大型語言模型在教案自動化評量的潛力：以 ChatGPT 為例

The Potential of Large Language Models in Automated Lesson Plan Assessment: A Case Study of ChatGPT

林青欣，張芷瑄，洪煌堯
國立政治大學教育學院
alwayshappyvivi@gmail.com

【摘要】 本研究採用個案研究，透過立意抽樣選取知識翻新教案（n=55），旨在探討 ChatGPT 在未經指令訓練及經指令訓練兩種情境下進行教案自動評量之潛力。研究結果顯示：（1）未經指令訓練時，ChatGPT 自動評量與專家評量的相關係數僅達 0.007，缺乏有效的評量能力；（2）經過指令訓練後，ChatGPT 展現良好的評量信度（$r = 0.766, p < 0.01$），且在效度方面與專家評量達高相關（$r > 0.803, p < 0.01$）；（3）精心設計的結構化指令（包含原則定義、評量尺規與具體例子）是提升 ChatGPT 評量品質之關鍵因素。本研究設計的結構化指令能有效協助教師運用 ChatGPT 評估教案設計，更為大型語言模型輔助教育評量建立實務應用基礎。

【關鍵字】 ChatGPT；教案評量；自動化評量；信效度檢驗；指令設計

Abstract: *This case study (n=55) aims to investigate ChatGPT's potential in automated lesson plan evaluation. The findings revealed: (1) Without prompt training, ChatGPT's automated evaluation showed a minimal correlation coefficient of 0.007 with expert evaluation, indicating insufficient evaluation capability; (2) After prompt training, ChatGPT demonstrated high evaluation reliability ($r = 0.766, p < 0.01$) and achieved high validity correlation with expert evaluation ($r > 0.803, p < 0.01$); (3) Well-designed structured prompts (including principle definitions, evaluation rubrics, and concrete examples) were identified as crucial factors in enhancing ChatGPT's evaluation quality. This study not only offers structured prompts for teachers to effectively evaluate lesson plans with ChatGPT, but also establishes a practical foundation for educational assessment aided by large language models*

Keywords: ChatGPT, Lesson Plan Evaluation, Automated Assessment, Reliability and Validity, Prompt Design

1. 前言

ChatGPT 等大型語言模型的出現，無疑對教學帶來重大影響，更是引起全球教育工作者的關注（Fütterer et al., 2023）。ChatGPT 不僅為教學帶來創新，更重新定義評量方式，提供更即時、彈性與個別化的評估反饋。在教育領域中，教案設計被視為教學的基石，良好的教案設計有助教師更有效運用時間、資源、教材與科技工具，確保教學方向符合教學目標，同時使課程具趣味性及吸引力，進而提升教師教學效能（Farhang et al., 2023；Iqbal et al, 2021）。然而，迫於教學進度壓力，教師在設計教案時往往獨自進行，難以確認教案是否符合教學目標，因此迫切需要一個教學工具協助教師客觀審視其所設計的教案。回顧現有文獻發現，現有研究多聚焦於探討 ChatGPT 在學生作業評量的潛力，對於其在教師教案評量方面的研究則相對缺乏。

有鑑於此，本研究旨在探討 ChatGPT 在教育評量的應用潛能，並聚焦於 ChatGPT-4 在教案評量的信效度表現，以及不同指令設計對評量結果的影響。故本研究的研究問題如下：（1）

未接受指令訓練的 ChatGPT 在教案評量上是否具有效度？（2）ChatGPT 在接受指令訓練後，其教案評量的信度如何？（3）ChatGPT 在接受指令訓練後，其教案評量效度為何？

2. 文獻探討

2.1.ChatGPT 的定義與背景

ChatGPT 是由 OpenAI 公司研發的大型語言模型，能生成近似人類的文本內容，且可準確回應各類型指令（prompt）（Floridi & Chiriatti, 2020）。隨著技術的進步，ChatGPT 背後的大型語言模型亦不斷優化。從初代 GPT-1 至現今 GPT-4，其模型規模與運算處理能力持續提升。GPT-4 為目前 OpenAI 最新推出的模型，具多模態輸入處理能力，可同時處理文本與圖像資料，且其指令理解與執行能力亦較前代 GPT-3.5 更為優異。基於上述特性，本研究將採用 ChatGPT-4 探討大型語言模型在教案自動評量之適用性，並進行指令訓練與信效度驗證。

