ChatGPT 於華語文試題生成之應用

馬千惠　陳品蓉　張純豪　陳柏熹

摘要

近年研究指出，ChatGPT 生成之英語聽讀測驗試題，在內容流暢度和語句自然度方面的表現與人工設計的試題相近（Shin 與 Lee，2023；沈葳，2024；Lin 與 Chen，2024），然而在選項設計上易出現正答過於明顯、誘答效果薄弱等問題（Shin 與 Lee，2023；沈葳，2024），進一步導致整體試題難度偏低（Lin 與 Chen，2024）。Aryadoust 等人（2024）的研究亦顯示，雖可藉由 ChatGPT 指令微調控制文本難度，但在選項設計上仍常出現語意重疊與邏輯一致性問題。Youn（2023）則強調，將生成式人工智慧應用於口語測驗命題，其延伸話題與追問能力尤為關鍵。綜上所述，人工智慧在試題生成方面具有一定潛力，但在題項效度與適用層級等面向仍需結合教師專業與系統性分析。

鑑於目前鮮有研究聚焦華語試題生成，本研究旨在探討如何應用 ChatGPT 產出符合華語文能力測驗設計原則，並涵蓋不同等級與能力指標的試題，同時評估其內容品質與應用成效，期望能分析 ChatGPT 在華語命題上的弱點，並梳理出系統化的指令模式。

本研究流程分為三個階段：

1. **試題初步生成**：研究者依據三種語言能力與對應題型（準備級聽力、進階高階級閱讀、流利精通級口語），撰寫具體指令輸入 ChatGPT 以產出初步試題。
2. **試題修正與調整**：參考官方命題原則，對初步生成試題給予進一步指令，讓 ChatGPT 加以改寫，針對選項設計、題幹結構、語言難度與圖片內容等進行細部調整，產出供審查的最終試題版本。
3. **專家審查與資料分析**：由六位華語命題專家填寫審查問卷，針對每道試題的難度設定、語言表現、圖片適切性、選項設計等面向給予評分並提供意見回饋。研究結果採用描述性統計與內容效度指標（CVI）進行分析。

本研究總計 220 個評估項目，整體效度表現良好，共計 202 項（91.8%）達 I-CVI ≥ 0.78 的內容效度接受標準。依各語言能力與對應題型觀之，準備級聽力試題（共 110 項）中，98 項（89.1%）的 I-CVI 達標，顯示多數題目能有效測量對應能力，然而未達標項目主要集中於圖片精確度及可能選項等面向。進階高階級閱讀試題（共 30 項）中，28 項（93.3%）I-CVI 達標，顯示閱讀文本與選項設計整體表現良好，但仍有選項誘答力不足的情形。流利精通級口語試題（共 80 項）中，76 項（95%）I-CVI 達標，證明 ChatGPT 在創建情境與設定角色立場方面表現優異，但部分題目在引導討論時，存在立場單一或觀點重複的問題，恐影響語料產出多樣性。

整體而言，ChatGPT 在研究者引導與修改下，所產出之試題能有效對應預設難度與能力指標，展現作為輔助命題工具的潛力。然而，為達測驗高信效度要求，選項設計、圖片準確性與口語試題多元性等面向仍需專業人員後續修訂與補強。

關鍵字：生成式 AI、華語文能力測驗、試題設計