「華語文口語能力測驗」評分者評分偏誤研究
-以入門基礎級為對象-
(A FACETS analysis of rater bias in measuring TOCFL-Speaking assessment)

廖才儀
(Liao, Tsaiyi)
國家華語測驗推動工作委員會
(The Steering Committee for the Test Of Proficiency-Huayu)
julis8814@gmail.com

摘要：本研究旨在探討長期接受持續性評分訓練，且其評分者間信度與評分者內信度表現皆良好的評分者，其內部一致性是否存在評分者偏誤（rater bias）的現象，以及可能造成評分偏誤的因素。因而，本研究收集華語文口語測驗（Test of Chinese as a Foreign Language-Speaking，簡稱 TOCFL-Speaking）於 2014 年 11 月舉辦的入門基礎級正式考試評分結果，採用多面向 Rasch 測量模式分析軟體 Facets 3.71.3 版的部分給分模式，針對八位評分者對該次測驗四道描述類試題和四道回答問題類試題的評分資料進行分析。研究結果顯示：1.以 bias size 值介於-0.5 至 0.5，t 值介於-2.0 至 2.0 作為判斷標準，僅有兩位評分者其內部一致性受試題類別影響，存在顯著的評分者偏誤；2.整體觀之，評分者偏誤於描述類題型的發生頻率高於回答問題類題型；3.各評分者在不同試題的標準誤皆介於 0.015 至 0.039 之間，顯示出各評分者於每一道試題給分時，內部皆具有一定穩定性。現今以華語為第二語言的口語測驗研究相當缺乏，尤其是評分者偏誤研究。本研究結果提供華語文口語測驗的實徵資料，希望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。

Abstract: This research aimed to investigate whether or not a rater who has long-term sustainability rating training, good inter-rater reliability and intra-rater reliability shows rater bias in his intra-rater reliability. We collect the rating results from the regular test of the TOCFL Speaking (held at Nov., 2014) at Band A. 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) Bias size values between-0.5 to 0.5,t values between-2.0 to 2.0 as standard, only two raters effect its internal consistency to a test category, shows significant rater bias; 2) Overall, rater bias to "Describing" task types of frequency than "Answering questions" task type; 3) individual raters' S.E. of the different items ranged from 0.015 to 0.039, shows that each rater have a certain stability in his intra-rater reliability, and individual raters are consistent in their own rating as most of the fit statistics fall in the acceptable range. So far, only little research on the speaking assessments which focus on using Chinese as

a second language (CSL) has been done, especially at “rater bias”, 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.

關鍵字：口語測驗、評分者偏誤、多面向 Rasch 測量模式、主觀式評分

1. 研究動機與目的

華語文口語測驗（TOCFL-Speaking），是專為母語非華語之人士所設計的外語／第二語言口語能力測驗，為一採取電腦化形式[1]的表現測驗（performance assessment），受測者需以口說的形式，完成各項口頭溝通任務，任務的類型皆屬於建構反應題（constructed response item），因而受測者成績之取得主要仰賴評分者實際進行人工評分，評分辦法相對主觀；緣此，評分者的「評分一致性」遂成為影響受測者分數之主要因素，同時也是影響測驗信度的重要因素。

正因如此，TOCFL-Speaking 自研發以來即相當注重評分一致性，每次評分前，研發人員皆會召開評分會議，透過線上評分系統，進行嚴謹的評分培訓，確保評分者具穩定且良好的評分者間一致性（inter-rater consistency）及評分者內一致性（intra-rater consistency），再交由評分者進行正式評分工作。即使如此，仍有一些疑問有待確認；例如，同樣具備良好評分者內一致性的評分者，有可能是該評分者對每一道試題的評分，不分題型皆能按照評分規準進行評分，不存在偏分偏誤（rater bias）；但也有可能是該評分者內部存在因不同題型而異的評分偏誤，對某些題型評分偏嚴或偏寬。此外，研發人員亦觀察到，過去幾年所有評分者於評分系統輸入的評分結果及其評分依據，以及每次評分會議的討論情況，有部分評分者似乎會在進行不同題型的評分時，受題型指向的口語表達能力難易不同的影響，出現部分題型給分偏寬或偏嚴的現象。

綜上所述，本研究意欲探討接受長期且持續性嚴謹評分訓練的評分者，其內部一致性是否存在評分者偏誤的現象，以及可能造成評分偏誤的因素。藉由收集 2014 年 11 月舉辦的入門基礎級口語測驗正式考試評分結果，採用多面向 Rasch 測量模式（Many-Facet Rasch Measurement，以下簡稱 MFRM）分析軟體 Facets 3.71.3 版的部分給分模式（partial credit model），針對八位評分者對該次測驗四道描述類試題和四道回答問題類試題的評分資料進行分析。觀察評分者內一致性的變異情形，包含評分者內是否存在著顯著性的評分者偏誤（rater bias），並進一步探討評分者與試題間是否存在交互作用，檢視各評分者內部評分的一致性是否受試題類別因素影響，而在不同試題間存在主觀因素造成的評分偏誤。

