生成式 AI 於 STEM 跨領域流程圖評量之評分一致性研究

謝雨蓁*

國立中興大學
校務發展中心
博士後研究員

摘要

隨著大型語言模型具備圖像理解與語意推理能力，生成式 AI 是否能協助教師對學生圖像化作品進行穩定評分，已成為教育評量研究之重要議題。本研究旨在探討生成式 AI 於 STEM 跨領域問題解決流程圖評量中的評分實務可行性。本研究以 81 份國中生 STEM 問題解決流程圖作品為研究樣本，比較生成式 AI 與人工評分之間的一致程度。研究結果顯示，生成式 AI 在結構性與進階認知面向中展現一定程度之評分一致性與實務應用潛力，在高階認知面向之概念整合與跨域推理辨識上仍存在侷限。這些發現表明，在明確評分規準與結構化流程支持下，生成式 AI 可作為輔助評量之可行工具。本研究成果為未來 STEM 跨領域問題解決能力評量設計與生成式 AI 自動化評分應用之參考基礎。

關鍵詞：生成式 AI、STEM 教育、流程圖評量

* 通訊作者：謝雨蓁，通訊方式：umi.sie168@gmail.com
收稿日期：2026/3/3；修訂日期：2026/5/22；接受日期：2026/6/8
DOI:10.6249/SE.202606_77(2).0013

壹、前言

近年來，生成式人工智慧（Generative artificial intelligence）（以下簡稱生成式 AI）之發展已從文字生成擴展至多模態整合與複雜推理應用。隨著大型語言模型具備圖像理解能力，生成式 AI 開始能處理包含文字、圖像與結構圖示在內之多元輸入形式。自 GPT-4 及其多模態版本之推出，使人工智慧不僅能處理自然語言文本，更能辨識手寫內容、圖表結構與視覺元素之間的空間關係（OpenAI, 2023）。此一技術突破，為教育領域長期面臨之表現性任務評分負擔提供新的可能性。更有研究指出，當生成式 AI 被應用於圖像或模型繪製任務之評分時，其在特定條件下可達到與人工接近之一致水準（Lee & Zhai, 2025）。此外，Kasneci 等人（2023）近期對生成式 AI 在教育情境中的系統性回顧亦指出，大型語言模型在形成性回饋與開放性任務評分上展現潛力，因此生成式 AI 如何應用於評量，仍需進一步實證驗證。

在國際政策層面，人工智慧已被視為影響教育體系轉型之關鍵科技。聯合國教科文組織（UNESCO）發布之《人工智慧與教育指引》強調，生成式 AI 應支持教育公平與教師專業，而非削弱人類判斷（UNESCO, 2021）。經濟合作暨發展組織（OECD）在多份報告中指出，未來能力導向課程與形成性評量將成為教育發展核心，科技工具可協助提供即時回饋與學習分析（OECD, 2023）。歐盟《人工智慧法案》（EU AI Act）亦將教育評量系統列為高風險應用領域，要求透明性、可解釋性與人類監督機制（European Parliament & Council of the European Union, 2024）。然而，現有研究多集中於文字型作文或開放式問答，針對圖像化 STEM 流程圖之生成式 AI 評分實證研究仍屬有限。

在 STEM 教育評量中，經常關注跨領域問題解決歷程。學生需透過分析問題、整合知識與反思修正等歷程，展現其跨學科統整能力。此種強調歷程與整體判斷之評量取向，符合表現評量之核心理念（Wiggins, 1998）。而流程圖作為一種結構化視覺表徵工具，在 STEM 教育情境中，學生常透過流程圖呈現其問題解決歷程。本研究所採「雞蛋降落傘」任務為廣泛採用之工程設計活動，學生需在有限材料與指定高度限制下，整合空氣阻力與緩衝原理（科學）、測量比較（數學）、材料選用與工具操作（科技）以及原型測試與迭代優化（工程）等跨域知識來解決真實問題。此類作品不僅呈現程序是否完整，也反映學生能否將不同學科知識形成有效連結。