2.2. ChatGPT 在教育上的應用

ChatGPT 不僅為教學帶來創新，更翻轉了傳統評量方式。Bucol 等人（2024）探討 ChatGPT 作為自動化英語寫作評量工具的潛力。研究顯示，在使用相同評分標準下，ChatGPT 多次評分結果與教師評分間具高度相關。研究亦指出，ChatGPT 不僅具備人性化介面、評分一致性和高效率等優點外，若能策略性運用，更可有效優化教師的評量工作。Kasneci 等人（2023）也認為，ChatGPT 可被應用於評閱作文、回答問題以及提供學生即時反饋，有助於降低教師工作負擔。

Klyshbekova 等人（2024）探討 ChatGPT-3 評量應用的潛力與限制。研究者首先要求 ChatGPT-3 撰寫一篇指定主題文章，並要求設計相應的評量尺規，最後檢視 ChatGPT-3 運用評量尺規的表現。結果顯示，ChatGPT-3 能快速有效完成指定任務。然而，在評量應用上仍有侷限，需由專家輔以判斷，因此 ChatGPT-3 尚無法完全取代專家評量。

Jauhiainen 等人（2024）探討 ChatGPT-4 在自動評量開放式答案的潛力。該研究分析 54 份英文書面答案，答題字數介於 24 至 256 字。在評量過程中，ChatGPT-4 依循五項評量指標進行評估，並採用六級評分制。研究發現，ChatGPT-4 的評分結果與教師評分間具有高度一致性，顯示 ChatGPT-4 能有效執行開放式答案的自動評分任務。

儘管現有研究肯定 ChatGPT-4 在自動評量上的潛力，尚缺乏針對其對教案自動評量方面的探討。故本研究將選用 ChatGPT-4 作為自動化教案評量工具，旨在檢驗其評量結果的準確性與穩定性，以期填補 ChatGPT 在教案自動評量的研究缺口。

2.3. 教案設計

教案設計是教師為課堂教學所做的書面規劃，教案設計不僅包含書面形式，還包含教師在準備課程時所需思考的各種教學元素（Brown, 2001; Ratnawati, 2017）。根據 Brown & Green（2018）提出設計一份完整良好的教案需要包含以下三種元元素，分別為教學目標（Teaching Objectives）、學習活動（Learning Activities）和學習評量（Learning Assessments）。隨著教育思潮的演進與社會需求的轉變，教案設計也在不斷創新與完善。目前的教案設計仍以教學目標、學習活動與學習評量等基本要素為基礎，但在實務運用中，已逐步因應教育改革、學生學習需求的變化和國際趨勢，發展出更多元且具針對性的設計模式。教案設計不再侷限於傳統的知識傳遞模式，而是朝向培養學生關鍵能力、深化學習理解、實踐永續理念等多元目標邁進。

2.4.知識翻新教案

知識翻新（knowledge building）的概念是由兩位加拿大教育學者 Carl Bereiter 和 Marlene Scardamalia 於 1970 年代提出，認為學習是一種社會化的歷程，並透過持續改進知識來促進

社群（community）的成長（Scardamalia & Bereiter, 2006）。因此，知識翻新被視為集體共構的過程，以「想法」（idea）為基本單位，透過社群成員持續的溝通，產生並翻新想法，逐步建構對社群具意義的新知識（Scardamalia, 2004）。學生於求知的過程中，並非被教師灌輸知識，而是作為知識工作者，積極主動地思考，持續修正與完善想法，從而建構更全面的知識（洪煌堯、蔡佩真、林倍伊，2014)。

