# 教育部教學實踐研究計畫成果報告

Project Report for MOE Teaching Practice Research Program

計畫編號/Project Number：PGE1090430

學門專案分類/Division：通識教育

執行期間/Funding Period：8/1/2020~7/31/2021

計畫名稱/Title of the Project

對話者之語言能力與評分嚴苛度對印尼語口語評量成績之影響

配合課程名稱/Course Name

初級印尼語

計畫主持人(Principal Investigator)：何德華

計畫助理(Research Assistant)：許婉儀

執行機構及系所(Institution/Department/Program)：

國立中正大學語言學研究所

成果報告公開日期：

☑立即公開 □延後公開(統一於 2023 年 9 月 30 日公開)

繳交報告日期(Report Submission Date)：8/2/2021

# 摘要

外語課堂以溝通式教學為目標者，需要搭配難易適中、真實可靠的口語評量，才能有效檢驗學生之口語能力，因此常見的口語評量模式是以二人一組搭檔對話的方式進行口試，並由評分者使用評分表檢定成效。然而學生在選擇口試搭檔時，可能因選擇不同對象而影響口試表現；而不同評分者在使用評分表時，也可能因個人評分嚴苛度有所差異，給出不同口試成績，因此教學者需要考慮是否需要規定口試對話搭檔之選擇標準，以及如何訓練助教團隊使用評分表以增進口試之公平客觀性。本研究以台灣一所國立大學通識教育中心之印尼語課程為研究場域，使用羅氏測量理論檢測：（1）評分者不同的嚴苛度在經過訓練之後能否達成口試評分的一致性？（2）學生在口試搭檔的選擇上，選擇與個人語言背景相當（初學者與初學者搭檔）或與個人語言背景不相同者（初學者與印尼華人搭檔）是否會影響其口試成績？本研究結果將能幫助教學者了解如何制定更完善的口試搭檔規定，並透過評分者嚴苛度分析給予評分者更有效的評分訓練，讓溝通式外語教學評量更臻完善。

關鍵詞：口語評量，印尼語，評分訓練，評分嚴苛度，對話搭檔

# The influence of interlocutor proficiency and rater severity in Indonesian oral assessment


## Abstract

The use of pair work in speaking assessment has been adopted as an authentic way of testing oral proficiency in a L2 communicative language teaching classroom; however, previous studies have controversial findings regarding whether interlocutor proficiency may influence the outcomes of oral assessment and whether rater training will achieve long lasting inter-rater reliability. It is, therefore, necessary to explore (1) whether providing training to raters to use the rubric will increase inter-rater reliability, and (2) whether the test taker will perform differently, paired with an interlocutor of the same or different proficiency level. This study aimed to investigate the oral assessment in the GEN ED Indonesian classes at a national university in Taiwan, using the Rasch analysis to measure to what extent interlocutor proficiency (Indonesian language learning beginners vs. Indonesian L1 speakers) will influence students' oral performance and to what extent rater severity of the Indonesian teaching assistants can be identified and controlled. Pedagogical implications are provided to inform how to choose pair work interlocutors and how to train novice teaching assistants to achieve inter-rater reliability.



Keywords: oral assessment, Indonesian, rater training, rater severity, interlocutor proficiency

# 目 錄

## 一、研究動機與目的(Research Motive and Purpose)

口語評量的客觀公正性一向是標準化測驗(standardized tests)最棘手的問題，因為牽涉到評分員的訓練，而且訓練的效力還無法一勞永逸。如果將口語評量機制使用在溝通式外語教學現場，則除了評分員嚴苛度不一致的問題以外，還要加上學生對話搭檔者之間語言能力不一恐怕也會影響學生的口語表現。學生在口試時常採取一些比較保守的策略:例如找尋程度好的同學搭檔，或背誦一些樣板對話，認為這些機制有加分效果。因此，對於有志推行溝通式外語教學的教師，必須找出提升口語評量信度的辦法，才能解除學生對此評量方式的焦慮。由於口試評量牽涉情境變項非常多，無法使用實驗方法一網打盡，必須根據教學課程之個案需求，以行動研究的方式予以處理，因此透過教學實踐計劃尋求有效解決之道，是非常合理的做法。

