以多面向 Rasch 測量模式分析 TOCFL 口語測驗評分者訓練效果

藍珮君

國家華語測驗推動工作委員會 / 研究員

摘要

本研究目的為探討在接受持續且長期的評分訓練後，評分者評分一致性的變化情形，包含與其他評分者之間的一致性，以及評分者自身的一致性。研究方法為收集 2009 年 2 月至 12 月五次華語文口語能力測驗(簡稱 TOCFL 口語測驗)基礎級預試評分結果，採用多面向 Rasch 測量模式(many-facet Rasch measurement)分析軟體 FACETS，針對評分者嚴格度(rater severity)、評分結果一致性(consistency)以及評分者面向之適配度(fit)進行分析。

研究結果顯示：1. 經過多次且密集的評分訓練後，評分者之間嚴格度差異逐漸縮小，但整體評分嚴格度達到顯著差異水準，顯示評分者之間的嚴格度仍有不同；2. 評分者本身的給分相當一致，適配度數值均在可接受之範圍內；3. 部分評分者間隔二個月後的評分嚴格度有所變化。

現今以華語為第二語言的口語測驗研究相當缺乏，尤其是採用多面向 Rasch 模式進行分析之研究。本研究結果提供華語文口語測驗的實徵資料，對於華語文口語能力測驗的評分訓練效果有初步的瞭解，也能做為未來進行華語文口語能力測驗評分訓練及測驗實施的參考。

關鍵字：多面向 Rasch 測量模式、口語測驗、評分者訓練

Using many-facet Rasch measurement to examining rater training effects of TOCFL Speaking

Pei-Jiun Lan
Steering Committee for the Test Of Proficiency-Huayu / Researcher

Abstract

This research aims to investigate how rater consistency varies after a continuing, long-term rater training by observing the inter-rater consistency and the intra-rater consistency. We collect the rating results from the five pilot tests of the TOCFL Speaking (Test of Chinese as a Foreign Language- Speaking, holding from February, 2009 to December) at beginner level. This research applies the many-facet Rash measurement by adopting the FACETS software to analyze the rater severity, rater consistency, and the fit statistics of the raters.

Three major findings discussed in this research are that: 1) after frequent rater training sections, the variation of inter-rater severity has reduced, but the overall rating severity has reached the significant differences, which indicates that the discrepancy of severity still exists among the raters; 2) individual raters are consistent in their own rating as most of the fit statistics fall in the acceptable range; 3) several raters reveal variations in their severity when they rerated two months later.

So far, only little research on the speaking assessments which focus on using Chinese as a second language (CSL) has been done, and those which applied many-facet Rasch measurement are even less. Empirical data from the results of this study will be provided to help gain a preliminary understanding toward the effects of rater training in CSL speaking tests, and as a reference for the future rater training.

Keywords: many-facet Rasch measurement, speaking test, rater training

壹、前言

華語文口語能力測驗(前稱為 Test of Proficiency-Huayu，現英文名稱更名為 Test of Chinese as a Foreign Language -Speaking，以下簡稱 TOCFL 口語測驗)為一種表現測驗(performance assessment)，考生需以口說的形式，完成考試的各種溝通任務。任務的類型均為建構反應題(constructed response item)，考生在觀看一段影片或一至數張圖片後，依照題目的說明或要求，以自己的方式表達與建構出答案。此種考試型態，與一般常見的選擇題，除了答題方式外，計分方式也有明顯不同，選擇題有標準答案，計分客觀；而建構式題型沒有標準答案，考生的成績是由評分

者根據評分規準，對考生的測驗表現進行評分而來。因此，與選擇題相較之下，影響考生在建構反應題表現的因素，除了考生自身能力以及試題難度以外，評分者也是一重要的影響因子。

正因如此，TOCFL 口語測驗自研發以來，即相當注重評分者的培訓。目的在透過嚴謹的評分訓練，使評分者充分理解評分規準並依據此原則給分，進而讓參與測驗的考生得到公平可靠的成績。在每次預試的評分工作完成後，也針對評分結果進行評分者間信度分析，瞭解評分者給分的一致情形，以確保評分者的評分品質。即使如此，仍有一些疑問難以釐清，例如，當二位評分者給分一致性高時，究竟是表示二位評分者均依據評分原則進行評分，或其實評分者恰巧都是給分偏嚴格或偏寬鬆，所以有看似良好的一致性。而評分者一致性低時，會不會是其中一位評分者依循評分原則給分，但另一位評分過於嚴格或寬鬆所致？以往的分析方式，除非事先在評閱的樣本中置入標準卷(standard rating)，否則很難看出端倪。

