教室中的成績等化食譜

➢ 林世華 / 謝佩蓉

教室中的成績等化食譜

林世華
國立臺灣師範大學副教授

謝佩蓉
國家教育研究院助理研究員

壹、緒論

常言道，總結性評量和形成性評量兩者不可偏廢。形成性評量（formative evaluation）意指評估學生於教學過程的學習進展，可視為教學過程的一部分，讓教師能調整其教學；其目的，乃在於協助教師與學生達成學習目標（Miller, Linn, & Gronlund, 2013）。而若要能適當地評估學生的學習進展，適切地運用某些統計方法是不可或缺的。本文透過淺顯的文字，簡單地介紹Rasch模式與等化設計的重要概念，接著便以實例示範，帶領讀者一步步操作測驗成績等化，使1至12年級的教師們能按圖索驥、掌握要訣，輕鬆連結學科測驗成績，記錄學生的學習進展。

一、具備物理學常用的測量量尺

量尺的議題，常常被簡化為「單位」的議題，二者也常被混用。不同的量尺，在物理學上就是不同的單位，例如：公分、吋，是長度的單位，公斤、磅，則是重量的單位。因為採用的量尺不同，使得公制和英制測量結果的數值不同，然而因為關係明確，彼此之間的關係是可以換算的，進行比較是沒有問題的。此外，由於量尺確立，我們對於體重數值所呈現的意義，很快就能有概念。譬如，四年級的小孩80公斤，就是太胖啦。

我們在對學生施測時，量尺會出現問題。期中考國語文和期末考國語文都是國語文測驗，施測後都產生一個數量，似乎隱含「分」這個單位，例如：期中考卷面分數88分，期末考卷面分數87分，然而此「分」非彼「分」，「分」這個單位並不具備物理學的特性。

$$\Pr(X_{ij}=1;\theta_i,b_j)=\frac{e^{\theta_i-b_j}}{1+e^{\theta_i-b_j}} \qquad \text{（公式1）}$$

丹麥數學家Rasch於1960年提出一個數學模式（公式1），而利用這個公式所發展出來的分析模式，稱為「Rasch模式」（Andersen & Olsen, 2001）。它具有很多良好的數學特性，也因此可以應用的情境很廣，其中一項便是適合用來作為發展性的測量。

暫時不看指數的話，分子和分母分別是 $\frac{e}{1+e}$；公式操弄的重點在於指數部分，也就是希臘字 θ（讀音：theta）與英文字 b。

θ　就如同學生的分數，b 是題目的難度，就好像是這題有百分之多少的人答對它。

θ_i　表示第 i 個學生的能力，每位學生的能力都不一樣。

b_j　表示第 j 個題目的難度，各題有各題的難度。

e　是自然底數，它是一個常數，大約等於 2.71828。

當學生完成一份測驗，我們會關心兩個向度：孩子考幾分？，另一面就是這一題被多少學生答對？我們會關心學生考幾分，是因為分數的背後代表了某種能力。這一題被多少學生答對，則是想知道這些測驗題項的難易程度。傳統的做法是人的能力與題目難度，單位分開建造，人的能力是「採取百分制的幾分」，題的難度則是「答對人數百分比」。既然公式1將學生能力值與題目難度放在同一個數學式中，而且能力值如果要能減難度值，表示能力與難度的單位相同（公斤和公分不能相減），意即「人的能力」與「題目難度」是同一量尺。

經由公式1所算出來的值，是個介於0至1的數值，用來表示：某個人帶著他的某項能力（例如：數學能力），作答一份測驗之中的某一題，答對的機率有多大？如果 $\theta_i - b_j > 0$，則答對這一題的機率大於 0.5；如果 $\theta_i - b_j = 0$，表示答對這一題的機率等於 0.5。用這個方式來描寫，當某位學生的能力越來越高，他答對某一題的機率是越來越大的；若是某位學生的能力非常低，則他答對某一題的機率就非常小。如要將學生區隔高下，最有用的試題是答對機率 0.5 的情況。

要注意的是，這個公式只適用於作答結果為「對」或「錯」兩種狀態的試題（例如：選擇題、是非題、填充題），其他給分方式試題的答對機率，要用其他的公式計算。關於 Rasch 模式更詳盡的介紹，可參閱本書第三章測驗理論與測驗分析技術。

二、生活中的等化情境

「等化」這個名詞，聽來陌生。事實上，你可能沒想過，「等化」和我們的生活息息相關。在正式進入等化議題之前，讓我們先來看兩個例子。

「物價高漲、薪水不漲」是近年社會上的熱門議題。根據行政院主計總處（2013）的統計指出，2000年工業及服務業的平均薪資為41,861元，2013年為47,881元，增加幅度很小。然而，2000年你可以用542元買到一桶家用液化石油氣，2013年

卻得花871元才能買到一桶（經濟部能源局，2013），是1.6倍的價格。臺北市信義區的房價變動就更大了，2000年的時候公寓每坪23.2萬元，2013年要價61.08萬元（內政部地政司，2013），幾乎是三倍價。這些現象告訴我們，數字的大小，其背後意義各不相同；2000年拿100元可以買到的東西，比2013年多！這個意思是說，2000年的物品價格和2013年的物品價格價是不能直接比較的。因此，經濟學家會透過公式，算出「消費者物價指數」來比較一般消費者在兩個不同時期購買相同商品組合所需付出的成本。

