方法是將 5 個 PV 個別的用來計算 PV 的平均後，再將此 5 個平均數進行估算；而 PV 法錯誤（wrong）的使用方法（PV_W）則是直接將針對每個受試者所抽出的 5 個 PV 直接加以平均進行估算。以上這兩種方法雖然在估算群體平均數時結果相當接近，但是在估算群體標準差時 PV_W 就會產生較大的偏差（von Davier, et al., 2009）

三、測驗等化

測驗等化是利用統計方法，將受試者在某測驗的分數轉換至另一測驗分數量尺，使得不同的測驗的結果能夠在同一個基準上做比較的一套流程，主要分為水平等化及垂直等化二種。水平等化係指利用測驗分數等化技術，將兩個或兩個以上測量相同特質、相同能力的測驗，其原始分數轉換之過程，在本研究中係指同年級間不同測驗間之等化；垂直等化（vertical equating）係指將兩組不同能力水平的受試者所獲得分數，利用測驗分數等化之技術將之置於同一個量尺上以使能互相比較。在本研究中係指不同年級間不同測驗間之等化。

測驗等化設計有許多種，諸如單組設計（single-group design）、等群組設計（equivalent-grwoup design）、定錨不等組設計（NEAT）、平衡不完全區塊（BIB）等。本研究主要以 BIB 與 NEAT 等化設計進行比較，NEAT 設計可參閱 von Davier, Holland, & Thayer（2004）之文獻；BIB 設計可參閱 van der Linden, Veldkamp & Carlson（2004）之文章。郭伯臣、王暄博（2008）以單向度三參數羅吉斯模式（three-parameter logistic model）為理論基礎探討利用 BIB 設計與 NEAT 設計，結果顯示 BIB 設計於試題參數估計精準度大致上優於 NEAT 設計；NEAT 設計受試者能力估計精準度較優於 BIB 設計。

參、研究方法

本研究係以單向度單參數試題反應理論為基礎，使用模擬資料探討不同等化設計與估計方法對於群體參數的估計成效。以下介紹本研究所使用之測驗等化設計與評估準則。

一、測驗等化設計

本研究為了讓 BIB 與 NEAT 設計的區塊數一樣以進行比較探討，本研究僅針對試題區塊數為 7 的 BIB 與 NEAT 等化設計。

（一）NEAT 等化設計

表 3-1 為低年級 NEAT 設計表，共包含 7 個題本、每題本 3 個區塊、每個區塊包含 12 個試題數，其中 M1 為定錨試題區塊，高年級亦為同樣的設計。

表 3-1 **低年級年級之 NEAT 設計表**

題本序號	區塊（K1）	區塊（K2）	區塊（K3）
S1	M1	M2	M3
S2	M1	M4	M5
S3	M1	M6	M7
題本（b）=3、區塊（t）=7、試題區塊數（k）=3、每個區塊試題數=12			

關於高低年級年級間的等化連結，係由低年級的 M1 區塊中，挑選出難度參數較高的數題做為定錨題，並將之放入高年級的 M1 區塊中，做為高年級 M1 區塊的的試題。排列方式如表 3-2 所示

表 3-2 **NEAT 高低年級垂直等化連結設計表**

低(L)年級		高(H)年級
L-M1(定錨區塊)		H-M1(包含由 L-M1 中的選出的定錨題)
L-M2		H-M2
L-M3		H-M3
	→	
L-M7		H-M7

（二）BIB 等化設計

表 3-3 為低年級 BIB 設計表，共包含 7 個題本、7 個試題區塊，每個題本包含 3 個試題區塊、每一試題區塊在題本中出現的次數為 3 次、以及成對試題區塊在題本中出現的次數只有 1 次。根據 BIB 設計之條件，每個題本中試題區塊的組合不重複。高年級亦為同樣的設計。

表 3-3 **低年級 BIB 等化設計**

題本序號	區塊 (k1)	區塊 (k2)	區塊 (k3)
S1	M1	M2	M4
S2	M2	M3	M5
S3	M3	M4	M6
S4	M4	M5	M7
S5	M5	M6	M1
S6	M6	M7	M2
S7	M7	M1	M3

BIB 垂直等化設計中不同兩年級的試題排列均依照 BIB 設計排列，在定錨題部分是將 H 年級中每個試題區塊中，放入 L 年級對應試題區塊中難度較難的試題，如表 3-4 中，H 年級的試題區塊 1（H-M1）中，包含 L 年級試題區塊 1 內試題難度較難的 g 題（L-M1-1～L-M1-g）定錨試題。

表 3-4 **BIB 高低年級垂直等化連結設計表**

低(L)年級		高(H)年級
L-M1		H-M1(包含由 L-M1 中的選出的定錨題)
L-M2		H-M2(包含由 L-M2 中的選出的定錨題)
L-M3		H-M3(包含由 L-M3 中的選出的定錨題)
	→	
L-M7		H-M7(包含由 L-M7 中的選出的定錨題)

二、模擬條件與估計方法設定

本研究係利用電腦模擬產生作答反應，並重覆進行 50 次的資料模擬後，分

別計算在不同等化設計及不同情境下各種估計方法所估計群體參數的估計誤差。變項設定如表 3-5，並分別說明如下：

表 3-5 **不同等化設計之變項設定**

實驗變項	變項設定
試題長度	每個題本施測題數 18 題及 36 題
受試者群能力分布	如表 3-7
每個年級施測人數	10920 人及 16128 人
試題難度參數分布(b)	截尾常態分布，高年級組：-2～4，低年級組：2～-4
估計方法	PV、PV_W、EAP_AV、EAP、MLE、WLE
等化設計	BIB、NEAT
每一情形模擬資料集個數	50 次

（一）試題長度

本次研究參考郭伯臣和王暄博（2008）之實驗設計，模擬每個題本施測題數為 36 題，因為試題區塊數為 3，故每個試題區塊之試題數 12 題。郭伯臣、王暄博（2008）的研究顯示，定錨題數愈多愈好；但在施測題本題數為 36 題的情形下，定錨題數 6 題和 9 題的精確度差異不大，故本研究設定定錨題數為 6 題。此外，同時模擬施測題數 18 題、定錨題 3 題之情形以做為參照比較，如表 3-6 說明。

表 3-6 **試題長度設定一覽表**

題本題數	試題區塊題數	定錨題數	施測總題數（NEAT）	施測總題數（BIB）
36	12	6	162	154
18	6	3	81	77

（二）受試者能力分布設定

本研究參考郭伯臣和王暄博（2008）的實驗設計，將高低年級群體能力平均數分別設定為 1 及-1，並同時參考 von Davier et al.（2009）設計，分別將高低年級各設定 2 組背景變項，分別假設是社經地位（Socioeconomic Status, SES）類別（高（S-H）低（S-L））及學校類別（（A）類及（B）類）。學校間群體能力平均差異為 0，高底社經地位分別差距群體能力平均各正負 0.707，不同群體之標準差皆為 0.707，而群體標準差為 1.000（0.7072＋0.7072）詳如表 3-7。