然而，多面向 Rasch 測量模式(many-facet Rasch measurement，以下簡稱 MFRM)的出現，解決了上述難題。延伸自單面向 Rasch 模式的 MFRM，能夠同時估計二個以上的面向(稱之為 facet)，除了試題難度以及考生的能力外，也能將評分者等其他相關因素納入估計，可以瞭解評分者評分的嚴格或寬鬆程度，甚至是評分面向的難易度，這對於測驗發展者來說不啻是一大福音。目前已有許多研究採用 MFRM 檢驗口語或寫作測驗評分者給分的品質，其中有研究發現，藉由評分訓練，能改善給分極端的評分者嚴格度，但無法使評分員的給分達到一致；此外，評分訓練可以提高評分員給分的信心，進而提升評分者內信度(Park, 2004；Weigle, 1998)。Lumley 與 McNamara(1995)的研究則指出，評分訓練的效果無法維持很久，評分者的嚴格度會隨時間產生變化。可惜這些研究多半是針對英語或德語等外語測驗，針對華語測驗所做的實徵研究相當缺乏，研究者透過搜尋引擎與各資料庫，最後僅在中國知識資源總庫查詢到二篇，均為中國大陸學者發表之期刊及論文(田清源，2007；羅丹，2008)。

綜上所述，本研究將採用 MFRM 對 TOCFL 基礎級口語測驗預試的評分結果進行分析，研究目的在探討接受持續且長期的評分訓練後，評分者評分一致性與嚴格度的變化，包括：

1. 評分者間評分一致性的變化。
2. 評分者內給分一致性的變化。
3. 評分者在二次評分訓練中評分嚴格度(rater severity)的變化。

貳、文獻探討

一、評量口語能力的測量模式

Engelhard 於 1992 年提出寫作評量計畫的測量模式，由於口語測驗的實施方式與計分流程與寫作測驗相似，本研究的口語能力測量模式參考 Engelhard 的模式加

以修改後如圖 1 所示。測量模式包含三個主要的面向[1]：口語能力(speaking ability)、評分者嚴格度(rater severity)以及口語任務難度(difficulty of the speaking task)。此模式將評分者與口語任務視為中介變項(intervening variables)。

因此從圖 1 可知，考生口語能力會受到考生特質所影響，而最後得到的觀察分數，除了考生的口語能力外，還受到評分者給分嚴格與否，以及口語任務難易度的影響。此外，評分量尺的架構(structure of the rating scale)也會影響考生獲得的成績。

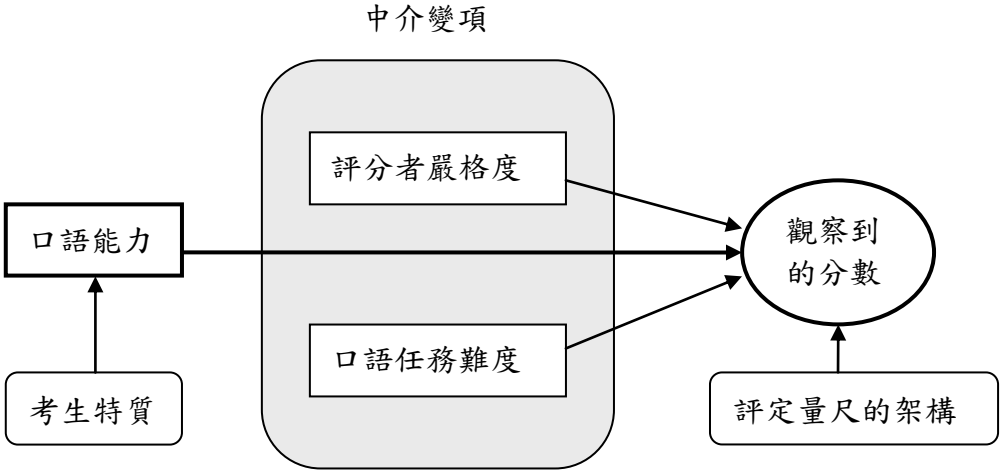

圖 1　口語測驗的測量模式(修改自 Engelhard, 1992)

除了 Engelhard 之外，亦有其他學者重視評分者因素對於考生測驗成績的影響。如 Cason 與 Cason(1984)的研究即指出考生得分不僅代表其真實能力水平，而是考生能力和評分者個人特質的函數(引自羅丹，2008)；而 Hoyt 和 Kerns(1999)對 79 篇概化(generalizability)研究進行後設分析(meta-analysis)發現，平均有 37%的成績變異來自於評分者主要效果以及評分者與考生的交互作用(引自 Eckes, 2005)。Engelhard(1994)採用 Saal、Downey 和 Lahey(1980)提出的四種主要評分者誤差類型(評分嚴格度、月暈現象、趨中效應以及全距限制)，來檢視 15 位評分者評閱寫作測驗的表現，結果顯示評分者的嚴格度達到顯著差異，其餘三種誤差類型也獲得證實。

