網上平台新進評判訓練比對傳統評判訓練的教學效益探究

馮樹勳　孔慶強

1. 前言

過去語言測試的研究中，誠然測試者為研究主體，但 Bachman (1990) 明確地指出，語言測試中，任何參與者，不單是受試者本身，也包括評核者，會對測試造成影響。有關評分者的研究，則與測試者同樣多樣化，有對評分者與測試者間的互動作研究，如 Wigglesworth (1994) 所提出，評核者本身的行為，如對測試者的態度及回應等，均對評核構成重大影響。也有對評分者在評核不同測試題目類型時準確度的研究 ，如 Stansfield (1991)、Stansfield 和 Kenyon (1992) 、Shohamy (1994) 針對直接式 (Direct test) 與半直接式測試 (Semi-direct test) 對評斷準確度的影響。

至於評核者及其所接受的訓練對評分的影響，早已引起學者關注（如 Barnwell，1989；Elder，1993；Fayer 和 Kraskinki，1987；Galloway，1980；Hadden，1991；Brown，1995）。在語言測試的研究中， Shohamy，Gordon 和 Kraemer (1992) 便對筆試的評核者的培訓程度及專業背景兩個項目的交互影響作了探討，並比對曾經接受訓練和無訓練的評核者，除發現評核者間的可信度較高，更發現作為一個變項，評分者的訓練較其經驗更為顯著。Weigle (1994) 發現評核者的訓練改善了評核者的可信度，卻不能完全去掉個人評分時的嚴厲或寬鬆。此後亦有對此兩個項目的研究，發現評核者即使進行特定的評核訓練，個人的嚴厲或寬鬆仍然存在 (Lumley 和 McNamara，1995；Weigle，1998；Barrett，2001）。

Luoma (2004) 指出，由於口語測試涉及互動性質，需要真實的評判，因而在評分過程中出現的變數較多，而考試委員會一般為確保評核者的可信度，最常使用方法是提供評核者的訓練。訓練通常在課堂裡，首先透過錄影的片段給予評核者評分，有些可以帶回家做評分，第二堂會討論給予分數。在口語測試研究中，早在九十年代以前，Fayer 和 Krasinski (1987) 便測試其母語為英語，與母語非英語，兩類評核者，對其評核英語的影響。然而，他們的研究中，評核者並沒有受過任何導師或評核者的訓練。Barnwell (1989) 比較母語為西班牙，與一位經 ACTFL (American Council on the Teaching of Foreign Languages) 訓練的評核者，在評核一些來自美洲其母語非西班牙者的西班牙話，發現經過訓練的評核者評分時仍會持續寬鬆。Wigglesworth (1994) 測試受過全面訓練的評核者，在評

核現場面試與錄影兩種不同類型的口試，亦會出現不同項目有個別不同嚴厲的偏向。Meiron (1998) 探究口語測試 SPEAK 一個單項測試中，評核者的行為。Meiron 發現即使有特定的評分準則，評核者依然會慣用自己一套標準，而這些標準並未有在評分準則中出現。Brown (2000) 則同樣為口語測試進行研究，並指出，即使在給予指定的準則下，評核者也有不同測量方法和標準。Derwing 等人 (2004)發現未經訓練的評核者，在評核第二語言口語測試時，會運用一套約定俗的標準去評定流利度。

本團隊在 2006-07 年度，曾就 13 位新進的評核者，受試者能力及比賽項目兩個範疇，其給分與專業評判的準繩度作對比。該文發現上述兩個不同範疇，各自對新進評核者評分的準繩度構成影響。但不能證明這兩個範疇間，有明確的互動關係。就受試者能力兩個分項：高能力組與一般能力組，對新進評核者評分準確度的影響而論，則有足夠顯證說明高能力組別的偏差比率較低。研究結果支持互動式的口語表達，較諸獨白式的口語表述，不但受試者覺得前者難度較高，對新進評核者而言，亦是較難準確評分的項目。此外，語評核較諸書面語評核 (例如作文) ，評核者需要在十分有限的時間內，完成評核工作，因此涉及的表達元素愈多，則愈難保持評分的準確程度。

