

# 生成式人工智慧時代的教育評量

Taiwan, 27 November 2025

## Generative AI in Educational Assessment and Future Challenges



Speaker:  
Ika Qutsiati Utami

Presented for  
Taiwan National Academy for Educational Research (NAER)

「生成式人工智慧時代的教育評量」專題演講簡報封面  
(測驗及評量研究中心提供)

### 【測驗及評量研究中心 張心洵】

本院測驗及評量研究中心於 2025 年 11 月 27 日辦理線上專題演講，邀請印尼艾爾朗加大學 (Universitas Airlangga) 資料科學技術系講師 Ika Qutsiati Utami。本次演講聚焦於生成式人工智慧 (Generative AI, GenAI) 在教育評量中的實務應用，以及透過實證研究方法檢驗生成式人工智慧產製評量之效度。

Utami 講師指出，隨著大型語言模型 (LLMs) 快速發展，教育評量已逐漸從傳統紙筆測驗與人工批改，轉向結合自動化內容生成、即時回饋與情境化任務的評量設計；然而，相關應用若缺乏嚴謹的實證研究支持，將難以回應評量效度、信度與公平性等核心議題，因此，她研發的評量皆透過多層次實證設計，系統性檢驗生成式人工智慧於評量情境中的實際效益。Utami 講師分享她近年在自動化生成數學題 (Mathematical Word Problems, MWPs) 生成方面之實證研究，該研究指出，傳統由教師人工編寫數學題，不僅耗時費力，也難以同時兼顧題目情境多樣性、難度層級調控與學習者背景差異，因此，Utami 講師研發結合真實情境辨識技術與生成式語言模型之自動化題目生成系統，嘗試將評量內容直接嵌入真實生活情境中，以提升評量的真實性。

除了開發自動化生成數學題系統外，她還透過多元實證評估架構檢驗題目品質與教學成效，評估方式涵蓋三個層次：第一，透過語意相似度 (Semantic Textual Similarity, STS) 與 BLEU (Bilingual Evaluation Understudy) 指標進行自動化評估，以檢驗生成題目與參考題目在語意與結構上的適切性；第二，邀請數學與語言領域專家進行人工評估，從語言流暢度、可理解性、數學適切性與難度合理性等面向進行質性判斷；第三，則以準實驗設計進行實地教學驗證，探索學生學習成效及行為的差異。

研究結果顯示，結合真實情境與生成式人工智慧所生成之數學情境題，對學生在學習表現與學習歷程上皆展現正向效果，使用自動化生成數學題的學生不僅成績明顯提升，也在課堂互動、問題解決投入度表現較佳，Utami 講師指出，這顯示生成式人工智慧若能與課程脈絡及評量目的高度符合，將有助於促進高層次思考與深度學習，而非僅成為產出答案的工具。

最後，Utami 講師亦分享其近期的研究，嘗試透過生成式人工智慧設計促進合作、協商與想法分享等社會互動能力的評量任務，進一步將評量焦點由個別解題能力，擴展至群體問題解決與社會化學習歷程。本次專題演講展現了生成式人工智慧於教育評量設計上的多元潛力，為未來評量的發展提供具體的參考方向。