國立臺灣師範大學教育心理與輔導學系
教育心理學報，2023，55 卷，1 期，25–46 頁
https://doi.org/10.6251/BEP.202309_55(1).0002

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

何德華[1]、張惠環[2]、許婉儀[1]

外語課堂以溝通式教學為目標者，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，給予不同口試成績，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。本研究以臺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 Rasch 模型檢測：（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？本研究結果發現不同評分者即使施予訓練仍無法完全達成評分一致性，因此目前由多位評分者共同擔綱，刪除離群值、取其平均數，或許是權宜之計。然而，根據多層面 Rasch 分析法檢測評分者嚴苛度，有助及早發現問題。其次，學生選擇與不同語言能力背景搭檔口試並不會影響其口試成績，因此應讓學生自由選擇對話搭檔，輔以鼓勵機制讓印尼華僑多跟初學者搭配，以達到雙贏的效果。

關鍵詞：多層面 Rasch 模式、口語評量、印尼語、評分嚴苛度、對話搭檔

[1] 國立中正大學語言學研究所
[2] 國立勤益科技大學語言中心

通訊作者：何德華，國立中正大學語言學研究所，Lngrau@ccu.edu.tw。
本研究獲 109 和 110 學年度教育部大專校院教學實踐研究計畫補助（計畫編號：PGE1090430, PGE1100891），特此致謝。

如何在大班通識教育外語教學的形成式評量中有效使用筆試和口試檢測學習成效，在教學現場至關重要。其中筆者之印尼語通識課程筆試層面已有效解決題庫問題（何德華，2019），改善其線上測驗機制，建立有效公平的評量方式。然而以溝通式教學理念建構的外語課堂另一最需解決的問題即開發有效可靠的口試評量工具，以達成溝通式教學的目標。

過去初、中級印尼語課堂使用傳統教科書時要求學生背誦課文對話作為口試，但學生往往因死背或課文過長而哀鴻遍野。自教學團隊以溝通式教學法理念開發了新教材（何德華等人，2019），學生仍需找尋對話搭檔、演出事先準備的簡短對話以展現真實溝通，而對話搭檔的口語能力和溝通技巧可能會影響彼此口試表現。即便受試者搭檔的選擇已不是問題，多位助教評分者若無一致評分標準或有不同嚴苛度，仍會影響口試的客觀公正性。因此如何制定完善的口試搭檔規定，給予評分者有效的評分訓練，是溝通式外語教學在口語評量上極待解決的問題。

眾所周知口語評量的客觀公正性一向是標準化測驗最棘手的問題，因牽涉到評分員的訓練，其效力還無法一勞永逸。除了評分員嚴苛度不一致的問題會影響評分結果，對話搭檔間能力不同恐怕也會影響學生表現。由於口試評量牽涉情境變項極多無法用實驗方法一網打盡，須根據教學課程個案需求以行動研究（Wallace, 1998）方式予以處理，因此透過教學實踐研究尋求有效解決之道。

有鑑於此，本研究目的為透過兩班「初級印尼語」通識課程探討評分助教使用口試評分表是否有嚴苛度差異、經過訓練之後評分者能否達成口試評分的一致性，以及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，以嘗試改善口語評量機制。

文獻探討

（一）Rasch 模型（Rasch model）

Rasch 模型具有客觀、等距的測量特性，對於估計及編製測驗研究有相當助益，且廣泛應用於教育、心理、醫學及管理等領域（王文中，2004），例如，運動評量（姚漢禱，2004；陸雲鳳，2016；曾盟堡，2002）和中小學學科教育實作評量（王佳琪，2020；吳昭容等人，2018；林小慧等人，2018；林怡君等人，2013；陳映孜等人，2017；陳建亨、楊凱琳，2021；謝如山、謝名娟，2013）。Rasch 模型的精髓在於將受試者能力和測驗題目的難度放在同個量尺上比較（Lee, 2012），一維 Rasch 模型可使用 Winsteps 軟體[1] 進行題目難易度與學生能力差異分析（莫慕貞，2019；E. V. Smith & R. M. Smith, 2004/2017），而維度較多的研究適用於多層面 Rasch 模式（multi-facets Rasch model，簡稱 MFRM）（張新立、吳舜丞，2008；謝名娟，2017，2020）。MFRM 統計模型是藉由從評分者因素（rater-mediated assessments）中擷取可靠、有效和公平的推論，提供一個有條理的框架（Eckes, 2015），而羅氏測量模式延伸的「多層面 Rasch 模式」（MFRM）即常用來做評分者嚴苛度的分析研究工具（Eckes, 2009），因此本研究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。

（二）口語評分嚴苛度

口試評量的信度一直是口試評分的重要議題之一（Luoma, 2004），評分員在其中更扮演關鍵角色。為確保受試者的評估具有公平性和可靠性，探討評分員之間評分成效的表現至關重要，而其中最重要的關鍵之一就是評分的嚴苛度（severity）（Eckes, 2015）。嚴苛的評分者過分嚴厲地遵守評估程序，可能會造成優秀的受試者成績低於他們的實際能力（Myford & Wolfe, 2003）。相反的，評分者也可能過度寬鬆（leniency）而給予比受試者實際應得的分數還高之成績（Wind, 2018）。評分員間若未能在評分嚴苛度上達到平衡點，可能導致評分差異過大，從而對考生成績產生負面影響。

以 MFRM 多層面 Rasch 模型為基礎探討第二語言口語能力評量嚴苛度的量化研究，明確證明口語評分者的嚴苛度是有其研究之必要性（例如 Bonk & Ockey, 2003; Eckes, 2005; Hsieh, 2011, McNamara, 1996; Sundqvist et al., 2020）。Bonk 與 Ockey（2003）針對學生口語能力、評分者以及包含發音、語法、詞彙和流利程度口語評量項目等進行多層面 Rasch 模型分析，評分者進行兩次英語口語測驗評分，主修英語的日本大學生受試者以同儕小組討論方式實施口語評量，結果發現整

體上評分者嚴苛度的差異大，有經驗的（連續兩次評分）評分員傾向給予更為嚴苛的評分。Hsieh（2011）則設計將 13 名 ESL 教師和 32 名美國大學生分成兩組不同組別的評分員，評分員使用評分量表評估考生的口語熟練程度、口音和可理解度性，對 28 位國際教學助教（ITA）進行英語口語能力評估測試（Speaking Proficiency English Assessment Kit, SPEAK）評量，檢驗分析其評分者的嚴厲度，結果表明兩組嚴厲度並無不同，但口音和可理解性評分卻有顯著的差異，大學生評分員在評估考生的口音（腔調）和可理解度評分項目比 ESL 教師評分員更嚴厲。Eckes（2005）以德語為外語的寫作與口語測驗（TestDaF）評量進行評分效應（rater effect）的相關研究，結果也呈現評分員對於測驗者評分嚴苛度有相當大的差異，雖然總體評分是一致，但相較於寫作，評分員對於學生口語評比標準項目能力評分有較高的一致性。Sundqvist 等人（2020）的研究中，資料分析來自 11 名瑞典小學教師評分員評量兩次小學六年級學生的英語口語測驗表現，藉此檢視兩次評分嚴格度的變化，分析顯示評分員可能因培訓使得評分一致性有所改善，但嚴苛度的差異性卻加大。

綜合上述相關文獻研究顯示，評分訓練固然可提高評分員信心，增加內在一致性（Davis, 2012, 2016; Huang et al., 2016; McNamara, 1996），但嚴苛度仍有差異（Eckes, 2005, 2009, 2015; Knoch, 2011; Sundqvist et al., 2020; Weigle, 1998），且訓練效果無法持久（Bonk & Ockey, 2003; Kim, 2011; Lumley & McNamara, 1995）。張可家等人（2011）、藍珮君（2012）及廖才儀（2016）以多面向羅氏測量理論探討華語文口語能力測驗評分員訓練效果也得出同樣的結論。雖然期待評分員完全公平、可靠，表面看似悲觀，但至少可發展有效訓練模式，增加評分員一致性。盱衡現階段口語評分仍須仰賴人工，甚至人工評分寫作仍比電腦自動評分一致性更高（Wang & Brown, 2008）。即使套用電腦評分寫作，仍需搭配專業人士給予學生回饋（O'Neill & Russell, 2019），可見訓練評分員正確使用評分表，應是維持口語評量一致性的不二法門。

（三）口試搭檔（paired oral assessment）

常見之高風險口語測驗，由於公平性考量，搭檔模式較多為一對一面試，多對多、或多對一團體面試等，鮮少由受試者二人搭檔。然而在課室形成性口語評量情境中，使用受試者二人搭檔對話有許多優點：包括反應語言溝通的真實性並有實用價值（Brooks, 2009; Ducasse & Brown, 2009; Galaczi & Taylor, 2018; Taylor, 2003; Van Moere, 2013）、不必聘用考官一對一面試較有經濟效益（Davis, 2009）、口試者較不緊張且能鼓勵學生合作學習（Együd & Glover, 2001; Jones, 2007; Rydell, 2019; Storch, 2001; Storch & Aldosari, 2013），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; Galaczi et al., 2011）、符合任務型語言教學的實際狀況（Long & Crookes, 1992）。

然而由受試者二人搭檔對話亦有其缺點。有學者（例如 Chuang, 2018; Foot, 1999）認為口試者緊張程度不減反增，會將彼此拖下水，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果。其次，由於口試目的在於評量口語溝通能力，若學生彼此個性相近，表現固然會比較好（Berry, 2007），但如果學生個性比較內向怯懦者分數會比較低；個性外向積極者分數會比較高（Nakatsuhara, 2011; Ockey, 2009）。