上述研究結果在在顯示出評分者因素對於表現測驗成績的影響不容忽視，若一份測驗沒有良好的評分品質，所得到的成績將沒有信度及效度可言。有鑑於此，測驗機構及測驗發展者會藉由一些方式或步驟，盡可能讓評分過程和結果更為可靠，例如：在評分前舉行評分者訓練，讓評分者熟悉評分的規準；或對評分者進行評分測驗，挑選出合格的評分者；評分結束後對結果進行分析，作為之後評分之參考……等等。傳統上，舉辦評分者訓練目的就是在減少整體評分嚴格度以及隨機誤差兩者的變異(Lumley & McNamara, 1995)。

[1] 原模式中介變項中，尚納入評分面向難度(domain difficulty)，因本研究不對此進行探討，故省略。

二、多面向 Rasch 測量模式

在過去，分析評分者間信度的方式，主要使用百分比一致性、皮爾森等級相關、肯德爾和諧係數、Kappa 一致性係數等相關分析。百分比一致性在瞭解評分者對考生評定成績完全相同的比例；相關分析的目的在檢視評分者之間給分是否存在相對應的次序關係。當參與測驗的考生人數眾多時，測驗機構為減輕評分者負擔，往往將二位評分者分為一組進行評分，當二位評分者給分有落差時，再由第三位評分者加入評閱，並以上述方式分析評分者間信度。然而，高的評分者間信度，並不一定代表評分結果是正確的，也可能是二位評分者都是較為嚴格或是較為寬鬆的評分者。在沒有置入標準卷或其他介入機制的情形下，得到的評分者間信度可能反而提供錯誤的訊息，使研究者判斷失當，也連帶影響考生權益。

多面向 Rasch 測量模式(Many-Facet Rasch Measurement，以下簡稱 MFRM)是 Rasch 測量模式的延伸，由 Linacre 於 1989 年提出，可以同時校準(calibrate)多個面向，並能分開呈現估計結果(引自 Engelhard, 1992)。MFRM 之下有幾種分析模式，一般常見的有評定量尺模式(rating scale model)和部分給分模式(partial credit model)。二者的差異在於前者假定每個試題的量尺架構相當；而後者每一個試題都有各自的量尺架構(Bonk & Ockey, 2003)。本研究由於每個題目均以相同的評分原則(rubric)進行給分，因此使用評定量尺模式，公式如下(Linacre, 2010)：

$$\log (P_{nmjk} / P_{nmj(k-1)}) = B_n - A_m - C_j - F_k$$

在此，

P_{nmjk} 是指考生 n 被評分者 j 在題目 m 評為 k 分的可能性

$P_{nmj(k-1)}$ 是指考生 n 被評分者 j 在題目 m 評為 k-1 分的可能性

B_n 是指考生 n 的能力

A_m 是指題目 m 的難度

C_j 是指評分者 j 的嚴格度

F_k 是指得到 k 分與 k-1 分之間的界線(barrier)，也有人稱為難度階(threshold difficulty)

由於 MFRM 是針對評分者嚴格度、題目難度以及考生能力等面向的資料，同時進行估計及校準至一個共同的量尺上，因此各個面向得到的數值是可以互相進行比較的。估計考生能力時，因將評分者嚴格度差異納入考量進行調整，故獲得的考生能力比原始分數更能代表考生的真實能力。

本研究使用 FACETS 統計軟體(3.67 版)進行分析，FACETS 提供一些指標呈現各個面向的分析結果，在本研究中，使用到的指標與評分者嚴格度有關，包括：分散指標(separation index)、信度(reliability)、卡方考驗，以及評分者給分一致性(rater consistency)的 infit 統計值。以下分別說明各指標的定義與數值代表的含意。

1. 分散指標(separation index)

分散指標是測量值的校正標準差(Adj.SD)與均方根誤差(RMSE)的比值。數值

越接近 0，表示評分者間的嚴格度越相近，數值越大，表示評分者間的嚴格度差異較大(Weigle, 1998)。

2. 信度(reliability)

FACETS 提供的信度數值是上述分散指標的可信度，也就是表示分析結果能區分不同能力水準考生的可信度。這裡所指的信度指標，就是眾所知悉的 KR20 信度指標的 Rasch 相似形(analogue)，只是改以能力量尺而非原始分數進行計算(Pollitt & Hutchison, 1987)。一般來說，信度數值越高越好，表示越能區分考生的能力，如在考生面向，分散指標的信度越高，表示測驗越能將考生能力區分為不同程度。不過在評分者面向來說，較低的信度數值是好的，因為理想上希望不同的評分者有相同的評分嚴謹度(Park, 2004)。