知識翻新的教案設計則是以知識翻新理論(Bereiter, 2002)和知識論壇(Knowledge Forum, KF）為基礎，其核心架構由三個要素組成，分別為「以想法為中心（idea-centered）」、「以社群為基礎(community-based)」和「以原則為導向(principle-based)」(Hong & Sullivan, 2009; Paavola, Lipponen, & Hakkarainen, 2004; Sawyer, 2006)。在此架構下，教案設計不預先設定學習活動，而是以十二項知識翻新原則作為引導（如圖 1 所示），使教師與學生等社群成員得以共同協作改進其知識作品（如教案設計），進而促進社群知識的整體發展。教師在運用原則導向進行知識翻新教案設計的過程中，持續反思與深化對知識翻新原則的理解，促進其專業成長，更有助於培養其成為具適應性的教師（adaptive teacher）。此外，知識翻新教案設計著重學生的自主學習與協作互動，並鼓勵學生在社群中建構對世界的理解，最終達到創造新知識（knowledge-creation）的目標。

圖 1 *知識翻新 12 項原則*

3. 研究方法

3.1. 研究樣本

本研究採用個案研究法，透過立意抽樣方式，選取以聯合國永續發展目標（Sustainable Development Goals, SDGs）為主題之 55 份知識翻新教案作為研究樣本。教案皆來自知識翻新國際教師培訓課程（Knowledge Building Teacher Professional Development, KB-TPD）之參與教師依循知識翻新十二項原則進行設計。

3.2. 研究工具

本研究基於 Scardamalia（2002）提出之知識翻新十二項原則編制評量尺規，採用李克特式五點量表進行評分，評分範圍從 1 分（最低分）至 5 分（最高分），分數愈高表示該教案愈符合知識翻新原則。為確保評量尺規的內容效度，研究者邀請兩位熟稔知識翻新理論與實務之教育專家進行專家效度檢核。研究者在整合專家建議後，對評量尺規進行修改，最終完成知識翻新評量尺規。

3.3. 研究設計

本研究使用 ChatGPT-4 作為評量工具，依據知識翻新評量尺規評估 55 份教案。研究分為兩個階段：第一階段採用無結構指令，僅要求 ChatGPT 依據原則定義進行評分，未提供評量尺規和明確例子。僅給予大型語言模型無結構指令，如下：「請你根據 Scardamalia 於 2002 年所定義的知識翻新 12 個原則，對這份教案進行逐項評估，為每個原則的達成程度打分，評分範圍從 1 分（最低）到 5 分（最高）並說明評分的理由」。

第二階段則使用經過精心設計的結構化指令，提供完整的評量框架，包含提供原則定義、評量尺規、具體例子，並要求扮演角色、說明得分原因。同時，提供高、中、低分的教案範例，以幫助模型學習如何準確識別和評估教案的質量。待 ChatGPT 生成評分和得分原因後，再次請 ChatGPT 對照得分原因與評量尺規，若兩者不符，請再次評量教案並修改分數。給予大型語言模型精心設計的結構化指令如圖 2。