由於搭檔對話方式利多於弊，在口語評量上已蔚為主流（East, 2015），但是否會因搭檔之語言能力差異影響彼此口試成績，到目前為止並無一致看法。Iwashita（1996）發現搭檔間有遇強則強現象，成績和話語量皆同時提升。Norton（2005）也提出若搭檔語言能力較高且彼此熟識，不但具加分效果，還能增加口試者話語量。Storch（2001）觀察到能力差異大的搭檔組別更能以合作的模式完成所交代的任務，而 Storch 與 Aldosari（2013）發現高低程度互相搭配的組合有助於降低 EFL 學生於英語口說的焦慮情緒，增加學生的自信心。然而，Galaczi（2008）發現語言能力較弱的搭檔不太參與互動對話，因此對應互動能力需要加強（Galaczi, 2014）。

Davis（2009）使用 Rasch 模型（Rasch model）分析中國學生英語口語對話搭檔對口語評分之影響，未發現搭檔語言能力對成績有任何統計上顯著差異，除了語言程度較低者和程度較高者搭檔時，話語量確實有所增加之外，「多言多語」現象其實沒有加分效果。為驗證 Davis 的結論，Son（2016）使用同方式分析韓國學生英語口試結果，發現搭檔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，當程度較低和較高者搭檔時話語量反而減少，但其話語量多寡不影響口試成績。

以上研究之受試者語言能力劃分，有的給予受試者單獨口試鑑定（Davis, 2009; Iwashita, 1996），有的採取自我評量問卷（Csépes, 2009），但本研究所採取的語言能力分類係根據前次研究結果（何德華，2019）。由於馬來西亞語（Bahasa Malaysia）和印尼語（Bahasa Indonesia）為兩種馬來語的變體（variety），好比英式英語和美式英語，或海峽兩岸華語的差異，因此本研究將印尼和馬來西亞華人（= 華裔學生）自述精通印尼／馬來語者劃分為語言能力最高的群體，初學者則是語言能力較低的群體。由於本課程只需要區分零起點和非零起點的學生，因此馬來西亞華人的語言能力不必再區分等級。

方法

（一）研究對象與場域

本研究以開在兩個學年度之兩班「初級印尼語」通識課程助教與學生為研究對象，以提供較長時間的研究觀察和採集不同學生群體的資料佐證。如表 1 簡介，班級一為 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（包含臺灣本地生 26 人，印尼華人 10 人，新加坡華人 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）；班級二為 110 學年度（2021 年 9 月至 2022 年 1 月）通識教育「初級印尼語」修課學生 38 名（包含臺灣本地生 17 人，印尼華人 14 人，馬來西亞華人 4 人，其他外籍學生 3 名；男性 18 人，女性 20 人）和 8 位印尼助教（4 位來自北蘇門答臘，4 位爪哇；男性 4 人，女性 4 人）。其中助教群因為畢業離校，只有 3 位印尼助教連續兩年續任。研究場域在 TEAL（Technology Enhanced Active Learning）教室中進行課室教學，在自然教學場域中、以不干擾教學的方式研究如何改進口語評量。期初已說明此課程為教學實踐計畫並取得所有參與者之研究倫理審查同意書。為瞭解學生之起始印尼語程度，本計畫採取自述的方式。根據第一週線上問卷其中一題：「請問你的印尼語程度為何？例如：「初學者完全從零開始、印尼華人、馬來西亞華人、印尼新住民、或其他……請說明。」由學生填答其印尼語程度。除了全部印尼華人和幾位馬來西亞華人以外，其他學生包括本地生或國際生均填寫其為從零開始的初學者。

表 1
「初級印尼語」課程評量次數及參與人員簡介

項目	「初級印尼語」課程	
	109 學年度 2020/9 ～ 2021/1	110 學年度 2021/9 ～ 2022/1
時間／內容	預備週：助教團隊評分訓練／評分表解說 ●→ W1：線上問卷／了解學生印尼語程度 ●→ W3：平時口試 1 ●→ W6：平時口試 2 ●→ W7：助教團檢驗討論所有評分者的嚴苛度（110 學年度） ●→ W9：期中考口試 ●→ W10：助教團檢驗討論所有評分者的嚴苛度（109 學年度） ●→ W12：平時口試 3 ●→ W15：平時口試 4 ●→ W18：期末考口試	
共同評分	期中考，期末考	平時口試 2，平時口試 4
學生人數	44 位（臺灣 = 26，印尼 = 10，新加坡 = 1，港澳 = 2，日本 = 2，俄羅斯 = 3；男性 = 10，女性 = 34）	38 位（臺灣 = 17，印尼 = 14，馬來西亞 = 4，其他外籍生 = 3；男性 = 18，女性 = 20）
印尼助教	7 位（北蘇門答臘 = 4，爪哇 = 2，蘇拉威西 = 1；男性 = 2，女性 = 5）	8 位（北蘇門答臘 = 4，爪哇 = 4；男性 = 4，女性 = 4）

該課程由具備南島語言學背景之教師帶領印尼籍助教團隊聯合授課。課程含四次平時口試和期中、期末口試評量共 6 次。口試內容為自編教材（何德華等人，2019）第一到四課之溝通活動，可參見課程之線上教材資源網以及期末考口試評量內容[2]。上課前一週，由教師召集助教團隊給予訓練，了解評分表使用（如表 2），並以四個程度之學生音檔範例，練習評分。五個評分項目包括：事先規定之對話內容、詞彙文法正確度、語言流利度、發音可接受度、與對話時展現一來一往的人際互動。教師給予助教培訓的評分表將各等級的表現明確寫出，搭配量尺和分數換算對照（例如：1 = 差 < 60%，1 表示差，60 分以下／不及格），並中英對照以利不同背景助教皆能理解。109 學年度每一項均為四段量尺（1 = 差，2 = 普通，3 = 好，4 = 優），本意為排除中間選項，但對於分數排序沒有影響，因此 110 學年度則為增強等距性，採用五段量尺（1 = 差，2 = 尚可，3 = 普通，4 = 好，5 = 優）。但此項改變不影響最後數據解讀。

表 2
助教培訓評分表

content 內容	accuracy 正確度	fluency 流利度	pronunciation 發音	interaction 互動
Follows the guided conversations, covering all required contents.	Uses accurate vocabulary and grammar to reflect culturally appropriate relationships.	Delivers the speech smoothly and effortlessly.	Uses intelligible consonants, vowels, stress, and intonation patterns.	Demonstrates turn-taking and conversation etiquette.
涵蓋所有規定內容	使用正確表達人際關係的詞彙語法	言語流暢不結巴	使用能清楚辨識的母音、子音、重音、和音調發音	展現二人互動輪流對話的樣貌
109 年度： 1 = poor 差 < 60%　2 = average 普通 60 ～ 79%　3 = good 好 80 ～ 89% 4 = excellent 優 90 ～ 100%				
110 年度： 1 = poor 差 < 30%　2 = fair 尚可 30 ～ 60%　3 = average 普通 60 ～ 79% 4 = good 好 80 ～ 89%　5 = excellent 優 90 ～ 100%				

（二）研究目標與研究問題

本研究以改進「初級印尼語」之口試評量為目標。學生在六次口語評量中自選對話搭檔，藉以比較初學者與初學者的搭檔和初學者與華僑搭檔之印尼語口試成績差異。全班分七組，每組由一位助教負責。平時由不同助教根據評分表評分，但 109 學年度期中、期末考口試則由所有助教組成裁判團共同評分；110 學年度則是選擇第二次及第四次平時口試由所有助教共同評分。於第一次助教團共同評分之後，先檢驗所有評分者的嚴苛度，並討論過於嚴苛或寬鬆（亦即超過＋／ -1 羅吉斯量尺者）之助教注意改進，再於期末第二次共同評分口試之後，比較其嚴苛度是否獲得改善。

根據以上研究目標，將研究問題聚焦為二：（1）不同評分者之嚴苛度有無差異，及不同印尼語口試評分項目之間的難易度有無差異？評分者不同的嚴苛度在經過訓練後能否達成口試評分的一致性？（2）不同背景學生的印尼語口說能力有無差異？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？

（三）研究步驟與分析工具

本教學實踐計畫為行動研究，蒐集資料含學生背景、評量成績，學生線上問卷填寫內容。根據助教口試評分成績，檢測評分者的不同嚴苛度在訓練後能否達成口試評分一致性，以及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。學生線上問卷指下課前 5 分鐘學生線上填寫開放式問題，給予授課者回饋，藉此得知學生每次評量後的反應，有助於解讀學生成績進步或退步的原因。除了九月開學時取得學生對印尼語先備語言能力的自評資料以外，在十月初第一次口試以及十一月初期中考口試完畢後亦調查自評口試表現[3]，以掌握學生對於口試方式的評價和

態度。期中口試結束後使用多層面 Rasch 模型（Linacre, 1989）FACETS 軟體[4] 分析學生的語言能力，讓助教了解題目的難易度和學生能力之間的關係，及評分者間對於不同評分項目的嚴苛度，進而更有動機改進自己的評分公正性和一致性，並使用此軟體最經典的 Wright map 呈現結果（參見圖 1-1、圖 1-2、圖 2-1、圖 2-2），將學生程度、評分者嚴苛度、以及評分項目難易度，一次完整呈現在同一張地圖中。學生則自由選擇搭檔進行 6 次口試，內容及方式請參考註釋 2，於學期結束時檢測成效。本研究使用 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。最後，使用線上問卷，進一步協助研究者理解口語評量分數所代表的意義。