3. 卡方考驗

為同質性考驗，目的在考驗所有評分者的嚴格度是否相等，虛無假設為所有評分者的嚴格度是相等的，若達到顯著水準，拒絕虛無假設，則表示至少有二位評分者的嚴格度是不同的。

4. 適配度(fit)

一般來說，適配度統計值是指 MFRM 模式中，觀察到的評分與預期評分結果的適配情形，分為 outfit 與 infit 二種。outfit 值對於偶發性的極端非預期評分較為敏感；而 infit 值則對於累積的非預期評分結果較為敏感(Eckes, 2009)。因此，許多學者認為 infit 數值較為適合作為判斷適配度的指標(Pollitt & Hutchinson, 1987；Park, 2004)。

當測量模式與觀察資料適配時，infit 的數值為 1.0。Lunz、Wright 和 Linacre(1990)建議可接受的範圍為 0.6 至 1.5(引自 Engelhard, 1992)；Linacre(2002)建議用 0.5 為低標，1.5 為高標，也有其他研究建議使用比較嚴格的標準(0.7-1.3)(McNamara, 1996；Bond & Fox, 2001；引自 Eckes, 2005)。Lunz 與 Stahl(1990)表示，若 infit 數值大於等於 1.5，表示評分者有太多非預期的給分，稱為 misfit；若小於等於 0.5 表示評分者給分的變異不足，稱為 overfit(引自 Weigle, 1998)。

Bonk 與 Ockey(2003)提到，比起考生的 misfit，評分者的 misfit 對於測驗的效度可能造成更嚴重的威脅，因為這表示評分者的給分偏離自身的標準，這對於其他面向的估計影響很大，misfit 也是 Rasch 分析無法進行校正的。

三、運用 MFRM 對於評分者及評分訓練的相關研究

Linacre(1989)提出 MFRM 後，許多學者開始採用此一測量模式對評分者因素及評分訓練的成果進行相關分析。Engelhard(1992)以接受了三天評分訓練，且通過評分測驗(評 20 篇作文，最少需達到 62%給分完全相同，其餘篇數僅能相差一級分)的 82 名合格評分者，對 1000 名學生的英語寫作測驗進行評分。結果顯示，這 82

名評分者中，給分最嚴格及最寬鬆的評分者嚴格度差距為 3.52 logits，嚴格度的卡方考驗達到顯著差異，而 separation index 的信度也達到 0.87，顯示評分者即使已經受過訓練且通過嚴格的測試，但正式評分時的嚴格度仍是有差異的。

田清源(2007)以漢語水平考試(HSK)高等作文進行分析，結果顯示 8 位評分者嚴格度相差 4.28 logits；羅丹(2008)對漢語水平考試(HSK)中級口語測驗的研究結果顯示，9 位評分者嚴格度差異為 2.72 logits，separation index 為 4.93，信度達到 0.96，卡方考驗亦達到顯著，顯示評分者整體的給分嚴格度並不一致；評分者內一致性方面，有 1 位評分者給分變異過大。

Weigle(1998)則針對有評分經驗與無經驗的評分者，比較兩者接受評分訓練前後的評分嚴格度。結果發現比起有經驗的評分者，無經驗的評分者給分較為嚴格，本身的給分也較不一致；而經過評分訓練後，二個群體間的嚴格度差異減少，但仍達到顯著差異水準，此外，大多數評分者內一致性已經提高。

Bonk 與 Ockey(2003)、Park(2004)以及 Eckes(2005)分別以英語口語測驗、英語寫作測驗(CEP writing test)與德語測驗(TestDaF)的寫作與口語測驗進行研究的結果也顯示，接受評分訓練後，評分者給分的嚴格度仍有顯著差異，然而，評分者內的一致性則較佳。McNamara(1996)從 McIntyre(1993)、Weigle(1994)與 Shohamy 等人(1992)的研究結果發現：1.評分者訓練能讓評分者在評分時更有信心，訓練的成效在於降低評分者給分的隨機誤差(random error)；2.評分者訓練能降低評分者整體嚴格度的變異，但無法完全消除。特別是能降低非常極端嚴格或寬鬆的評分傾向，但是明顯的評分者差異依然存在。

其他研究則有不同的發現，Du、Brown 與 Rogers(1997)以學生能力、評分嚴格度、文本難度以及評分標準四個面向對寫作測驗的評分結果進行分析。30 位評分者嚴格度的差異不到 1 logits，顯示評分者嚴格度沒有非常極端嚴格或寬鬆，不過由於作者未提供其他指標結果，所以無法得知整體評分嚴格度有沒有差異。Liu 與 Wen(2007)邀請學生與教師對一批參加口語演說比賽的學生進行二次評分，前後間隔二個月。結果發現學生評分嚴格度的差異雖然縮小，但卡方考驗結果仍達到顯著差異；而教師的評分嚴格度差異極小，二次評分結果，卡方考驗均未達到顯著水準，顯示教師整體評分嚴格度沒有不同。唯該研究將教師分為男女二組進行分析，分別為 3 人和 2 人，評分者人數較少。