第二個例子和貨幣有關。出國旅遊前，我們常到金融行庫兌換當地貨幣。究竟手中的新臺幣可以換到多少當地貨幣呢？倚賴「匯率」決定。例如，要到香港旅遊，在不考慮手續費的情況下，每3.85元新臺幣可以換到1元港幣（表1）。但若是要到澳門旅遊，澳門幣並非國際貨幣，無法在臺灣直接兌換。這時我們就可以透過港幣作為新臺幣和澳門幣的共同量尺，將換得的1元港幣再換為1.0306元澳門幣（表1）。也就是說，有了港幣這個量尺基礎，我們便可知道澳門幣比新臺幣的幣值大。這樣的轉換也可以應用在測驗分數上，如果我們知道乙校期中考80分的學生能力，相當於甲校期中考100分的學生能力；又知道丙校期中考90分的學生能力，相當於甲校期中考100分的學生能力，那我們就可以知道甲校期中考最簡單，丙校次之，乙校最難。

值得注意的是，消費者物價指數的誤差甚大，而貨幣轉換是受經濟供需的法則所支配，亦受眾多因素所影響。同樣地，我們在進行測驗分數等化之際，也得留意其中所包含的誤差和可能的影響因素。

表1　匯率牌價範例

港幣	新臺幣	澳門幣
1	3.85	
1		1.0306

三、等化的意義、設計以及限制

等化（equating）的定義是：對同一群學生而言，一份新測驗的分數和一份舊測驗的分數，用來代表其中某位學生的相對位置時，兩者是等值的（Livingston, 2004）。其目的是，透過統計對於測驗分數進行轉換，以校準不同測驗間的難度（Kolen & Brennan, 2004）。等化並沒有辦法校準內容，純粹就「難度」這個議題做校準。等化可依其應用情境分為水平等化（horizontal equating）、垂直等化（vertical equating）以及分數連結（score linking）（van der Linden, 2000）。托福網路測驗（TOEFL iBT）讓世界各地的考生在不同時間點作答不同的測驗卷，成績仍然可以相互比較，便是水平等化的一例。Rasch當年在發展模式時，曾經檢驗某個世代學生閱讀能力的進展，該種設計便需採用垂直等化，連結學生在不同年齡所測量之相同構

念（閱讀表現）的分數（Peter, Cieza, & Geyh, 2013）。分數連結的典型例子為ACT和SAT兩種美國大學常採用的入學考試成績之間的連結，它們是兩種不同的測量工具，但彼此需要有分數之間的對應以便用於入學申請（Dorans, 1999）。

等化的設計和方法有很多種，本文介紹的是「共同題等化設計」（common-item equating），也就是新編測驗（new form）的題項之中，包含了一組參照測驗（reference form）之中的試題，而這些重複使用的試題便被稱為「共同題」。通常，我們會將第一份卷作為參照測驗，例如：期初考、期中考，並從中搬一些題目到下

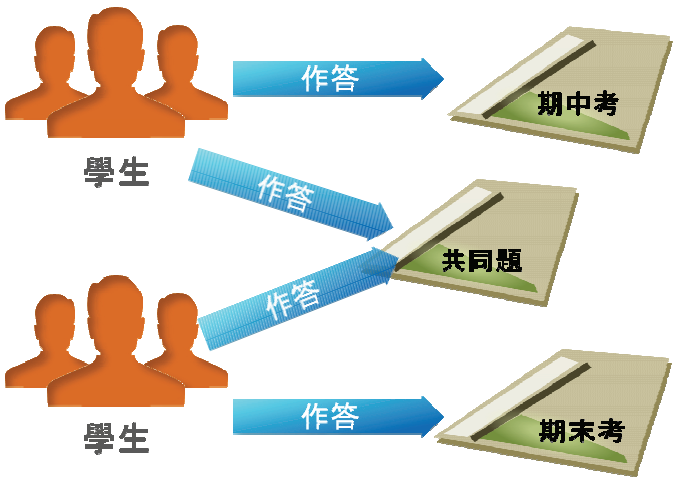


圖1　共同題等化設計示意

一次的考試，以檢視學生於教學之後的進展（圖1）。

共同題等化是國內外大型評量採用的設計，像國際數學與科學教育成就趨勢調查（Trends for International Mathematics and Science Study, TIMSS）、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）等，也被應用於校園之中（Baghaei & Amrahi, 2011; Meyer & Zhu, 2013; Stewart & Gibson, 2010）。共同題等化共同題等化設計的好處是，較其他等化設計更有彈性，可適用於各種情境。有的等化設計，同一位學生需要於施測日作答兩份試卷；而共同題等化設計，同一位學生只需於施測日作答一份試卷。最重要的優點是，除了共同題內容不能曝光，其餘的非共同題都可以公告，讓老師和學生能於考試結束後一同檢討試題內容，符合教室中的實務需求。

可想而知，共同題等化設計成敗的關鍵點，在於「共同題」。共同題的試題品質要好，不能太難也不要太簡單。共同題要有內容代表性，最好是該份測驗所涵蓋內容的迷你版。一般建議，一份40題左右的測驗，共同題的題數至少應該占測驗題數的20%（Kolen & Brennan, 2004），亦即40題之中至少應含8題為共同題，但可以更

多。

凡有測量必有誤差，凡有等化當然也有誤差。等化技術存在著許多限制，是我們在使用這項技術時需要格外謹記在心的。首先，等化無法精準地校正每一位受測學生的個別差異。假如A學生恰好對於參照測驗的題項格外精熟，而B生對於新編測驗的題項格外精熟，那麼對這兩位學生而言，兩份測驗的難度很可能會恰好相同。因此，等化較為適用於校正群體的測驗分數。此外，若我們以卷面分數進行等化，再轉換為量尺分數，數值常常不是整數，便衍生「進位誤差」，而此類誤差會隨著學生的卷面分數離散程度上升而增加（Livingston, 2004）。

垂直等化的誤差來源就更多了。垂直等化意欲比較學生在不同時間點的成長進展，題項難度的設計是否真能和學生成長相符，是一項很大的挑戰。再者，同一群學生於不同年級接受測驗時，所經歷的題項本質和評量程序都可能會產生變化，使得垂直等化可能混雜了題項內容與評量方式的變化，準確估計試題難度的挑戰更大（Lissitz & Huynh, 2003）。