然而，流程圖評量需同時判斷結構完整性與高階思維品質，其評分難度與時間成本遠高於傳統客觀測驗。教師在評閱流程圖時，不僅需辨識步驟是否完整、順序是否合理，更需判斷學生是否展現策略性思考與跨領域統整能力。當 STEM 專題課程採取開放式成果呈現形式時，教師往往需針對多項構念評分標準進行質性判讀與整體評估。此類高度仰賴專業判斷之評量方式，不僅需要長時間逐一分析學生作品，亦涉及評分標準校準與一致性維持之壓力。

尤其在跨領域問題解決情境中，缺乏標準答案之特性進一步提高評分複雜度，進而形成實務執行上的挑戰。在國內教育現場，12 年國民基本教育課程綱要強調素養導向教學與多元評量，STEM 跨領域課程亦日漸普及，但教師在實施表現評量時所面臨之評分負擔與一致性挑戰，相關支持策略仍有持續強化之空間。

先前研究指出，多模態生成式 AI 在結構性項目之評分上可達中高度一致，但在高階推理品質判斷上仍存在限制（Lee & Zhai, 2025）。此結果顯示，生成式 AI 在可觀察結構特徵辨識方面具優勢，但在整體品質評估與多概念關聯判斷上仍待進一步驗證。此外，Al Zubaer 等人（2025）指出 GPT-4 在排序型任務中可維持與人工評分者相當之信度；Heimberg 與 Bernhard（2025）則指出生成式 AI 在低層次認知任務中表現較佳，而在高層次任務中一致性下降。綜合上述研究，可觀察到一項共同趨勢：生成式 AI 之評分表現具有任務結構依賴性。然而，目前多數研究集中於文字型作文或開放式問答，針對圖像化 STEM 流程圖之 AI 評分實證研究仍屬有限。

因此，本研究旨在填補此研究缺口，探討生成式 AI 於 STEM 跨領域問題解決流程圖評量中的可行性與限制。本研究目的為檢驗生成式 AI 在 STEM 跨領域問題解決流程圖評量中的評分一致性。本研究問題如下：

一、在結構性評分面向（問題解決步驟呈現程度、程序邏輯連貫性）上，生成式 AI 評分是否與人工評分達一致？

二、在進階認知歷程（策略檢視與調整表現）與高階認知歷程（跨學科整合品質）之評分面向上，生成式 AI 與人工評分之一致性是否隨構念抽象程度提升而下降？

三、生成式 AI 與人工評分之間是否存在系統性偏差？

貳、文獻探討

一、流程圖作為問題解決歷程之評量工具

在 STEM 教育情境中，學生問題解決歷程往往無法僅透過最終答案呈現，其思考順序、決策節點與修正策略需透過外顯化形式加以觀察與評量。流程圖等圖像化表徵工具，被視為揭露學生知識結構與推理歷程的重要媒介（Ruiz-Primo & Shavelson, 1996）。此類圖像產出能呈現節點之間的連結關係與組織方式，使評量者得以從結構配置推論學生之理解模式與思維路徑。

因此，在強調歷程取向與素養導向評量之脈絡下，流程圖作為認知歷程之外顯化工具，有助於將抽象思考轉化為可觀察之結構表徵。學生透過繪製流程圖，需將問題解決過程中的各項步驟加以組織，明確標示步驟之間的因果關係、條件判斷與迴路結構，此過程本身即為

一種後設認知活動。Cañas 與 Novak（2014）指出，圖像化表徵工具不僅能外顯化既有知識結構，更能促進學生對自身理解之覺察與反思。在 STEM 跨領域情境中，流程圖更可呈現學生如何整合不同學科知識以解決問題，為評量者提供觀察跨學科統整能力之管道。

然而，當學生以流程圖形式呈現其問題解決歷程時，評分者不僅需檢視圖像本身的結構是否完整，也需判斷其中概念之間的連結是否合理。前者包括節點配置、箭線方向與分支安排等具體特徵，較易觀察與確認；後者則涉及整體邏輯是否連貫、策略是否一致，以及概念是否形成有效整合。整體品質之判斷則高度依賴評分者之專業知識、教學經驗與對構念內涵之理解深度。