上述 FACETS（Linacre, 2022a）軟體，為印尼語學習者的口說能力、評分者嚴苛度和評分項目三層面（facets）建立客觀的量化指標進行參數估計分析，將原數據轉換成羅吉斯（logit）為單位的連續等距尺 （interval scale），受測者能力與問項放置於同一個「羅吉斯量尺」（logit scale）以衡量各個向度的強弱與各組別之間的比較（例如評分者嚴苛度差異），且藉以分析 MFRM 執行結果所呈現的統計數據包括參數估計值（logit scale measure）、信度（reliability）、適合度統計（訊息加權 the information-weighted mean-square fit statistics，簡稱 Infit）、偏離反應（the outlier sensitive mean-square fit statistics，簡稱 Outfit）、分離度（separation）與卡方值等。若適合度 Infit 和 Outfit 介於 0.5 與 1.5 之間的理想範圍內，表示 Rasch 模型產生的測量結果適合進行估計分析（Linacre, 2022b）。對於適合度統計值的檢驗，Infit 比 Outfit 更適合作為判斷適配度的指標（藍珮君，2012；Park, 2004; Pollitt & Hutchinson, 1987），因為 Infit 的數值對於非預期的評分更能提供精確統計數據（Eckes & Jin, 2021; Linacre, 2002; Myford & Wolfe, 2003; Wright et al., 1994），所以當 Infit 及 Outfit 均方值互有高低時，本研究以 Infit 均方值（0.5—1.5）為適配度篩檢的指標。藉由可信度（reliability）檢驗資料的穩定度，若信度值越接近 1，代表愈能分辨學生能力的不同程度，對評分者面向而言，高信度值代表不同評分者有不同的評分嚴苛度。顯著的卡方值則表示評分者之間的判斷存在差異。分離指數（separation index）是檢驗所測面向在測量變項的分散程度，如數值愈高表示有愈多的分層，評估評分者嚴苛度時評分者的分離指數愈低表示其一致性愈高。

結果與討論

以下針對評分者嚴苛度、學習者能力及搭檔的語言能力差異對口試成績影響的研究問題所作統計分析，呈現其結果與討論。

（一）評分者嚴苛度與學習者能力之檢測

以下我們分別呈現兩次初級印尼語課程的結果。第一次為 109 學年度課程。圖 1-1 與圖 1-2 Rasch 變數分布圖（variable map）顯示兩次印尼語口試之三層面變數分布的狀況：學生印尼語口語表現能力、評分者嚴苛度及口語能力表現評分項目。最左邊是以 logit 為單位刻度的等距尺度欄位，logit 值越高，代表學生口試表現越好、評分者越嚴厲、或是印尼語口說能力成就表現項目越困難。

第二欄為學習者口語成就表現能力估計分布，透過 Rasch 分布圖顯示其變數有明顯差異存在，印尼語背景的學習者（I = 印尼）表現能力無庸置疑最優異。初學者（T = 臺灣、J = 日本、R = 俄羅斯、H 和 M = 港澳、S = 新加坡）的能力估計分布差異性很大；外籍初學者一般位於中間地帶，兩次表現最差的則是臺灣的初學者，分別為期中考 T20（-0.20 logits）及期末考 T23（-1.33 logits），logit 值均小於零。經過兩個月的訓練，臺灣初學者 T20 口語表現能力顯著進步，從 -0.20 logits 提升到幾近 + 3 logits 的位置。但也有一位港澳生 M01 的口說能力變差，從期中高達 +5 logits 落到期末的 + 0.5 logits。變差的原因可從即時回饋中找到蛛絲馬跡，該生表達不習慣教學方法是從印尼語對話著手，比較偏好詞彙翻譯法，也可能因打工無暇參與每週 2 小時課外助教口語訓練所致。

第三欄是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口說表現判別時比較一致，AA 最嚴厲（1.03 logits），評分員 AI 的 logit 值最低（= -2.42 ），代表評分最為寬鬆，之間相差 3.45 logits；評分嚴苛度與評分者屬拘謹或輕鬆的個性亦有關聯[5]。課程教師於期中口試後根據評分結果圖 1-1 與助教們討論，期望期末評分有所改善。期末 Rasch 變數分布顯示，PR 變得比

較嚴苛（0.97 logits），MA 最寬鬆（-0.45 logits），其他評分者 logit 值大約一致介於 0 與 -1。由此可見，影響評分員的變數頗多，即使助教同樣認真負責，但口語評分訓練仍難一致。因此，將多位評分員的成績平均後作為學生的最終口試成績，絕對比「雞蛋放在同個籃子裡」安全可靠且是較公平公正的方式。存在於口試評分當中的評分者效應（rater effects）問題，不但不容忽視且應予重視及研究（余民寧，2013；Farrokhi & Esfandiari, 2011; O'Brien & Rothstein, 2011）。

第四欄為口語能力表現評分項目，分為：內容、正確、流利、互動、及發音 5 項，每一項評量尺規均使用 1 ～ 4 李克量表（1 = 差，4 = 優）。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是因為口試內容已事先規範，學生不但可事先預備，甚至能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「各自努力」，反映出學習者之間呈現口語溝通聽覺辨識和模仿功力之高低。

另外本研究也發現評分者與評分項目之間也有交互作用，亦即，不同評分者針對不同項目的評分嚴苛度也有差異。交互作用偏差分析的主要檢驗評斷指標是偏差值 *t* 值，*t* 值須介於 -2.0 至 2.0（Engelhard, 2002; Engelhard & Myford, 2003），以及偏誤量值（bias size）介於 -0.5 至 0.5 作為判斷標準，若超出此標準表示存在顯著的評分偏差（McNamara, 1996）。表 3 的結果報告分析顯現期中考有四位評分者（MA, PR, PU, MU）出現自身評分會因口試評分項目不同而產生偏寬鬆或偏嚴苛之評分，而期末考只有兩位有評分偏差現象。期中考的口試評量 MA 對使用表達人際關係的詞彙語法正確度（accuracy）（*t* = 2.76）表現項目評分寬鬆，但對於對話展現二人輪流對話的樣貌互動能力（interaction）卻是嚴苛（*t* = -2.11）；然而 PR 對正確度（accuracy）表現項目評分嚴苛（*t* = -2.18），但對於對話互動能力（interaction）（*t* = 2.73）及口語涵蓋所有規定內容（content）卻是寬鬆（*t* = 2.45）；PU 對口語內容（content）表現項目評分嚴苛（*t* = -3.35）；MU 卻是寬鬆評分發音（pronunciation）（*t* = 2.40）。在期末考方面，評分者 AA 似乎相當重視是否能使用清楚辨識的母音、子音、重音、和音調發音（pronunciation）（*t* = -4.25），但對於言語流利度（fluency）卻給予寬鬆之評分（*t* = 2.55）；MA 期中與期末評量分析都有評分偏差現象，MA 在期末口試評量不僅在對話互動（interaction）項目評分嚴苛（*t* = -2.21），且對口語內容（content）也採取偏嚴苛（*t* = -2.03）的評分態度。

表 3
109 學年度課程交叉分析評分者和口試評分項目考驗達顯著結果報告表

評分者	口試評分項目	偏差量	標準誤	*t* 值	自由度	機率
期中考						
MA	互動（interaction）	-0.79	.37	-2.11	38	.0419
MA	正確度（accuracy）	1.13	.41	2.76	38	.0089
PR	正確度（accuracy）	-0.73	.33	-2.18	38	.0358
PR	互動（interaction）	1.98	.72	2.73	38	.0096
PR	內容（content）	2.29	.94	2.45	38	.0192
PU	內容（content）	-1.82	.54	-3.35	38	.0019
MU	發音（pronunciation）	0.88	.37	2.40	38	.0212
期末考						
AA	流利度（fluency）	1.18	.46	2.55	37	.0152
AA	發音（pronunciation）	-1.37	.32	-4.25	37	.0001
MA	內容（content）	-1.13	.56	-2.03	37	.0496
MA	互動（interaction）	-0.98	.44	-2.21	37	.0333

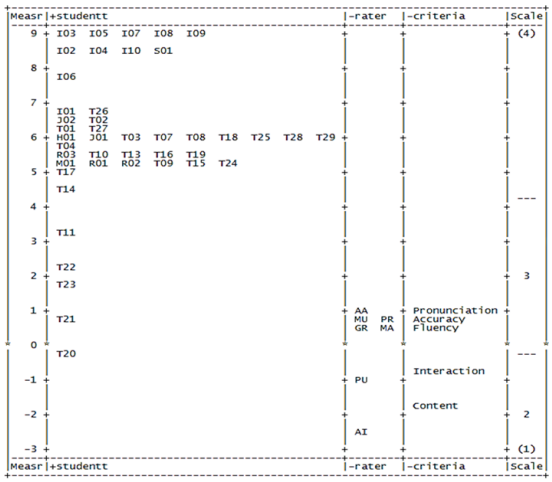
註：僅列出達顯著水準者。

$p < .05$.

