「華語文口語能力測驗」評分者間信度檢視

國家華語測驗推動工作委員會

摘要

本研究旨在檢視「華語文口語能力測驗」評分規準（rubric）的使用成效，做為後續研發工作的進行基礎，進而達到提高評分教師信度、並做為修正評分規準的參考。華測會口語能力測驗的評分方式是以整體式（holistic）為主，分析式（analytic）為輔，提供評分者三大分項（內容組織、表達能力、語言運用）為基準進而評定考生口語能力的整體表現。本研究藉由肯德爾（the Kendall coefficient of concordance）和諧係數與斯皮爾曼等級相關（Spearman rank correlation）以基礎級和進階級五月及七月兩次預試的評分資料，求得評分者間信度（Inter-rater reliability）數據，以了解評分者的評分表現。肯德爾統計結果顯示兩次結果均落在0.7-0.8之間，表示評分者間的評分信度介於良好與相當良好之間；斯皮爾曼等級相關統計結果亦顯示評分者與本會的一致性也相當高。這說明了本會評分規準實際使用的狀況相當良好。然而，為使口語評分工作更臻完美，後續工作包括挑選更多各等級樣本、分析評分者內信度（Intra-rater reliability）、蒐集更多口語樣本以供分析以及口語規準的修訂。期盼藉由上述規劃，讓口語評分品質更穩定。

一、研究動機與目的

本研究旨在檢視「華語文口語能力測驗」評分規準（rubric）的使用成效。評分為口語測驗的重點工作之一，口語評分為主觀評分，評分者的訓練以及評分規準的使用都是影響評分品質的重要因素。

本會所研發的「華語文口語能力測驗」是專為華語學習者所設計。本測驗基於華語學習者的實際口語需求，以「溝通任務」為導向，在命題方面，力求貼近於真實情境中需要達成的各種溝通任務；在評量方面，著重於考察考生能否在特定語境下，藉由口語表達，有效地傳遞訊息。適用對象相當廣泛，包括想要瞭解自己華語文口語程度，或是想要在華語地區求學或工作之人士。考生可藉由參加本測驗，了解自身的華語口語程度，並作為進一步學習的參考。

整套測驗共分四個等級，分別為「基礎級」、「進階級」、「高階級」和「流利級」。及至目前，「基礎級」和「進階級」已舉辦過六場預試；「高階級」和「流利級」則尚在資料蒐集階段。基礎級題目著重經驗的描述以及對事物的喜好，並回答與日常生活有關的話題；進階級題目著重情感表達，並且能根據所提供的資料說出自己的意見和想法。各等級皆採用1-5級分之整體式評分法（holistic scoring），3分為通過門檻。在評量方面，著重於考察考生能否在特定語境下，藉由口說有效地傳遞訊息。

由於測驗目前尚在研發階段，為確保明年舉辦正式口語能力測驗的評分品質，為此，本研究藉由實際的評分工作，將蒐集到的資料加以分析，希望做為後續研發工作的基礎，進而提高穩定評分教師信度，亦做為修改評分規準的參考。

二、文獻探討

在建構評分規準之前，必須對主觀式評分有相當的認識，主觀式評分主要分為兩大類：一為分析式評分，一為整體式評分。分析式評分顧名思義就是將考生的口語表現分為數個項目進行給分；整體式則恰好相反，整體式評分是對考生的口語表現給予一個單一的分數。這兩種方式各有優缺點。

在語言測驗中，對於整體式和分項式評分方式的優劣，許多學者皆有精闢的分析解說與討論（Bachman, 1988; Bachman & Savignon, 1986; Douglas & Smith, 1997; Fulcher, 1997; Ingram & Wylie, 1993; Underhill, 1987; Weir, 1990）。綜觀目前各個口語能力測驗，多以整體式評分方式評定考生的口語能力。其原因除了整體式評分方式較能評定考生整體的口語表達溝通能力之外，Weir特別提到整體式評分方式較分項式評分方式更為省時，且對於評分者而言，整體式評分方式較容易依循。然而，整體試評分方式卻也隱藏許多問題。

首先，整體式評分原則所提及的口語能力的定義往往是不明確的。此外，評分者對於評分原則中包含的不同評分重點常常有不同的比重。Brown（1995）指出，不同評分者因為自身背景的不同以及對整體口語溝通能力的解讀不一，為造成評分信度不佳原因之一。而Bachman (1996: 209)亦指出global scoring的三個主要問題：（一）單一分數難以反應包含多種面向的口語表達能力；（二）評分者難以決定考生程度；（三）評分者心中對口語表達能力的評分比重不同。有鑑於上述學者提出的見解，以及配合目前各大考試的口語規準，這些考試包括全民英檢、拖福、多益等。