表 3-7 **受試者能力分布設定一覽表**

年級	**社經地位**		
	學校類別	**S-L**	**S-H**	**平均**
G-L	A	-1.707(0.707)	-0.293(0.707)	**-1(1.000)**

	B	-1.707(0.707)	-0.293(0.707)	**-1**(**1.000**)
G-H	A	0.293(0.707)	1.707(0.707)	**1**(**1.000**)
	B	0.293(0.707)	1.707(0.707)	**1**(**1.000**)
	平均	**-0.707**(**1.000**)	**0.707**(**1.000**)	**0**(**1.000**)

（三）受試者人數設定

參考 TASA 之受測人數（郭伯臣等人，2012）並配合等化設計要求，分別設定高低年級受試者各 8064 人（總和 16128 人）及各 5460 人（總和 10920 人），每個獨立群體之施測人數為 2016（8064÷4）人與 1365（5460÷4）人。

（四）試題難度參數設定

高年級組試題難度參數設定平均數為 1，低年級組為-1，標準差皆為 1，並採平均數上下 3 個標準差之截尾常態分布，故高年級組之範圍界定於 -2～4，記為 N（1,1）；低年級組之範圍界定於 2～-4，記為 N（-1,1）。

（五）估計方法

本研究主要為探討不同估計方法對於群體參數估計之效果，估計方法分計有單向度納入輔助變數之可能值方法（PV）、加入輔助變數之期望後驗估計法（EAP_AV）、期望後驗估計法（EAP）、最大概似估計法（MLE）以及加權最大概似估計法（WLE）五種。

有關 PV 法估計個體能力的部份係用 5 個可能值平均數來表示受試者個體能力值，此種方法於 von Davier, et al.（2009）的研究中以 PV_W（“W”為 wrong）表示，這是常見的錯誤使用 PV 法，如公式 3-1：

$$\mathbf{PV}_j = \frac{\mathbf{PV}_{j1} + \mathbf{PV}_{j2} + \mathbf{PV}_{j3} + \mathbf{PV}_{j4} + \mathbf{PV}_{j5}}{5} \qquad \text{（公式 3-1）}$$

其中，j 表示受試者，$j = 1,2,3,...,N$ ；

$\mathbf{PV}_{j1} \sim \mathbf{PV}_{j5}$ 為第 j 位受試者抽取的 5 個可能值。

三、研究工具與評估準則

（一）研究工具

本研究使用的工具有 MATLAB 軟體、Acer ConQuest 3.0 軟體，茲分述如下。MATLAB 2009 主要模擬產生受試者的能力、背景變項與題本、試題的難度參數，然後進行模擬作答反應，並計算個體能力與群體參數的估計誤差，Acer ConQuest 3.0 軟體進行能力與試題參數估計。

（二）評估準則

本研究係將模擬產生之受試者能力參數視為真值，模擬 50 筆資料，計算不同等化設計下不同的估計方法之估計誤差。本研究是使用根均方差（root mean square error, RMSE）計算估計誤差誤差， RMSE 較小即表示該情境之估計誤差小，有較好的估計結果；反之則表示估計結果較差。

本研究中分別探討受試者之群體能力平均值與標準差兩個部分。受試者群體能力平均值之 RMSE 如公式 3-2，受試者群體能力標準差之 RMSE 如公式 3-3。

$$\mathbf{RMSE}(\mu,\hat{\mu})=\frac{\sqrt{\sum_{m=1}^{50}(\mu_m-\hat{\mu}_m)^2}}{50} \quad \text{（公式 3-2）}$$

其中 m 表示每一情形模擬資料集個數，$m=1,2,3,...,50$

$\hat{\mu}=(\hat{\mu}_1,\hat{\mu}_2,\hat{\mu}_3,...,\hat{\mu}_m)$：在第 m 個模擬資料集之群體能力平均估計值

$\mu=(\mu_1,\mu_2,\mu_3,...,\mu_m)$：在第 m 個模擬資料集之群體能力平均真值

$$\mathbf{RMSE}(\sigma,\hat{\sigma})=\frac{\sqrt{\sum_{m=1}^{50}(\sigma_m-\hat{\sigma}_m)^2}}{50} \quad \text{（公式 3-3）}$$

$\hat{\sigma}=(\hat{\sigma}_1,\hat{\sigma}_2,\hat{\sigma}_3...,\hat{\sigma}_m)$：在第 m 個模擬資料集之群體能力標準差估計值

$\sigma=(\sigma_1,\sigma_2,\sigma_3...,\sigma_m)$：在第 m 個模擬資料集之群體能力標準差真值

肆、研究結果與討論

本節分為三個部分：第一部份是 NEAT 與 BIB 水平等化設計下，不同的背景變項、不同的題數與人數於不同估計方法之群體參數估計成效。第二部份是 NEAT 與 BIB 垂直等化設計，不同的題數與人數於不同估計方法之群體參數估計成效。第三部分則是二種等化設計間的綜合比較。

一、水平等化群體能力參數估計結果

本研究設計兩個群體（低年級與高年級），為節省篇幅，僅呈現低年級之水平等化設計估計效果，以下呈現不同背景變項於不同等化設計在不同方法之群體參數估計結果。

（一）NEAT 水平等化設計

1.平均數之估計

圖 4-1 顯示社經地位變項的群體能力平均數的估計結果，分別針對不同估計方法、施測人數與題數進行說明。圖 4-1 之橫軸 18_5460_L 表示題數 18 題，人數 5460 人，社經地位是 L 之群體；縱軸是 RMSE，以此類推。圖 4-1 顯示：不同的估計方法方面，納入輔助變數的估計方法（PV、PV_W、EAP_AV）的 RMSE 低於其他的方法（EAP、MLE、WLE），後者是常被用於估計個體能力的估計方法，顯見並不適合用於估計群體能力之參數，特別是本研究設計中，是以低年級的資料探討水平等化設計，其群體平均數是設定為-1，而在 EAP 的估計方法中，對於先驗分佈的假設是標準常態分佈，此種假設與資料型態並不符合，而在 MLE 和 WLE 中並未考慮先驗分佈，此三種方法均造成較大的平均數估計誤差；不同的人數估計方面，隨著人數的增加，RMSE 有降低的趨勢，但不是很明顯，原因可能是本研究的人數設定是以 TASA 之受測人數為參考依據，兩種人數都已是屬於大樣本之設計，故人數的影響不是那麼明顯；不同的題數方面，隨著題數的增加，納入輔助變數的估計方法（PV、PV_W、EAP_AV）的 RMSE 有降低的趨勢；然而，沒有納入輔助變數的估計方法（EAP、MLE、WLE）之 RMSE 並沒有隨著題數的增加而降低。