二、學習表現評量與評分標準之理論基礎

表現評量強調在真實或模擬情境中觀察學生運用知識與策略的能力，其核心在於對複雜能力之整體判斷（Wiggins, 1998）。相較於傳統客觀測驗評量知識記憶與簡單應用，表現評量能深入評估學生之高階認知歷程，包括分析、統整、評鑑與創造等能力。然而，此類評量通常涉及多重規準與開放性產出，因此評分一致性與效度證據成為重要議題（Messick, 1987）。當評分標準包含高階整合或品質判斷時，不同評分者可能因構念理解差異而產生分歧，此為表現評量長期以來之核心挑戰。

評分標準（Rubric）作為結構化評分工具，可透過明確指標與等級描述提升評分一致性。良好的評分標準應具備兩項基本功能：一為描述性功能，透過具體的等級描述使評分者能辨識不同表現水準之特徵；二為規範性功能，透過統一規準減少評分者之間的主觀差異。Jonsson 與 Svingby（2007）在對評分標準信度研究的統合分析中指出，分析型評分標準（analytic rubrics）相較於整體型評分標準（holistic rubrics）通常能達到較高之評分者間信度，尤其在評分構念具明確操作型定義時效果更為顯著。

近年來，評分標準在生成式 AI 自動評分中的角色亦受到關注。Liu 等人（2025）指出，在 GPT-based 自動評分系統中，若提供明確評分標準說明與標準示例，可顯著提升模型與人工評分之相關程度。Kim（2025）亦發現，當提示要求模型逐項分析並說明評分理由時，其評分者內信度顯著提升。然而，Al Zubaer 等人（2025）指出，模型在給分任務中可能受到文本長度或表面特徵影響，顯示即便在評分標準已明確界定的情境下，模型仍可能產生偏誤。此結果指出，僅有清楚的評分規準並不足以確保評分品質，模型是否能準確掌握構念之深層意涵，仍與其推理能力與訓練資料基礎密切相關。

三、生成式AI於教育評量之能力與限制

自動評分之研究發展可追溯至早期寫作評分系統（Attali & Burstein, 2004）。Williamson

等人（2012）提出之自動評分效度評估架構，進一步系統化自動評分工具在教育情境中的品質檢核程序。Shermis 與 Burstein（2013）指出在涉及高階推論與整體品質判斷時，仍需審慎檢驗其推論合理性。

隨著大型語言模型與多模態模型的發展，生成式 AI 之評分能力已逐漸擴展。Kasneci 等人（2023）指出，生成式模型在教育應用中具有回饋生成與初步評分之潛能，但需審慎處理信度與效度議題。Zhai 等人（2020）發現機器學習自動評分系統在科學解釋任務中可達中高度一致，且構念越具體、操作型定義越明確，自動評分之表現越穩定；反之，當構念涉及高階推理或整體品質判斷時，一致性下降。

在多模態評分方面，Lee 與 Zhai（2025）以 GPT-4V 評分學生手繪模型圖，發現生成式 AI 在結構性項目上與人工評分達中高度一致，但在高階推理品質判斷上出現下降。Ray（2024）指出，多模態大型語言模型在涉及高度專業判斷與脈絡理解時，模型仍存在穩定性挑戰。此外，Caraeni 等人（2024）、Heimberg 與 Bernhard（2025）、Al Zubaer 等人（2025）也都指出，生成式 AI 在判斷高層次認知任務仍存在不穩定性。

綜整前述研究結果，生成式 AI 於評分任務中的表現並非單純受模型能力所決定，而與任務本身之結構特徵密切相關。當評分標準具高度規則化與可量化屬性時，模型較能維持穩定判讀；然而，當評分涉及整體品質之統整判斷、多重概念之關聯建構，或高度抽象構念之詮釋時，其評分結果則呈現較高波動性。此一現象顯示，構念愈趨抽象，模型評分的一致性可能愈受挑戰。

參、研究設計與實施

一、評分面向建構

本研究依據流程圖表徵特性（Ruiz-Primo & Shavelson, 1996），並參照問題解決歷程理論（George et al., 2021）、STEM 課程跨域整合需求（Roehrig et al., 2021），建構四項評分構念，各採 0–4 分五等順序尺度。本研究將評分構念依認知層次由低至高排列為 P1 至 P4，使本研究得以系統性檢視生成式 AI 在不同認知投入深度之構念上的評分表現差異。