2. 文獻探討

[1] 受測者於考試時，透過本會研發的口語考試系統聽到、看到事先錄製好的真人影音畫面，模擬真實口語情境，考生再透過耳機麥克風將口語回答內容直接上傳至考試系統。

前人研究（Eckes, 2008; Edward Schaefer, 2008; Johnson & Lim, 2009）指出，在表現型測驗中，即便是富有經驗，經嚴謹訓練過，且評分者間信度與評分者內信度表現皆良好的評分者，於其自身一致性內仍可能存在著評分者偏誤（rater bias），其偏誤情形值得探討。此外，針對此類評分者偏誤研究，多數研究者採用 MFRM 模式探討評分者與評分向度（category）、主題（topic）等的交互作用。

MFRM 模式，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）。

本研究欲探討評分者在面對不同試題時，是否存在不同的評分嚴格度，因而此次採用 MFRM 的部分給分模式進行分析，可供觀察評分者信度的統計指標包括分散指標（separation index）、信度（reliability），以及評分者給分一致性（rater consistency）的適配度統計值 Infit MNSQ。其中，適配度統計值的意義為觀察到的評分與預期評分結果的適配情形，也是運用 MFRM 模式以觀察評分者信度時最常使用的統計指標之一。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）。

3. 研究方法

3.1 研究對象

本研究於 2014 年 11 月至 12 月初進行 2014 年 11 月入門基礎級口語測驗八道試題的評分，參與此次評分的評分者共計 8 位（含一位研發人員），皆為長期參與口語測驗評分工作的華語教師。分為三組，除了研發人員需評閱所有受測者音檔外，每位評分者分配到的受測者人數為 49 至 50 位不等，每位受測者都有八道計分題。

3.2 測驗簡介

TOCFL-Speaking 係對應 CEFR，專為母語非華語之人士研發的一種電腦化外語/第二語言口語能力測驗，分為三等六級，目前所有等級皆已規畫完畢。在入門基礎級測驗中，題目分為三大部分，第一部份為暖身題，共有 2 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、三部分皆為計分題，前者為回答問題類題型，後者為描述類題目，共計 8 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。

計分方式採用 0 至 3 級分之整體式評分法，評分者需依照一套評分規準進行給分，評分規準內將口語表達分為內容組織、表達能力以及語言運用等三大

向度，評分者需同時考量這三個向度，再給予考生回答一個分數；因此，考生在每一題的回答均會得到一個成績；最後得分則為每題計分題分數的加總。

3.3 研究方法

透過線上評分系統收集八位評分者於 2014 年 11 月入門基礎級口語測驗八道試題的評分資料後，採用可分析 MFRM 模式之 Facets 3.71.3 版的部分給分模式進行分析，探討評分者與試題間是否存在交互作用，檢視各評分者內部評分的一致性是否受試題類別因素影響，而在不同試題間存在評分者偏誤。

4. 研究結果與討論

4.1 評分結果分析

4.1.1 評分者內信度分析

此部分採用 Facets 3.71.3 版的部分給分模式對資料進行分析，檢視評分者內信度，可由嚴格度標準誤（standard error；簡稱 S.E.）和統計指標 Infit MNSQ 觀察評分者自身給分穩定性。由表 1 分析結果可知，各評分者的嚴格度標準誤介於 0.057 至 0.093 之間，顯示出八位評分者給分均具有自身的穩定性，其中 A16 可能由於評閱人數較多，標準誤較其他七位評分者小。再由 Infit MNSQ 值介於 0.5 到 1.5 的標準，評估評分者自身給分一致性是否如模式所預期，結果顯示，所有評分者之評分者內一致性均佳，自身評分穩定性良好。

表 1：評分者內信度

評分者編號	評閱人數	平均值	嚴格度標準誤（S.E.）	Infit MNSQ
A10	49	0.94	0.093	1.09
A04	50	1.25	0.088	0.97
A08	49	1.05	0.091	0.84
A11	50	1.31	0.085	0.94
A02	50	1.35	0.084	0.90
A05	50	1.36	0.087	1.28
A17	49	1.11	0.089	0.80
A16	109	1.43	0.057	0.88

註：觀察的平均值表示評分者平均給分成績；Infit MNSQ 表示訊息加權適配度均方差。A16 為研發人員。

4.1.2 評分者內評分偏誤分析

此部分參考 McNamara（1996）研究作法，對試題與評分者交互作用分析結果的偏誤量（bias size）進行 t 考驗，以 bias size 值介於-0.5 至 0.5 之間，t 值介於-2.0 至 2.0 作為判斷標準，若超出此標準表示存在顯著的評分偏誤，將八位評分者與入門基礎級八道試題評分結果間的偏誤頻率統計如表 2；表格內，「S」表示「評分者於該試題給分比八道試題的整體性給分低，也就是內部給分偏嚴格」，同理，「L」則表示內部給分偏寬鬆；內部細格則以「數值/數值」的方式呈現給分偏嚴或偏寬的情形，舉例來說，編號 A04 評分者於第一題細格內所對應的數值為「1/0」，即表示 A04 第一題的給分標準相較於自身給分的整體標