至於電子化平台，對評核訓練的影響，Campell(2005)指出，有關網上評核研究的文獻並不常見，現存研究多集中運用量表作網上評核。至於如何運用網上量表訓練新進評核員更是如鳳毛麟角，有待進一步探討。本文的目的之一，便是要探究利用網上平台協助新進評核者，能較否較傳統平台，有更佳的評核準繩度。

2. 研究方法

是次研究的要旨，主要包括以下兩方面：其一、延續前述 2006/07 年度對新進評核者評判準繩程度的研究，由於該研究的採樣較少 (n=13)，而且亦沒有作歷時性的趨勢對比，因此，研究團隊決定蒐集 2007/08 及 2008/09 兩個年度，分別對 41 及 78 位新進評核者的 1,179 筆和 1,494 筆評核紀錄進行研究，尤其著重於對互動式 (interactive) 與非互動式 (non-interactive) 項目，對評核準繩程度的影響。其二、團隊於 2008/09 年度採用電子平台作為評核訓練的媒界，對比於 2007/08 年度在特定的時間和地點，集中對新進評核者進行培訓，對評核的準繩度有沒有影響。

本研究計劃人員，分別在 2007-08 及 2008-09 年度，為「善言巧論：全港學生口語溝通大賽」訓練了一批新進的助理評核人員[1]。這些助理裁判的受訓者，大部分都是香港教育學院的學生。其性別比例分佈如下：

[1]該批評核人員將在比賽協助資深裁判評分，不會單獨進行裁判工作。

年度	男性人數	女性人數	實際出席人數	應出席人數
2007/08	9	22	31	41
2008/09	15	57	72	78

為免影響受試者，新進評判訓練不會以本年度受試者作為評核對象，本計劃分別採用 2006/07 及 2007/08 年，「善言巧論：全港學生口語溝通大賽」參賽者的現場錄影片段。為了使評核範圍適度收窄，評分採用九品制 (上上至下下) ，並以 9-1 分來表示 9-1 評級[2]。這些樣本片段的標準評級，參考比賽仲裁委員會的專家及資深中文科教師 (4 位中文學者及 20 位具經驗的中國語文科老師) 。

由於這批新進評核者，其背景都是教育學院學生，亦是初次受訓擔任助理裁判工作，因此排除了 Brown (1993)、Elder (1993) 和 Shohamy (1994) 等人述及的裁判個人及專業背景差異問題。同時，受試者與評核者，皆是以粵語作為母語者，亦避免了 Fayer 與 Krasinski (1987) 提出，評核者因並非以測試語言為母語者，可能產生的評核偏差。

針對使用錄像片段的問題，Shohamy (1994) 、O' Loughlin (1997)和 Kathryn (1997) 的研究都指出，使用間接的口語評核方式，與直接的口語評核模式，在針對語言及非語言互動部分，對評分準確度有實質的影響。然而，十年前的科技是以錄音片段為主，本計劃現時的取樣片段，則是以錄像片段為本。科技的進步，不會出現錄音帶不能傳達影象的問題，使間接評核者對受試者的語言反應及身體語言等，都有較全面的掌握。此外，本計劃選取的測試項目，縱有互動性項目，例如：討論，但評核者完全是旁觀者，不像面試一般，考官必須介入與考生溝通，也沒有面試中介人的存在，自然可以避免 Wigglesworth (1994) 提出考官個人行為，對考生表現可能構成的影響。

至於評核項目的類型，我們延續 2006/07 年度的研究，把基本類型分為互動式與非互動式兩類，但從前的研究，只針對兩種非互動式的評核項目：看圖說故事及口頭報告；及一種互動式的評核項目：小組討論，作出對比。但 2007/08 及 2008/09 年度的訓練中，團隊把非互動式的項目展至包括：即席演講、備稿演講、口頭報告、看圖說故事、備稿說故事等，較少即時性互動交流的項目；亦把互動式項目延伸至包涵：小組討論、生活交談、主題面試等，考生必須與其他參與者交流，並且有互動性溝通的項目。由於 2006/07 年度的研究已發現，對