本計劃探討對話者之語言能力與評分嚴苛度對印尼語與口語評量成績之影響，使用羅氏測量模式(Rasch model)分析對話者之語言能力差異對於口語評量成績之影響並探討對於評分者加以訓練後之效果，以發展出更具公信力的口語評量機制。研究目的有二: 首先探討學生在口試搭檔的選擇上是否會因搭檔的語言能力差異影響彼此的口試成績，尤其在印尼語通識教育課堂有三分之一的學習者為印尼/馬來西亞華人和新住民(即所謂的祖語學生)，當初學者和祖語學生搭檔時，初學者之口語表現是否會優於和同為初學者搭檔之表現。其次，評分者也是影響口試成績的重要關鍵，因此需要探討評分助教是否能準確使用評分表，如何發掘評分者不同的嚴苛度，以及如何訓練評分者以達成口試評分的一致性。

## 二、 文獻探討(Literature Review)

### 2.1 口試搭檔對話方式

口語評量採取受試者二人搭檔之對話者方式具有許多優點: 包括反應語言溝通的真實性並有實用價值(Taylor, 2003)、不必聘用考官一對一面試較有經濟效益(Davis, 2009: 368)、口試者比較不緊張，且能鼓勵學生合作學習，具有考試領導教學的正面效果(Saville & Hargreaves, 1999)、兩人對話相較於由考官面試在語言風格上較多變化(ffrench, 2003)、並符合任務型(task-based)語言教學的實際狀況(Long & Crookes, 1992)。但是也有學者(例如 Foot, 1999)認為口試者緊張程度不減反增，會將彼此拖下水，而且如果研究者對於口試者背景並不完全了解，則無法片面接受其研究結論。其次，由於口試的目的在於評量口語溝通能力，如果學生彼此個性相近，表現會比較好(Berry, 2007)。如果學生的個性比較內向怯懦，分數會比較低；個性外向積極，分數會比較高(Ockey, 2009; Nakatsuhara, 2011)。

由於搭檔對話方式(paired test format)優點多於缺點，在口語評量上目前已蔚為主流(East, 2015)，但是否會因搭檔之語言能力差異影響彼此的口試成績，研究結果到目前為止並無一致的看法。Iwashita (1998)發現搭檔之間有『遇強則強』的現象，成績和話語量皆同時提升。Norton (2005)也提出如果搭檔語言能力較高、而且彼此之間熟識、不但具有加分效果，而且能增加口試者話語量。但是 Davis (2009)使用羅氏測驗法，分析 20 位中國學生的英語口語對話搭檔對於口語評分之影響，則未發現搭檔的語言能力對於口試成績有任何統計上顯著的差異，除了語言程度較低的學生和程度較高者搭檔時，話語量確實有所增加之外，『多言多語』現象其實並未有任何加分效果。為了驗證 Davis 的研究結果，Son (2016)使用多面向羅氏測量模式分析 24 位韓國學生的英語口試結果，基本上也發現搭檔的語言能力高低不影響口試成績，比較特別的是韓國學生『遇強則縮』，當語言程度較低的學生和程度較高者搭檔時，話語量反而減少，但是話語量的多寡也不影響口試成績。

以上的研究在語言能力的劃分上，有的採取給予受試者單獨口試鑑定(Davis, 2009; Iwashita, 1998)、有的採取自我評量問卷(Csépes, 2009)，但是本研究所採取的語言能力分類則

以根據前次本人教學實踐研究結果(何，2019)，將印尼/馬來西亞華人(=華裔學生) 區分為語言能力最高的群體，而初學者則是語言能力較低的群體。此項研究將印尼語能力為第一語言使用者和初學者明顯區分開來，探討這種搭檔之組合，填補了文獻上缺乏如此分類的空缺。

2.2 評分嚴苛度

口試評分員在口語評量上扮演了關鍵角色，但是根據口語評分員訓練結果之相關文獻顯示，雖然評分訓練可以提高評分員的信心，增加評分員內在一致性 (McNamara, 1996)，但嚴苛度仍有差異(Weigle, 1998)，且訓練效果無法持久(Bonk & Ockey, 2003; Lumley & McNamara, 1995)。藍(2010)以多面向羅氏測量模式(many-facet Rasch measurement)分析探討華語文口語能力測驗(TOCFL)評分員訓練效果也得出同樣結論。如此結論表面看似負面悲觀，但至少可以發展出一套有效的訓練模式，以增加評分員針對同一評分法彼此間的一致性，並剔除經過訓練之後仍然過於嚴苛或寬鬆的評分員。盱衡現階段口語評分仍須完全仰賴人工評分，甚至連寫作評分時，也是人工評分方式比電腦自動評分的一致性更高(Wang & Brown, 2008)，即使套用電腦自動評分，仍無法判斷語言表達字裡行間的奧妙，因此仍需要搭配專業人士給予學生回饋才行(O'Neill & Russell, 2019)，可見投資訓練評分員正確使用評分表，仍是口語評量成敗的不二法門。