貳、實例操作示範

以下透過實例操作過程，說明如何設計期中考與期末考卷，並以固定試題參數量尺化（fixed common item parameter calibration, FCIP calibration）方法，進行兩次考試分數之連結校準。

一、成績等化設計

某科目修課學生數為58人。期中考題數35題，題型為選擇題，期末考題數50題，亦為選擇題。每題均含4至5個選項，其中只有一個選項是正確答案。測驗後，並未將試卷發回給學生，也沒有公告試題。為連結期中考與期末考測驗分數，兩次測驗之間有7題共同題，由期中考試題之中挑選而來。鑒於共同題挑選原則「試題品質要好，不能太難也不要太簡單。內容要具代表性，最好是該份測驗所涵蓋內容的迷你版」，因而此例7題共同題分別來自期中考所涵蓋的7個章節。教師於學期初便告知學生，期末考的測驗範圍部分涵蓋期中考的範圍，讓學生能事先有所準備。

二、成績等化材料

1. 期中考逐題作答反應[1] 與試題正確答案
2. 期末考逐題作答反應與試題正確答案
3. 期中考與期末考共同題對照表與共同題參數檔

[1] 本文實例所使用之期中考與期末考逐題作答反應下載網址：https://www.dropbox.com/s/nnnlsdyany2bb1z/1011.zip

4. 試題分析軟體 ACER ConQuest[2]
5. 試算分析軟體 Microsoft Office Excel

三、成績等化做法

（一）期中考學生能力值與試題難度分析

1. 整理期中考逐題作答反應

將每位學生期中考的逐題作答反應鍵入電腦，並整理存檔為ACER ConQuest所需的檔案格式；也就是資料與資料之間緊密相連、沒有空白。檔案內容由左至右分別是：學生座號（共四碼）、第1題至第35題的逐題作答反應（圖2）。整理好之後，

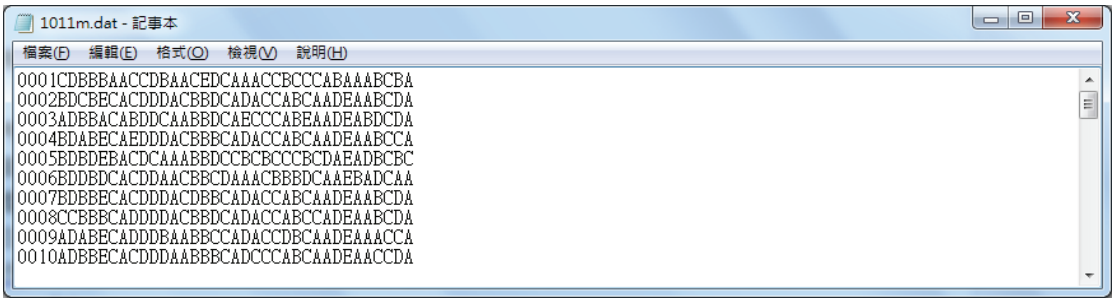
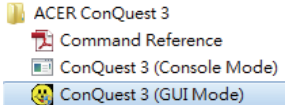


圖2　期中考資料格式實例

將檔案存檔並命名；此例，我們將之命名為「1011m.dat」。

2. 準備試題分析指令檔



於程式集中點選執行 。點選「開新檔案」（圖3）後，於左側 Input Window 輸入 ACER ConQuest所使用的指令檔（圖4），並存檔為「D:\1011m.CQC」。

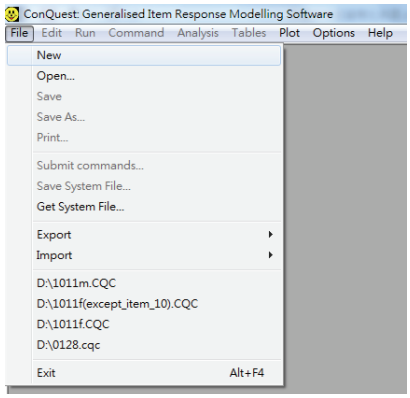


圖3　開新檔案實例

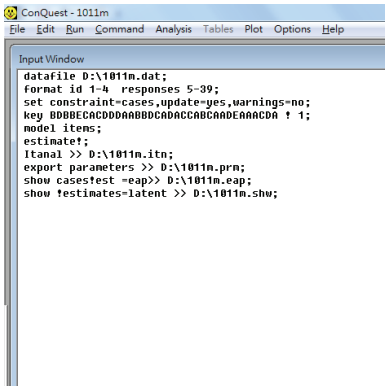


圖4　輸入指令實例

[2] ACER ConQuest 試用版和操作手冊下載網址：http://conquest-sales.acer.edu.au/index.php?cmd=collect&e=g05DzYDarjM%3D

ACER ConQuest所使用的指令檔非常容易上手，只需10列即可完成所需要的分析，逐列詳細說明如表2。

表2　ACER ConQuest 指令與說明

指令	說明
datafile **D:\1011m.dat;**	讀取資料檔「1011m.dat」，檔案存放在「D:\」。
format id **1-4** responses **5-39**;	資料檔的第1格至第4格資料為學生座號，第5格至第39格為作答反應。
set constraint=cases,update=yes,warnings=no;	以人為定位，設定能力值的總和為0；估計出來能力值為0的學生，表示站在正中間。
key **BDBBECACDDDAABBDCADACCABCAADEAAACDA** ! 1;	依照題號順序輸入標準答案，共35個正確答案。
model items;	
estimate!;	
Itanal >> **D:\1011m.itn**;	輸出傳統的試題分析報表於「D:\」，檔名為「1011m.itn」。
export parameters >> **D:\1011m.prm**;	輸出試題難度於「D:\」，檔名為「1011m.prm」。
show cases!est =eap>> **D:\1011m.eap**;	輸出學生能力值於「D:\」，檔名為「1011m.eap」。
show !estimates=latent >> **D:\1011m.shw**;	輸出總表於「D:\」，檔名為「1011m.shw」。