另外，也有研究更進一步針對評分者評分嚴格度的變化進行探討。Lumley 和 McNamara(1995)檢視 13 名評分者對同樣 10 名考生二次評分嚴格度的變化，評分時間間隔 18 個月；其中有 6 位評分者，隔 2 個月後，再對另一群考生(73 人)進行評分，持續觀察嚴格度的改變。結果發現，有些評分員的嚴格度會隨著時間改變，且這個改變並不穩定，有的評分者變得比較嚴，有的則變得較為寬鬆，顯示評分訓練的結果不必然能持久。Bonk 與 Ockey(2003)的研究也有類似的結果，13 位評分者在連續二年的評分嚴格度變化很大，有 8 位評分者變得較為嚴格，3 位評分比之前寬鬆，另 2 位變化的幅度較小。

綜合上述研究結果及 McNamara(1996)的看法，無論是何種語言測驗，評分訓

練對於提升評分者間信度(interrater reliability)的效果似乎不盡理想，但對於提升評分者內信度(intrarater reliability)的幫助較大。畢竟評分者有其個人的人格特質、專業知識、教學經驗等不同背景，即便接受評分訓練，共同理解了評分原則與規準，掌握了大致的評分方向，但對於一些細部的解讀上可能仍有差異，特別是一些比較難評閱的學生表現，很難透過培訓達到完全一致的評分結果。但是透過評分訓練，評分者在短時間練習評閱了為數不少的樣本，對於考生整體能力開始產生較為具體的認知，並逐漸歸納出不同級分考生的能力表現；可能因而使評分者在培訓後的正式評閱，更有依據進行給分，達到個人的評分一致性。

先前的研究，二次評分訓練的間隔期間多半為一年左右，次數最多為三次，若採取較為密集的評分訓練模式，是否能使評分者的嚴格度有更為明顯的改善，達到一致的評分嚴格度，是研究者感興趣的議題。故本研究採取多次且密集的評分訓練模式，希望藉此瞭解評分者的評分嚴格度與一致性是否與過去研究發現相同，亦或是因為密集的評分訓練而能達到更為一致的嚴格度。

參、研究設計

一、研究參與者

本研究於 2009 年 2 月至 12 月間共陸續舉辦五次基礎級口語測驗預試，在每次預試結束後，即緊接著開始進行評分訓練及評分工作，考試時間、樣本大小以及參與評分老師分佈如表 1 所示。五次評分訓練間隔約二至三個月不等，第一次評分訓練有 3 位評分者參加，第二次則有 5 位評分者參加，其餘三次評分者均為 4 人。若從評分者參加的次數來看，以評分者 A 和 B 共五次最多，評分者 E 參加二次，評分者 C 與 D 均參加四次。此五位評分者包含四位華語教師，以及華測會口語測驗研發專員一名(評分者 A)，華語教師平均教學年資均超過十年。

考生人數方面，由於 10 月及 12 月預試使用同樣題目，研究者欲藉此進行評分者嚴格度的比較，故於 12 月評分時，隨機選取 20 位 10 月預試之考生錄音檔案，一併納入評分，為避免評分者憑先前一次評分的印象給分，事先未告知評分者。12 月實際預試人數為 63 人，加上重複評閱之 20 名考生，共計為 83 人。

表 1　TOCFL 基礎級口語測驗預試時間、考生人數以及評分者參與次數一覽表

考試時間	人數	評分者 A	評分者 B	評分者 C	評分者 D	評分者 E
2009/02	60	V	V	V		
2009/05	28	V	V	V	V	V
2009/07	49	V	V		V	V
2009/10	69	V	V	V	V	
2009/12	83	V	V	V	V	

二、測驗簡介

TOCFL 口語測驗是專為母語非華語之人士研發的一種外語/第二語言口語能力測驗，目前規劃有基礎、進階、高階以及流利四個等級。測驗形式為電腦化測驗，口語考試題目透過電腦螢幕與耳機播放，題目播放完畢後，給予考生準備時間，準備時間結束後，考生以麥克風回答問題，錄音檔案將記錄在電腦裡。

在基礎級測驗，題目分為二大部分，第一部份為暖身題，共有 3 題，目的讓考生熟悉測驗介面，故考生在此部分的回答不納入計分；第二部分為正式題目，共有 5 題，命題方向著重於描述個人經驗、表達對事物的喜好，以及回答與日常生活有關的話題等，考試時間約 30 分鐘。