2.3 羅氏測量模式(Rasch Model)

目前從事語言評量之教育心理學家常使用根據丹麥統計學家 Georg Rasch 於 1960 年代發展之羅氏測量模式(Rasch Model)所設計之電腦軟體。主要的精髓在於將受試者之能力和測驗題目的難度放在同一個量尺上做比較(Lee, 2012)。使用羅氏測量表，可以使用 Winsteps 軟體[1](一維 Rasch 模型)進行以下三種分析(莫, 2019；莫、張, 2017): (1) 題目分析: 包含題目難度、信度、效度、鑑別度、適合度、資料偵誤；(2) 學生分析: 包含學生能力、信度、效度、不尋常反應、近側發展區、個別學習地圖；(3) 評分者分析: 主要探討評分者嚴苛度。

維度比較多的研究常使用 Multi-Facets Rasch Model (MFRM) (張、吳 2008; 謝, 2017)。雖然測驗的統計方法可以使用 SPSS 等軟體執行，但是王(1997)認為因素分析不宜當作測驗建構分析的工具，而使用羅氏測驗模式則更為簡單直接。此外，測驗所需要用到的統計工具也已經設計在羅氏測量模式當中，因此，目前香港初等與中等學校教師均已使用羅氏測驗模式作為教師分析評量結果的好幫手。

雖然羅氏測量模式在香港中小學各級學校應用極廣，但是台灣應用語言學界外語教學評量領域對此軟體並不熟悉，因此本研究希望能嘗試將此工具應用在印尼語口試評量上，偵測學生個別學習能力和助教評分嚴苛度。

## 三、研究問題(Research Question)

本計畫研究問題目在於改進溝通式外語教學之口語評量方式，首先從學生的角度出發，探討不同的學生如何在印尼語溝通式教學課堂「實務群體」中學習印尼語，如何自我定位、選擇搭檔、如何參與口語評量、以達成學會印尼語的目標；其次從印尼語助教的角度出發，探討助教群如何在印尼語團隊教學之「實務群體」中自我定位，如何透過評分訓練機制逐漸調整嚴苛度，以達成教學團隊對助教工作的期待。

在此大架構下，將研究問題聚焦為二: (1)學生在口試搭檔的選擇上是否會因搭檔的語言能力差異影響彼此的口試成績?(2)評分者不同的嚴苛度在經過訓練之後能否達成口試評分的一致性?

[1] https://www.winsteps.com/index.htm

## 四、研究設計與方法(Research Methodology)

4.1 研究對象介紹

本研究以 109 學年度（2020 年 9 月至 2021 年 1 月）通識教育「初級印尼語」44 位修課學生（包含台灣本地生 26 人，印尼華人 10 人，馬來西亞華人 1 人，港澳生 2 人，日本學生 2 人，俄羅斯學生 3 人；男性 10 人，女性 34 人）和 7 位印尼助教（4 位來自北蘇門答臘，2 位爪哇，1 位蘇拉威西；男性 2 人，女性 5 人）為研究對象，研究場域在 TEAL 教室中進行的印尼語課室教學。期初在課堂上說明此課程為教學實踐計畫，並取得所有參與者符合研究倫理審查要求的知情同意書。

4.2 採用之研究方法與研究流程

本計畫為行動研究(action research)，在自然教學場域中、以不打擾正常教學的方式研究如何改進口語評量。

每一次口試後立刻蒐集學生口試成績和助教評分表，即能做測量分析，檢驗對話搭檔之口語能力高低是否影響彼此口語評量成績，並檢驗助教口語評分之嚴苛度。因此能在不打擾正常教學的原則下，蒐集到一整學期的資料，從事資料三角檢驗(triangulation)，增強研究的可信度(trustworthiness)。

4.3 研究資料蒐集工具與評量工具

本計畫蒐集之資料包括學生背景、評量成績，線上問卷填寫內容、助教口試評分成績、以及助教線上教學日誌內容。

口語評量結果根據羅氏測量架構，使用 FACETS 軟體觀測學生的程度和試題的難易度/鑑別度之間的關係、分析學生能力，檢測信/效度、學生不尋常反應(detention of unexpected student response)、找出學生近測發展區(Zone of Proximal Development)、畫出個別學習地圖(Wright map)、以及各個評分者對於不同評分項目的嚴苛度。