註：Acer ConQuest所使用的指令檔非常容易上手，只需依照實例，將粗體字部分依照資料真實情況進行修改，便能應用於教室中的真實情境。

3. 執行試題分析

按下「執行全部」（圖5），程式便開始進行分析。分析結束後，便可以在「D:\」找到「1011m.eap」、「1011m.itn」、「1011m.prm」以及「1011m.shw」四個檔案。

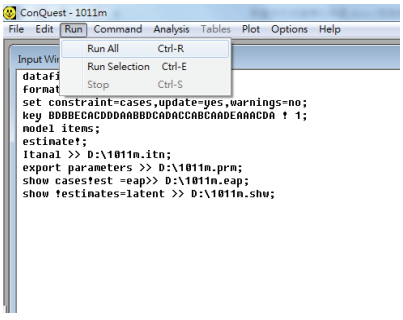


圖5　執行試題分析實例

4. 詮釋試題分析結果

（1）「1011m.itn」報表重點說明

這個檔案呈現的是傳統的試題分析結果，可用來和學校既有的試題分析報表相互印證。如果出現全部學生都答對與全部學生都答錯的題目，則是沒有功能的題目，並沒辦法將學生區分高下。

以第五題為例（圖6），作答此題的學生共有58人（Cases for this item）。此題得分與測驗總分之間的相關是0.37（Item-Total Cor.），為鑑別度的指標；當試題的鑑別度越佳，越能區隔學生的能力。此題的正確答案為E（選擇E者，Score為1，其餘為0），58位學生中有37位答對，通過率63.79%，屬於中間偏易的試題。

```
Item 5
------
item:5 (5)
Cases for this item      58   Item-Rest Cor.  0.26   Item-Total Cor.  0.37
Item Threshold(s):    -0.64   Weighted MNSQ   1.00
Item Delta(s):        -0.64
-------------------------------------------------------------------------------
 Label    Score     Count   % of tot  Pt Bis     t  (p)   PV1Avg:1 PV1 SD:1
-------------------------------------------------------------------------------
   A       0.00         6      10.34   -0.16   -1.25(.216) -0.23     0.63
   B       0.00        11      18.97   -0.16   -1.24(.221) -0.55     0.57
   C       0.00         1       1.72    0.03    0.20(.844)  0.04     0.00
   D       0.00         3       5.17   -0.07   -0.55(.584) -0.35     0.80
   E       1.00        37      63.79    0.26    2.05(.045)  0.23     0.71
===============================================================================
```

圖6　古典測驗理論試題分析實例

（2）「1011m.prm」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的試題難度，也就是該題答錯（得0分）跨到答對（得1分）的閾值。就定義而言，試題難度指的是擁有50%機率可以答對該題的學生能力值（Verhelst, 2004）。如果學生的能力值呈現常態分配，那麼68%學生能力值介於-1至+1之間，95%學生能力值介於-2至+2之間；試題難度達到2以上為高難度的題目，試題難度落於-2以下為極簡單的題目。

同樣以第五題為例（圖7），試題難度為-0.63852，表示能力值為-0.63852的學生，有50%的機率可以答對此題。由於我們已經在指令界定「估計出來能力值為0的學生，表示站在正中間」，能力值為-0.63852的學生能力值未達整體學生能力值一半，反應出此題屬於中間偏易的試題，和此題通過率63.79%相互呼應。再看第10題，試題難度為-4.32683，表示能力值為-4.32683的學生即有50%的機率可以答對此題，因而此題顯得非常非常容易。

```
 1   -0.89911   /* item 1 */
 2   -1.50364   /* item 2 */
 3   -0.55523   /* item 3 */
 4   -3.60308   /* item 4 */
 5   -0.63852   /* item 5 */
 6   -0.99060   /* item 6 */
 7   -2.59618   /* item 7 */
 8   -0.31258   /* item 8 */
 9   -2.59618   /* item 9 */
10   -4.32638   /* item 10 */
```

圖7　Rasch模式分析所得試題難度實例

（3）「1011m.eap」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的學生能力期望值。第一欄為識別碼，依照期中考資料順序排列，識別碼1代表學生座號0001、識別碼2代表學生座號0002、識別碼3代表學生座號0003，可供辨識出是哪位學生的資料。

第二欄即為學生能力值，如果學生的能力值呈現常態分配，那麼68%學生能力值介於-1至+1之間，95%學生能力值介於-2至+2之間。第三欄為學生能力值變異數（Wu, Adams, Wilson, & Haldane, 2007）。圖8所呈現的10位學生之中，識別碼2（座號0002）能力值0.69792最高、識別碼5（座號0005）能力值-1.54555最低。

```
 1    -1.17822    0.35749    0.78445
 2     0.69792    0.46151    0.64076
 3    -0.34681    0.38430    0.75090
 4     0.12747    0.41385    0.71113
 5    -1.54555    0.34141    0.80340
 6    -1.17822    0.35749    0.78445
 7     0.49339    0.44317    0.66875
 8    -0.03889    0.40210    0.72730
 9    -0.03889    0.40210    0.72730
10     0.49339    0.44317    0.66875
```