圖 2-1、2-2、表 5 及表 7 顯示第二次初級印尼語課程八位評分者嚴苛度差異。小考 2 中評分員 PR 最嚴厲（0.86 logits），KE 的 logit 值最低（= -0.91），代表評分最為寬鬆。就如同 109 學年度課程，助教團檢驗討論小考 2 所有評分者的嚴苛度結果後，接下來的小考 4 轉變為評分員 NI、KE 最為嚴

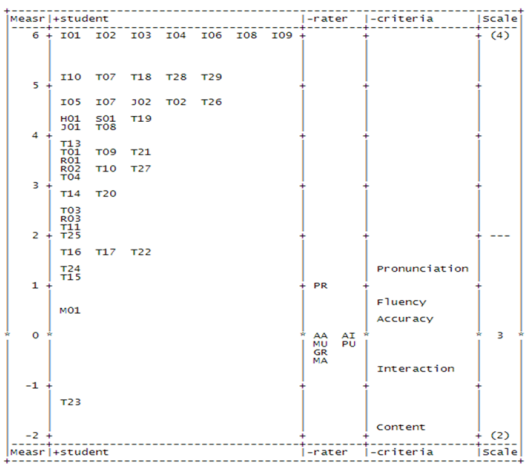
苛（0.44 logits），PU 最寬鬆（-1.33 logits）。評分者的分離度從小考 2 到小考 4 從 3.00 降為 2.68，信度值也從 0.90 降為 0.88，顯示評分者在看過、檢討過小考 2 的結果後，雖然整體嚴苛度差異逐漸縮小，但是卡方值仍顯著（$p < .05$），表示評分者之間的嚴苛度仍有不同。

圖 1-1
109 期中口試變數分布圖



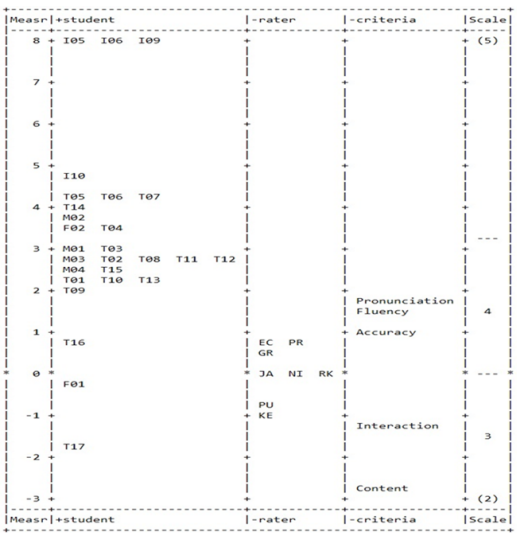
註：Measr 為參數值的刻度，student 為學生口說能力值的分布位置（I = 印尼；T = 臺灣；J = 日本；R = 俄羅斯；M = 港澳 1；H = 港澳 2；S = 新加坡），rater 為評分者的嚴厲度的分布位置，criteria 則為口說表現評分項目的難度分布。

圖 1-2
109 期末口試變數分布圖



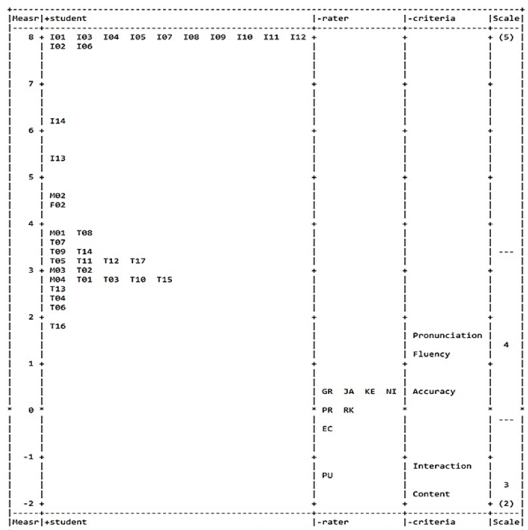
註：Measr 為參數值的刻度，student 為學生口說能力值的分布位置（I = 印尼；T = 臺灣；J = 日本；R = 俄羅斯；M = 港澳 1；H = 港澳 2；S = 新加坡），rater 為評分者的嚴厲度的分布位置，criteria 則為口說表現評分項目的難度分布。

圖 2-1
110 小考 2 口試變數分布圖



註：Measr 為參數值的刻度，student 為學生口說能力值的分布位置（I = 印尼；T = 臺灣；M = 馬來西亞；F = 其他外籍生），rater 為評分者的嚴厲度的分布位置，criteria 則為口說表現評分項目的難度分布。

圖 2-2
110 小考 4 口試變數分布圖



註：Measr 為參數值的刻度，student 為學生口說能力值的分布位置（I = 印尼；T = 臺灣；M = 馬來西亞；F = 其他外籍生），rater 為評分者的嚴厲度的分布位置，criteria 則為口說表現評分項目的難度分布。

表 4 及表 5 為兩次初級印尼語課程各面向分析參數整理，藉此探討學生口說能力、評分者、以及口說表現項目三個層面於兩門課程四次考試的差異性。結果呈現所有的固定效果卡方檢定（fixed chi-square）均達顯著水準（$p < .05$），拒絕虛無假設，表示不同學生印尼語口說能力、不同評分者之嚴苛度、及不同印尼語口說表現項目具有差異。本研究參考 Eckes（2015）的研究，將學生口說能力設定為浮動未固定（non-centered）的變數面向，亦即同時將評分者與口說表現項目的估計參數皆設為 0（$M = 0$），以便根據 Rasch model 設計的理論基礎，建立所要估計參數的參照依據，藉此觀察分析學生印尼口說能力與評分者嚴苛度，以及與口說表現項目難度之間關係。

表 4
109 學年度期中與期末考 Rasch 各層面估計數據

參數值（logits）	學生口說能力		評分者		口說表現項目	
	期中考	期末考	期中考	期末考	期中考	期末考
平均（*M*）	6.01	3.93	0.00	0.00	0.00	0.00
標準誤（*SE*）	.65	.68	.20	.19	.17	.17
Infit MNSQ	0.98	0.97	0.97	1.01	1.01	1.02
Outfit MNSQ	1.26	1.09	1.40	1.09	1.40	1.09
分離度	2.54	2.09	5.31	2.00	5.91	6.22
信度	.87	.81	.97	.80	.97	.97
χ^2	756.4*	476.1*	123.9*	41.5*	145.8*	169.9*
N	44	44	7	7	5	5
df	43	43	6	6	4	4

註：Infit MNSQ 為訊息加權適配度均方值；Outfit MNSQ 為極端值加權適配度均方值。
* $p < .05$.

表 5
110 學年度小考 2 與小考 4 Rasch 各層面估計數據

參數值（logits）	學生口說能力		評分者		口說表現項目	
	小考 2	小考 4	小考 2	小考 4	小考 2	小考 4
平均（*M*）	3.34	5.13	0.00	0.00	0.00	0.00
標準誤（*SE*）	.47	.80	.19	.20	.16	.16
Infit MNSQ	0.97	0.98	0.99	0.96	1.08	0.99
Outfit MNSQ	1.15	1.76	1.15	2.29	1.15	2.29
分離度	3.65	2.38	3.00	2.68	10.97	8.26
信度	.93	.85	.90	.88	.99	.99
χ^2	542.6*	274.2*	78.8*	56.1*	493.7*	317.9*
N	27[6]	36[7]	8	8	5	5
df	26	35	7	7	4	4

註：Infit MNSQ 為訊息加權適配度均方值；Outfit MNSQ 為極端值加權適配度均方值。
* $p < .05$.

109 學年度的期中、期末考與 110 學年度的小考 2、小考 4 的評分者嚴苛度與口說表現項目的難易度平均估計參數（logit）皆為 0，學生口說能力的平均估計參數分別為 6.01、3.93、3.34 以及 5.13，顯示整體學生能力平均水準遠高於評分者嚴苛度與口說表現項目的難度，評分者嚴苛度似乎偏鬆，應加強對於不同學生水準對應到口說表現項目應有適當評分標準的訓練；這也表示經過一學期的評分者訓練及學員的印尼語口語訓練，使得差異性確實有所降低。109 學年度及 110 兩個課程的第二次學生口說能力分離度數值皆呈現下降，分別從 2.54（期中），降低至 2.09（期末），以及從 3.65（小考 2）降至 2.38（小考 4）；然而雖有降低，但學生印尼語口說能力還是呈現出差異性，因為兩次的分離度數值皆顯示這兩門課程學生至少各可區分成兩個以上具有統計意義的印尼語口說能力群體，且皆具有 0.8 以上的穩定信度係數值。再者，兩門課程的口說表現項目的分離度與信度皆呈現較高的數值，卡方檢定亦達顯著（$p < .05$），109 學年度期中、期末考分離指數分別為 5.91

（信度 = .97）及 6.22（信度 = .97），110 學年度小考 2、小考 4 則分別為 10.97（信度 = .99）及 8.26（信度 = .99），顯示口說項目難度能有效的區隔這些學習者能力，但口說正確、流利、互動、及發音表現項目之間具有顯著性的不同，在判別上有很大的差別。最後，109 學年度課程之評分者分離度從期中考的 5.31 在期末考時減少為 2，嚴苛度差異從至少分成 5 個層級的類別區隔下降成至少只分成 2 個層級，表示評分者的判斷歧異性減少，逐漸趨於一致性，且估計分離信度數值也從 .97 降到 .80。110 學年度的課程亦是如此，該課程評分者的分離度從 3.00（小考 2）降為 2.68（小考 4），信度值也從 .90 降為 .88。對於評分者趨向一致性而信度卻降低的部分，須注意 Rasch 信度在檢驗不同面向時有不同的意義。一般而言，對於學生口說能力面向，信度數值愈高愈能區分學生的表現；亦即，分離信度數值越高，表示此測驗越能將學生口說能力區分為不同程度。然而，對於評分者面向而言，較低的信度卻是被預期的，因為低的評分者信度數值表示不同的評分者有相同的評分嚴苛度（Eckes, 2015; Linacre, 2022b; McNamara, 1996; Park, 2004; Weigle, 1998）。因此，可能受益於培訓，兩個課程的評分者嚴苛度不但逐漸趨於一致性，而且不同評分者也趨向有相同評分嚴苛度。

綜合上述，可見期初的評分訓練固然必要，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。但仍有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，而應採取多位評分員共同評定的方向努力。

表 6
109 學年度評分者嚴苛度估計結果

評分者	參數估計值 Measure		嚴苛度標準誤（*SE*）		Infit MNSQ	
	期中	期末	期中	期末	期中	期末
AA	1.03	0.06	.17	.19	0.89	0.99
PR	0.63	0.97	.17	.17	0.94	0.87
MU	0.63	-0.24	.17	.20	0.98	0.89
MA	0.60	-0.45	.17	.21	1.06	1.22
GR	0.48	-0.28	.18	.20	1.23	1.04
PU	-0.95	-0.09	.23	.19	0.98	1.11
AI	-2.42	0.02	.33	.19	0.75	0.96