期中和期末口試評量由 8 位助教團成員根據發音、詞彙、句型、風格、與人際關係五項評分表(表二)給分，並輸入 Excel 檔案。將所有助教的評分製成 Excel 檔案，使用 FACETS 軟體分析學生能力與評分者嚴苛度分佈。

除了使用 Rasch model 檢驗學生能力與評分者嚴苛度以外，並使用 SPSS 之 t-test 檢測初學組與印尼華僑組程度差異，以 Kruskal Wallis test 測試學生搭檔選擇對口試成績的影響。線上問卷與助教教學成果報告則使用言談分析法歸納出學生和助教針對口語評量分別從受試者和評分者的角度所提供的回饋。

## 五、教學暨研究成果(Teaching and Research Outcomes)

5.1 學生學習成果評估

印尼語背景的學習者表現能力一致最優異是無庸置疑。而初學者的能力估計分布差異性很大，且兩次評量中表現最差的是來自初學者，但是經過兩個月的課程訓練，顯示出初學者的口語表現能力進步。

期中與期末兩次考試皆顯示口語內容(content)及對話互動(interaction)能力表現項目較簡單，正確度(accuracy)與流利度(fluency)居中，而發音(pronunciation)最為困難。內容和互動之所以容易拿高分是因為口試內容事先已規範，學生二人搭檔不但可事先預備，甚至還能背誦，並且在助教課外練習時間預演排練，如此充分準備大大提升詞彙語法正確和流利度。唯獨發音需「兄弟登山、各自努力」，反映學習者口語溝通聽覺辨識和模仿功力之高低。

5.2 教學歷程之評估

七位評分者嚴苛度差異，依其期中考變數分布顯示有五位評分員進行學生印尼語口說表現判別時比較一致，AA最為嚴厲(1.03 logits)，評分員AI的logit值最低 ( = -2.42 )，代表評分最為寬鬆，其次為PU(-.95 logits)；評分嚴苛度與評分者屬於拘謹或輕鬆的個性亦有關聯。當教學者將評分結果呈現給助教們觀看過後，期末所有的評分者差異都介於+1與-1之間。由於影響評分員的變數頗多，即使訓練也難以一致，因此將多位評分員的成績平均之後作為學生的最終口試成績，絕對比「雞蛋放在同一個籃子裡」安全可靠。

5.3 研究結果之分析評估

值得注意的是第一次評分的分離度5.31在第二次評分減少為2，嚴苛度差異從至少分成5個層級的類別區隔已下降至少只分成 2 個層級，表示評分者的判斷歧異性減少，似乎有趨於一致性的現象。由此可見期初的評分訓練固然必要，然而評分者也要有實際評分的經驗之後，才能摸索出箇中奧妙，逐漸趨於穩定。但是仍然有個性和情緒等不可控制的人為因素影響，讓口試評分僅由單一評分員為之的風險大增，而應該採取由多位評分員共同評定的方向努力。

學生各次考試的搭檔「與高配」平均值皆無高於與「同程度」搭檔的平均值，「與低配」平均值沒有一次是低於「同程度」搭檔的平均值，而「與高配」平均值也沒有高於「同程度」的平均值。換言之，「高攀」不能麻雀變鳳凰；「低就」口試也不會被拖下水。初學者不論是和與自己同程度的初學者搭檔，或是和比自己能力好的印尼華僑搭檔，對成績並沒有太大的影響。因此，印尼語初學者找比自己能力要好的學生搭檔並不會影響口試成績。

## 六、建議與省思(Recommendations and Reflections)

6.1 對教學所遭遇實務問題之省思

本研究團隊把握上學期初級印尼語課程 44 位學生修課的機會，蒐集了完善的資料，做了行動研究，並且利用寒假期間寫出了研究論文，投稿至教學實踐研究期刊。因此雖然下學期受到疫情影響，中級印尼語課程僅有 11 位學生修習，而且印尼華僑就佔了 8 位，且從 5 月中改成遠距教學，皆不利於繼續作為本計畫的觀察研究場域，但是我們在上學期的研究超前部署，讓我們的計畫不受影響，得以順利進行，並且獲致不錯的研究成果。

6.2 未來應用於教學實務現場之反思與建議

未來在印尼語通識課程仍應讓學生自由決定他們的口試搭檔，只要測試的機會夠多，評分者的嚴苛度相當且依照公平公正原則評分，學生並不因為搭檔的能力，而對個人口試成績產生太大的影響。本研究建議語言課程中應該鼓勵能力較好的學習者，如本研究中的印尼華僑，和初學者搭檔，以期產生雙贏的效果。另外，學生對於自己的表現，雖然「幾家歡樂幾家愁」，但對自己選擇的口試搭檔表現都感到滿意，正因為口試方式是二人搭檔，在準備口試過程中，發展出同舟共濟的革命情感，並沒有人表達任何遇弱則強或與強則縮的感覺。

## 參考文獻(References)


Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.