圖8　Rasch模式分析所得學生能力值實例

（4）「1011m.shw」報表重點說明

這個檔案呈現的是以Rasch模式進行分析後，所得到的總表，內容包含數個部分。首先，呈現的是試題的適配指數MNSQ及其95%信賴區間，用來檢視試題是否符合Rasch模式的前提假設。理論上，MNSQ數值的虛無假設等於1，若分析所得之MNSQ數值超出其95%信賴區間，表示該題並不符合Rasch模式「鑑別度 = 1」的前提假設，此時，該試題之T值的絕對值也會大於 2（Wu, Adams, Wilson, & Haldane, 2007）。當原始 T 值大於2，表示和理論模式相較，試題之鑑別度較差；當原始 T 值

小於 -2，表示和理論模式相較，試題之鑑別度更高。

此外，「未加權MNSQ」（UNWEIGHTED FIT）較容易受到極端值的影響，Bond與Fox（2007）建議使用者，首要以「加權MNSQ」（WEIGHTED FIT）作為判斷試題是否適配的指標。

同樣以第五題為例（圖9），試題難度為-0.639（ESTIMATE），加權MNSQ為1.00並未超出其95%信賴區間，T 值的絕對值為0，表示此題符合Rasch模式「鑑別度=1」的前提假設。再以第八題為例（圖9），試題難度為-0.313（ESTIMATE），加權MNSQ為1.21略超出其95%信賴區間，而T值的絕對值為2；顯示此題鑑別度較低，但還不致於非常差，在沒有更合適試題的情況下，可以考慮保留。在試題適配度皆為可接受的情況下，亦顯示此份測驗符合Rasch模式單向度（unidimensionality）的前提假設，可視為測量同一特質「統計學能力」。

	VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
	item	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	1	-0.899	0.318	1.22	(0.64, 1.36)	1.1	1.18	(0.74, 1.26)	1.3
2	2	-1.504	0.356	1.23	(0.64, 1.36)	1.2	1.16	(0.63, 1.37)	0.9
3	3	-0.555	0.305	1.12	(0.64, 1.36)	0.7	1.13	(0.79, 1.21)	1.1
4	4	-3.603	0.737	0.30	(0.64, 1.36)	-5.2	0.84	(0.00, 2.23)	-0.1
5	5	-0.639	0.308	0.99	(0.64, 1.36)	0.0	1.00	(0.78, 1.22)	-0.0
6	6	-0.991	0.322	0.87	(0.64, 1.36)	-0.7	0.91	(0.73, 1.27)	-0.7
7	7	-2.596	0.493	1.38	(0.64, 1.36)	1.9	1.11	(0.31, 1.69)	0.4
8	8	-0.313	0.300	1.24	(0.64, 1.36)	1.3	1.21	(0.81, 1.19)	2.0
9	9	-2.596	0.493	0.52	(0.64, 1.36)	-3.1	0.86	(0.31, 1.69)	-0.3
10	10	-4.326	1.021	0.30	(0.64, 1.36)	-5.3	0.89	(0.00, 2.82)	0.2

圖9　Rasch模式分析所得試題適配指標實例

其次，解讀「試題與受試者關係圖」（圖10）所呈現的訊息。垂直的虛線將圖一分為二，虛線左側「×××」圖示，表示受試者；虛線右側「數字」表示試題題號，最左側「下至 -3 上至 +2」表示能力值與試題共用的刻度。傳統的試題分析結果是以百分比呈現，學生與試題有各自的計算基準，兩者無法畫在同一張圖；Rasch 模式使得學生與試題使用相同量尺，便能發揮優勢，以一張圖呈現兩者之間的相互關係。

整體而言，受試者能力值分佈介於 -2 至 +2 之間，試題難度分佈介於 -3 至 0 之間，唯獨第 32 題難度最高，介於 1 至 2 之間；顯示這 35 題試題相對於受試者而言，乃如指諸掌。

細部觀察可知，對於能力值介於 -2 至 0 的受試者而言，尚有足夠的試題可以區辨其能力；反觀能力值介於 0 至 2 的受試者而言，幾乎沒有試題能夠區辨他們。此

外，為數不少的試題難度介於 -3 至 -2 之間，卻沒有能力值相對應的受試者。顯示期中考的命題為中間偏易，大多數學生均駕輕就熟、輕鬆作答。

另一方面，這個圖隱含了「適性」概念在其中，意即困難的題目應該給能力比較高的學生作答、簡單的題目給能力比較低的學生，這樣對兩者都有挑戰性。試想，若是簡單的題目給能力比較高的學生、而困難的題目給能力比較低的學生，那麼前者輕鬆答完、後者無法理解題意，亦非評量所欲達成之目的。

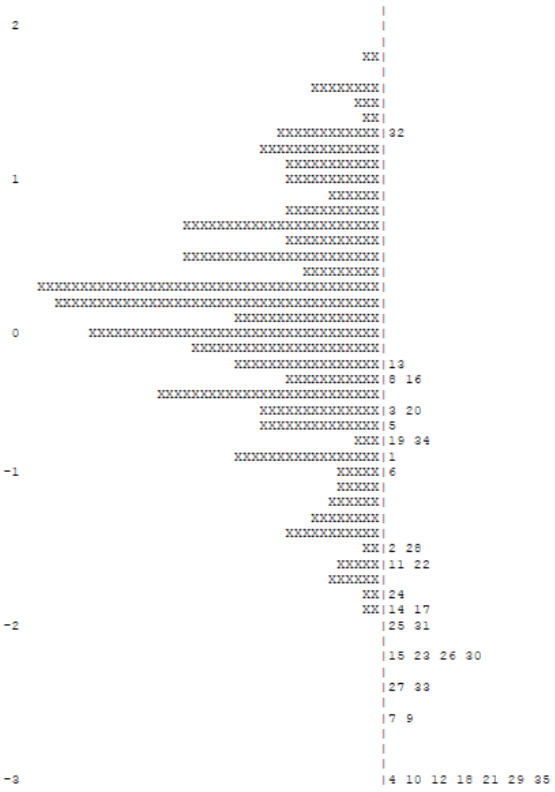


圖10　試題與受試者關係圖實例

（二）期末考學生能力值與試題難度分析

1. 整理期末考逐題作答反應

和期中考資料檔作法相同，將每位學生期末考的逐題作答反應讀進電腦，並整理存檔為ACER ConQuest所需的檔案格式。檔案內容由左至右分別是：學生座號（共四碼）、第1題至第50題的逐題作答反應。整理好之後，將檔案存檔並命名；此例，我們將之命名為「1011f.dat」。