[2]即把評分限制於由上上至下下的九等評級以內。最高評級：上上（以 9 分代表）、上中（以 8 分代表），如此類推，以至最低評級：下下（以 1 分代表）。

新進評核者而言，對互動式的項目的評核準繩度，低於非互動式項目。因此，研究團隊仍會朝著這個方面進行研究。

針對提高評核者個人可信度和不同評核者之間的信度，訂立評核準則和雙重評核人員無疑是可行的方法。如果要面對大量評核的習作，繁複的文件往來、整理和量表分數運算，卻是十分費時，而評核的質量和信度便較難控制 (Campell，2005) 。為了增加評核習作的數量，和提昇多個評核者的評核效能和信度，減去不必要的文書往來和數字運算，運用資訊科技設立網上評核平台已是刻不容緩。

網上評核平台的設計是希望達到兩個目的：

其一、增加評核效能：透過先進資訊科技，打破地域和時間的限制，讓評核者可隨時隨地通過網上平台評核中學生的口語能力。所有評核結果會數碼化，自動化計算量表分數及儲存分檔，讓評核者專注於評核。

其二、訓練新進評核人員，增加評核信度：面對大量的要評核的習作和大量的新進評核人員時，專業評核員可選取樣本習作，利用網上量表作出評核，並儲存於網上評核平台。其後新進評核人便可選取相同樣本和量表樣本進行評核。評核完畢後便與網上資料庫的專家評核結果對照，比較差異，自行調節，增加信度。

網上評核平台是由香港教育學院數碼錄像資源庫演化而成的。數碼錄像資源庫的創立，是源於學習組件 (learning objects) (Littlejohn，2003) 的概念所引發。理念是將教學數碼錄像片段視作基本學習組件，透過統一格式於網上發放，讓教學人員或學生於網上觀看、下載、剪裁並上載回資源庫共享 (Hung，So 和 Yip，2006) 。

經過數年的演變，數碼錄像資源庫加入學習社群的概念，支援教學實習。教育學院學生於實習期間，自組學習社群，將試教片段上載到資源庫作互享和討論，共同建構好的教學例子 (Hung，So 和 Yip， 2006)。自二零零七年起，數碼錄像資源庫開始支援全港中學生口語測試片段，並加入網上量表功能，使數碼錄像資源庫便轉化為網上評核平台。

源於學習組件的概念和數碼錄像資源庫的經驗，得知統一錄像格式的重要，故所有中學生口語測試錄像皆統一以視窗格式 wmv 儲存和發佈。評核者無論身處何地都能以微遠視窗作業系統觀看。每一條中學生口語測試錄像片段附有評核量表，評核員觀看完錄像片段後便可即時填寫量表，網上評核平台會自動計算該量表分數，並儲存於資料庫中。

數碼錄像資源庫經過了數年的運作，技術人員已累積了寶貴的數據管理經驗，亦由此寶貴經驗落實推行新進評核員的網上訓練。網上評核平台其中特色是專家評核資料庫的建立。經驗的評核人員會預先選取部份參賽學生口語測試錄像片段作為樣本，給予評核並存儲於專家評核資料庫中。如新進評核員初次進入網上評核平台，便會被指派樣本錄像片段進行評核，評核完畢便與專家評核結果對照，比較差異，自行調節。新進評核員經過反覆以樣本錄像片段進行評核和與專家評核結果反覆對照，不斷調節，新進評核員的信度便會加強。

由於研究涉及兩個不同的方面：其一、互動項目與非互動項目，對評核準繩度的影響，這可以使用 2007/08 及 2008/09 年度的資料，針對不同類型項目作對比研探；其二、2008/09 年度，採用電子化培訓平台，對評核的準繩度有無影響。由於 2008/09 年度較傳統評核訓練，增加了一項電子平台的因素，因此在處理研究數據時，我們把前述研究方向的順序，倒轉過來，即先探討 2008/09 年度，採用電子化評核培訓，相較 2007/08 年度傳統評核培訓有沒有影響。然後，在考量這些影響對數據可能做成的干擾，對不同類型項目作對比。

在第二部分的研究中，團隊除了蒐集評核的數據以外，尚引入了對 2008/09 年度的新進評核者，作了問卷調查（問卷詳附件一）。問卷內容主要包括：網上評核是否較方便；對網上平台的熟練程度；網上平台對執行評核的幫助；即時專業意見參考有助掌握評核尺度等。雖然，調查問卷以表述主觀印象為主，但以之印證客觀數據，有助對研究課題有更全面的把握。