2.標準差之估計

圖 4-2 顯示社經地位變項的群體能力標準差的估計結果，顯示 PV 的方法對

於回復群體標準差，效果特別的好，優於其他的方法，特別是題數少的情境，結果與 Wu（2005）的研究是一致的。原因可能是 PV 的方法是從受試者的後驗分佈中，隨機抽取 5 個變數，相較於其他方法，相當於使用 5 倍的資料估計群體的分散量數—標準差，故其估計精準度較佳。隨著人數的增加與題數的增加，不同的估計方法的估計誤差（RMSE）也隨著降低。

如以 PV 法和 PV_W 法兩種方法比較，PV_W 是將 PV 法所抽取的 5 個可能值直接加總再平均作為個體能力值再計算群體平均數或標準差，其概念近似於 EAP 法，在估計群體平均數時，兩者（PV 法和 PV_W 法）的誤差差異不大，但在估計標準差時，PV_W 法的誤差高於 PV 法，對於次級資料分析者，常需使用大型測驗之資料進行各項假設檢定，如誤用可能值方法，會造成較大的標準差估計誤差，計算標準誤時恐會高估或低估，而造成錯誤的結果推論。學校變項在水平等化設計之研究結果與社經地位變項類似，為節省篇幅，不再贅述。

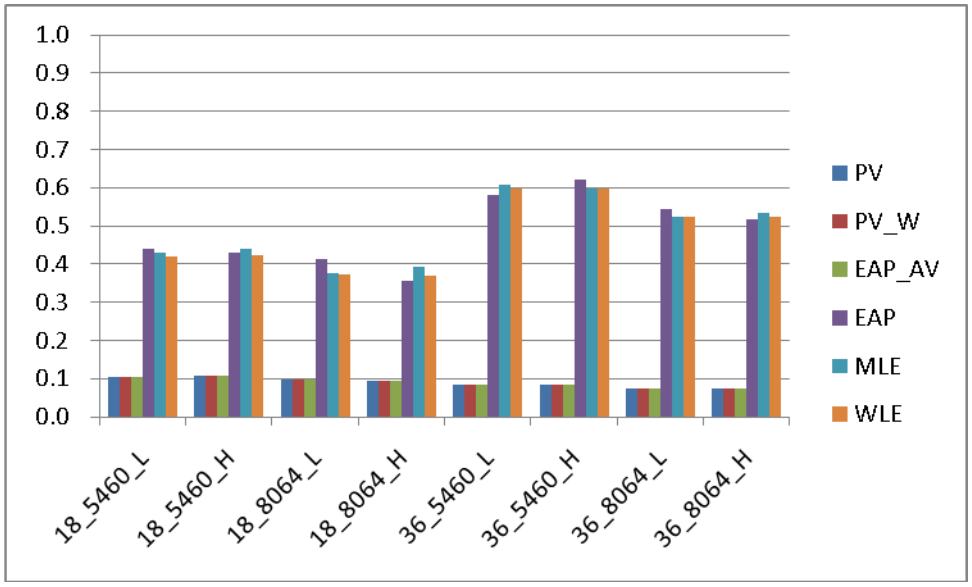

圖 4-1 NEAT 水平等化設計社經地位變項群體平均數不同估計方法結果

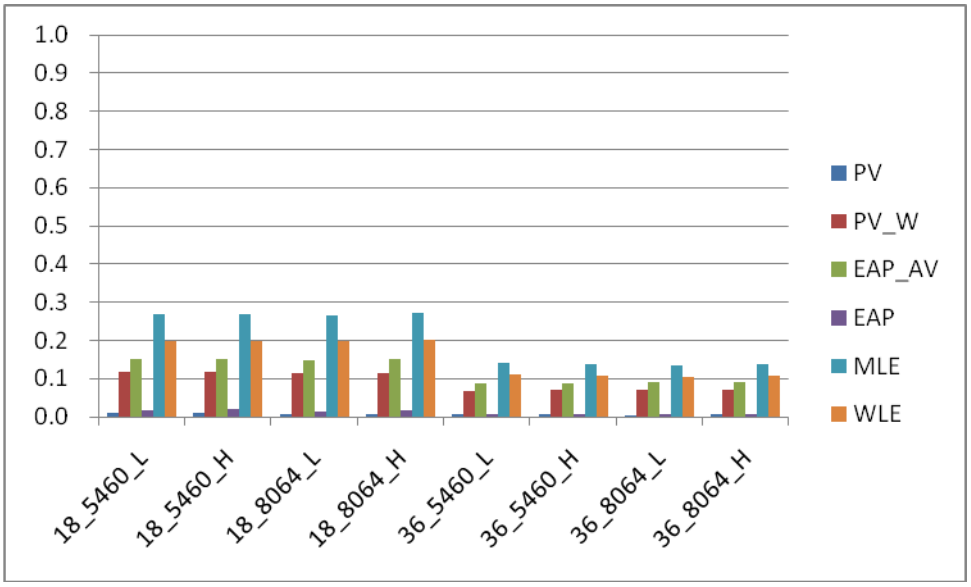

圖 4-2 NEAT 水平等化設計社經地位變項群體標準差不同估計方法結果

（二） BIB 水平等化設計

1.平均數之估計

圖 4-3 顯示社經地位變項的群體能力平均數的估計結果，不同的估計方法方面，納入輔助變數的估計方法（PV、PV_W、EAP_AV）的 RMSE 低於其他的方法（EAP、MLE、WLE），此結果與 NEAT 水平等化設計之結果相似，如前所述，可能的原因為估計群體參數時，先驗分佈對於群體參數是一個相當重要之影響因素。不同的人數估計方面，隨著人數的增加，RMSE 有降低的趨勢，但亦不是很明顯，可能是本研究的人數設定是以 TASA 之受測人數為參考依據，兩種人數都已是屬於大樣本之設計，故人數的影響不是那麼明顯；不同的題數方面，隨著題數的增加，納入輔助變數的估計方法（PV、PV_W、EAP_AV）的 RMSE 有降低的趨勢；然而，沒有納入輔助變數的估計方法（EAP、MLE、WLE）之 RMSE 並沒有隨著題數的增加而降低，此種結果跟 NEAT 水平等化設計類似。

2.標準差之估計

圖 4-4 顯示社經地位變項的群體能力標準差的估計結果，顯示 PV 的方法對於回復群體標準差，效果特別的好，優於其他的方法，特別是題數少的情境，此結果與 Wu（2005）的研究結果一致。隨著人數的增加與題數的增加，不同的估計方法的估計誤差（RMSE）也隨著降低。雖然 EAP 的方法於 BIB 設計對於回覆群體標準差有良好的估計成效，然而對於回覆群體平均數確有較高的 RMSE。學校變項在 BIB 水平等化設計之研究結果與社經地位變項類似。