最後本會採用以整體式為主，分析式為輔的評分量表建立評分規準。主要乃是希望保留這兩種量表的優點，盡量避免這兩種量表的缺點。以下是融合這兩種量表的優點簡述：（一）能評定考生整體的口語表達溝通能力。現在無論是教學還是測驗，溝通式或是任務型的題目是主要命題方向，我們希望學生能夠完成某項任，希望能夠評估他們的整體表現，所以整體式的評分方式便能夠達到這個目的。（二）評分者容易依循。研究顯示使用整體式評分量表評分的評分教師在使用上容易上手。（三）方便省時。因為評分教師使用整體式容易上手，因此也就能讓評分工作更有效率。（四）可提供使用者更多資訊。這裡的使用者包括考生、使用本測驗的機關學校等。分析式之前介紹過，就是將考生的表現分項給分，這樣的方式能夠提供考生更完整的資訊，假設某考生希望能通過本會的基礎級測驗，但是考了兩次都沒有通過，若我們提供考生更詳細的成績分析，該名考生便可以對自己較弱的部分加強練習，又或者是某學校使用我們的測驗，目的是希望能將選修華語會話課程的學生依照不同的能力分班，便可以依照學生考完本會考試的成績中的詳細分析資訊進行分班作業。（五）避免評分者在評分時標準不一。也就是說，譬如A評分教師與B評分教師對同一個語音樣本都給予3級分的成績，表面上這兩位給分是一致的，不過在評分過程中，A教師著重的是考生選詞用字的部分，而B教師則是著重語音部分，所以這兩位評分教師的評分信度表面上很漂亮，但仔細探究之後，兩位的評分標準是很不一致的，也因此，融合分析式評分方式，便可以在這個部分做一個把關，讓評分教師能意識

到評分時應該盡量不偏向任何一個部分。

所謂以整體式（holistic）為主，分項式（analytic）為輔的評分方式意思是評分教師依據考生在內容組織、表達能力、語言運用三項的表現，給出一個整體口語表現成績。詳細評分規準請參考附錄一。

選定評分規準的形式之後，接下來就是進行評分規準內容的撰寫。根據Fulcher（2003：91, 92）提到口語測驗評分原則（rating scale）的建立主要有兩種方式：一為直觀式（Intuitive methods），一為實證式（Empirical methods）。本會依照Fulcher提到的直觀式評分量表建立流程進行本會的評分規準建立工作。首先經由專家的判斷，草擬出一套量表，接著邀請學者專家組成委員會進行初步修訂，接著進入實際使用的階段，經過實際的使用，逐步修正。以下是較詳細的說明：

- 專家的判斷（Expert judgment）－所謂的專家包含有經驗的老師或是語言測驗研發人員，藉由他們的專業判斷，依照該測驗的性質，參考其他測驗的評分原則、教學計畫或是需求分析等資訊，草擬出評分原則。
- 委員會（Committee）－邀請數位專家學者組成諮詢委員會，一起討論評分原則用字遣詞的適當性。
- 經驗累積（Experiential）－評分原則建立之後，評分者經由使用之後，慢慢加以修改。此方式為目前最常見的評分原則建立的方式。

本會草擬出的評分原則根據諮詢委員的建議做了初步的修正，並且實際運用到上述六次的預試評分工作當中。因此藉由本研究分析評分教師間的評分信度，可以對此套評分規準的使用狀況作初步評估，做為進行後續修改的依據。

三、研究方法

擬定評分原則之後，接著進行實際的口語測驗和評分工作，接著舉辦評分會議挑選各級分樣本，然後請評分教師進行評分工作。本研究對象為華語文口語能力測驗的評分教師，目前本測驗聘請8位資深華語教師擔任口語評分核心教師，主要工作包含與本會研究員共同挑選各級分標準樣本，實際使用評分原則進行評分工作，合作相關研究，以便於日後擔任口語評分教師工作的帶領者。此外，評分教師皆為華語教學經驗豐富的華語教師。

評分流程的進行方式簡述如下：我們將評分教師分為評定基礎級與進階級兩組，各四位教師。每次評分工作會舉辦兩次評分會議，第一次主要是做norming的工作，藉由讓評分教師試評，統一大家的評分標準，接著評分教師進行評分之後，再舉辦第二次評分會議，進行討論，一方面再次調整不一致的部分，一方面再次確定各級分樣本。

本研究所分析的資料是今年五月和七月的考試，基礎級共有77位學生，進階級共有82位學生。基礎級試題包含兩個部分，第一個部分為暖身題，共三題，做為考生練習用，不予計分；第二部分為描述題，共五題。進階級試題包含三個部分，第一部分為暖身題，共三題，做為考生練習用，不予計分；第二部分為描述題，共三題；第三部分為說明題，共三題。簡表如下：