根據研究前述次序，我們首先建立下述的空虛假設：

- 虛空假設 (出席率) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的出席率，並沒有改變；
- 虛空假設 (資料缺失率) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的資料缺失率，並沒有改變；
- 虛空假設 (綜合評核準繩度) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的綜合評核準繩度，並沒有改變；
- 虛空假設 (年度不同項目評核準繩度 1) ：2007/08 年度新進評核者，對互動項目與非互動項目的評核準繩度，並沒有不同；
- 虛空假設 (年度不同項目評核準繩度 2) ：2008/09 年度新進評核者，對互動項目與非互動項目的評核準繩度，並沒有不同；
- 虛空假設 (分年分項評核準繩度 1) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的互動項目評核準繩度，並沒有改變；
- 虛空假設 (分年分項評核準繩度 2) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的非互動項目評核準繩度，並沒有改變。

為了更為直觀地表述研究數據，對於出席率及資料缺失率的情狀，團隊將直接以百分率表列出來。至於涉及準繩度的探究，團隊邀請了專家對不同的項目及個別參賽者的表現，作了綜合性的評級，並把評分開列為以下九個不同的等級：

等級	分數表述
上上	9
上中	8
上下	7
中上	6
中中	5
中下	4
下上	3
下中	2
下下	1

這種九品制在本港公開考試的口試項目中，也有廣泛採用。把專家組的標準評分與新進評判對同一項目同一參賽者的表現作對比，即可得出評核準繩度的差距。為方便統計，凡是新進評核者的評分與專家評分完全相同的均以 0 來表述，凡與專家評分差距 1 個等級 (即+1 或-1) 均以 1 來表述，如此類推。此準繩度的差距，可以由最少距離的 0，到最大距離的 8 不等。2007/08 紀錄了 41 位新進評核者的 1,179 筆評核紀錄，並由之而有相同筆數的評核差距資料；至於 2008/09 年度，則紀錄了 78 位新進評核者的 1,494 筆評核紀錄，並由之而有相同筆數的評核差距資料。把這些綜合及不同分項準繩度的差距，團隊以 t-test 作顯著性的比較。

3. 研究結果報告

虛空假設 (出席率) ：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的出席率，並沒有改變。研究結果參考下表：

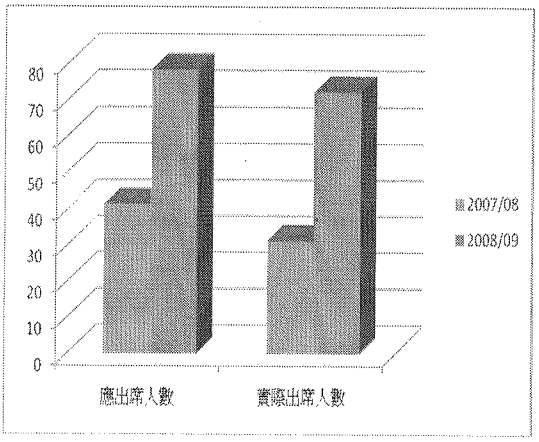
	應出席人數	實際出席人數	比率
2007/08	41	31	75.61%
2008/09	78	72	92.31%

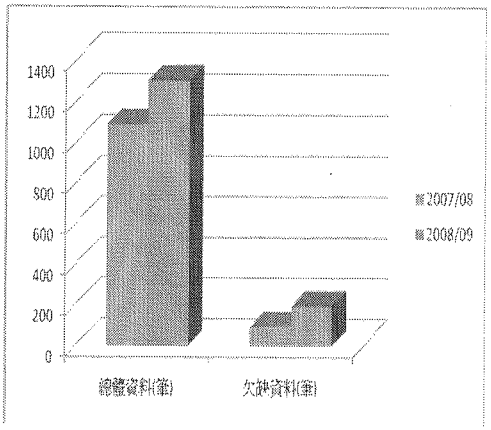
從上表看來，2008/09 年度的新進評核者的出席率，較 2007/08 年度的新進評核者的出席率，明顯地提高了 16.7%。

虛空假設（資料缺失率）：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的資料缺失率，並沒有改變。研究結果參考下表：