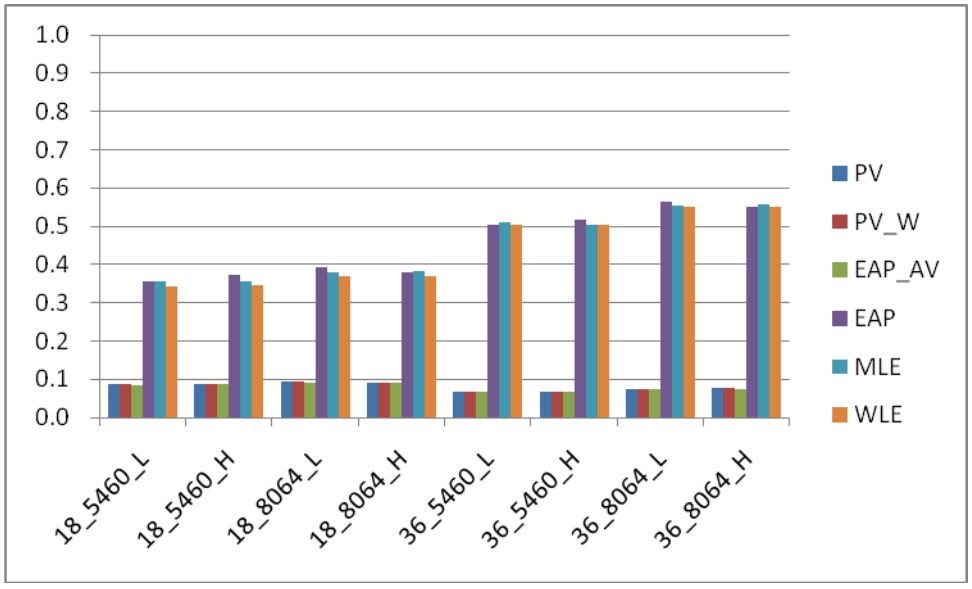

圖 4-3 BIB 水平等化設計社經地位變項群體平均數不同估計方法結果

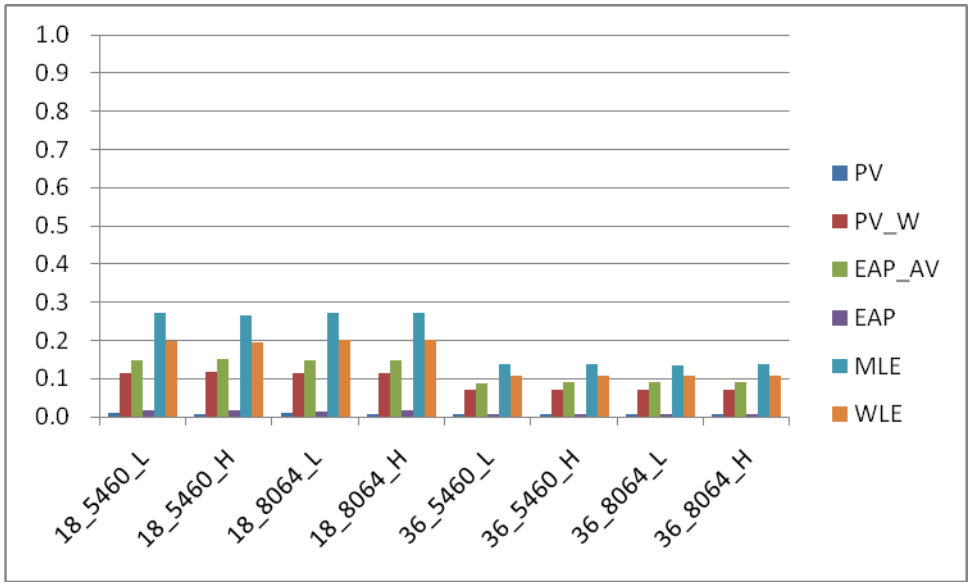

圖 4-4 BIB 水平等化設計社經地位變項群體標準差不同估計方法結果

二、垂直等化群體能力參數估計結果

本研究設計兩個能力之群體：低年級（GL）與高年級（GH）。探討不同方法於於垂直等化設計下，群體能力平均數與標準差估計結果，以下呈現不同年級不同等化設計在不同方法之群體參數估計結果，圖 4-5 之橫軸 18_10920_GL 表示題數 18 題，人數 10920 人，低年級之群體；縱軸是 RMSE，以此類推。

（一）NEAT 垂直等化設計

1.不同估計方法間之比較

群體能力平均數與標準差在不同情境中估計結果之 RMSE 如圖 4-5 及圖 4-6 所示。就群體能力平均數的部份來說，其中 PV、PV_W 及 EAP_AV 三種方法的 RMSE 低於 EAP、MLE 及 WLE 三種估計方法；不管是高年級還是低年級，PV、PV_W 及 EAP_AV 方法估計群體能力平均數的 RMSE 皆低於 EAP、MLE、WLE 等三種估計方法。在 PV、PV_W 及 EAP_AV 這部份而言，這三種估計法的差距非常小，最高及最低差距僅有 0.004。就 EAP、MLE、WLE 這三種估計方法而言，三種估計法的 RMSE 的差異不大。以不同年級層面而言，在群體平均數之估計，以 PV、PV_W 及 EAP_AV 方法估計低年級群體平均數的 RMSE 略高於高年級，但若以 EAP、MLE、WLE 方法估計低年級和高年級的群體平均數，兩者的 RMSE 差異是相當大的。在群體標準差之估計，則無呈現此種現象，不同的估計方法估計低年級和高年級的標準差之 RMSE 差異不大。以群體能力標準差的估計情形來看， PV 法的 RMSE 遠低於各種估計法，表示 PV 的估計效果最好，而在 EAP、MLE、WLE 這三種估計方法的部份，EAP 的估計效果最佳，MLE 的估計效果則較差。

2.不同施測人數間之比較

就群體能力平均數的估計結果而言，首先，在施測題數為 18 題的情境中，就 PV、PV_W 及 EAP_AV 此三種估計方法而言，施測人數 10920 的群體能力平均數的 RMSE 略低於 16128 人的情境，但差異不大；而在 EAP、MLE、WLE 等另外三種方法在估計群體能力平均數方面，就低年級群組方面，此 3 種估計方法的 RMSE 亦是施測人數 10920 的群體能力平均數的 RMSE 略低於 16128 人的情

境，但是在高年級組部份，則是 16128 人的情境的 RMSE 略低，但差異皆不大。

在施測題數為 36 題的部份，六種估計方法的 RMSE 皆是施測人數 10920 人的情境略低於 16128 人的情境，除了在低年級群組中 EAP、MLE、WLE 此三種方法的 RMSE 在 2 種施測人數情境間差異較大外，其餘情境差異不大。

