發表於 2013 年第二屆標準本位評量國際研討會，國立台灣師範大學，2013 年 11 月 14-15 日

Yes/No Angoff 法在華語文口語測驗的應用

藍珮君[1]、陳柏熹[2]、張可家[1]、施泰亨[1]、林玲英[1]
[1]國家華語測驗推動工作委員會
[2]國立台灣師範大學教育心理與輔導學系

摘要

標準設定為一透過合理的、規範的程序和步驟建立測驗通過分數的方法，因而逐漸廣為許多大型語言測驗或學習成就評量資料庫所使用，藉此將學習者的學習成果或成就做適當的分類。

華語文口語測驗為建構反應題型之測驗，目前用於此題型之標準設定方法較不常見，本研究應用 Yes/No Angoff 法的概念，對華語文口語測驗進階高階級進行三回合的標準設定，分別制訂出進階級與高階級的切截分數，以對應至歐洲共同語文參考架構(CEFR)之 B1 與 B2 等級。

完成華語文口語測驗進階級與高階級之通過門檻設定後，本研究對標準設定結果進行效度的程序性證據，以及內部證據的檢核，結果顯示於前述二項效度證據皆得到支持。

關鍵詞：華語文口語測驗、Yes/No Angoff 法、標準設定

一、緒論與研究目的

由國家華語測驗推動工作委員會(簡稱華測會)研發之華語文口語測驗(Test of Chinese as a Foreign Language: Speaking，簡稱 TOCFL Speaking)，自 2007 年開始籌劃研發並陸續舉行多次預試。華語文口語測驗設計理念為，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，著重於考察應試者能否在特定語境下，藉由口語表達，有效地傳遞訊息。

華語文口語測驗於 2011 年 10 月在台灣地區第一次舉行正式考試，測驗等級為基礎級與進階級。2012 年華語文能力測驗調整測驗架構，並於 2013 年推出新版華語文能力測驗，將語言能力分成三等六級，三等分別為入門基礎級、進階高階級及流利精通級；而每一等又可再依據測驗成績細分為兩級，分別為入門級、基礎級、進階級、高階級、流利級、精通級，共六級。在口語測驗方面，目前實施之正式考試等級為進階高階級，預計明年推出入門基礎級測驗。

由於測驗架構已做了改變，過去的通過門檻勢必無法適用於新版測驗，因此訂定新的通過門檻乃有其必要性。本研究目的在應用標準設定方法中的 Yes/No Angoff 法(Impara & Plake, 1997)對華語文口語測驗進階高階級進行通過門檻的設定，使通過門檻符合 CEFR B1 與 B2 等級之能力描述，並對標準設定結果之效度進行檢核。

二、研究方法

本研究於 2013 年 1 月邀請 14 位華語文領域教授以及資深華語教師參加華語文口語測驗進階高階級之標準設定研究，由於口語測驗各題給分為 0 至 5 級分的多點計分制，與單選題非對即錯的概念不同，因此在標準設定方法的操作上，研究者討論後決定參考 Yes/No Angoff 法之概念，並加以調整。

華語文口語測驗標準設定會議分為四個階段，首先說明本次會議目的與流程，接著講述進階級與高階級學習者口語能力的最低表現描述，確認與會專家皆理解後，再進行第三與第四階段。第三階段為進階高階級測驗描述題題型的標準設定，第四階段則為進階高階級測驗說明題題型的標準設定。

每個題型標準設定又分為以下四個步驟：1. 題型與評分原則簡介，讓專家了解該題型欲測量之口語能力面向以及評分重點。2. 播放 0 至 5 級分範例音檔，藉由考生實際的答題反應，具體化評分原則，以利後續的配對工作。3. 進階級、高階級最低能力描述與評分原則之配對。在此步驟，與會專家根據會議帶領者提供的進階級與高階級最低能力描述，分別與評分原則(1-5 級分)進行配對，決定進階級和高階級最低能力表現最為接近評分原則的哪一級分，並寫下判斷依據。提供判斷口訣為：「以 OO 級最低能力考生的口語表現，能否在此題型得到 O 級分？」。4. 進階級、高階級最低能力描述與 CEFR 等級之配對。專家聽完 10 名考生音檔後，參照進階級和高階級口語最低能力描述，分別判斷每個音檔的等級(高階級、進階級、不到進階級)，並寫下判斷依據。