	總體資料(筆)	欠缺資料(筆)	比率
07/08	1086	93	8.56%
08/09	1299	195	15.01%

從上表看來，2008/09 年度的新進評核者的資料缺失率，較 2007/08 年度的新進評核者的出席率，提高了 6.45%。綜合前述兩項，可開列成圖表如下：





針對 2008/09 年度新進評核者，調查問卷的結果如下（累積百分比）：

題目內容	極度同意	同意	不同意	極不同意
本人認為網上評核訓練較面授評核訓練方便。	10.1%	71%	97.1%	100%
本人已經完全熟習地使用網上評核訓練平台。	10.1%	72.5%	97.1%	100%
本人認為網上口語訓練平台即時提供的專業意見，有助本人掌握評核尺度。	4.3%	85.5%	100%	100%
本人認為網上口語評練平台，能有效提昇本人的口語評鑑能力。	4.3%	85.5%	100%	100%

虛空假設（綜合評核準繩度）：採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的綜合評核準繩度，並沒有改變；採用獨立樣本、單尾的 t-test 作年度綜合評核準繩度對比，以下為研究結果摘要表：

2007/08 與 2008/09 年度綜合評核準繩度對比

	平均數(Mean)	範圍(Range)	數目(N)	有效程度(Sig.)
2007/08	1.2399	0-8	991	0.11

2008/09	1.3468	0-8	991

註：*表示中位數 p<.05 *** p<.01

從上表的數據看來，虛空假設並不能予以否證，即採用電子化評核培訓(2008/09)年度，相較傳統定時定點評核(2007/08)年度的綜合評核準繩度，並沒有顯著差異。

由於 2007/08 及 2008/09 綜合評核準繩度，並無顯著差異。我們進入下一層次的驗證，即就不同年度的互動項目與非互動項目的評核準繩度，作出對比。首先，我們察看，虛空假設 (年度不同項目評核準繩度 1) ：2007/08 年度新進評核者，對互動項目與非互動項目的評核準繩度，並沒有不同。採用獨立樣本、單尾的 t-test 作年度綜合評核準繩度對比，以下為研究結果摘要表：

2007/08 年度互動式與非互動式項目評核準繩度對比

	平均數(Mean)	範圍(Range)	數目(N)	有效程度(Sig.)
互動式	1.2225	0-8	309	0.44
非互動式	1.1387	0-8	309	

註：*表示中位數 p<.05 *** p<.01

從上表的數據看來，虛空假設並不能予以否證，即就 2007/08 年度的數據而論，互動式與非互動式項目間評核準繩度，並沒有顯著差異。

接著，我們驗證虛空假設 (年度不同項目評核準繩度 2) ：2008/09 年度新進評核者，對互動項目與非互動項目的評核準繩度，並沒有不同。採用獨立樣本、單尾的 t-test 作年度綜合評核準繩度對比，以下為研究結果摘要表：

2008/09 年度互動式與非互動式項目評核準繩度對比

	平均數(Mean)	範圍(Range)	數目(N)	有效程度(Sig.)
互動式	1.1674	0-8	261	0.00***
非互動式	0.8931	0-8	261	

註：*表示中位數 p<.05 *** p<.01

從上表的數據看來，虛空假設已被否證，即就 2008/09 年度的數據而論，互動式與非互動式項目間評核準繩度，具有統計上的極度顯著差異 (p<0.01)。

再從另一組虛空假設而論，探討的是分年分項的評核準繩度。首先檢視的是互動性項目方面，兩個不同年度的評核準繩度。其虛空假設 (分年分項評核準繩度 1) ：採用電子化評核培訓 (2008/09) 年度，相較傳統定時定點評核 (2007/08) 年度的互動項目評核準繩度，並沒有改變。採用獨立樣本、單尾的 t-test 作年度綜合評核準繩度對比，以下為研究結果摘要表：

2007/08 及 2008/09 年度互動式項目評核準繩度對比

	平均數(Mean)	範圍(Range)	數目(N)	有效程度 (Sig.)
2007/08	1.2852	0-8	681	0.84
2008/09	1.2713	0-8	681	

註：*表示中位數 p<.05 *** p<.01

從上表的數據看來，虛空假設並不能予以否證，即就 2007/08 年度及 2008/09 年度的數據而論，互動式項目的評核準繩度，並沒有顯著差異。

接著，我們針對非互動式項目，進行同類型的檢證。虛空假設 (分年分項評核準繩度 2) ：採用電子化評核培訓 (2008/09) 年度，相較傳統定時定點評核 (2007/08) 年度的非互動項目評核準繩度，並沒有改變。採用獨立樣本、單尾的 t-test 作年度綜合評核準繩度對比，以下為研究結果摘要表：