2. 備妥共同題對照表與共同題參數檔

為了將期末考試題難度和期中考試題難度置於相同量尺，採用固定試題參數量尺化（fixed common item parameter, FCIP）方法，校準期末考試題參數；也就是估計期末考試題難度時，匯入已知之共同題參數而不再估計共同題難度。

透過共同題對照表（表2）得知，期中考的第1題為期末考的第1題、期中考的第7題為期末考的第3題、期中考的第14題為期末考的第6題，以此類推。考量共同題品質對於期末考試題難度與學生能力值估計影響甚鉅，在分析期末考資料之前，再次檢查「1011m.shw」中，共同題的MNSQ值，確認它們皆符合Rasch模式前提假設，才進行下一個步驟。

表3　共同題對照表實例

期中考題號	期末考題號
1	1
7	3
14	6
19	9
24	12
30	15
34	18

開啟「1011m.prm」，刪除非共同題，僅保留第1、7、14、19、24、30、34題（圖11），並將最左方「題號」變更為1、3、6、9、12、15、18（圖12），試題難度不變，另存新檔為「D:\1011a.prm」即完成。

```
 1  -0.89911   /* item 1 */
 7  -2.59618   /* item 7 */
14  -1.88579   /* item 14 */
19  -0.81016   /* item 19 */
24  -1.74923   /* item 24 */
30  -2.19878   /* item 30 */
34  -0.81016   /* item 34 */
```

圖11　共同題於期中考之題號與參數

```
 1  -0.89911
 3  -2.59618
 6  -1.88579
 9  -0.81016
12  -1.74923
15  -2.19878
18  -0.81016
```

圖12　共同題於期末考之題號與參數

3. 準備試題分析指令檔

於程式集中點選執行ConQuest 3，開啟「D:\1011m.CQC」（圖13）小幅度修改指令，也就是小部分修改期中考試題分析指令檔，即可完成期末考試題分析，並達成測驗成績等化之目的。修改完成後，另存新檔為「D:\1011f.CQC」。

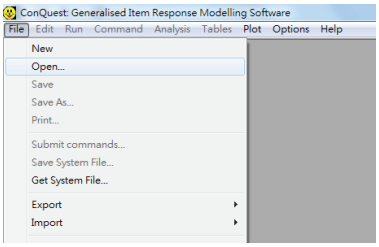


圖13　開啟舊檔實例圖

測驗分數等化所使用的指令檔非常容易上手，只需11列即可完成所需要的分析，逐列詳細說明如表。和表1的主要差異在於「以共同題為定位」和「匯入共同試題參數」兩列指令。

表4　ACER ConQuest測驗分數等化指令與說明

指令	說明
datafile **D:\1011f.dat**;	讀取資料檔「1011f.dat」，檔案存放在「D:\」。
format id **1-4** responses **5-54**;	資料檔的第1格至第4格資料為學生座號，第5格至第54格為作答反應。
set constraint=**none**,update=yes,warnings=no;	以共同題為定位，需將set constraint=改為「none」。
key **BAADABAADCBBCAADBDDCBABABBEBCDCBDDCDBDBDACAAAACBCB!**1;	依照題號順序輸入標準答案，共50個正確答案。
model items;	
import anchor_parameters << **D:\1011a.prm**;	新增一列指令，匯入共同題試題參數。
estimate!;	
Itanal >> **D:\1011f.itn**;	輸出傳統的試題分析報表於「D:\」，檔名為「1011f.itn」。
export parameters >> **D:\1011f.prm**;	輸出試題難度於「D:\」，檔名為「1011f.prm」。
show cases!est =eap>> **D:\1011f.eap**;	輸出學生能力值於「D:\」，檔名為「1011f.eap」。
show !estimates=latent >> **D:\1011f.shw**;	輸出總表於「D:\」，檔名為「1011f.shw」。

註：Acer ConQuest所使用的指令檔非常容易上手，只需依照實例，將粗體字部分依照資料真實情況進行修改，便能應用於教室中的真實情境。

4. 執行試題分析

按下「執行全部」，程式便開始進行分析。分析結束後，便可以在「D:\」找到「1011f.eap」、「1011f.itn」、「1011f.prm」以及「1011f.shw」四個檔案。

5. 詮釋試題分析結果

（1）「1011f.itn」、「1011f.prm」以及「1011f.eap」報表

「1011f.itn」、「1011f.prm」以及「1011f.eap」三個報表的詮釋方式，和期中考試題分析產生的輸出檔相同，不再贅述。

（2）「1011f.shw」報表重點說明

試題適配指標的部分，和「1011f.shw」報表略有出入：7題共同題的試題難度旁邊，出現「*」註記，並且沒有顯示標準誤！（圖14）乃因這個例子採用FCIP進行試題參數校準，7題共同題的試題難度已經被固定不再估計，也就不會有估計標準誤。謹慎起見，再次核對7題共同題的試題難度，確認和「1011a.prm」之中的數值相同，顯示程式讀取資料正確。