註：Infit MNSQ 為訊息加權適配度均方值。

表 7
110 學年度評分者嚴苛度估計結果

評分者	參數估計值 Measure		嚴苛度標準誤（*SE*）		Infit MNSQ	
	小考 2	小考 4	小考 2	小考 4	小考 2	小考 4
PR	0.86	0.07	.18	.20	1.14	0.79
EC	0.64	-0.37	.18	.20	0.78	0.74
GR	0.41	0.41	.18	.19	0.94	1.08
RK	-0.04	-0.01	.19	.20	0.67	0.55
JA	-0.07	0.34	.19	.19	1.32	0.87
NI	-0.07	0.44	.19	.19	0.63	1.18
PU	-0.83	-1.33	.20	.23	1.45	1.43
KE	-0.91	0.44	.20	.19	0.98	1.01

註：Infit MNSQ 為訊息加權適配度均方值。

表 6 與表 7 展現兩個課程個別的評分者嚴厲度估計結果。透過 fit 適合度統計數值分析評分者自身一致性程度，亦即檢視評分者自身給分的穩定度。如前所述，當 Infit 及 Outfit 均方值互有高低時，本研究以 0.5—1.5 Infit 均方值為適切指標。兩個課程四次考試 Infit MNSQ 適配度統計值皆在理想範圍內，即本研究評分者適合 Rasch 模式，在評分者的測量上具有效度，且信度值都在 0.8 以上，更表示此數據在估計上是穩定的。簡言之，結果顯示雖然各評分者自身給分一致性似乎是穩定且良好，但如前述表 3 所顯示，不同評分者針對不同項目的評分嚴苛度仍有差異。目前由多位評分者共同擔綱，刪除離群值、取其平均數，或許是權宜之計。未來應深入了解原因，以進行評分標準之調整。

（二）搭檔的語言能力差異對口試成績的影響

為了進一步分析搭檔的語言能力差異對兩個課程 12 次口試成績的影響，尤其探討檢驗是否「高攀」有優勢，而「低就」是否會受到連累。本研究將學生依搭檔的選擇分成四個組別，分別是 LL[8]（初學者—初學者）、LH（初學者—非初學者）、HL（非初學者—初學者）、HH（非初學者—非初學者），其中 LH 是指一位初學者搭檔一位印馬華人學生，即所謂高攀組，而 HL 是指一位印馬華人學生搭配一位初學者，即所謂低就組。再者，為因應探討組別搭檔對於口試成績是否有所差異的問題，此部分數據資料排除口試時臨時交換搭檔而導致組別不一致性的學員，以期達到可信賴的統計分析，詳細的學生數量請參閱註釋 9 與 10。

四組組數不平均，因此採用 Kruskal Wallis test 多樣本中位數差異檢定。以下依序呈現兩次初級印尼語課程的結果。如表 8 所示，109 學年度的 6 次口試卡方值皆達顯著（$p < .05$），表 9 所示的 110 學年度口試則是在期中考、小考 3、小考 4 和期末考達顯著（$p < .05$）。卡方值達顯著表示不同搭檔的組別在成績上有顯著差異，110 學年度未達顯著的小考 1 和小考 2 係因新冠肺炎疫情，有 10 位「非初學者」學生尚未進班上課，未參與口試。

首先，檢驗「高攀」是否有優勢：初學者若與印馬華僑搭檔是否比較佔優勢？檢驗表 8 與表 9 兩個學年度課程的 12 次口試統計資料，只有 110 年度的期末考「初學者與高配 LH」的平均值 91.4 高於「初學者同程度互相搭配 LL」的平均值 89.6，其他的 11 次考試都是同為「初學者同程度互相搭配 LL」比「初學者與高配 LH」的平均分數高。換言之，從平均值初步檢驗似乎看出「高攀」並沒有優勢。隨後使用無母數 Dunn 事後比較，也並未發現任何一次口試 LH—LL 有任何顯著的差異，因此，本研究的統計結果證實「高攀」並沒有優勢。

其次，檢驗「低就」是否會吃虧：印馬華僑與初學者搭檔成績是否會被拖下水？若比較各次考試的各組平均成績，「印馬華僑與低配 HL」的平均成績幾乎都低於「印馬華僑同程度互相搭配的 HH 」的平均成績，除了 110 年度（表 9）的小考 3 例外（HL = 94.3, HH = 93.2）。此統計結果是否可解讀為：若學生本身的能力較好，但與程度較弱的學生搭檔，其成績反而被拖累。緊接著，本研究進一步使用無母數 Dunn 事後比較，結果顯示 110 學年度並沒有任何一組呈現 HL 和 HH 之間有顯著的差異；然而 109 年度包含小考 4、期中考以及期末考的成對比較統計顯示 HL 與 LH 或 LL 均達顯著差異（$p < .05$）。亦即，「印馬華僑與低配 HL」的平均成績，無論期中考或期末考均高過「初學者與高配 LH」以及「初學者同程度互相搭配 LL」的組合，小考 4「印馬華僑與低配 HL」的平均成績也高過「初學者與高配 LH」：HL > LH（期中考：HL = 99.0, LH = 91.7；期末考：HL = 99.0, LH = 90.7；小考 4：HL = 100.0, LH =85.0），HL > LL（期中考：HL = 99.0, LL = 92.2；期末考：HL = 99.0, LL = 92.2）。換言之，統計結果也證實「低就」也不會吃虧。

總而言之，本研究統計結果驗證了初學者不論是找同程度的初學者搭檔，或是和能力比自己好的印尼華僑搭檔，對成績並沒有太大的影響。而程度高的印尼／馬來西亞華僑學生無論是和同樣高程度的夥伴搭檔，或是與初學者搭檔，也不會影響其口試成績。此結果呼應 Davis（2009）分析中國學生英語口語對話搭檔和 Son（2016）分析韓國學生英語口試結果，均未發現搭檔語言能力對成績有任何統計上顯著差異。相較於韓國學生遇強則縮，當程度較低和較高者搭檔時話語量反而減少，本研究並未發現臺灣學生有此現象。

表 8
依搭檔組合分類之統計量與檢定分析彙整表（109 學年度）

口試	搭檔組合	學生數量[9]	平均值	標準差	最小值	最大值	Kruskal-Wallis 檢定	
							等級平均數	檢定統計量
小考 1	LL	28	86.3	10.1	65	100	18.27	$\chi^2(2) = 6.307$ $p = .043$*
	LH	6	85.0	8.9			16.92	
	HL	5	97.0	6.7			31.70	
	HH	0	–	–			–	

（續下頁）

表 8
依搭檔組合分類之統計量與檢定分析彙整表（109 學年度）（續）

口試	搭檔組合	學生數量[9]	平均值	標準差	最小值	最大值	Kruskal-Wallis 檢定	
							等級平均數	檢定統計量
小考 2	LL	28	87.1	9.4	70	100	19.89	$\chi^2(2) = 8.761$ $p = .033$*
	LH	6	85.0	11.0			17.58	
	HL	6	95.8	4.9			31.75	
	HH	4	97.5	2.9			34.25	
期中考	LL	26	92.2	5.5	74	100	16.90	$\chi^2(2) = 18.036$ $p = .000$*
	LH	6	91.7	4.8			15.33	
	HL	6	99.0	1.1			35.75	
	HH	2	99.5	0.7			37.00	
小考 3	LL	28	89.8	9.2	65	100	19.86	$\chi^2(2) = 12.700$ $p = .005$*
	LH	6	86.7	9.3			15.25	
	HL	6	97.5	2.7			31.75	
	HH	4	100.0	0.0			38.00	
小考 4	LL	26	91.0	5.6	75	100	15.90	$\chi^2(2) = 11.652$ $p = .009$*
	LH	3	85.0	10.0			9.67	
	HL	3	100.0	0.0			30.50	
	HH	2	100.0	0.0			30.50	
期末考	LL	26	92.2	7.4	66	100	17.48	$\chi^2(2) = 15.284$ $p = .002$*
	LH	6	90.7	7.1			14.75	
	HL	6	99.0	0.9			33.33	
	HH	2	100.0	0.0			38.50	

* $p < .05$.

表 9
依搭檔組合分類之統計量與檢定分析彙整表（110 學年度）

口試	搭檔組合	學生數量[10]	平均值	標準差	最小值	最大值	Kruskal-Wallis 檢定	
							等級平均數	檢定統計量
小考 1	LL	10	84.2	5.7	64	100	12.65	$\chi^2(2) = 3.544$ $p = .170$
	LH	7	77.7	10.3			8.86	
	HL	7	87.1	13.5			15.93	
	HH	0	–	–			–	
小考 2	LL	10	84.0	3.5	60	100	11.55	$\chi^2(2) = .665$ $p = .717$
	LH	7	82.0	11.7			12.07	
	HL	7	87.7	9.6			14.29	
	HH	0	–	–			–	
期中考	LL	10	90.4	5.1	80	100	12.00	$\chi^2(2) = 18.682$ $p = .000$*
	LH	7	89.1	5.0			9.64	
	HL	7	94.9	5.0			19.07	
	HH	10	99.2	1.7			27.40	
小考 3	LL	10	86.4	7.8	72	100	13.95	$\chi^2(2) = 10.612$ $p = .014$*
	LH	7	82.6	6.2			9.71	
	HL	7	94.3	5.1			23.07	
	HH	10	93.2	9.6			22.60	
小考 4	LL	10	88.1	1.9	82	100	11.40	$\chi^2(2) = 21.094$ $p = .000$*
	LH	7	86.7	2.8			8.71	
	HL	7	94.1	6.0			20.43	
	HH	10	99.5	1.1			27.70	
期末考	LL	10	89.6	6.9	80	100	12.30	$\chi^2(2) = 8.271$ $p = .041$*
	LH	7	91.4	4.9			13.93	
	HL	7	94.9	4.5			19.93	
	HH	10	96.8	4.1			23.50	

* $p < .05$.