就群體能力標準差的估計結果而言，如圖 4-6 所示，PV 的估計結果遠優於其它五種估計方法。而不管是在題數為 18 題或是 36 題的情境中，不同施測人數間相對之各種估計法的 RMSE 均無明顯差異。

3.不同施測題數間之比較

在群體能力平均數的 RMSE 部份，如圖 4-5 所示，就 PV、PV_W 及 EAP_AV 三種估計方法而言，不管是 10920 人的情境抑或是 16128 人的情境下，36 題的施測情境的 RMSE 皆低於 18 題的施測情境。而對於 EAP、MLE、WLE 等三種估計方法的 RMSE 而言，除了在 10920 人低年級群組是 36 題情境低於 18 題情境外，其餘情境的 RMSE 則是 18 題情境低於 36 題情境。

就群體能力標準差而言，不管是在 10920 人的情境還是 16128 人的情境中，五種估計方法的 RMSE 都是 36 題的的情境低於 18 題的情境。

在 NEAT 等化設計下，群體能力平均數估計在本研究的不同施測人數情境中的估計結果相差不大。而在不同施測數之比較下，則 PV、PV_W 及 EAP_AV 這三種納入背景變項的估計法皆隨著受試題數的增加而估計愈精準。從不同估計方法間的比較來看，PV、PV_W 及 EAP_AV 這三種納入背景變項的估計法的估計效果優於其它三種未納入背景變項的估計法。

就群體能力標準差而言，不同施測人數情境下的 RMSE 均無明顯差異；隨著受試題數的增加，則估計愈精準，從不同估計方法間的比較來看，PV 的估計效果優於各種估計方法。此外，就 PV 及 PV_W 這二種估計方法而言，在群體能力平均數的估計上，PV 及 PV_W 幾乎沒有任何差別，而在群體能力標準差的估計上，PV_W 的 RMSE 高於 PV，表示 PV_W 在群體能力標準差的估計上產生了偏誤，此一研究結果與 von Davier 等人（2009）的研究結果相近。

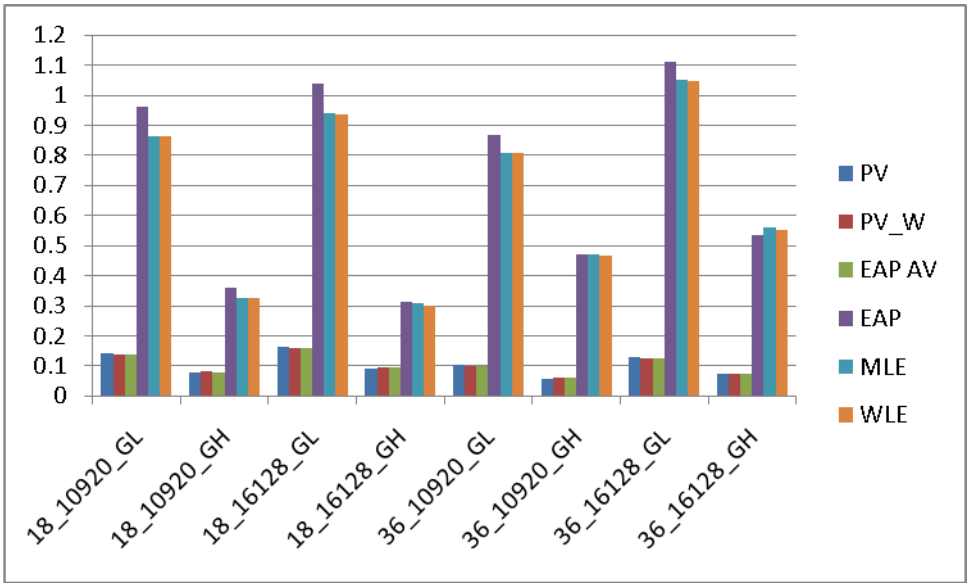

圖 4-5 NEAT 垂直等化設計群體能力平均數不同估計方法之結果

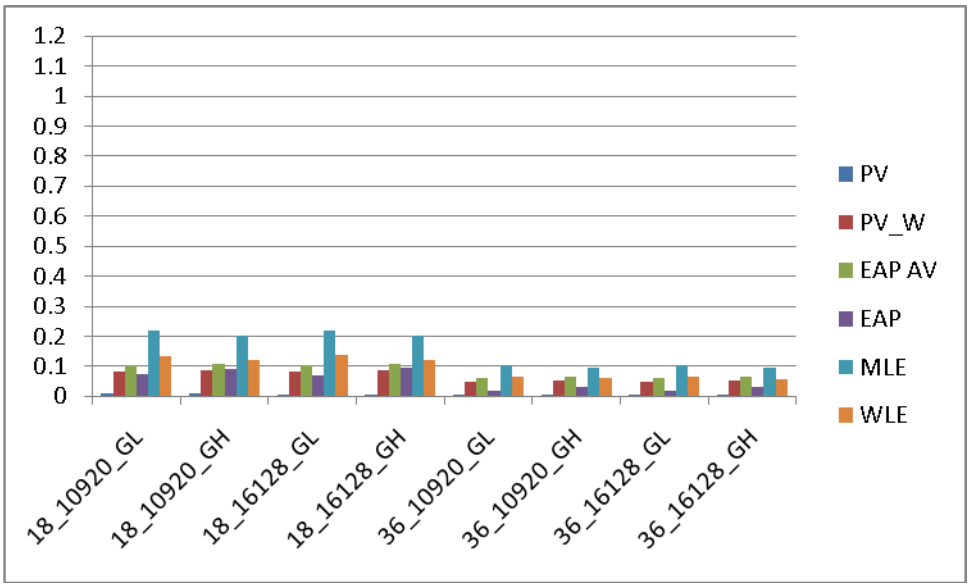

圖 4-6 NEAT 垂直等化設計群體能力標準差不同估計方法之結果

（二）BIB 垂直等化設計

圖 4-7 結果顯示：在 BIB 等化設計下，群體能力平均數估計在本研究的不同施測人數情境中的估計結果相差不大。而在不同施測題數比較下，則 PV、PV_W 及 EAP_AV 這三種納入背景變項的估計法隨著受試題數的增加而估計愈精準，從不同估計方法間的比較來看，PV、PV_W 及 EAP_AV 這三種納入背景變項的估計法的估計效果優於其它三種未納入背景變項的估計法。

圖 4-8 顯示：群體能力標準差而言，同樣是不同施測人數情境下的 RMSE 均無明顯差異；隨著受試題數的增加，則估計愈精準，從不同估計方法間的比較來看，PV 的估計效果優於各種估計方法。此外，就 PV 及 PV_W 這二種估計方法而言，在群體能力平均數的估計上，PV 及 PV_W 幾乎沒有任何差別，在群體能力標準差的估計上，PV_W 的 RMSE 遠高於 PV，表示 PV_W 在群體能力標準差的估計上產生了偏誤，此一研究結果與 von Davier 等人（2009）的研究結果相近。