VARIABLES			UNWEIGHTED FIT			WEIGHTED FIT		
	item	ESTIMATE	ERROR^	MNSQ	CI	T	MNSQ	CI	T
1	1	-0.899*		0.60	(0.64, 1.36)	-2.5	0.68	(0.74, 1.26)	-2.8
2	2	-3.262	0.622	0.87	(0.64, 1.36)	-0.7	0.98	(0.02, 1.98)	0.1
3	3	-2.596*		1.67	(0.64, 1.36)	3.1	1.32	(0.32, 1.68)	1.0
4	4	-3.704	0.745	1.18	(0.64, 1.36)	1.0	1.04	(0.00, 2.25)	0.3
5	5	-0.259	0.313	1.32	(0.64, 1.36)	1.6	1.24	(0.81, 1.19)	2.3
6	6	-1.886*		0.98	(0.64, 1.36)	-0.0	1.10	(0.54, 1.46)	0.5
7	7	-1.034	0.335	1.08	(0.64, 1.36)	0.5	1.03	(0.72, 1.28)	0.3
8	8	-1.034	0.335	0.89	(0.64, 1.36)	-0.5	0.93	(0.72, 1.28)	-0.5
9	9	-0.810*		0.90	(0.64, 1.36)	-0.5	0.93	(0.75, 1.25)	-0.5
10	10	-4.434	1.028	0.25	(0.64, 1.36)	-5.9	0.92	(0.00, 2.85)	0.2
11	11	0.466	0.317	1.08	(0.64, 1.36)	0.5	1.08	(0.80, 1.20)	0.8
12	12	-1.749*		1.41	(0.64, 1.36)	2.0	1.40	(0.58, 1.42)	1.7
13	13	-2.939	0.551	0.85	(0.64, 1.36)	-0.8	1.05	(0.18, 1.82)	0.3
14	14	-2.939	0.551	0.78	(0.64, 1.36)	-1.2	0.97	(0.18, 1.82)	0.1
15	15	-2.199*		0.96	(0.64, 1.36)	-0.1	0.88	(0.46, 1.54)	-0.4
16	16	-2.463	0.468	0.61	(0.64, 1.36)	-2.4	0.81	(0.37, 1.63)	-0.5
17	17	0.302	0.314	1.02	(0.64, 1.36)	0.2	1.05	(0.81, 1.19)	0.6
18	18	-0.810*		1.11	(0.64, 1.36)	0.6	1.16	(0.75, 1.25)	1.3
19	19	-2.274	0.442	1.01	(0.64, 1.36)	0.1	1.06	(0.43, 1.57)	0.3
20	20	-2.680	0.503	0.69	(0.64, 1.36)	-1.8	0.89	(0.29, 1.71)	-0.2

圖14　FCIP所得試題適配指標實例

至於「試題與受試者關係圖」的詮釋方式，亦和期中考試題分析產生的輸出檔相同，不再贅述。

（三）期中考與期末考學生能力值差異分析

1. 彙整期中考與期末考能力值

分別開啟「1011m.eap」和「1011f.eap」，將第二欄學生能力值複製貼上至Microsoft Excel（圖15），可以看出，和期中考相較，座號0001、0004、0006、0008以及0010五位學生，期末考能力值是增加的；而且可以確定的是，期中考的能力值與期末考的能力值是建造在同一量尺，兩者可以相互比較。

	A	B	C
1	座號	期中考能力值	期末考能力值
2	0001	-1.17822	-1.10513
3	0002	0.69792	0.65959
4	0003	-0.34681	-0.91384
5	0004	0.12747	0.2475
6	0005	-1.54555	-1.83439
7	0006	-1.17822	-0.81529
8	0007	0.49339	0.12379
9	0008	-0.03889	0.2475
10	0009	-0.03889	-0.11013
11	0010	0.49339	0.99052

圖15　期中考與期末考能力值彙整實例

2.全班學生期中考與期末考能力值差異分析

除了可以看出個別學生能力值的發展，也可以透過「成對母體平均數差異檢定」（paired-samples T Test）分析全班學生能力值的發展。

首先，點選Microsoft Excel「資料分析」功能（圖16）。

圖16　Microsoft Excel資料模組介面

接著，選擇「t檢定：成對母體平均數差異檢定」，按「確定」（圖17）。

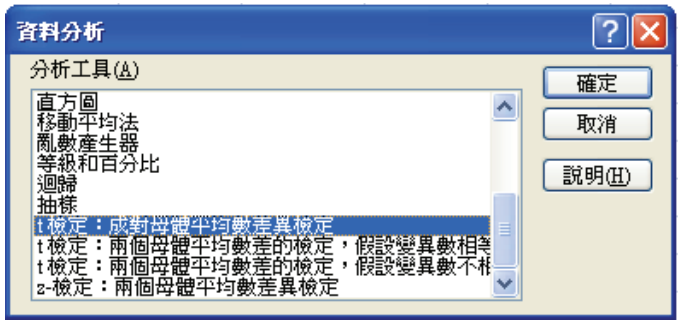


圖17　Microsoft Excel資料分析模組介面

利用滑鼠選取期中考能力值資料範圍，讓Excel讀入「變數1的範圍」，本例為「$B$2:$B$59」；再利用滑鼠選取期末考能力值資料範圍，讓Excel讀入「變數2的範圍」，本例為「$C$2:$C$59」；最後，按「確定」（圖18）。