綜上所述，雖然影響成績之變數頗多，但讓學生自由決定口試搭檔，只要測試次數夠多，評分者的嚴苛度相當，且依公平公正原則評分，學生並不會因為搭檔的能力，而對個人口試表現產生太大的影響。另外，分析學生對口試方法的課堂即時回饋後，發現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，並沒有表達任何遇弱則強或遇強則縮的感覺。雖然搭檔的口說能力並不會影響個人的口試成績，然而當臺灣的大學階段初級印尼語通識課程有超過二成以上學生均為印尼華僑時，教學者應妥善運用僑生得天獨厚的語言能力，賦予他們在初學者中擔任同儕助教的角色，不但能促進彼此文化交流，且能減低僑生取得營養學分的刻板印象。

結論

此項行動研究發現與多數 MFRM 多層面 Rasch 模式研究結果一致，如張可家等人（2011）、藍珮君（2012）、廖才儀（2016）、Eckes（2005, 2009, 2015）、Knoch（2011）、Sundqvist 等人（2020）及 Weigle（1998）等，不同評分者之嚴苛度確實有所差異，即便經過訓練，尚需實際執行評分後才能發現其差異。即使事先給予訓練有助於未來評分的一致性，但仍有其他無法掌控的個人情緒因素可能影響下一次評分的嚴苛度。因此面對高風險的口試，若能由多位受過訓練的評分者共同擔綱，刪除離群值、取其平均數，或許是權宜之計。本研究雖然發現不同評分者針對不同項目的嚴苛度仍有差異，但並未對此深入探討。未來研究應在培訓過程針對嚴苛度之標準做深入說明與試評，探討哪些項目不易改變嚴苛度的培訓，據以修正評分量表。

其次，評分項目之間的難易度會因口試方式有所差異。此課程的口試方式有明確範圍，鼓勵事先準備，因此從受試者的角度而言，對話內容和詞彙語法的正確度最容易掌握，人際互動與發音則最難拿分。從評分者的角度而言，最難和最容易的項目和受試者吻合，但由於評分者容易聽出初學者的語誤，因此針對詞彙語法正確性的評分敏銳度自然比初學的受試者來得犀利。

最後，統計結果也顯示選擇與不同語言能力背景搭檔口試並不會影響其口試成績，此呼應 Davis（2009）及 Son（2016）的第二語言口語能力研究結果。其實學生喜歡從一而終，滿意並包容自己所選的搭檔。因此，由學生自行選擇搭檔，不僅人性化且有穩定軍心的效果。然而由於本研究並非隨機分派，且組數落差很大，再加上學生都是自由選組，且自由演練對話腳本，其中同學間彼此的熟悉度、腳本的難易度都可能影響表現，因此，很難就此完全定論不同程度搭檔對成績的影響。

本研究雖有未盡事宜，然而在口語教學評量上仍有啟發與應用。透過比較兩次課程學生搭檔的選擇和評分者嚴苛度對於四次平時口語評量和期中、期末成績的影響，得知日後口語評量應讓學生自由選擇對話搭檔，輔以鼓勵機制讓印尼華僑多跟初學者搭配，以達到雙贏的效果。此外，將助教的評分訓練從傳統耳提面命進化至從做中學，根據多層面的 Rasch 分析方法找到降低評分者嚴苛度不一致問題的方案，有助於「及早發現、及早診斷」，以提升評分者的一致性。

註釋

[1] Winsteps 軟體 https://www.winsteps.com/ministep.htm

[2] 平時 4 次口試和期中期末口試形式相同，內容根據數位教材網站（何德華等人，2019），將課文內容濃縮成幾個情境，二人搭檔對話，不超過 3 分鐘，可事先演練。以下是期末印尼語口試範例四情境內容：

（1）A 的寵物把公園的草地踩壞了。警衛告誡 A 要看管好自己的寵物。Hewan peliharaan A merusak rumput di taman. Penjaga taman memberitahu A untuk mengawasi hewan peliharaannya.

（2）公園警衛提醒 A 遠離蜂窩。Penjaga taman memperingatkan A untuk menjauh dari sarang lebah.

（3）A 牙齒疼痛難耐，B 帶 A 去看牙醫。Gigi A terasa sakit dan A mengeluh tentang itu. B membawanya untuk periksa ke dokter gigi.

（4）演出一場牙醫診所醫病對話。例如：牙醫告知病人需要如何治療，牙醫囑咐病人

如何保健牙齒、何時需要回診。Percakapan pasien-dokter di klinik dokter gigi. Dokter gigi memberikan arahan kepada pasien tentang apa yang harus dilakukan.

[3] 學生口試自評：「你對今天自己的口試表現是否滿意？下面五點你是否都做到了？（1）涵蓋所有指定內容、（2）使用對話中合宜的詞彙和句子、（3）事先有準備地流利互動、（4）清楚的印尼語發音、（5）自然的一來一往對話」。

[4] Minifac 為免費 Facets 展示版本軟體 https://www.winsteps.com/minifac.htm；目前最新版本為 Facets 3.84.0。

[5] 本文第一作者曾在私人聚會中詢問熟識助教的同儕，請他們猜測誰會是嚴苛或寬鬆的評分者，結果得到評價與本研究結果不謀而合，AA 最嚴苛，AI 最寬鬆。

[6] 110 學年度的全體修課人數是 38 人，但有一位學員並無意願參與本研究，因此數據執行時其資料都是被排除的。因 COVID-19 新冠肺炎疫情，小考 2 時有 10 位印尼華僑尚在隔離，來不及參與考試，小考 2 人數為 27 人；38 − 10（隔離）− 1（不參與研究的學員）= 27。

[7] 110 學年度小考 4 人數為 36 人，全體修課人數是 38 人，去除未參與組別研究的兩位學員（包含 1 位未參與全程研究及一位參與嚴格度研究但未參與組別研究學員）：38 − 2（不參與組別研究的學員）= 36。

[8] 第一個英文字母表示學生的程度，第二個字母為其搭檔的程度。

[9] 109 學年度的全體修課人數是 44 人，因 COVID-19 新冠肺炎疫情，小考 1 時有 5 位印尼華僑尚在隔離，小考 1 人數為 39 人：44 − 5（隔離）= 39；小考 4 人數為 34 人，1 位缺考，1 位印尼華僑分飾兩角而不符合兩人對話，8 人因組別改變不列入計算：44 − 1（缺考）− 1（不符合兩人對話）− 8（組別改變）= 34；期中考及期末考人數各為 40 人，各有 4 人因組別改變不列入計算：44 − 4（組別改變）= 40。

[10] 如註釋 6 與 7 所提及，110 學年度的全體修課人數是 38 人，此部分資料統計因顧及組別一致性而排除了無意願參與組別研究的學員兩名及其搭檔，共 4 位；小考 1 和小考 2 有 10 位印尼華僑因新冠肺炎疫情隔離缺考的。

參考文獻

Smith, E. V., Jr., & Smith, R. M.（2017）：《羅氏測量：應用與導讀》（莫慕貞、張權主編）。一豐印刷有限公司。（原著出版年：2004）[Smith, E. V., Jr., & Smith, R. M. (2017). *Introduction to Rasch measurement: Theory, models and applications* (M. M. C. Mok & Q. Zhang, Eds. & Trans.). Yi feng yinshua youxian gongsi. (Original work published 2004)]

王文中（2004）：〈Rasch 測量理論與其在教育和心理之應用〉。《教育與心理研究》，*27*，637–694。[Wang, W.-C. (2004). Rasch measurement theory and application in education and psychology. *Journal of Education & Psychology, 27*, 637–694.]

王佳琪（2020）：〈科學想像力圖形測驗之驗證〉。《教育心理學報》，*51*，341–367。[Wang, C.-C. (2020). Validation of the scientific imagination test-figural. *Bulletin of Education Psychology, 51*, 341–367.] https://doi.org/10.6251/BEP.202003_51(3).0001

余民寧（2013）：〈口試在國家考試應用之再檢討與改進〉。《國家菁英季刊》，*9*（2），87–107。[Yu, M.-N. (2013). Re-examination and improvement of oral exam on the application of National Official Examination. *National Elite, 9*(2), 87–107.]