2007/08 及 2008/09 年度非互動式項目評核準繩度對比

	平均數(Mean)	範圍(Range)	數目(N)	有效程度 (Sig.)
2007/08	1.1221	0-8	261	0.02*
2008/09	0.8931	0-8	261	

註：*表示中位數 p<.05 *** p<.01

從上表的數據看來，虛空假設已被否證，即就 2007/08 及 2008/09 年度的數據而論，非互動式項目間評核準繩度，具有統計上的顯著差異(p<0.05)。

4. 討論

是次研究所蒐集材料，2007/08 紀錄了 41 位新進評核者的 1,179 筆評核紀錄，並由之而有相同筆數的評核差距資料；至於 2008/09 年度，則紀錄了 78 位新進評核者的 1,494 筆評核紀錄，並由之而有相同筆數的評核差距資料。較之團隊採用

2006/07 年度新進評核者 13 人共 39 筆評核資料，在資料量數方面有明顯的增加。因之，從統計角度來看，是次研究的可推擴性，亦當較前為增加了。

針對第二節的虛空假設及第三節的相關數據，團隊認為就出席率、資料缺失率、互動與非互動項目的評核準繩度等，數個不同方面當作更詳細的討論。

首先，採用電子化評核培訓 (2008/09) 年度，出席率為 92.31%，相較傳統定時定點評核 (2007/08) 年度，出席率為 75.61%，高出 16.7%。其增長程度十分明顯，考其原因，大抵傳統的定時定點訓練方式，在繁忙的學習生活中，較難同一時間聚集所有的新進評核者，進行集中訓練。以電子化方式作培訓，則有較大的彈性，使參與評核者的「出席率」，得以提高。從問卷調查中，取得的數據亦有 71%的新進評核者，同意或極度同意「網上評核訓練較面授評核訓練方便」的說法，則從方便評核者的角度看來，電子平台確有較傳統評核方式為優。

其次，採用電子化評核培訓 (2008/09) 年度，資料缺失率為 15.01%，相較傳統定時定點評核 (2007/08) 年度，資料缺失率為 8.56%，竟然高出 6.45%。就常識而論，使用電子化平台，應當比使用傳統的紙筆方式，能保存更多的資料。易言之，資料缺失比率，應是較低而非較高。2008/09 年度參與的新評核者，人數達 78 位之眾，很難把現象歸因於個別評核者的不小心行為。團隊仔細覆檢整個流程及數據庫內其他資訊，發現 2007/08 的傳統紙筆模下，新進評核者幾乎沒有提供個人評語，而採用電子化評核培訓的 2008/09 年度，個人評語則有 81 則，雖然僅佔評核項的 5.5%左右，但較前已有一定進步。是以，最有可能是由於新進評核者對電子化平台的操作，尚未全然熟習，以致有較高的資料缺失率。再以問卷調查的數據作印証，雖然認為自己「已經完全熟習地使用網上評核訓練平台」的新進評核者，有 71%之多，但認為極度熟悉者，則僅達 10.1%而已。這些數據說明，在未來使用電子評核平台時，可能需要更長的訓練時間。

再者，就提高新進評核者的個人信心方面，電子平台提供的即時專業意見，應有助提昇新評判們的表現。累計有 85.5%的評判，同意或極度同意「網上口語訓練平台即時提供的專業意見，有助本人掌握評核尺度」。同樣地，累計有 85.5%的評判，同意或極度同意「網上口語評練平台，能有效提昇本人的口語評鑑能

力」。是故，從主觀印象而論，則對新進評判而言，電子平台有助提昇他們的專業評鑒能力。主觀印象良好，未必即表示能客觀地提高評核的準繩度，團隊試以下述數據，說明使用電子平台以後，新進評核者的客觀表現，是否與主觀印象吻合。

就綜合評核準繩度而論，我們以 t-test 對比 2007/08 及 2008/09 年度，綜合評核準繩度，發現 p=0.11，虛空假設並不能予以否證，即就數據而論，兩個年度的綜合評核準繩度，並沒有顯著差異。由是我們認為，一般而言電子化平台對評核準繩度，應無重大影響。這也方便團隊，延續探討 2006/07 年以來，即予關注的互動式與非互動式項目，對評核準繩度的影響問題。