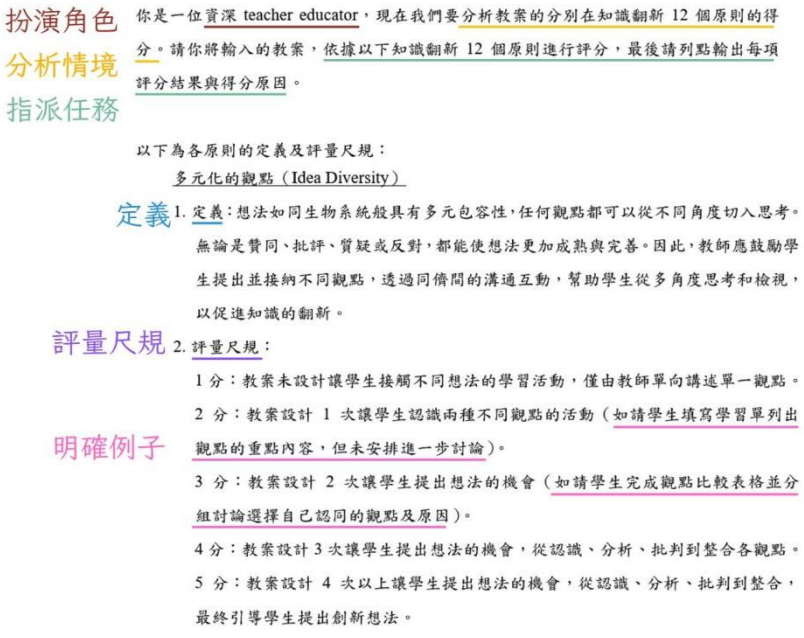

圖 2 *精心設計的結構化指令示例*

4. 資料分析

為驗證 ChatGPT 教案自動評量的信度與效度，本研究採取下列檢測方法：首先，在信度檢驗部分，研究者使用相同指令對 55 份知識翻新教案進行二次評量，評量時間相隔一周。接著，採用 Spearman 相關係數檢測兩次評量結果的一致性。效度檢驗方面，從 55 份知識翻新教案中隨機抽取 20 份樣本，由 ChatGPT 與專家依據知識翻新評量尺規進行評分。最後，透過 Spearman 相關係數比較 ChatGPT 與專家的評分結果，以檢視 ChatGPT 教案自動評量的準確程度。

5. 研究結果

5.1. ChatGPT 未接受指令訓練的效度分析

在未經指令訓練的階段，ChatGPT 自動評量與專家評量的 Spearman 相關係數僅達 0.007。研究者推測造成此結果的主要原因有二：首先，由於未提供明確的核心素養三面九項和知識翻新十二項原則，導致 ChatGPT 僅能依據模糊概念進行判斷，因而產生偏差或不完整的理解，影響其客觀評估教案中原則實踐的能力。以知識翻新十二項原則中的「自主的知識追求者（Epistemic Agency）」原則為例，模型呈現明顯的理解不足；其次，缺乏明確的評量尺規使評分標準不一致，模型評分呈現單一傾向，僅固定給予 5 分，反映其缺乏有效的區辨能力（如圖 3）。由此可知，在未經指令訓練的情況下，使用無結構指令的 ChatGPT 無法達到與專家相當的教案評量水準。

圖 3 *模型針對初期無結構指令的回應*

5.2. ChatGPT 接受指令訓練的信度檢驗

研究結果顯示，在對 ChatGPT 實施指令訓練後，間隔一週進行再次使用相同結構化指令評估教案，兩次評量結果呈現顯著正相關（$r = 0.766$，$p < 0.01$），且 ChatGPT 在兩次教案評量中的得分判別理由一致（如圖 4）。結果表明，當 ChatGPT 經過精心設計的結構化指令進行模型訓練後，ChatGPT 能在多次評量中保持穩定的一致性，具有良好的信度表現。

圖 4 *模型針對結構化指令的回應*

5.3. ChatGPT 接受指令訓練的效度檢驗

在效度檢驗階段，研究採用知識翻新評量尺規作為評分標準，並將 ChatGPT 的自動評量結果與專家人工評量進行對照分析。結果顯示，兩種評量方式具顯著正相關（$r = 0.803$, $p < 0.01$）由此可知，經過適當的指令訓練後，ChatGPT 不僅能達到與專家相當的教案評量水準，且在使用結構化指令的情況下，其評分結果與專家評量具有高度相關性，支持 ChatGPT 作為教案自動評量工具的潛力。