（一）問題解決步驟呈現程度（P1）：檢視流程圖是否涵蓋問題辨識、構想、測試、修正等核心步驟。對應問題解決流程之外顯化（George et al., 2021）。屬最易觀察之結構性面向。

（二）程序邏輯連貫性（P2）：檢視步驟間順序、因果與條件分支是否合理，箭線方向是否一致。反映程序規劃與邏輯關係（George et al., 2021）。屬程序性面向，仍可透過圖像特徵判讀。

（三）策略檢視與調整表現（P3）：檢視是否呈現檢核點、替代方案或迭代歷程。反映後設認知與設計修正能力（Chinofunga et al., 2025）。抽象程度介於結構與高階之間。

（四）跨學科整合品質（P4）：檢視學生是否將不同學科概念形成有效連結並支持設計決策。屬高階認知與整體品質判斷（Roehrig et al., 2021）。最可能成為 AI 評分的困難構念。

二、研究樣本

本研究樣本為臺灣 81 份國中學生於「雞蛋降落傘」跨域課程中完成之問題解決流程圖作品。本任務於 2 節課內由學生個別完成，任務情境為自 3 公尺高度投放降落裝置以保護生雞蛋不破裂。教師提供任務情境與材料說明，後由學生自主規劃並完成任務，最終以流程圖記錄其問題解決歷程，作為本研究分析資料。所有作品皆於相同教學活動下完成，並於去識別化後掃描為數位影像檔，以確保資料處理之一致性與隱私保護。

本研究參考 Lee 與 Zhai（2025）之評分研究流程，將 81 份作品分為「校準、驗證、測試」三組樣本（9 + 9 + 63）。

（一）標準與提示語建立樣本（N = 9）：由 2 名人工評分者就此 9 份作品建立各構念之評分共識（例如：「P3 何種修正才算策略調整」、「P4 是否需同時出現學科概念與設計理由」），所得共識作為發展 AI 提示語之依據，並以人工評分結果與評分理由嵌入提示語中作為範例。

（二）提示驗證樣本（N = 9）：待上一步驟之提示語確立後，以本組 9 份作品交由 AI 評分，並依驗證結果逐步修訂提示語，至判讀方向穩定後定稿。

（三）正式測試樣本（N = 63）：提示語定稿後不再調整，以本組 63 份作品進行 AI 正式評分，並與人工評分結果進行一致性分析。

三、人工評分流程及範例

人工評分由 2 名具 STEM 教學與評量經驗之評分者參與。2 名評分者均具備國中自然科學或科技領域教學經驗，且曾參與校內或區域性之學生作品評量工作。正式評分前，2 名評分者先接受評分標準之詳細說明，包括各構念之定義、等級描述與評分範例，並以標準與提示語建立樣本（N = 9）進行試評，針對理解分歧之處進行校準，以提升評分穩定性。為檢驗評分者間一致性，本研究自正式測試樣本 63 份作品中隨機抽取 32 份，由 2 名評分者依據相同評分標準獨立評分，並計算評分者間一致性指標。

結果顯示，各向度之 κ 值介於 .74 至 .86 之間；依 Landis 與 Koch（1977）之 κ 解釋標準，κ 值 .41–.60 為「中度一致」（moderate）、.61–.80 為「良好一致」（substantial）、.81–1.00 為「近乎完美」（almost perfect），本研究各向度 κ 值（.74–.86）均達「良好一致」以上水準，

整體均達良好以上一致水準。其中，向度三之一致性最高（κ = .86），其次是向度一（κ = .84）；相較之下，向度二（κ = .78）與向度四（κ = .74）略低，但仍屬於良好一致範圍。整體而言，評分者間在不同向度之評分結果具穩定一致性。在評分者間信度確認後由其中 1 名主要評分者依據相同標準獨立完成全部作品之評分，作為人工評分基準值，用以與生成式 AI 評分進行比較。