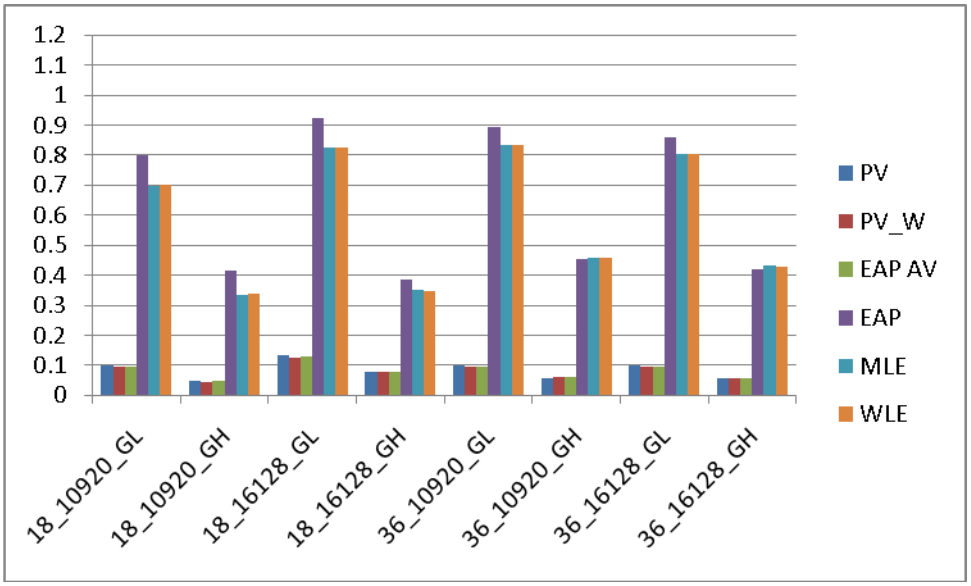

圖 4-7 BIB 垂直等化設計群體能力平均數不同估計方法之結果

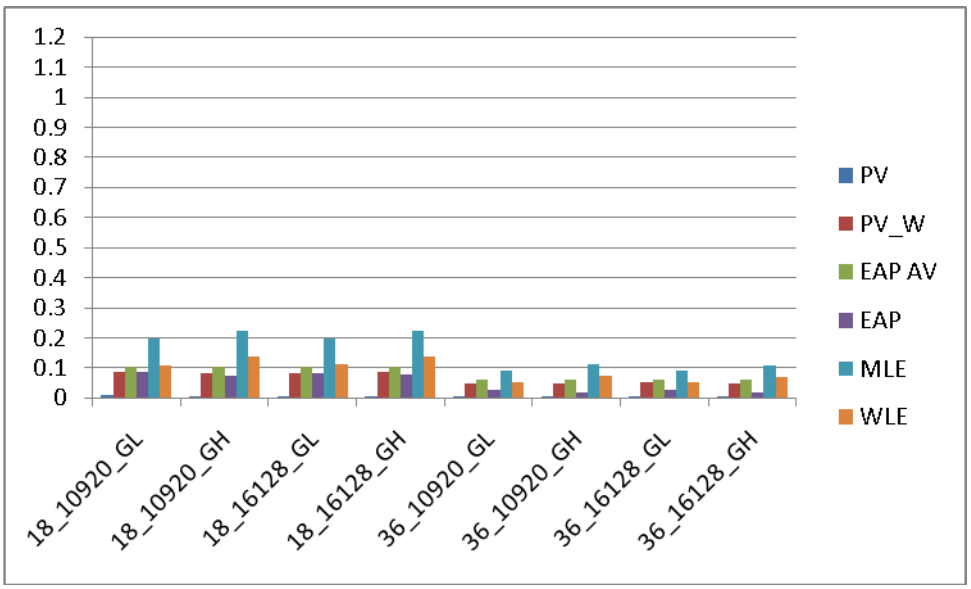

圖 4-8 BIB 垂直等化設計群體能力標準差不同估計方法之結果

三、等化設計方法之比較

本研究使用兩種不同的等化設計，NEAT 和 BIB 等化設計，以下探討不同方法於不同垂直等化設計下，群體能力平均數與群體能力標準差參數估計結果，圖 4-9 之橫軸 18_10920_GL 表示題數 18 題，人數 10920 人，低年級之群體；縱軸是 RMSE，為縮小圖片大小以節省篇幅，縱軸量尺並未統一。

1.平均數之估計

如圖 4-9 所示，在 PV、PV_W 及 EAP_AV 中，NEAT 設計的 RMSE 皆比 BIB 設計的為高，而在 EAP、MLE 及 WLE 三種估計方法中，除了 10920 人－18 題－高年級、16128 人－36 題－高年級及 10920－36 題－低年級這 3 個群組是 NEAT

設計低於 BIB 設計外，其它皆是 NEAT 設計高於 BIB 設計。

2.標準差之估計

如圖 4-10 所示，在 PV 及 EAP_AV 中，NEAT 設計和 BIB 設計的 RMSE 值雖然是高低互見，但兩者之間的差異非常小，可以說是非常接近。而在 EAP 法中，所有低年級群體的情境中皆是 NEAT 設計低於 BIB 設計；而所有高年級群體則是 NEAT 設計高於 BIB 設計。在 MLE 及 WLE 中，則是所有低年級群體的情境中皆是 NEAT 設計高於 BIB 設計；而所有高年級群體則是 NEAT 設計低於 BIB 設計。

群體能力平均數估計上，在 PV、PV_W 及 EAP_AV 這三種納入背景變項的估計法上，也是 BIB 設計的估計結果優於 NEAT 設計。而在群體能力標準差的估計上，在 PV 及 EAP_AV 中，NEAT 設計和 BIB 設計的 RMSE 的差異非常小，可說是完全一致。

