快篩聽力和閱讀測驗效標關聯效度研究

一、前言與目的

本研究旨在評估「快篩聽力和閱讀測驗」(TOCFL-Speedy Screening Listening and Reading Tests)的效標關聯效度(criterion-related validity)，以作為快篩系統的效度證據。為此，本研究透過分析考生於快篩聽力和閱讀測驗與華語文電腦化適性測驗(TOCFL Computerized Adaptive Test: Listening and Reading, TOCFL-CAT)的通過等級，包括百分比一致性(Percent Agreement, PA)和 Kappa 係數，以及測驗量尺分數的皮爾森積差相關係數(Pearson's product-moment correlation coefficient)，來說明兩者在聽力和閱讀能力評量上的一致性和可比性，俾使外界了解快篩系統於華語文能力評量上的準確性。

二、測驗簡介

「華語文電腦化適性測驗」以試題反應理論(item response theory, IRT)為基礎，電腦會根據考生每一題的作答結果，即時估計考生能力，並選擇最適合該能力水準的下一道試題。由於每道題目皆依據考生能力進行動態調整，因此與一般測驗相比，只需要較少的題目就可以達到相同的精準度。又因為每位考生的作答速度不同，每個人實際作答總題數略有差異，各項測驗通常介於 25~40 題之間，平均作答題數約為 35 題。

「快篩聽力和閱讀測驗」則採用與美國教育測驗服務社(Educational Testing Service, ETS)研發之 GRE 相仿的多階段適性測驗(multistage adaptive testing)架構，將測驗分成三個階段：第一階段先提供混合各種難度的試題；系統依據第一階段的答對題數，於第二階段給予符合考生程度的試題；接著在第二階段作答結束後，再依據其答題表現選擇第三階段的試題。最後，綜合三個階段的作答反應，估計考生能力水準並呈現測驗結果。聽力和閱讀測驗採分開施測，題目的配置為第一階段 10 題、第二階段 8 題、第三階段 7 題，共計 25 題。

兩者同屬於適性測驗，主要差異在於華語文電腦化適性測驗為「逐題適性」，考生作答完每題後立即根據表現挑選下一道試題，在能力估計上的精準度較高，但對電腦硬體效能與網路連線穩定度的要求也更高。而快篩測驗為「階段適性」，題數較少且僅需在階段與階段之間進行調整，對電腦設備和連線環境的要求較低，使用上更為便捷，但同時仍能維持相當水準的能力等級評估，適合教學單位用於快速篩選與評估學習者的華語文聽力與閱讀程度。

在能力分級上，兩種測驗皆參考歐洲共同語文參考架構(Common European Framework of Reference, CEFR)進行能力分級判定，快篩測驗的分級由低至高為 A1、A2、B1、B2、C1；華語文電腦化適性測驗大致相同，但在 C1 等級之上還設有 C2 等級，為便於比較及後續分析，本研究將 C1 和 C2 合併為「C1 以上」。

三、資料來源

本會自 112 學年度起，承接教育部「專科以上學校外國學生華語文能力檢測實施計畫」，配合靜宜大學專案辦公室之查核業務，由本會派員前往教育部指定之學校，分為初測、複測和普測，以「快篩聽力和閱讀測驗」作為評量工具，進行實地檢測，評估外國學生之華語能力，俾利全面了解各校輔導外國學生華語學習之成效，以維繫外國學生受教權益。同時，這些受測學生也會參加 TOCFL-CAT 的正式考試，因此本會針對參加計畫的學生，有系統地蒐集其正式考試成績。

一般而言，成人語言學習者在初期需要適應新的語言系統，並經由大量聽說讀寫練習的過程逐漸累積熟練度，故多數教學者觀察到，約在兩、三個月之內若僅以普通強度學習，學習者的進步幅度可能不明顯。本研究據此先篩選出「快篩測驗」與「電腦化適性測驗」施測日期間隔在兩個月內的樣本，於 112 學年度上、下學期蒐集到的資料中，共有 65 名符合條件的考生。以下分析將以這些受測者之測驗成績為基礎，探討「快篩聽力與閱讀測驗」與「華語文電腦化適性測驗」之間的效標關聯效度。

四、分析方法

1. 百分比一致性

百分比一致性通常用於衡量兩次分類結果(如等級或通過與否)的符合程度。本研究計算考生在快篩聽力和閱讀測驗與華語文電腦化適性測驗通過等級一致的比例，並區分為「同級分類一致性」(等級完全相同)與「相鄰分類一致性」(通過等級相同或僅差一個等級)，以了解兩份測驗在能力分級上的一致情況。

2. Kappa 係數

Kappa 係數亦為常見的分類一致性指標，能凸顯實際觀察到的資料分類一致性與「機率性一致」之間的差異。依據 Landis 與 Koch 於 1977 年提出的判斷標準(Everitt, 1992)[1]，Kappa 係數可分為以下等級：

- 0.21~0.40：普通一致(fair)
- 0.41~0.60：中度一致(moderate)
- 0.61~0.80：相當一致(substantial)
- 0.81~1.00：幾乎完全一致(almost perfect)