計分方式採用 0 至 5 級分之整體式評分法，評分者依照內容組織、表達能力以及語言運用三大方向對考生回答內容進行評分，考生在每一題的回答均會得到一個成績，若考生沈默未作答或回答離題，則評為 0 級分，3 級分以上表示通過。

三、研究程序

TOCFL 口語測驗研發人員於 2008 年底舉行第一次評分研習，邀請教學經驗豐富之華語教師進行培訓，培訓後初步挑選出數名評分一致性較高，且能配合進行後續評分訓練及評分工作之華語教師做為核心評分者，開始長期的合作。2009 年五次評分訓練期間，研發人員亦陸續邀請曾於 2008 年參加過評分研習的教師擔任評分者，因有時在時間上難以配合，故部分評分者未能每次皆參加，評分者 E 因個人因素於 2009 年 7 月第三次評分訓練後，退出評分工作。

評分訓練流程可分為三個階段，第一階段測驗由研發人員簡介 TOCFL 口語測驗發展概況，以及說明基礎級考生的口語能力表現。第二階段測驗研發人員向評分者展示該次預試題目，以及各題各個級分的考生回答錄音檔範例(sample)，同時提供評分細則，說明各範例給分依據。第三階段給予評分者每題各數個考生錄音檔案進行評分練習，再公佈練習檔案之得分，並與評分者針對給分較不一致的考生表現進行討論，以確認評分者對於評分原則的掌握程度相當。2009 年 5 次評分訓練，皆依照上述流程進行。

三個階段評分訓練時間合計約為 4 至 5 小時，評分者於評分訓練後取得考生錄音檔案進行正式評閱，評分時間為一個月。待收齊評分者繳交之評分結果，研發人員隨即整理評分結果，挑選出給分歧異性較大的錄音檔案，加以反覆聆聽，再召集評分者針對這些檔案再作討論。如所有評分者給分相差 2 級分以上，或是分數介於通過與未通過臨界點的考生表現，進一步微調評分者給分的標準，目的在達成評分的一致性，整個討論約費時 4 至 5 小時。

肆、研究結果與討論

一、整體模式適配

資料與模式的整體適配度可以由非預期反應的次數得知，根據 Linacre(2010)，

大約 5%的標準化殘差等於或大於±2，大約 1%的標準化殘差等於或大於±3，表示資料與模式適配。由於本研究五次評分訓練的試題和考生並非完全相同，故評分結果均採用獨立分析的方式。五次分析資料與模式的適配情形，標準化殘差等於或大於±2 的比例介於 4.7%至 5.5%之間；等於或大於±3 的比例介於 0.3%至 0.9%之間，大致上均符合適配的標準，顯示資料與模式的達到良好適配。

二、評分者嚴格度與一致性

表 2 至表 6 為五次評分訓練後評分者的嚴格度估計結果，在嚴格度欄位，數值為正表示偏嚴格，負值表示偏寬鬆，數值越大表示給分越嚴格，越小則表示越寬鬆。五次評分訓練後評分者嚴格度的差異分別為 1.09 logits、0.93 logits、0.31 logits、0.32 logits 以及 0.31 logits，顯示經過幾次評分訓練後，評分者之間嚴格度的差異呈現出逐漸縮小的趨勢，然而由於五次分析分別為獨立的估計，每次估計得到的量尺(logit)未必相同，需再參考其他的指標。從 separation index 則可以得知，從第一次評分的 6.08，逐步降低到 3.30，甚至是 1.15、1.57 與 1.28，顯示在長期且密集的評分訓練之下，評分者之間整體的嚴格度越來越接近；此外，信度數值也從剛開始的 0.97、0.92，下降到 0.57、0.71 以及 0.62。不過在評分者整體嚴格度的卡方考驗，除了 2009 年 7 月未達到顯著外(p=0.08)，其餘四次的卡方考驗結果仍達到顯著水準，顯示整體來說，評分者的嚴格度雖然已趨於接近，評分者間一致性雖有所提升，但仍然未到達理想的水準。

在評分者內給分一致性方面，可以發現，即使採用較嚴格的標準(0.7-1.3)進行審視，五次評分結果，所有評分者的 infit 數值均落在適配的範圍之內(0.94-1.07、0.84-1.10、0.83-1.25、0.91-1.19、0.81-1.18)，表示評分者內給分一致性良好，評分者參與評分訓練後，實際進行給分時能保持相當的穩定性，不會過於偏離自身的標準。

上述分析結果與 Engelhard(1992)、Weigle(1998)、Bonk 與 Ockey(2003)、Park(2004)以及 Eckes(2005)的研究發現相同，支持評分訓練對於給分極端的評分者能改善其嚴格度，但不能完全降低評分嚴格度的差異，評分訓練無法使評分者間的嚴格度達到一致，但是有助於提升評分者內給分一致性的論點。