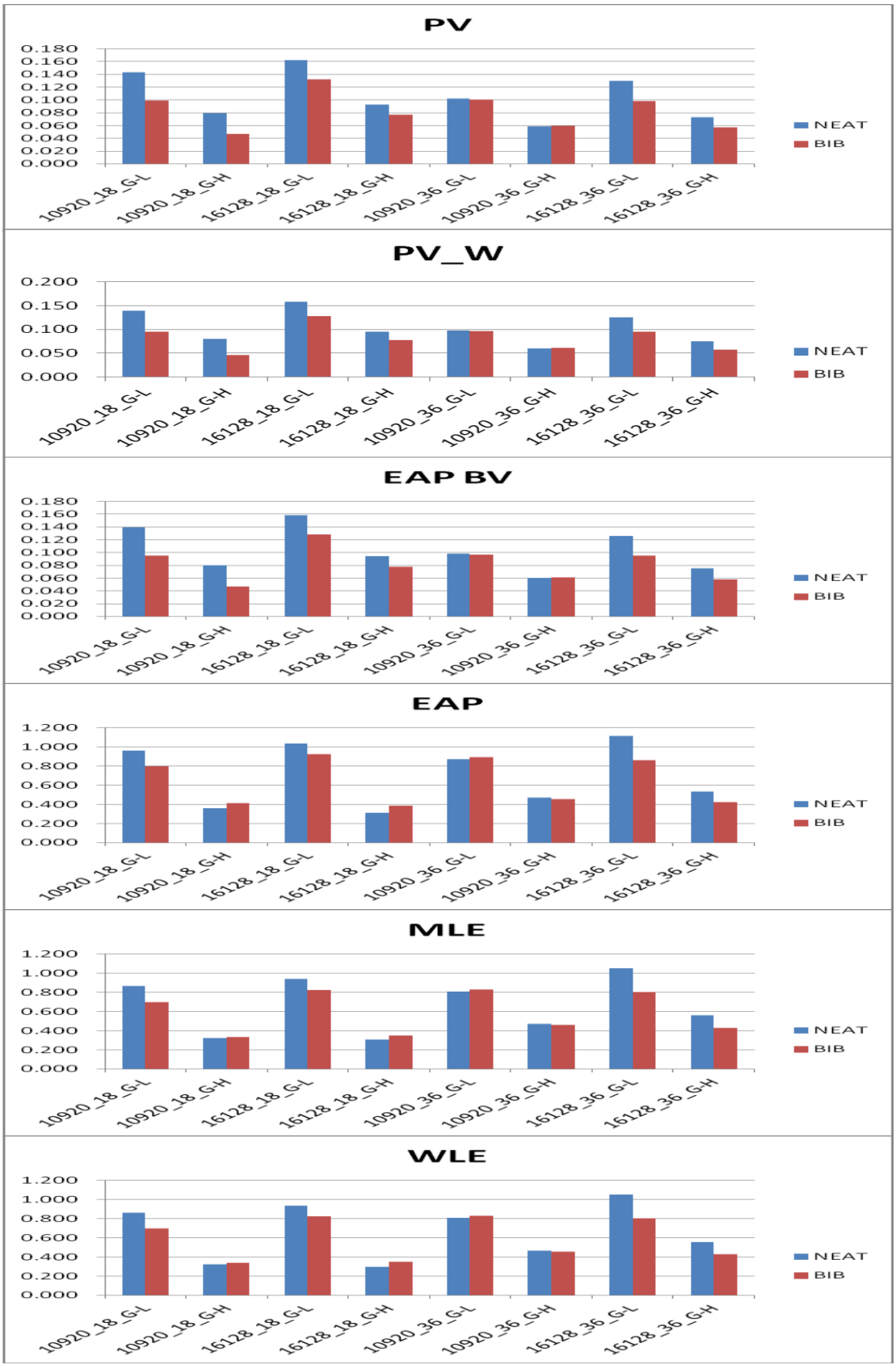

圖 4-9 不同等化設計群體能力平均數各種估計方法之結果

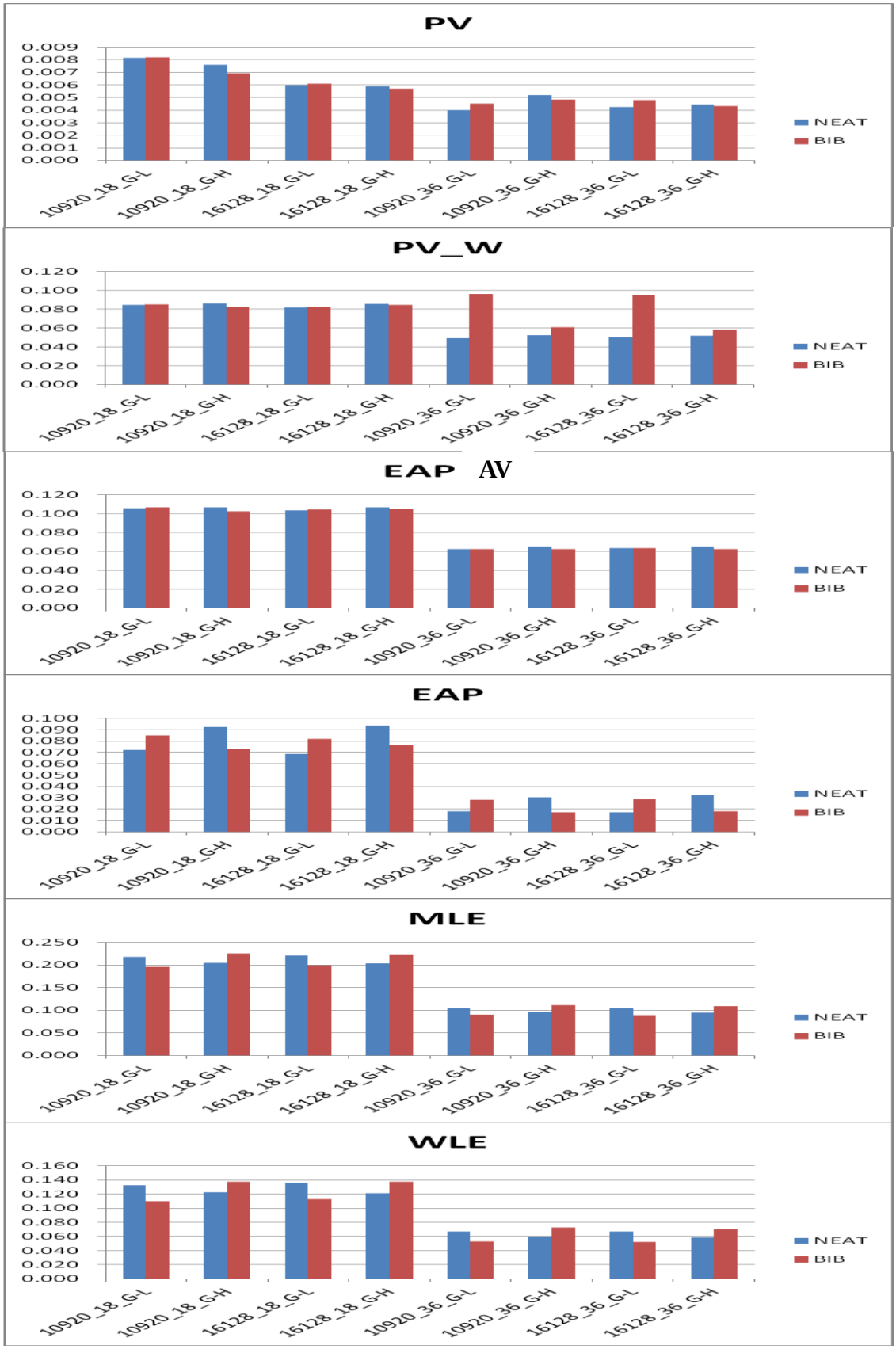

圖 4-10 不同等化設計群體能力標準差各種估計方法之結果

伍、 結論與建議

本研究是以模擬實驗探討在水平和垂直等化設計，不同估計方法及等化設計對於群體能力參數的估計效果。以下分別估計方法、施測人數、施測題數、及等化設計等得出如下結論：

一、納入背景變項的估計方法估計效果較佳

在群體能力平均數估計或是群體能力標準差估計上，不論是水平等化或是垂直等化，PV、PV_W 及 EAP_AV 這三種納入背景變項的估計方法的估計效果都比 EAP、MLE、WLE 等三種未納入背景變項的估計方法來的好，而所納入背景變項的平均數高或低對於估計的成效影響並不大；對於群體能力標準差的估計，PV 的估計效果遠優於各種估計方法。