表 1 *ChatGPT 教案自動評量信效度*

階段	指令設計	信度	效度
未接受指令訓練	無結構指令	略	0.007
接受指令訓練	結構化指令	0.766**	0.803**

**p < .01

註：在效度明顯不足的情況下，即便具備高信度，亦無法支持評量工具的實用性。故本研究省略對無結構指令 ChatGPT 進行信度檢驗之步驟。

6. 結論與建議

本研究為首篇探討 ChatGPT-4 在教案自動評量工具應用潛力的研究。研究結果顯示，指令訓練對 ChatGPT 的評量品質有關鍵影響。在未經指令訓練的情況下，使用無結構化指令的 ChatGPT 與專家評量結果呈現近乎零相關（$r = 0.007$），主要受限於模型對評量原則的理解不足以及評分標準不一致所致。然而，經過適當的指令訓練後，使用結構化指令的 ChatGPT 不僅展現出良好的評量信度（$r = 0.766$，$p < 0.01$），在效度方面更達到與專家評量高度一致的水準（$r > 0.803$，$p < 0.01$）。由此可知，精心設計的結構化指令（包含原則定義、評量尺規與具體例子）是提升 ChatGPT 評量品質的關鍵，能有效協助模型建立穩定的評量標準。基於此發現，建議教師在使用 ChatGPT-4 檢視其教案設計時，可參考本研究所設計的指令集。

透過本研究所發展的結構化指令集，即使教師在獨立進行教案規劃時，亦可藉由 ChatGPT 所提供的客觀評量回饋，有效且即時地優化其教案內容，不僅能促進教師提升教案品質，更可強化其教學效能，進而增進學生的學習成效，使教師具備自主精進教學設計的專業能力。

然而，本研究僅聚焦於知識翻新教案的評量效能探討，研究範圍仍有其侷限性。未來研究建議可朝向探討不同類型教案的評量效能差異、持續優化指令設計，以及擴大研究樣本範圍，以進一步驗證 ChatGPT 作為教案自動評量工具的一致性與穩定性。

參考文獻

洪煌堯、蔡佩真、林倍伊（2014）。透過知識創新教學理念與學習平臺以培養國小學生自然課合作學習與翻新想法的習慣。科學教育學刊，**22(4)**，413-439。

Bereiter, C. （2002）. *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 1-16.

Brown, H. D. （2001）. *Teaching by Principles: An Interactive Approach to Language Pedagogy 2nd Edition*. New York: Addison Wesley Longman, Inc.

Brown, Abbie & Green, Tim. (2018). Beyond teaching instructional design models: exploring the design process to advance professional development and expertise. *Journal of Computing in Higher Education.*

Farhang, A. P. Q., Hashemi, A. P. S. S. A., & Ghorianfar, A. P. S. M. (2023). Lesson plan and its importance in teaching process. *International Journal of Current Science Research and Review*, *6*(08), 5901-5913.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds & Machines*, *30*(4), 681–694.

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: global reactions to AI innovations. *Scientific reports*, *13*(1), 15310.

Hong, H.Y. & Sullivan, F.R. （2009）. Towards an Idea-Centered, Principle-Based Design Approach to Support Learning as Knowledge Creation. *Educational Technology Research and Development, 57(5)*, 613-627.

Jauhiainen, J. S., & Garagorry Guerra, A. (2024). Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innovations in Education and Teaching International*, 1-18.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*

Klyshbekova, M., & Abbott, P. (2024). ChatGPT and Assessment in Higher Education: A Magic Wand or a Disruptor? *Electronic Journal of e-Learning*, *22*(2), 30-45.

Paavola, S., Lipponen, L., & Hakkarainen, K. （2004）. Models of innovative knowledge communities and three metaphors of learning. *Review of Educational Research, 74(4)*, 557-576.

Ratnawati, Ratnawati. （2017）Developing a Lesson Plan for Teaching English for Specific Purposes to Adult Learners, *University.Journal of Applied Linguistics and Literacy*

Sawyer, R. K. 2006. Introduction: The new science of learning. In R. K. Sawyer, Ed., The Cambridge handbook of the learning sciences, 1-16. Cambridge University Press.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. *Liberal education in a knowledge society*, *97*, 67-98.

Scardamalia, M., & Bereiter, C. （2006）. Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* 97-11. New York: Cambridge University Press.

Scardamalia, Marlene. (2004). CSILE/Knowledge Forum. Education and technology: An encyclopedia. 183-192.