何德華（2019）：〈印尼語 TEAL 創意互動教學測驗與評量〉。《通識教育學刊》，*24*，79–131。[Rau, D. V. (2019). Large class assessment of Indonesian language proficiency. *Taiwan Journal of General Education, 24*, 79–131.] https://doi.org/10.6360/TJGE.201912_(24).0003

何德華、李萍、賴思悅、潘家貝、阿芬達（2019）：〈印尼旅蛙來電了〉。YouTube。

https://www.youtube.com/playlist?list=PLQn99bzkJv9yDZbCZaQE4Sj23guoQ9AVu [Rau, D. V., Pulungan, P. L. S., Lase, A., Panggabean, G. C., & Samosir, A. (2019). *Indonesian travel frog called.* YouTube. https://www.youtube.com/playlist?list=PLQn99bzkJv9yDZbCZaQE4Sj23guoQ9AVu]

吳昭容、曾建銘、鄭鈐華、陳柏熹、吳宜玲（2018）：〈領域特定詞彙知識的測量：三至八年級學生數學詞彙能力〉。《教育研究與發展期刊》，*14*（4），1–40。[Wu, C.-J., Cheng, C.-M., Cheng, C.-H., Chen, P.-H., & Wu, Y.-L. (2018). The measurement of domain-specific vocabulary knowledge: The mathematical vocabulary ability of third to eighth grade students. *Journal of Educational Research and Development, 14*(4), 1–40.] https://doi.org/10.3966/181665042018121404001

林小慧、林世華、吳心楷（2018）：〈科學能力的建構反應評量之發展與信效度分析：以自然科光學為例〉。《教育科學研究期刊》，*63*（1），173–205。[Lin, H.-H., Lin, S.-H., & Wu, H.-K. (2018). Developing and validating a constructed-response assessment of scientific abilities: A case of the optics unit. *Journal of Research in Education Sciences, 63*(1), 173–205.] https://doi.org/10.6209/JORIES.2018.63(1).06

林怡君、張麗麗、陸怡琮（2013）：〈Rasch 模式建置國小高年級閱讀理解測驗〉。《教育心理學報》，*45*，39–61。[Lin, I.-C., Chang, L., & Lu, I.-C. (2013). The development of reading comprehension test for 5th and 6th graders using the Rasch model. *Bulletin of Educational Psychology, 45*, 39–61.] https://doi.org/10.6251/BEP.20121128

姚漢禱（2004）：〈利用線性 logistic Rasch 模式估計排名賽的成績表現—以 34 屆世界盃棒球賽為例〉。《國立體育學院論叢》，*15*（1），149–158。[Yau, H.-D. (2004). Valid estimation performances of rank ordering matches for using linear logistic Rasch measurement-2001 Baseball World Cup. *Journal of Physical Education and Sports, 15*(1), 149–158.] https://doi.org/10.6591/JPES.2004.10.11

張可家、施泰亨、藍珮君（2011，6 月 25 日）：〈「華語文口語能力測驗」評分者一致性探討〉（口頭發表論文）。華語文能力測驗成果發表會，臺北市。https://tocfl.edu.tw/assets/files/SP_research.pdf [Chang, K.-C., Shih T.-H., & Lan, P.-J. (2011, June 25). *'Huayuwen kouyun engli ceyan' pingfenzhe yizhixing tantao* (Paper presentation). Test of Chinese as a Foreign Language Chengguo Fabiaohui, Taipai. https://tocfl.edu.tw/assets/files/SP_research.pdf]

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*，105–128。[Chang, H.-L., & Wu, S.-C. (2008). A multi-facet Rasch analysis on rating the academic scientific papers. *Psychological Testing, 55*, 105–128.] https://doi.org/10.7108/PT.200804.0105

陳映孜、何曉琪、劉昆夏、林煥祥、鄭英耀（2017）：〈從教師自編科學成就測驗之 Rasch 分析看教與學〉。《教育科學研究期刊》，*62*（3），1–23。[Chen, Y.-T., Ho, H.-C., Liu, K.-H., Lin, H.-S., & Cheng, Y.-Y. (2017). Glimpse into teaching and learning using Rasch analyses of a teacher-made science achievement test. *Journal of Research in Education Sciences, 62*(3), 1–23.] https://doi.org/10.6209/JORIES.2017.62(3).01

陳建亨、楊凱琳（2021）：〈題型對學生數學表現水準之影響—以相似形為例〉。《教育科學研究期刊》，*66*（3），247–277。[Chen, C.-H., & Yang, K.-L. (2021). Effects of item type on student mathematics performance: Similar figures as an example. *Journal of Research in Education Sciences, 66*(3), 247–277.] https://doi.org/10.6209/JORIES.202109_66(3).0008

陸雲鳳（2016）：〈利用 Rasch 測量分析桌球女單優秀個案比賽技術分析〉。《臺灣體育學術研究》，*61*，139–150。[Lu, Y.-F. (2016). Using Rasch measurement to analyze the competition techniques for excellent cases of women's singles. *Taiwan Journal of Sports Scholarly Research, 61*, 139–150.] https://doi.org/10.6590/TJSSR.2016.12.08

莫慕貞（2019，11 月 8—9 日）：〈精進教學工作坊：Rasch 可觀測量在大學甄試檔案評量之應用〉（工作坊）。國立中正大學，嘉義。https://reurl.cc/edQbyW [Mok, M. M. C. (2019, November 8–9). *Jingjin jiaoxue gongzuofang: Rasch keguanceliang zai daxue zhenshi dangan pingliang zhi yingyong* (Workshop). National Chung Cheng University, Chiayi. https://reurl.cc/edQbyW]

曾盟堡（2002）：〈是誰評判不公〉。《測驗統計年刊》，*10*，121–133。[Tseng, M.-P. (2002). Who is unfair? *Journal of Research on Measurement and Statistics, 10*, 121–133.] https://doi.org/10.6773/JRMS.200212.0121

廖才儀（2016，5 月 28 日）：〈華語文口語能力測驗」評分者內評分偏誤研究—以入門基礎級為對象〉（口頭發表論文）。第九屆國際電腦漢語教學研討會（TCLT9），澳門。https://tocfl.edu.tw/assets/files/4EST688_CaiyiLiao.pdf [Liao, T.-Y. (2016, May 28). *A FACETS analysis of rater bias in measuring TOCFL Speaking assessment* (Paper presentation). The 9th International Conference and Workshops on Technology and Chinese Language Teaching (TCLT9), Macau. https://tocfl.edu.tw/assets/files/4EST688_CaiyiLiao.pdf]

謝如山、謝名娟（2013）：〈多層面 Rasch 模式在數學實作評量的應用〉。《教育心理學報》，*45*，1–18。[Hsieh, J.-S., & Hsieh, M.-C. (2013). An application of many-facet Rasch model to evaluate mathematics performance assessment. *Bulletin of Educational Psychology, 45*, 1–18.] https://doi.org/10.6251/BEP.20121101.1

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*，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*, 551–566.] https://doi.org/10.6251/BEP.20160801

謝名娟（2020）：〈從多層面 Rasch 模式來檢視不同的評分者等化連結設計對參數估計的影響〉。《教育心理學報》，*52*，415–436。[Hsieh, M.-C. (2020). Investigating the effects of rater equating designs on parameter estimates in the context of preservice principal oral performance. *Bulletin of Educational Psychology, 52*, 415–436.] https://doi.org/10.6251/BEP.202012_52(2).0008

藍珮君（2012）：〈以多面向 Rasch 測量模式分析 TOCFL 口語測驗評分者訓練效果〉。國家教育研究院（主編），《永續教育發展創新與實踐論文集：2010 年國際學術研討會—測驗及評量論文專輯》，頁 125–139。國家教育研究院。 [Lan, P. J. (2012). Using many-facet Rasch measurement to examining rater training effects of TOCFL speaking. In National Academy

for Educational Research. (Ed.), *2010 NAER Conference: Education for sustainable development-Innovation and implementation* (pp. 125–139). National Academy for Educational Research.]

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341–366. https://doi.org/10.1177/0265532209104666

Chuang, E. (2018). *EFL students' anxieties towards paired-oral testing* [Unpublished master's thesis]. Tunghai University.

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance.* Peter Lang. https://doi.org/10.3726/978-3-653-01227-9

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396. https://doi.org/10.1177/0265532209104667

Davis, L. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* [Unpublished doctoral dissertation]. University of Hawai'i at Mānoa.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443. https://doi.org/10.1177/0265532209104669

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing, 32*(1), 101–120. https://doi.org/10.1177/0265532214544393

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. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H, pp. 1–52). Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

Eckes, T., & Jin, K.-Y. (2021). Measuring rater centrality effects in writing assessment: A Bayesian facets modeling approach. *Psychological Test and Assessment Modeling, 63*(1), 65–94.

Együd, G., & Glover, P. (2001). Readers respond. Oral testing in pairs-secondary school perspective. *ELT Journal, 55*(1), 70–76. https://doi.org/10.1093/eltj/55.1.70

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Lawrence Erlbaum Associates Publishers.

Engelhard, G., Jr., & Myford, C. M. (2003). Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition program with a many-faceted Rasch model. *ETS Research Report Series, 2003*(1), i-60. https://doi.org/10.1002/j.2333-8504.2003.tb01893.x

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of raters. *Theory and Practice in Language Studies, 1*(11), 1531–1540. https://doi.org/10.4304/tpls.1.11.1531-1540

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. 367–471). Cambridge University Press.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*(1), 36–41. https://doi.org/10.1093/elt/53.1.36

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the first certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119. https://doi.org/10.1080/15434300801934702

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E. D., ffrench, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice, 18*(3), 217–237. https://doi.org/10.1080/0969594X.2011.574605

Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Hsieh, C.-N. (2011). *Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency* [Unpublished doctoral dissertation]. Michigan State University.

Huang, H.-T. D., Hung, S.-T. A., & Hong, H.-T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly, 13*(4), 283–301. https://doi.org/10.1080/15434303.2016.1236111

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51–66.