等級	基礎級	進階級
第一大題	暖身題三題	暖身題三題
第二大題	描述題五題	描述題三題
第三大題		說明題三題

經由統計算出基礎級四位評分教師，和進階級四位評分教師之間的評分信度。接著又將每位評分教師的評分結果與本會預設的評分結果加以算出一致性。也就是說透過Kendall和諧係數分析（the Kendall coefficient of concordance），檢測評分教師間評分信度（inter-rater reliability）。透過斯皮爾曼相關分析(Spearman rank correlation)，檢測評分教師與華測會訂定的標準間的信度。

四、研究分析與討論

表1和表2是五月份基礎級和進階級的統計分析結果。八位評分教師分別以A~H英文字母為代號。參閱對於口語測驗有專門研究的學者著作,尚無研究顯示確切的評分者間信度的標準為何，故此次便以分析後所得數據相互比較，檢視本會口語測驗評分者間信度之成果。

此兩個表格左邊的欄位是以Kendall和諧係數分析評分教師間的評分信度。「基礎級」四位評分教師間的評分信度，統計結果為0.899，這顯示基礎級四位評分教師的一致性相當高；而「進階級」的統計結果為0.803則為評分教師評定考生描述題的評分信度，也算是很不錯，然而說明題的統計結果0.718，則還有提升的空間。

而右邊的欄位是以Spearman相關係數分析四位評分教師分別與華測會預設的評分標準的一致性。「基礎級」整體來說統計分析結果都在0.7以上；「進階級」除了G教師在評定說明題的一致性差強人意之外，也都落在0.7以上。

表1　五月份基礎級

	Kendall和諧係數分析 評分教師間信度	Spearman相關係數分析 四位評分教師與華測會間信度			
描述題	0.899**	A	B	C	D
		0.865**	0.865**,	0.796**,	0.862**

表2　五月份進階級

	Kendall和諧係數分析 評分教師間信度	Spearman相關係數分析 四位評分教師與華測會間信度			
描述題	0.803**	E	F	G	H
		0.728**,	0.822**,	0.864**,	0.886**
說明題	0.718**	E	F	G	H
		0.728**,	0.706**,	0.693**,	0.748**

表3和表4是七月份的統計分析結果，A~C為「基礎級」三位評分教師代號，E~G為「進階級」三位評分教師代號。此次評分工作，兩等級各有一位評分教師因故無法參與。

無論是「基礎級」或是「進階級」評分信度大都落在0.7以上，整體評分信度算是相當不錯。

兩次考試的評分者間信度絕大部分都在0.7以上，本會預期目標是希望能將評分者間的信度未達到0.85的，提升至0.85，已經落在0.85的，維持0.85以上的水準。

表3　七月份基礎級

	Kendall和諧係數分析 評分教師間信度	Spearman相關係數分析 四位評分教師與華測會間信度
描述題	0.833**	A　　B　　C 0.715**, 0.865**, 0.681**,

表4　七月份進階級

	Kendall和諧係數分析 評分教師間信度	Spearman相關係數分析 四位評分教師與華測會間信度
描述題	0.712**	E　　F　　G 0.733**,　0.720**, 0.674**
說明題	0.778**	E　　F　　G 0.856**,　0.757**, 0.710**

五、小結

根據本研究的統計結果顯示目前口語評分整體表現大多在0.7以上，希望藉由日後更多的評分經驗，將本會評分教師評分信度保持在0.85以上。為了達到此目標，本會將此研究結果請益諮詢委員，委員主要建議如下：

首先我們要做的是增加各級分的樣本數，樣本的挑選不只是典型的，還必須包括非典型的。口語測驗有別於以選擇題形式施測的測驗，考生於口語測驗中的回答包羅萬象，所以樣本的選擇應該包含典型與非典型，才能夠提供評分教師更多資訊，進而達成評分共識，例如對於所謂答非所問的標準該如何統一。

再者，應著手進行評分者內評分信度的分析，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。例如「進階級」評分教師G，0.693的結果顯示該教師與本會的評分信度不是很理想，所以針對該教師的資料，我們進一步分析每一題的評分信度，由表5至表7發現該教師於說明題第一題和第二題的評分信度皆有0.7以上的表現，而第三題評分信度則較不理想。此數據便可做為下次評分會議對該教師訓練的重點。

表 5　第一題斯皮爾曼等級相關

第一題	教師 G
華測會	0.787**

**表示 p＜0.01

表 6　第二題斯皮爾曼等級相關

第二題	教師 G
華測會	0.700**

**表示 p＜0.01

表 7　第三題斯皮爾曼等級相關

第三題	教師 G
華測會	0.529**

**表示p＜0.01

而累積更多的口語評分樣本絕對能使統計分析結果提供更有意義的資訊。最後，影響評分信度的變數之一也包含評分原則的敘述精確與否，故我們也將同時進行評分原則的修訂工作，期盼提升本會口語評分工作品質。