二、本次實驗情境中施測人數多寡不影響估計結果

對於群體能力平均數估計或是群體能力標準差估計上，在本次的實驗情境中，不同施測人數情境下的 RMSE 均無明顯差異，可能是本研究所設定的人數已是屬於大樣本，無法顯示樣本大小之因素對於估計結果所造成之影響。

三、施測題數愈多估計愈精準

群體能力平均數估計或是群體能力標準差估計上，在本次的實驗情境中，有納入背景變項之估計方法，皆隨著試題數的增加而 RMSE 降低，表示施測題數愈多估計愈精準。

四、BIB 垂直等化設計的估計效果略優 NEAT 垂直等化設計

在群體能力平均數的估計上，就納入背景變項的二種估計法上，皆是 BIB 設計的估計結果優於 NEAT 設計。而在群體能力標準差的估計上，在 PV 及 EAP_AV 中，NEAT 設計和 BIB 設計的 RMSE 的差異非常小，可說是完全一致，整體而言，BIB 垂直等化設計的估計效果略優於 NEAT 垂直等化設計。

五、誤錯的 PV 使用方法將導致估計偏誤

在群體能力平均數的估計上，PV 及 PV_W 幾乎沒有任何差別，而在群體能力標準差的估計上，PV_W 在群體能力標準差的估計上產生了偏誤。

本研究結果顯示：PV 法及 EAP_AV 法因納入了背景變項做為輔助變數計算，使得不管是群體能力平均數或是標準差，在各種情形下之估計結果皆比其它 3 種未納入背景變項的估計方法為佳。茲就本研究結果提出下列研究建議及未來研究建議。

一、群體參數之估計宜考慮納入背景變項之分析方法

納入背景變項的估計法對於群體能力參數的估計，優於傳統的點估計方法；尤其是 PV 法在群體能力標準差的估計上，更是遠優於各種估計法。建議大型測驗要進行群體參數之估計，不論是垂直或水平等化，應使用納入背景變項的估計法進行。

二、增加施測之題數

本研究結果顯示，對於納入背景變項之估計方法，施測題數愈多、估計愈精準，建議在大型測驗在進行等化設計時，能在適度的範圍內增加施測題數，以增加估計精確度。

三、正確使用可能值方法進行次級資料分析

本研究發現，錯誤的 PV 使用方式，在垂直等化設計上亦會造成估計上的嚴重偏誤，是以在使用 PV 進行次級資料分析時，必須依據正確方式來使用，否則

將造成估計上的偏誤。

四、未來研究建議

本研究結果顯示，BIB 設計估計結果優於 NEAT 設計，可能是因為 BIB 垂直等化設計係由低年級組的每個試題區塊中各挑選出難度最高的題目做為高低年級間的定錨題。而 NEAT 垂直等化設計卻是完全由低年級組中的定錨試題區塊直接挑選出難度最高的數題做為高低年級的定錨題，建議以後可設計不同的等化情境進行更深入的研究

參考文獻

中文部分

洪碧霞、林素微、林娟如（2006）。認知複雜度分析架構對 TASA-MAT 六年級線上測驗試題難度的解釋力。**教育研究與發展期刊，2**（4），69-86。

郭伯臣、王暄博 （2008）。大型測驗中同時進行垂直與水平等化效果之探討。**教育研究與發展期刊，4**（4），87-120。

郭伯臣、曾建銘、吳慧珉主編（2012）。**大型標準化測驗建置流程應用於 TASA 之研究**。新北市：國家教育研究院。

英文部分

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Allen, N. L., Carlson, J. E., Johnson, E. G., & Mislevy, R. J. (1999). *The NAEP 1998 technical report.* Educational Testing Service.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report.* Washington, DC: National Center for Educational Statistics.

Andrew, R. W. & Terry, L. S., (2001). *The NAEP 1998 Technical Report (NCES 2001-509).* National Assessment Governing Board, U.S. Department of Education.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques*. Basel, N. Y.: Marcel Dekker, Inc.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* New York: Chapman & Hall. (Distributed by Halsted Press, New York)

De la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, 33, 465-485.

Foy, P., Galia, J., & Li, L. (2008). Scaling the data from the TIMSS 2007 Mathematics and Science assessments. In John F. Olson, Michael O. Martin ,Ina V.S. Mullis. (Eds). *TIMSS 2007 Technical Report.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. Studies in Educational Evaluation, 35, 83–88.

Graham J. R., Christine, Y. O'S., Alka, A., & Ebru, E. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Lee, J., Grigg, W., & Dion, G. (2007). *The Nation's Report Card: Mathematics 2007*.

National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education, Washington, D. C.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report.* TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mislevy, R. J. (1991). Randomization-based inference about latent variable from complex samples. *Psychometrika, 56*(2), 177-196.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

Mislevy, R. J., & Sheehan, K. M. (1989). Information matrices in latent-variable models. *Journal of Educational Statistics, 14*, 335-350.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics form sparse matrix samples of item response. *Journal of Educational Measurement, 29,* 133-161.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 International Mathematics Report.* TIMSS & PIRLS International Study Center, Lynch School of Education,Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff C. (2009). *TIMSS 2011 Assessment Framework.* Finding from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

OECD (2005). *PISA 2003 Technical Report*. OCED, Paris.

OECD (2009). *PISA 2006 Technical Report*. OCED, Paris.

Rasch, G. (1960). *Probabilistic models for some Intelligence and attainment tests.* Chicago: University of Chicago Press.

Tianyou, W. (2005). An Alternative Continuization Method to the Kernel Method in von Davier, Holland and Thayer's (2004) Test Equating Framework.

von Davier M., Gonzalez, E., & Mislevy, R. J. (2009).What are plausible values and why are they useful? *IERA Monograph Series：Issues and Methodologies in Large-Scale Assessment, 2*, 9-36.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika , 54*, 427–450

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31* (2-3), 114-128.