以下以學生作品為例（詳見圖 1），說明 2 名人工評分者之判讀歷程。本作品流程圖共 10 步驟，呈現「了解原理、確認製作條件、找資料、設計草稿、製作、測試、檢討、修正、完成、交換想法」之完整問題解決歷程，箭線方向與步驟順序清楚；步驟中提及「降落傘的大小」、「材料」等詞彙，但未說明物理因果（如空氣阻力、緩衝原理），亦未解釋「降落傘的大小」如何影響落下歷程。2 名評分者於 P1 均判定步驟涵蓋問題辨識、構想、測試與修正，給予 4 分；於 P2 均判定箭線方向與步驟順序合理，給予 4 分；於 P3 均辨識出檢討測試與修正歷程之呈現，給予 3 分；於 P4 均嚴格檢視因果連結，認為作品尚未展現跨域概念整合，給予 0 分。此案例顯示 2 名評分者於各構念之判讀方向一致，特別於 P4 構念上同步採取「需具備可辨識之因果連結」之嚴格判準，反映前述評分者間一致性指標（κ = .74–.86）於實際判讀情境中之具體展現。

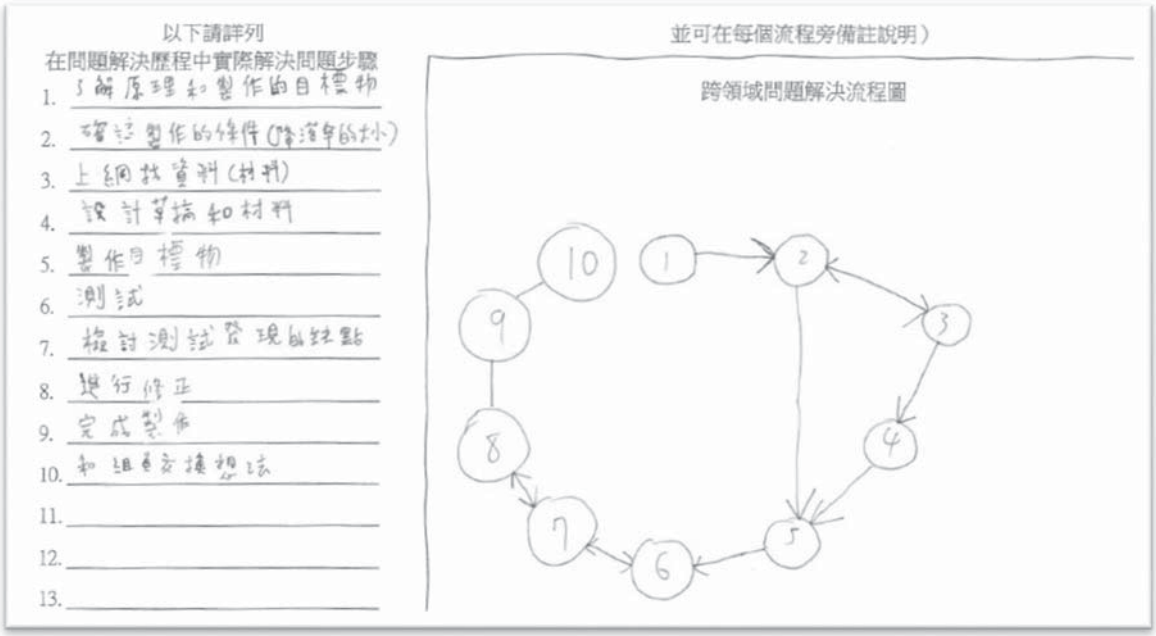

圖 1　STEM 問題解決流程圖作品範例

四、生成式AI評分流程及範例

本研究採用 GPT-5.2 多模態模型進行圖像化作品評分。所有流程圖作品均以統一格式之數位影像檔輸入模型，並使用固定之提示內容與參數設定，以確保評分條件之一致性。本研究參考 Lee 與 Zhai（2025）之 NERIF 框架編寫提示語，包含「角色」、「任務」、「問題情境」、「範例」與「規準」等 5 項組件。模型參數方面，溫度值（temperature）設定為 0，以降低隨機生成變異。

以下為提示內容：

【問題情境】
本研究評分之作品來自「雞蛋降落傘」STEM 跨域任務：學生自主規劃並完成課堂任務，最終以流程圖記錄其問題解決歷程。
【任務角色】
你是一位協助研究者分析國中學生 STEM 問題解決流程圖的評分助手。請根據我提供的學生流程圖圖片，依照以下步驟進行分析。
【任務目標】
針對每一張學生流程圖，完成以下三部分：
擷取圖片中的文字內容
依四個評分向度分別分析並給分
整合成總結結果
【分析流程】
第一部分：文字擷取
請盡可能完整擷取流程圖中的所有可辨識文字，包括：
標題
步驟名稱
箭頭連接中的文字
補充說明
圖中標示的條件、原因、結果、修正方式等