準而言，為一偏嚴格的情形；若評分者於該試題給分未出現偏嚴或偏寬的情形，對應的細格內將不會出現任何數值資料。

表 2：評分者與入門基礎級八道試題評分偏誤頻率之統計表

題號（S/L） 評分者	自身評分偏嚴/偏寬情形統計								
	回答問題類題型				描述類題型				
	第 1 題	第 2 題	第 3 題	第 4 題	第 5 題	第 6 題	第 7 題	第 8 題	共計
A02									0/0
A04	1/0				0/1				1/1
A05			0/1	1/0					1/1
A08									0/0
A10			0/1	0/1	1/0	1/0	1/0	1/0	4/2
A11					0/1		1/0	1/0	2/1
A16									0/0
A17								0/1	0/1
共計	1/0	0/0	0/2	1/1	1/2	1/0	2/0	2/1	8/6

由表 3 的分析結果可知，八位評分者中，有五位評分者出現自身評分一致性會隨著試題類別而偏寬或偏嚴的情況，其中，評分者評分偏誤頻率較高者為 A11 與 A10 兩位評分者，A10 評分者八道試題中有六道試題隨著個別試題的類別或難度不同，而出現內部一致性偏嚴或偏寬的情形，A11 評分者則出現八道試題中有三道試題內部一致性偏嚴或偏寬的情形；A04、A05 和 A17 三位評分者的偏誤情形次之，各僅有 1 至 2 題出現偏嚴或偏寬的情形。再由試題類別對評分者評分偏誤的影響來看，在回答問題類題型中給分偏寬的情形有 3 例，偏嚴的情形有 2 例，共計 5 例；在描述類題型中給分偏寬的情形有 3 例，偏嚴的情形則有 6 例，共計 9 例，顯示出，評分者偏誤的情形似乎在描述類題型中較易發生。

另外，由各評分者在不同試題嚴格度的標準誤來看，各評分者對八道試題的標準誤介於 0.015 至 0.039 之間，顯示出八位評分者於進行每一道試題給分時，內部均具有一定穩定性，其中 A16 評分者可能由於評閱人數較多，八道試題的標準誤皆較小，介於 0.153 至 0.177 之間，而 A10 的標準誤較其他評分者大，介於 0.230 至 0.390 之間，其他六位評分者自身對八道試題評分的穩定性大致相當，皆介於 0.210 至 0.279 之間。再由統計指標（Infit MNSQ）介於 0.5 到 1.5 的標準，評估評分者於各道試題給分一致性是否如模式所預期，結果顯示，多數評分者之評分者內一致性均佳，自身評分穩定性良好。

4.2 討論與建議

由上述分析結果可知，2014 年入門基礎級口語測驗正式考試之評分者偏誤研究顯示，以 bias size 值介於-0.5 至 0.5 之間，t 值介於-2.0 至 2.0 作為判斷標準，若超出此標準表示存在顯著的評分偏誤，八位評分者中僅有兩位評分者的評分者內一致性受試題類別因素影響，而在不同試題間存在評分者偏誤。這兩位評分者分別為 A10 與 A11，研發人員推測，A10 評分者因近年教學經驗以高級班為主，故對入門基礎級口語能力的掌握較不穩定，導致該評分者整體內部一致

性雖適配，但自身變異性過大，八道試題中有六道試題給分標準不一致；另一位評分者 A11 自身給分一致性的不穩定情況，則集中在描述題類型的試題部分，試題難度較高的描述類試題給分偏嚴，試題難度較低的描述類試題給分則略為寬鬆，在回答問題類試題部分的自身評分一致性部分表現則穩定；研發人員據 A11 評分者提供的評分依據推測，該評分者於評分過程，在考量受測者內容組織向度中關於段落層次、任務完成度部分的表現時，會不自覺依試題難度調整給分。此外，評分者偏誤於描述類題型的發生頻率高於回答問題類題型；據研發人員推測，這幾位受題型類別影響而產生評分偏誤現象的評分者，於評分期間前後皆參與過其服務單位的學生成就測驗評量工作；成就測驗與能力測驗評量本質上的區別，可能導致評分者受這兩類題型所指向的口語能力、溝通任務的複雜度有明顯難易程度之別的影響，而較難穩定操作評分規準，以至於產生評分誤差。

未來，將針對評分者個別評分偏誤的情形，進一步與各評分者溝通，並做為下次評分會議對該評分者訓練的重點，藉由提高評分者內信度，便能穩定評分教師自己本身的一致性，這樣也有助於提升評分者間信度。此外，培養一個優良評分教師是一件很不容易的事情，我們將藉由後續工作觀察該教師自身的評分信度，或是進一步分析該教師每一題個別的評分信度，再作進一步的評分工作調整。

參考文獻

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.

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. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155-185.

Edward Schaefer. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

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

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

Gillian Wigglesworth. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

McNamara, T. F.（1996）. Measuring second language performance. New York：Longman.