完成上述四個步驟後，即結束第一回合的判斷，研究者收集專家的判斷結果進行計算，在評分原則的配對方面，提供給專家進階級與高階級 1 至 5 級分的判斷人數，與結果的平均數和標準差；在音檔的等級配對方面，則提供每個音檔被判定為高階級、進階級與不到進階級的人數，由會議帶領者邀請專家就自己的判斷說明原因，進行討論以凝聚共識。經過討論後再進行第二回合的判斷，第二回合討論中，在音檔配對方面另外提供音檔實際級分，藉此讓專家們檢視自身判斷結果與級分是否一致，最後進行第三回合。第三回合結束後進行問卷調查，了解與會成員在整個標準設定會議過程中的想法，並作為未來標準設定會議流程調整的參考。

標準設定結果的檢核部分，Kane(1994)提出可分為三個方向進行：效度的程序性證據、內部證據以及外部證據。本研究提供前二項效度證據，藉由詳實記錄標準設定流程，與會議後的問卷調查作為程序性證據；成員三回合判斷結果的一致性，以及評分原則配對與音檔配對二項結果的一致性則作為內部證據。

三、結果與討論

華語文口語測驗標準設定結果如下表 1 所示，在進階級部分，二種題型的平均數隨著討論的進行，均呈現微幅的降低；在高階級部分，二種題型的平均數則是逐漸提高。第三回合結束後，進階級描述題與說明題通過門檻接近 2 級分，高階級則均是 4 級分。

表 1　華語文口語測驗標準設定各回合結果

等級	回合	描述題		說明題	
		平均數	標準差	平均數	標準差
進階級	1	2.14	0.535	2.07	0.267
	2	2.14	0.363	2.00	0.000
	3	2.07	0.267	2.00	0.000
高階級	1	3.57	0.646	3.93	0.267
	2	3.86	0.363	4.00	0.000
	3	4.00	0.000	4.00	0.000

標準設定結果的檢核分為程序性效度與內部效度二部分進行說明。首先，程序性效度方面，標準設定會議按照既定流程進行，且在各回合間給予專家們充分的分享與討論時間。會議後的問卷調查結果也顯示，成員均同意會議帶領者對會議目的/任務解釋清楚、對標準設定方法的操作流程說明得很清楚、能了解最低能力者在標準設定方法的涵義、每回合後團體討論和分享，有助於進行下一回合的判斷、對於自己所設定的切截分數有信心.....等等，可做為本研究之程序性效度來源。

內部效度證據則由每一回合通過門檻的標準差，每一回合音檔等級判斷與實際級分之斯皮爾曼等級相關，以及每一回合音檔等級判斷的 Kappa 係數作為依據。通過門檻標準差部分，從表 1 可知，在進階級通過門檻部分，描述題的標準差在第一回合最大，然後逐漸降低，說明題可能經過早上描述題的討論後，專家們判斷原則漸趨一致，在第二回合 14 位專家的判斷已達到完全一致，標準差為 0，高階級通過門檻的標準差也呈現同樣的情形。

描述題與說明題各 10 個音檔等級判斷與實際級分的斯皮爾曼等級相關分析，將不到進階級、進階級、高階級分別編碼為 0、1、2，與音檔實際級分求相關的結果，描述題三個回合的相關係數均介於.703 至.949 之間(p<.01)，說明題三個回合的相關係數依序為.730 至.949(p<.01)、.791 至.949(p<.01)、.783 至.949(p<.01)，顯示專家們對於音檔通過等級的判斷與實際得分之間具有中高度或高度的正相關存在，判斷結果與實際得分頗為一致。

以 Randolph (2008)的 Kappa 計算程式得到 14 位專家在描述題三個回合音檔等級判斷的 Kappa 係數分別為.543、.692 以及.685；說明題三個回合的 Kappa 係數則是.674、.745 以及.745。依照 Landis 和 Koch(1977)提出的判斷標準，數值介於 41~.61 為中度一致(moderate)，介於.61~.80 為相當一致(substantial)，可見 14 位專家在音檔的等級分類表現上相當一致。

上述程序性與內部效度證據結果皆支持本研究以 Yes/No Angoff 法對華語文口語測驗進階高階級進行標準設定，所訂出之進階級與高階級通過門檻的有效性得到驗證，此一結果是可靠的。通過進階級與高階級門檻的學習者，其口語能力可分別對應到 CEFR 的 B1 以及 B2 等級。

四、參考文獻

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment* (chap.1 & chap.4). Retrieved January 17, 2007, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Impara, J. C., & Plake, B. S. (1997). Standard-setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Landis J.R., & Koch G.G. (1977). A one-way components of variance model for categorical data. *Biometrics, 33,* 671-679.

Randolph, J. J. (2008). *Online Kappa Calculator*. Retrieved August 28, 2013 , from http://justus.randolph.name/kappa