第二部分：四向度分析與評分
請根據流程圖內容，從以下四個向度分別分析並評分。每一向度請先簡要說明判斷理由，再給分。

向度 1：問題解決步驟呈現程度
評估學生是否呈現出明確且相對完整的問題解決步驟。
請給 0–4 分。

向度 2：程序邏輯連貫性
評估流程前後是否具有合理順序、因果或決策邏輯，步驟銜接是否清楚。
請給 0–4 分。

向度 3：策略檢視與調整表現
評估學生是否展現出對策略的檢查、修正、失敗後調整、重新嘗試等歷程。
請給 0–4 分。

向度 4：跨學科整合品質
評估學生是否具體運用多個 STEM 領域知識，並非只是零散提到名詞，而是有實際結合在問題解決過程中。
請給 0–4 分。

【評分原則】
評分必須依圖面證據判斷，不要過度推論。
若學生只是模糊提到概念，但沒有具體展現於流程中，分數不宜過高。
若流程圖內容不足，應誠實給低分。
不要因為文字多就自動高分，重點是結構、邏輯、調整歷程與跨學科整合品質。

【範例】
以下提供 9 份學生作品之人工評分結果與評分理由，請依此判讀標準進行評分。

第三部分：整合結果
請整理成以下格式輸出：
【向度分析與評分】

問題解決步驟呈現程度：X 分
程序邏輯連貫性：X 分
策略檢視與調整表現：X 分
跨學科整合品質：X 分

【整合分數】
總分：X/16

以圖 1 為例，生成式 AI 依提示語依序輸出：（1）擷取流程圖之 10 步驟文字（含「瞭解原理」、「確認製作條件（降落傘的大小）」、「上網找資料（材料）」、「和組員交換想法」等）；（2）四向度評分為 P1 = 4、P2 = 4、P3 = 3、P4 = 2，其中 P4 之判斷理由為「作品提及『降落傘的大小』、『材料』等學科相關詞彙，判定具備初步跨學科整合」；（3）整合結果為總分 13/16。

五、資料分析方法

本研究首先以生成式 AI 與人工評分在各構念之描述統計（平均數與標準差），呈現兩者評分分布之基本統計特徵。在一致性分析方面，本研究以加權 Cohen's κ（Cohen, 1968）作為主要一致性指標。為補充說明差異方向與程度，本研究同時計算 2 項輔助指標：（一）完全一致率（exact agreement rate），即生成式 AI 與人工評分完全相同之比率，用以呈現精確吻合之程度；（二）平均差距（mean difference, AI − Human），計算生成式 AI 評分減去人工評分之平均值，用以檢視是否存在系統性偏差。正值表示生成式 AI 傾向高估，負值表示傾向低估，接近零表示無系統性偏差（Williamson, Xi, & Breyer, 2012; Al Zubaer et al., 2025）。

肆、研究結果

本研究目的為檢驗生成式 AI 在 STEM 跨領域問題解決流程圖評量中的評分一致性，以下為本研究各構念一致性分析與系統性偏差分析結果：

生成式 AI 與人工評分在各構念之描述統計與一致性分析（詳如表 1）。描述統計方面，在結構性面向（P1、P2）和進階認知面向（P3）之生成式 AI 評分與人工評分差距不大。然而，在高階認知面向（P4），生成式 AI 評分之平均數（M = 1.25）明顯高於人工評分（M = 0.52）。此外，生成式 AI 評分之標準差在所有構念中，均低於人工評分，顯示生成式 AI 之評分分布較為集中，變異範圍明顯受限。此一現象在 P4 構念尤為明顯，反映生成式 AI 在高階認知面向之評分辨識精細度不足，傾向給出趨近之分數。