Kappa 係數的計算方式可分為未加權(unweighted Kappa)與加權(weighted Kappa)兩種，由於快篩測驗和電腦化適性測驗的通過等級屬於次序變項(ordinal data)，即等級之間具有順序關係，不同等級間的分類差異影響程度並不相同，而未加權 Kappa 的計算方式會將「相鄰等級的不同分類」與「跨多個等級的不同分類」視為相同程度的差異，可能低估測驗分類的一致性(Cohen, 1968)[2]。

[1] Everitt, B. S. (1992). The Analysis of Contingency Tables (2nd ed). London: Chapman & Hall.
[2] Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

為更準確地反映測驗結果的分類一致性，本研究採用平方加權 Kappa (quadratic weighted Kappa)，此方法能根據通過等級之間的距離調整權重，對等級接近的分類結果給予較高的一致性，更貼近實際應用情境(Gwet, 2014)[3]。

3. 皮爾森積差相關

本研究採用皮爾森積差相關係數評估快篩聽力與閱讀測驗量尺分數與華語文電腦化適性測驗之聽力及閱讀量尺分數之間的關係。一般而言，相關係數介於 0.40~0.70 可視為中度相關，若超過 0.70 通常視為高度相關。

五、結果與討論

1. 百分比一致性

表 1 與表 2 分別呈現考生在快篩測驗與華語文電腦化適性測驗之聽力和閱讀測驗通過等級的交叉摘要表。由下表可知，65 名考生中，有 36 人兩項測驗的聽力通過等級完全相同，同級分類一致性為 55.4%；另有 59 人的通過等級相同或僅相差一個等級，相鄰分類一致性達到 90.8%。閱讀測驗方面，有 39 人通過等級完全相同，同級分類一致性為 60.0%；若再加上僅相差一個等級者，共計 60 人，相鄰分類一致性高達 92.3%。

綜上可知，快篩測驗與電腦化適性測驗在能力分級判斷上表現大致相符，尤其是當以相差不超過一個等級為標準時，兩者的分類一致性皆達到九成左右。

表 1　聽力測驗通過等級交叉摘要表

		快篩測驗						
		A1 以下	A1	A2	B1	B2	C1 以上	總計
電腦化適性測驗	A1 以下	6	0	0	0	0	0	6
	A1	5	15	5	2	0	0	27
	A2	0	6	7	4	0	0	17
	B1	0	3	1	8	0	0	12
	B2	0	0	0	1	0	0	1
	C1 以上	0	0	0	1	1	0	2
	總計	11	24	13	16	1	0	65

[3] Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics, LLC.

表 2　閱讀測驗通過等級交叉摘要表

		快篩測驗						
		A1 以下	A1	A2	B1	B2	C1 以上	總計
電腦化適性測驗	A1 以下	9	4	1	0	0	0	14
	A1	8	23	3	0	0	0	34
	A2	0	4	1	1	0	0	6
	B1	0	2	1	4	0	0	7
	B2	1	1	0	0	1	0	3
	C1 以上	0	0	0	0	0	1	1
	總計	18	34	6	5	1	1	65

2. Kappa 係數

平方加權 Kappa[4]的分析結果顯示，聽力測驗為 0.700，閱讀測驗為 0.631，均達到「相當一致」的範圍，顯示快篩測驗與電腦化適性測驗在通過等級的判斷上具有良好的一致性，且測驗結果的分類差異主要發生在相鄰等級，亦即大部分快篩測驗考生的通過等級與電腦化適性測驗僅相差一級，而非出現大範圍的等級落差。

3. 皮爾森積差相關

快篩系統與電腦化適性測驗的聽力量尺分數之間呈現.749 的高相關(p<.01)，閱讀量尺分數的相關係數亦達.720 (p<.01)，顯示兩項測驗在評估受測者能力表現的趨勢是一致的，兩項測驗之間具有高度的正相關。

六、結論與建議

本研究透過百分比一致性、Kappa 係數與皮爾森積差相關三種指標，評估「快篩聽力和閱讀測驗」與「華語文電腦化適性測驗」的分級一致性和量尺分數關聯性。綜合上述分析結果，兩項測驗在聽力與閱讀通過等級的相鄰分類一致性皆達到九成左右，且平方加權 Kappa 數值落在「相當一致」的範圍，顯示兩項測驗的分類結果具備一定程度的一致性。此外，量尺分數的皮爾森積差相關係數均達到 0.70 以上，表示兩者在能力測量上的關聯性高，支持快篩系統在華語文聽力與閱讀能力評量上的效標關聯效度。

為強化快篩系統的效度，未來將規劃收集更多不同等級的受測者，特別是通過快篩測驗中高級以上(B1、B2 或 C1 以上)的華語學習者，以增加等級分布的多樣性。隨著樣本涵蓋能力範圍的擴展，預期測驗在分類一致性和相關分析的表現將進一步提高。此外，若能配合教學端的課堂學習成效或其他外部指標進行比對，

[4] 加權 Kappa 分析語法請參見 https://www.ibm.com/support/pages/weighted-kappa-kappa-ordered-categories

將能為快篩聽力和閱讀測驗提供更多元的效度證據。整體而言，本研究驗證了快篩系統在快速評估華語文聽力與閱讀能力上的成效，可有效協助教學單位及學習者掌握實際的華語文能力表現。