圖18　Microsoft Excel成對母體平均數差異檢定介面

從分析結果報表可知（圖19），期中考全班學生能力值平均數為-0.0008、期末考全班學生能力值平均數為-0.02548，似乎微幅退步？然而，再進一步檢視「P(T<=t)雙尾」為0.739071，大於社會科學常用的判斷標準「0.05」，表示期中考全班學生能力值平均數與期末考全班學生能力值平均數差異，並未達到統計上顯著水準，也就是和期中考相較，期末考全班學生能力值平均數，沒有進步也沒有退步，維持平盤。由於一般國小學生數學基本能力的年進步量0.5至0.7（任宗浩、譚克平、張立民，2011），大學一年級學生統計能力的年進步量約為0.04（Lin & Hsieh, 2013），顯示學習年段越高，一年能進步的能力值越少。合理推論，此例期中考和期末考僅相距二個月，致使能力值的改變不明顯。

t 檢定：成對母體平均數差異檢定

	變數 1	變數 2
平均數	-0.0008	-0.02548
變異數	0.429169	0.63463
觀察值個數	58	58
皮耳森相關係數	0.717059	
假設的均數差	0	
自由度	57	
t 統計	0.334712	
P(T<=t) 單尾	0.369535	
臨界值：單尾	1.672029	
P(T<=t) 雙尾	0.739071	
臨界值：雙尾	2.002465	

圖19　Microsoft Excel成對母體平均數差異檢定實例

參、結語與展望

測驗成績等化，當然不是只侷限於選擇題型，是非題、填充題、甚至問答題，也都可以進行遵循這一套操作程序，略加修改執行指令即可達成。教師們一旦學會操作成績等化，教務處便可以建立新版校內成績冊。未來，期中考試過後，學生會有兩欄成績，第一欄是原有的卷面分數，就是滿分一百分而學生得幾分。第二欄則是將作答反應轉化成能力值，期中考呈現的是直接估計所得的能力值，期末考呈現的則是和期中考連結等化後的能力值，供老師判斷學生學習的進展和演變。

等化技術的應用層面是很廣的，不但能了解學生於期中考至期末考的能力變化，也可以了解補救教學、差異化教學等各式教學方案實施前後，學生能力的變化。如果設計得宜，甚至可以用於了解學生從1年級升至12年級的能力變化。要注意的是，校準後的能力值仍有其使用上的限制，不適合作為繁星計畫等升學方案之參考指標。

本文所演示之實例，連結等化的重要設計在於「測驗後，並未將試卷發回給學生，也沒有公告試題」，學生不會因為期中考試後反覆練習舊題目，而使得期末考時表現得更好。然而，這個做法卻不符合現今1至12年級教學實務所需。

建議未來由各學科中心主導建構「各學習單元的共同題題庫」，確保試題品質優良，沒有試題參數漂移（item parameter drift, IPD）的現象（IPD意指共同題的參數因為時間的不同而產生變化致使測驗分數效度受到威脅）（Goldstein, 1983），且內容亦具備該單元的代表性。這樣不但能於試前獲悉共同題參數並確保試題功能無

虞，也能於試後回收共同題，避免學生背誦共同試題干擾評量結果。意即期中考與期末考的試題，一部份來自學校教師，一部份來自學科中心。學校教師命題的部分仍維持傳統做法，試後發回並檢討考卷；但共同題的部分則收回，不公開試題。

期中考之前，各校教師自共同題題庫中，挑選合適試題融入其中；期末考的時候，共同題仍然融入題目中再測驗一次。教師需要及早準備學生，告知期末考的測驗範圍將會有部分涵蓋期中考的範圍，讓他們在心理上和時間上，都有足夠的機會能事先準備。

在更長遠的未來，可由教育部建置「雲端等化系統」，教師們僅需依照格式匯入學生作答反應，並輸入共同題題號與難度參數，系統即能自動完成學生能力值估計並輸出報表，供教師與學生們了解學習趨勢，是不是相當便利又有效率呢！

參考文獻

內政部地政司（2013）。**房地產交易價格**。檢索自http://www.land.moi.gov.tw/chhtml/new_quehl.asp

任宗浩、譚克平、張立民（2011）。二階段分層叢集抽樣的設計效應估計。**教育科學研究期刊，56**（1），33-65。

行政院主計總處（2013）。**受僱員工薪資調查統計**。檢索自http://www.dgbas.gov.tw/ct.asp?xItem=1135&ctNode=3253&mp=1

經濟部能源局（2013）。**家用液化石油氣大母體區平均價格年報表(零售價)**。檢索自http://web3.moeaboe.gov.tw/oil102/

Andersen, E. B., & Olsen, L. W. (2001). The life of georg rasch as a mathematician and as a statistician. In A. Boomsma, M. A. J. van. Duijn & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 3-24). New York, NY: Springer-Verlag.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*, 192-211.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores (College Board Rep. No. 99-1).* New York, NY: College Entrance Board.

Goldstein, (1983). Measuring changes in educational attainment over time: Problems and

possibilities. *Journal of Educational Measurement, 20*(4), 369-377.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices (2nd ed.). New York, NY: Springer-Verlag.

Lin, S.-H., & Hsieh, P.-J. (2013, August). *Longitudinal study of undergraduate students' use of learning strategies in introductory statistics.* Paper presented at the Pacific Rim Objective Measurement Symposium 2013, Kaohsiung, Taiwan.

Lissitz, R.W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability *Practical Assessment, Research & Evaluation, 8*(10). Retrived from http://pareonline.net/getvn.asp?v=8&n=10

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: Educational Testing Service.

Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research and Practice Assessment, 8,* 26-39.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2004). *Measurement and assessment in teaching* (11th ed.). Upper Saddle River, NJ: Pearson.

Peter, C., Cieza, A., & Geyh, S. (2013). Rasch analysis of the general self-efficacy scale (GSES) in spinal cord injury (SCI). *Journal of Health Psychology.* Advance online publication. doi:10.1177/1359105313475897

Stewart, J. & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *JALT Testing & Evaluation SIG Newsletter, 14*(2), 11-18.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika, 65*(4), 437-456.

Verhelst, N. D. (2004). *Reference supplement to the preliminary pilot version of the manual for "relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment" sectiong: Item response theory.* Strasbourg, France: Council of Europe.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software.* Victoria, Australia: ACER.