表 1
生成式 AI 與人工評分一致性統計結果（N = 63）

評分構念	人工 M (SD)	生成式 AI M (SD)	加權 κ	完全一致率（%）	平均差距（AI – Human）
P1 問題解決步驟呈現程度	3.13 (1.04)	3.29 (0.63)	.47	52.38	+0.16
P2 程序邏輯連貫性	3.62 (0.97)	3.35 (0.85)	.42	50.79	−0.27
P3 策略檢視與調整表現	2.78 (0.98)	2.67 (0.72)	.44	53.97	−0.11
P4 跨學科整合品質	0.52 (0.91)	1.25 (0.44)	.22	12.70	+0.73

在一致性指標方面，本研究以加權 Cohen's κ 作為主要一致性指標，搭配完全一致率與平均差距進行分析（Williamson, Xi, & Breyer, 2012; Al Zubaer et al., 2025）。

在「問題解決步驟呈現程度」（P1）構念，加權 $\kappa = .47$，達中度一致水準。此構念主要檢視流程圖是否涵蓋問題解決之核心步驟，屬結構性特徵之辨識，評分規準相對具體。完全一致率為 52.38%，表示超過半數作品之生成式 AI 與人工評分完全相同。平均差距為+0.16，接近零且微幅偏正，顯示生成式 AI 在此構念未呈現明顯系統性偏差。

在「程序邏輯連貫性」（P2）構念，加權 $\kappa = .42$，亦達中度一致水準。完全一致率為 50.79%，約半數作品評分完全吻合。平均差距為−0.27，顯示生成式 AI 在此構念微幅低估，但幅度有限。此構念涉及步驟間因果與順序關係之判斷，仍屬可觀察結構特徵之範疇，生成式 AI 之評分表現維持穩定。

在「策略檢視與調整表現」（P3）構念，加權 $\kappa = .44$，同樣達中度一致水準。完全一致率為 53.97%，為四構念中最高。平均差距為−0.11，接近零，顯示生成式 AI 能依據此類可辨識特徵維持一定穩定性。

然而，在「跨學科整合品質」（P4）構念方面，人工評分在 P4 之平均數僅為 0.52（SD = 0.91），顯示多數學生得分集中於量尺低端，呈現明顯的地板效應。相較之下，AI 平均數為 1.25（SD = 0.44），評分變異範圍明顯受限，未能充分反映作品間在跨域整合品質上的實際差異。差異主要來自 AI 將人工評為低分之作品系統性地評為較高分，當學生流程圖中僅出現「材料」、「測試」等表面性學科詞彙時，AI 可能將這些片段線索過度推論為跨域整合之證據，而人工評分者則更嚴格地檢視學生是否真正理解概念間的因果關聯。因此 P4 構念之加權 $\kappa = .22$，僅達一般水準，且完全一致率驟降至 12.70%，顯示生成式 AI 與人工評分在此構念之判斷差異高於其他構念。平均差距為+0.73，為四構念中最高，顯示生成式 AI 系統性地高估學生在此構念之表現。

為進一步剖析 P4 構念之差異成因，本研究自正式測試樣本中選取三類典型作品進行案例對照（詳如圖 2、圖 3、圖 4 與表 2），藉以呈現生成式 AI 與人工評分在不同作品類型下之判讀差異模式：