Jones, L. (2007). *The student-centered classroom.* Cambridge University Press.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* [Unpublished doctoral dissertation]. Teachers College, Columbia University.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28*(2), 179–200. https://doi.org/10.1177/0265532210384252

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). Cambridge University Press.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2022a). *Facets computer program for many-facet Rasch measurement* (Version 3.84.0) [Computer Software]. Winsteps.com. https://www.winsteps.com/facets.htm

Linacre, J. M. (2022b). *A user's guide to WINSTEPS® MINISTEP Rasch-model computer programs. Program Manual 5.2.5*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Long, M. H., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly, 26*(1), 27–56. https://doi.org/10.2307/3587368

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Luoma, S. (2004). *Assessing speaking.* Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508. https://doi.org/10.1177/0265532211398110

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal, 59*(4), 287–297. https://doi.org/10.1093/elt/cci057

O'Brien, J., & Rothstein, M. G. (2011). Leniency: Hidden threat to large-scale, interview-based selection systems. *Military Psychology, 23*(6), 601–615. https://doi.org/10.1080/08995605.2011.616791

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. https://doi.org/10.1177/0265532208101005

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology, 35*(1), 42–56. https://doi.org/10.14742/ajet.3795

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4*(1). https://doi.org/10.7916/salt.v4i1.1602

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*(1), 72–92. https://doi.org/10.1177/026553228700400107

Rydell, M. (2019). Negotiating co-participation: Embodied word searching sequences in paired L2 speaking tests. *Journal of Pragmatics, 149*, 60–77. https://doi.org/10.1016/j.pragma.2019.05.027

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53*(1), 42–51. https://doi.org/10.1093/elt/53.1.42

Son, Y. A. (2016). Interaction in a paired oral assessment: Revisiting the effect of proficiency. *Papers in Language Testing and Assessment, 5*(2), 43–68. https://doi.org/10.58379/LZZZ5040

Storch, N. (2001). How collaborative is pair work? ESL tertiary students composing in pairs. *Language Teaching Research, 5*(1), 29–53. https://doi.org/10.1177/136216880100500103

Storch, N., & Aldosari, A. (2013). Pairing learners in pair work activity. *Language Teaching Research, 17*(1), 31–48. https://doi.org/10.1177/1362168812457530

Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology, 8*(1), 3–29. https://doi.org/10.46364/njmlm.v8i1.605

Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes, 13*, 2–4.

Van Moere, A. (2013). Paired and group oral assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–4). Wiley-Blackwell.

Wallace, M. J. (1998). *Action research for language teachers.* Cambridge University Press.

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

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Wind, S. A. (2018). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement, 43*(2), 159–171. https://doi.org/10.1177/0146621618789391

Wright, B. D., Linacre, J. M., Gustafson, J.-E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measure-ment Transactions, 8*(3), 370.

收稿日期：2022 年 03 月 15 日
一稿修訂日期：2022 年 03 月 21 日
二稿修訂日期：2022 年 05 月 23 日
三稿修訂日期：2022 年 06 月 20 日
四稿修訂日期：2022 年 08 月 01 日
五稿修訂日期：2022 年 08 月 09 日
接受刊登日期：2022 年 08 月 16 日

Bulletin of Educational Psychology, 2023, 55(1), 25–46
National Taiwan Normal University, Taipei, Taiwan, R. O. C.

Influence of Interlocutor Proficiency and Rater Severity in Indonesian Language Oral Assessment

D. Victoria Rau[1], Hui-Huan Chang[2], and Wan-Yi Hsu[1]

The use of pair work in speaking assessment has frequently been adopted as an authentic manner of testing oral proficiency in second-language communicative language classrooms; however, the findings of studies regarding whether interlocutor proficiency influences the outcomes of oral assessment and whether rater training enables long-term interrater reliability have been inconclusive or contradictory. Studies have indicated that if one of a pair of interlocutors exhibits higher proficiency than the other or if the individuals know each other well, they may collaborate to produce more speech and achieve higher performance in oral assessments (Iwashita, 1996; Norton, 2005; Storch, 2001). However, a higher volume of speech is not always associated with higher overall performance scores (Davis, 2009). Other studies (Galaczi, 2008, 2014) have found that weaker language users might be more reluctant to contribute in oral interactions when paired with more proficient interlocutors. Son (2016) reported that Korean students of English as a foreign language spoke less when paired with more proficient interlocutors, although their overall oral performance did not necessarily decrease. The outcomes of oral assessments may also be influenced by the reliability of the ratings of assessors. Rater severity can be identified by applying the many-facet Rasch model (MFRM; Eckes, 2009, 2015). Although rater training can theoretically increase the confidence and consistency of raters (Davis, 2012, 2016; Huang et al., 2016; McNamara, 1996), differences in rater severity often persist after training (Eckes, 2005, 2009, 2015; Knoch, 2011; Sundqvist et al., 2020; Weigle, 1998) but the results of training are not necessarily long-lasting (Bonk & Ockey, 2003; Chang et al., 2011; Kim, 2011; Lan, 2012; Liao, 2016; Lumley & McNamara, 1995). Because second language assessment generally involves more than one assessor, providing on-the-job rater training is necessary to increase interrater reliability in oral assessments.

Therefore, the following must be explored: (1) Whether training raters in the use of assessment rubrics increases interrater reliability, and (2) whether test takers perform differently when paired with interlocutors of different proficiency levels. This study investigated oral assessment in two General Education Indonesian language classes at a national university in Taiwan that was conducted in the fall semesters of 2020 and 2021. The study used Rasch analysis to measure to what extent interlocutor proficiency (Indonesian language learning beginners vs. speakers of Indonesian as a first language) influenced the students' oral performance and to what extent the severity of the Indonesian teaching assistants (TAs) could be identified and controlled for. The 2020 class comprised 44 students (Taiwanese individuals = 26, Chinese Indonesian individuals = 10, individuals of other nationalities = 8; men = 10, women = 34) and 7 Indonesian TAs (TAs from North Sumatra = 4, TAs from Java = 2, TA from Sulawesi = 1; men = 2, women = 5), and the 2021 class comprised 38 students (Taiwanese individuals = 17, Chinese Indonesian individuals = 14, Chinese Malaysian individuals = 4, individuals of other nationalities = 3; men = 18, women = 20) and 8 Indonesian TAs (TAs from North Sumatra = 4, TAs from Java = 4; men = 4, women = 4). The data comprised six oral assessments performed throughout the semester for each class that were scored by the trained TAs according to a rubric

[1] Institute of Linguistics, National Chung Cheng University
[2] Language Center, National Chin-Yi University of Technology

Corresponding author:
D. Victoria Rau, Institute of Linguistics, National Chung Cheng University. Email: Lngrau@ccu.edu.tw

containing five categories: Content, accuracy, fluency, pronunciation, and interaction. The participants self-assessed their Indonesian language proficiency at the beginning of the semester. Generally, the Chinese Indonesian and Chinese Malaysian students rated themselves as native speakers of Indonesian and Malay, respectively, whereas the Taiwanese students and those of other nationalities identified themselves as true beginners. The participants selected their partners for the oral exams from among their classmates. The data were analyzed using Facets (Linacre, 2022a) to investigate the oral performance of each student pair, the severity of their assessor, and the difficulty of the criteria in the scoring rubric. The scores were transformed into a logit scale for comparison. Analysis based on the MFRM was used to obtain the following information for interpretation: logit measurements, the information-weighted mean-square fit statistic (infit), the outlier sensitive mean-square fit statistic (outfit), the separation index, reliability of separation index, and Chi-square tests for homogeneity. The results were represented using a variable map for each semester, divided into sections for each of the aforementioned three facets. A higher logit value in the three facets indicated higher student pair performance in oral exams, more severe rating, and more difficult criteria for high scores.

The results indicate that even after training, rater consistency was low. In the 2020 class, Chinese Indonesian students had the highest scores, as expected. Performance ranged widely among the Taiwanese students and those of other nationalities. Among the seven TAs, five provided similar ratings and two provided ratings that were either excessively high (logit = -2.42) or excessively low (logit = 1.03) for the midterm oral assessment. After further training was provided before the final exam, two different TAs provided markings that were either excessively high (-0.45 logits) or excessively low (0.97 logits); however, the rater severity among the seven TAs for the final exam was within 1 and -1 logits, the acceptable range. The rater variable interacted with the rating criteria. One TA rated accuracy favorably (t = 2.76) but rated interaction (t = -2.11) severely. Another rated fluency favorably (t = 2.55) but rated pronunciation severely (t = -4.25). In the 2021 class, although the eight TAs were fully trained to use the rubric consistently, variables beyond our control that influenced rating consistency, especially the interaction between the rater and criteria, remained. Therefore, using average scores after outliers are removed may be a viable alternative method of grading until a superior solution is identified. Nonetheless, identifying rater severity variability was helpful as a basis for further rater training.

Different Indonesian proficiency levels between assessment partners did not influence individual student scores in the oral assessments. The students from the 2020 and 2021 classes were categorized into four groups, LL, LH, HL, and HH (L = true beginner, H = proficient Indonesian/Malaysian speaker). Their mean scores were analyzed using Kruskal–Wallis tests. We first investigated whether beginners paired with proficient speakers (LH) scored higher than did those paired with other beginners (LL). However, the scores of these groups did not differ significantly. Next, we determined whether proficient speakers paired with beginners (HL) would score lower than did those paired with other proficient speakers (HH). The scores of these groups did not differ significantly. Our results support the findings of Davis (2009) and Son (2016). We did not demonstrate that interlocutor proficiency positively or negatively affected the students' oral performance. However, based on the comprehensive analysis of students' feedback on the oral examination method, the students seemed to prefer to select partners and remain in their partnerships throughout the semester. Because they were allowed to prepare their scripts and practice their oral exams before the exams, the students developed a sense of solidarity and camaraderie with their partners. The amount of speech they used appeared to not be influenced by differences in interlocutor proficiency. The students were also tolerant of mistakes made by their partners and exhibited patience. Thus, allowing students to choose their own partners and encouraging local students to pair with Chinese Indonesian students would increase their intercultural experiences.

The research site had two unique features that may not be present in other second language classrooms. One was team instruction conducted by a linguist and 7–8 TAs. The other was the presence of a considerable number of proficient speakers of Indonesian/Malay as students attending class with true beginners. Nonetheless, these unique features provide valuable information in this case study with multiyear data.

Keywords: many-facet Rasch model, oral assessment, Indonesian, rater severity, interlocutor proficiency