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.

Csépes, I. (2009). *Measuring Oral Proficiency through Paired-Task Performance. Vol. 14.* Frankfurt: Peter Lang.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *26*(3): 367-396.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120.

ffrench, A. (2003). The change process at the paper level. Paper 5, Speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge proficiency in English examination 1913–2002* (pp. 367–446). Cambridge: UCLES/Cambridge University Press.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal*, *53*(1): 36 – 41.

Iwashita, N. (1998). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, *5*, 1–65.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.

Lee, Y. J. (2012). Software to facilitate language assessment: Focus on Quest, Facets, and Turnitin. In C. Coombe, P. Davidson, B. O'Sullivan, & S., Stoynoff (Eds.), *The Cambridge Guide to second language assessment* (pp. 280-288). New York, NY: Cambridge University Press.

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, *26*, 27–56.

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

McNamara, T. F. (1996). *Measuring second language performance*. Addison Wesley Longman.

Morita, N. (2004). Negotiating participation and identity in second language academic communities. *TESOL Quarterly*, *38*(4), 573-603.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing 28*(4), 483–508.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT journal*, *59*(4), 287-297.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(2), 161–186.

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australian Journal of Educational Technology, 35*(1), 42-56.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE: 42 – 51.

Taylor, L. (2003, August). *The Cambridge approach to speaking assessment*. University of Cambridge Local Examinations Syndicate Research Notes, 2–4.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: a correlational study. *Contemporary Issues in Technology and Teacher Education, 8*(4), 310-325.

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

王文中(1997)。<測驗的建構；因素分析還是 Rasch 分析?> 《調查研究-方法與應用》，3, 129-166。

何德華(Rau, D. V.) (December 2019). <印尼語 TEAL 創意互動教學測驗與評量> Large Class Assessment of Indonesian Language Proficiency. 《通識教育學刊》*Taiwan Journal of General Education* 第 24 期: 79-132 頁。 https://www.airitilibrary.com/Publication/alDetailedMesh?docid=19993331-201912-202002150001-202002150001-79-131

何德華、李萍、賴思悅、潘家貝、阿芬達(Rau, D. Victoria, Priska Lydia S. Pulungan, Apriliana

Lase, Ganda Christian Panggabean, and Afrinda Samosir) (2019) 《印尼旅蛙來電了》(Indonesian Travel Frog CALLed)(Petualangan katak di Indonesia)。手稿，尚未出版。

莫慕貞 (2019)。 <精進教學工作坊: Rasch 可觀測量在大學甄試檔案評量之應用>講義。嘉義: 國立中正大學. 2019/11/8~9。

莫慕貞、張權(主編) (2017)。《羅氏測量:應用與導讀》(英文版原編者: Everett V. Smith, Jr., & Richard M. Smith (Eds.) Introduction to Rasch Measurement)。Maple Grove, MN: JAM Press.

張新立、吳舜丞 (2008)。<多層面 Rasch 模式於學術研討會論文評分之應用>。《測驗學刊》，55（1），105-128。[Chang, H. L., & Wu, S. C. (2008). A multi-facet rasch analysis on rating the academic scientific papers. *Psychological Testing, 55*(1), 105-128.]

謝名娟 (2017)。<誰是好的演講者? 以多層面 Rasch 來分析校長三分鐘即席演講的能力>。《教育心理學報》，48(4)，551-566。[Hsieh, M. C. (2017). Who is a good speaker? Applying multifaceted Rasch model to analyze principal three-minute impromptu speech. Bulletin of Educational Psychology, 48(4), 551-566.]

藍珮君 (2010)。<以多面向 Rasch 測量模式分析 TOCFL 口語測驗評分者訓練效果> 第九屆海峽兩岸心理與教育測驗暨 2010 NAER 「永續教育發展-創新與實踐」國際學術研討會論文。10/21-23/2010. 台北: 國立台灣師範大學、國家教育研究院籌備處。

## 附件(Appendix)

- 成果報告 10 分鐘影片檔: https://youtu.be/ZqpL-_6GWNo