附錄一

評分原則

基礎級

級分	內容組織	表達能力	語言運用
5	內容與題目要求相關，內容豐富，話語多有組織	語速適中，偶有停頓；詞語重複次數甚少；發音大致清楚，偶有偏誤，聽者都能理解	已能掌握基本詞彙和語法，仍有少許偏誤
4	回答與題目要求相關，內容尚稱豐富，話語尚有組織	語速適中，常有停頓；詞語重複次數少；語音尚稱清楚，偶有偏誤，聽者幾乎都能理解	已能大致掌握基本詞彙和語法，仍偶有偏誤
3	內容與題目要求相關，內容稍嫌不足，話題無法擴展，組織較差	語速緩慢，常有停頓；詞語重複次數多；部份語音不正確，聽者尚能理解	詞彙量仍有限，使用大致適當，尚能掌握基本語法結構，仍常常有偏誤
2	內容與題目要求相關但不完整，內容不足，組織甚差	語速過於緩慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解	詞彙量有限，且多不適當，尚未掌握基本語法結構
1	內容與題目要求不相關	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解	詞彙量極有限，且不適當，無基本語法結構
0	考生靜默，沒回答		

評分原則

進階級

描述題

級分	內容組織	表達能力	語言運用
5	內容與題目要求相關，內容豐富，話語有組織	語速適中，偶有停頓；詞語重複次數甚少；語音清楚，較少偏誤，聽者都能理解	具備足夠的詞彙量和語法，能適當地使用，仍有少許偏誤
4	內容與題目要求相關，內容尚稱豐富，話語多有組織	語速適中，偶有停頓；詞語重複次數少；語音大致清楚，偶有偏誤，聽者幾乎都能理解	具備足夠的詞彙量和語法，能大致適當使用，偶有偏誤
3	內容與題目要求相關，但豐富性尚不足，部份話語缺乏組織	語速稍慢，常有停頓；詞語重複次數多；語音尚稱清楚，不時有偏誤，聽者尚能理解	尚有足夠的詞彙量和語法，尚能適當使用，不時有偏誤
2	內容與題目要求相關但不完整，內容不足，話題無法擴展，組織較差	語速過慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解	詞彙量有限，且多不適當 ，僅能掌握基本語法結構
1	內容與題目要求不相關	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解	詞彙量極有限，且不適當 ，僅能掌握部份基本語法結構
0	考生靜默，沒回答		

評分原則

進階級

說明題

級分	內容組織	表達能力	語言運用
5	說出自己的選擇或意見，並能提出理由且詳細解釋，以支持自己的論點，話語有組織	語速適中，偶有停頓；詞語重複次數甚少；語音清楚，較少偏誤，聽者都能理解	具備足夠的詞彙量和語法，能適當地使用，仍有少許偏誤
4	說出自己的選擇或意見，並能提出多個理由支持自己的論點，偶有重複說明的情況，但已能進一步解釋，話語多有組織	語速適中，偶有停頓；詞語重複次數少；語音大致清楚，偶有偏誤，聽者幾乎都能理解	具備足夠的詞彙量和語法，能大致適當使用，偶有偏誤
3	說出自己的選擇或意見，並能提出至少一個理由支持自己的論點，常有重複說明的情況，以致於無法進一步加以解釋，部份話語缺乏組織	語速稍慢，常有停頓；詞語重複次數多；語音尚稱清楚，不時有偏誤，聽者尚能理解	尚有足夠的詞彙量和語法，尚能適當使用，不時有偏誤
2	說出自己的選擇或意見，但無法清楚地提出理由支持自己的論點；話語多不連貫，組織較差	語速過慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解	詞彙量有限，且多不適當 ，僅能掌握基本語法結構
1	沒說出自己的選擇或意見	語速非常緩慢，說話非常費力，停頓次數極多且時間極長；詞語重複次數極多；語音不正確，聽者無法理解	詞彙量極有限，且不適當 ，僅能掌握部份基本語法結構
0	考生靜默，沒回答		

參考文獻

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A Critique of the ACTFL Oral Interview. *The Modern Language Journal*, 70(4), 380-390.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition*, 10(2): 149-64

Bachman, L. F. (1997). *Language Testing in Practice.* Oxford: Oxford University Press.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series No. 9). Princeton, NJ: ETS.

Fulcher, G. (2003). *Testing Second Language Speaking.* Pearson Education.

Goulden, N. R. (1994) Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education, 27, 73-82.*

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation and Research.* Heinle & Heinle.

Lado, R. (1961). *Language testing.* London: Longman.

Underhill, N. (1987). *Testing Spoken Language.* Cambridge: Cambridge University Press.

Weir, C. J. (1993). *Understanding and developing language tests.* London: Prentice Hall.