表 2　2009/02 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.38	0.10	0.98	0.99
B	0.33	0.10	0.94	0.92
A	-0.71	0.10	1.07	1.06

RMSE 0.10　Adj S.D. 0.61　Separation 6.08　Reliability 0.97
Fixed (all same) chi-sqaure: 76.8　d.f.: 2　sig: 0.00

表 3　2009/05 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
C	0.46	0.12	0.92	0.90
B	0.33	0.12	1.10	1.06
A	0.12	0.12	0.84	0.84
E	-0.44	0.12	1.07	1.07
D	-0.47	0.12	1.09	1.09

RMSE 0.13　Adj S.D. 0.41　Separation 3.30　Reliability 0.92
Fixed (all same) chi-sqaure: 47.2　d.f.: 4　sig: 0.00

表 4　2009/07 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
B	0.22	0.10	0.83	0.87
A	-0.04	0.10	0.86	0.84
D	-0.09	0.09	0.94	0.96
E	-0.09	0.09	1.25	1.30

RMSE 0.10　Adj S.D. 0.11　Separation 1.15　Reliability 0.57
Fixed (all same) chi-sqaure: 6.8　d.f.: 3　sig: 0.08

表 5　2009/10 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.18	0.08	1.19	1.17
C	0.05	0.08	0.92	0.92
A	-0.09	0.08	0.96	0.95
B	-0.14	0.08	0.91	0.91

RMSE 0.08　Adj S.D. 0.12　Separation 1.57　Reliability 0.71
Fixed (all same) chi-sqaure: 10.4　d.f.: 3　sig: 0.02

表 6　2009/12 評分者嚴格度估計結果

評分者	嚴格度	標準誤(S.E.)	INFIT MNSQ	OUTFIT MNSQ
D	0.14	0.08	1.14	1.14
B	0.03	0.08	0.82	0.82
A	0.00	0.09	0.81	0.81
C	-0.17	0.08	1.18	1.18

RMSE 0.08　Adj S.D. 0.10　Separation 1.28　Reliability 0.62
Fixed (all same) chi-sqaure: 8.6　d.f.: 3　sig: 0.04

三、評分者嚴格度的變化

由於 2009 年 10 月與 12 月的預試使用相同一套試題，研究者隨機挑選 20 名 10 月份考生口語錄音檔案，安插至 12 月份的評閱資料中，給評分者進行評分，藉由共同考生達到資料連結的目的。再將四位評分者二次的評分結果視為八位不同

評分者的評分結果(如：評分者 A 分為 A_10 與 A_12)，對 10 月與 12 月二次評分結果進行同時估計，以比較四位評分者前後二次評分的嚴格度變化情形。評分者嚴格度的變化如表 7 及圖 2 所示，四位評分者中，評分者 A 與 B 評分嚴格度較先前一次嚴格，評分者 A 嚴格度由原先的-0.11 提高為 0.01，評分者 B 由-0.16 提高為 0.04；評分者 C 則是變得較為寬鬆，嚴格度由 0.04 降為-0.15；評分者 D 嚴格度的變化最小，僅降低 0.04。

此結果和 Lumley 與 McNamara(1995)以及 Bonk 與 Ockey(2003)的發現一致，部分評分者嚴格度會隨時間改變，且變化的方向不同，有些評分者變得較嚴格，有些評分者則變得較寬鬆。本研究最後二次評分訓練與評分僅間隔二個月，評閱同一套試題，仍然有部分的評分者嚴格度產生變化，雖然變化幅度並非很大，但仍突顯出評分者很難維持相同的嚴格度。

表 7　評分者二次評分嚴格度變化

評分者	2009/10	標準誤	2009/12	標準誤	變化情形
A	-0.11	0.08	0.01	0.08	0.12
B	-0.16	0.08	0.04	0.07	0.20
C	0.04	0.08	-0.15	0.07	-0.19
D	0.18	0.08	0.14	0.07	-0.04