就數據顯示，2007/08 年度互動式與非互動式項目間評核準繩度對比 p=0.44，兩種項目並沒有顯著差異，即虛空假設並不能予以否證。這與 2006/07 年度的發現，有一定的差別。但就 2008/09 年度的數據而論，互動式與非互動式項目間評核準繩度對比 p=0.00***。從這項數據看來，虛空假設已被否證，即具有統計上的極度顯著差異 ($p<0.01$)。是則，團隊建議就這個項目的研探，作更長久的歷時性研究，當可獲得更為準確的結果。

至於就互動與非互動分項的評核準繩度，作跨年度的對比方面。就互動性項目而言，2007/08 年度及 2008/09 年度的數據而論，互動式項目的評核準繩度 p=0.84，並沒有顯著差異，即虛空假設並不能予以否證。另一方面，就非互動性項目而論，2007/08 及 2008/09 年度的數據顯示 p=0.02*，從這項數據看來，虛空假設已被否證，即具有統計上的顯著差異 ($p<0.05$)。從這些統計事實看來，則電子化平台以後，似乎對非互動項目評核的準繩度，有一定程度的提昇。當然，團隊於未來，作更多的歷時性研究，以確定這方面的理解。

5. 参考書目

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Barnwell, J. (1989). "Naive" native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

Brown, A. (1993) The effect of rater variables in the development of an ocupation-specific language performance test. Paper presented at the annual Language Testing Research Colloquium, Cambridge, U.K., August 2-5.

Brown, A. (1995). The effect of rater background variables in the development of an occupation-specific language performance test. *Language testing 12*(1), 1-15.

Brown, A. (2000). An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (ed.), *Research reports* (1999, Vol. 3, pp. 49-85). Canberra, Australia: IELTS Australia.

Campbell, A. (2005). Application of ICT and rubrics to the assessment process where professional judgement is involved: the features of an e-marking tool. *Assessment & Evaluation in Higher Education, 30*(5), p529-537.

Derwing, T.M., Rossiter, M.J., Munro, M.J. & Thomson, R.I. (2004). Second Language Fluency: Judgments on Different tasks. *Language Learning 54*(4), pp. 655-679.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing 10*(3), 235-254.

Fayer, J.M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313–326.

Galloway, V. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-33.

Hadden, B.L. (1991). Teacher and non teacher perceptions of second language communication. *Language Learning, 41*(1), 1-24.

Hung, H.K., So, W.M. & Yip, Y.M. (2006). "數碼錄像資源庫—從單打獨鬥到網上建構". Proceedings of the Hong Kong International IT in Education Conference. Hong Kong. February 6-7, 2006.

Littlejohn, A. (2003). Issues in reusing online resources. In Littlejoin, A. (eds.), *Reusing online resources: a sustainable approach to e-learning* (pp.1-8). Creative Print and Design.

Lumley, T., & McNamara, T.F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing 12*(1), 54-71.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Meiron, B.E. (1998). Rating oral proficiency tests: A triangulated study of rater thought processes. Unpublished master's thesis, University of California at Los Angeles.

O'Loughlin K. 1997. Test-taker performance on direct and semi-direct versions of the oral interaction module. In Brindley G. & Wigglesworth, G. (eds.) *Access: Issues in language test design and delivery (pp117-145).* Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Shohamy, E. (1994). The validity of direct versus semi–direct oral tests. *Language Testing, 11*(2), 99–123.

Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The effect of rater' background and training on the reliability of direct writing test. *The Modern Language Journal, 76*(1), 27–33.

So, W.M., Hung, H.K. & Yip, Y.W. (2008). The Digital Video Database: A virtual learning community for teacher education. *Australasian Journal of Educational Technology 24*(1), 73-90.

Stansfield, C. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (Eds.), *Current developments in language testing* (pp.199–209) .Singapore: RELC.

Stansfield, C., & Kenyon, D. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*(3), 347–364.

Weigle, S. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

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

Wigglesworth, G. (1994). Patterns of rater behavior in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17*(2), 77-103.

馮樹勳〈2008〉:〈口語評核類型和受試者能力對新進評核者評斷準繩度的影響〉，香港大學：9th ALA Conference，2008 年 6 月。