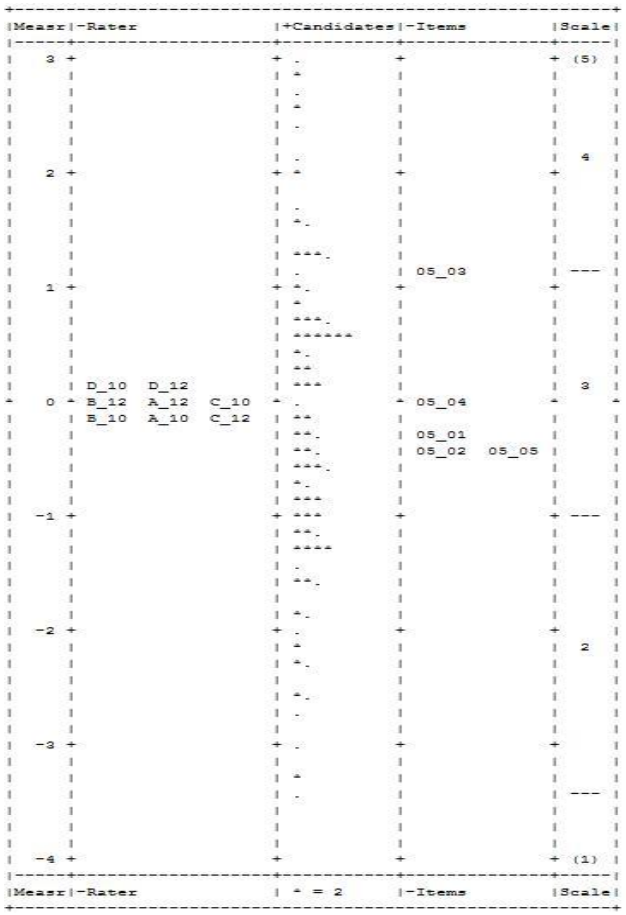

圖 2　FACETS 各面向對照圖

伍、結論

本研究結果顯示，密集且長期的評分訓練模式雖然有助於降低整體評分者嚴格度的變異，但無法完全消除評分者個別的差異，評分者間的一致性仍有待提升。但在評分訓練後，評分者內的一致性均符合 0.7-1.3 的範圍，顯示評分訓練對於評分者維持自身給分標準是有幫助的。此外，評分者在不同時期的嚴格度可能產生變化，不一定會維持同樣的嚴格度。上述研究發現顯示出評分者因素對於考生成績確實有影響，未來宜朝向以 MFRM 估計出考生能力參數後，再進行分數的轉換，取代傳統的原始分數，以提供更可靠的成績。因為謹慎挑選評分者以及密集的評分訓練仍然不能充分達到使評分者給分完全相等的目標，使用 MFRM 對於評分員給分的嚴格度進行統計上的調整，應是更為適當的作法。

本次研究參加的評分者人數較少，且因長期培訓，評分教師有時因個人因素無法每次皆參與，是較為可惜的地方。未來若能再增加評分者的人數，或許可以更明顯看出評分訓練對於評分者嚴格度的影響，當人數增加時，嚴格度的落差仍然與先前相當，或是變得更大。

此外，研究者觀察到 9 月第三次評分訓練之後接連三次的評分結果，無論是嚴格度差異、separation index 或是 reliability index 的數值，均達到一個較為穩定的狀態，這也許表示評分者參加二到三次的評分訓練後，對於評分原則的掌握有了更清楚的瞭解，也更勝任評分工作。若是如此，未來新進評分者，可能需要先具備二到三次的評分訓練經驗，才較為適合擔任正式評閱的工作。而這些評分者若再繼續參與評分訓練，彼此之間的嚴格度是否會更接近，或是維持現狀，亦是研究者所好奇的，未來可持續加以觀察。

本研究結果提供華語文口語測驗的實徵資料，對於華語文口語能力測驗的評分訓練效果有初步的瞭解，也能做為未來進行華語文口語能力測驗評分訓練及測驗實施的參考。

陸、參考文獻

田清源(2007)。HSK 主觀考試評分的 Rasch 實驗分析。**心理學探新，第 27 卷**，第 1 期，65-69 頁。民國 99 年 7 月 5 日，取自「中國期刊全文數據庫」 (DOI: CNKI:ISSN:1003-5184.0.2007-01-013)。

羅丹(2008)。**多面 Rasch 模型在 HSK(中級)口語評分檢驗的應用**。北京語言大學課程與教學論碩士論文，未出版。民國 99 年 7 月 5 日，取自「中國優秀碩士學位論文全文數據庫」(DOI: CNKI:CDMD:2.2010.046217)。

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion. *Language Testing, 20*(1), 89-110.

Du, Y., Brown, W. L., & Rogers, C. (1997, March). *Raters and single prompt-to-prompt equating using the Facets model in a writing performance assessment*. Paper presented at the Ninth International Objective Measurement Conference, Chicago,

IL.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages; Learning, teaching, assessment (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Engelhard, G. (1992). The measurement of writing ability with a Many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171-191.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*(2), 93-112.

Linacre, J. M. (2010). *A user's guide to Facets: Rasch-model computer program (Version 3.67).* Chicago: Winsteps.com

Liu, Y. L., & Wen, S. M. (2007). *Rating Reliability on the Assessment of Speaking Performance.* Proceedings of English Education and Inter-Discipline Learning, Shih Chien University, Taipei, 408-427,April, 28-29.

Lumely, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54-71.

Lynch, B. K., and McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4*(1), 1-21.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*, 72-92.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.