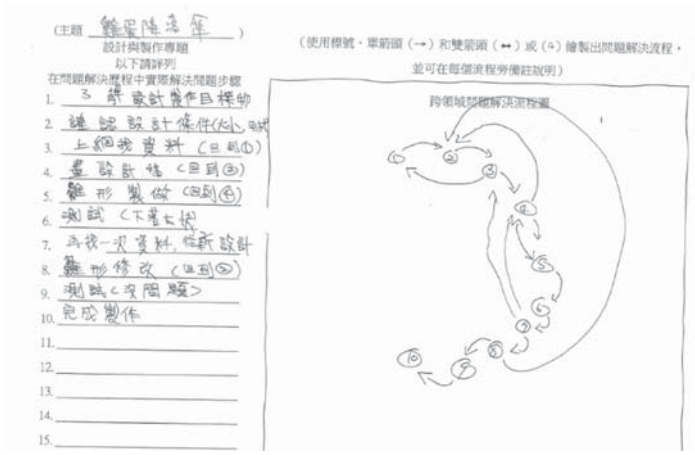

圖 2　表面詞彙誤判

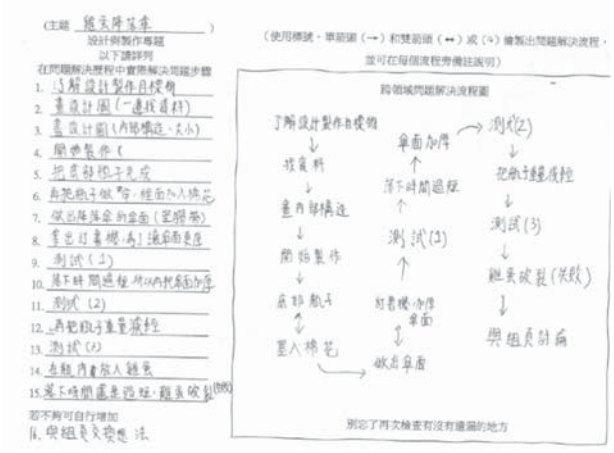

圖 3　因果連結識別不足

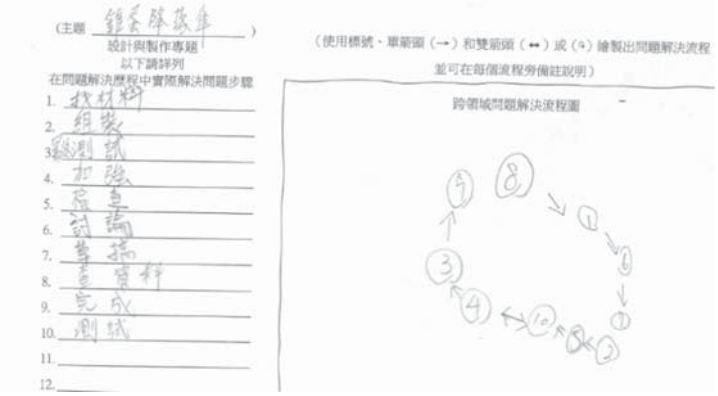

圖 4　無學科內容微幅高估

表 2
P4 跨學科整合品質構念之人工與生成式 AI 評分案例對照

案例類型	學生流程圖特徵	人工／AI	差異成因
A. 表面詞彙誤判	流程圖共 10 步驟，含 4 次「回到前面再修改」之循環。步驟有「找資料」、「雛形」、「下落太快」、「修改」等詞，但沒有說明物理原因（例如為什麼會下落太快）	0／2（差距+2）	AI 看到「設計稿」、「修改」等步驟詞彙，就判斷學生已具備跨學科整合；人工則要求明確說出科學原理
B. 因果連結識別不足	學生依測試結果反覆調整傘面大小：「傘面加棉花→測試→太慢→傘面縮短→再測試→太快→再剪小一點」。學生實際運用了空氣阻力概念（傘面大小影響下落速度），亦以棉花作為緩衝	3／2（差距−1）	學生明確展現「傘面大小影響下落速度」之物理關係，人工給高分；AI 沒看出這層因果，只當作一般測試
C. 無學科內容微幅高估	10 步驟（查資料、找材料、討論、草稿、組裝、檢查、測試、加強、最後測試、完成），第 4 步「加強」與第 10 步「測試」之間有迭代。除「材料」外，未提及任何科學概念或物理原因	0／1（差距+1）	學生只有零散程序詞彙，沒有實質跨學科整合；人工給 0 分，AI 仍給 1 分

綜合表 2 三類案例可知，P4 構念之系統性高估主要源自 2 種判讀傾向：一為 AI 對表面詞彙誤判（案例 A、C），二為 AI 對明確因果連結之識別精細度不足（案例 B）。前者導致對缺乏實質整合之作品給予偏高評分，後者則造成對真正具備跨域整合之作品給予偏低評分；兩種傾向疊加，使 AI 評分分布明顯向量尺中段收斂（SD＝0.44），亦呼應前段所述評分辨識精細度不足之現象。

生成式 AI 會將「一起出現的詞彙」做判斷，而非真正理解概念之間的因果關係，因此把表面詞彙當成跨學科整合的證據，同時也無法辨識學生實際寫出的因果連結。這也顯示 P4 構念會同時出現一致性下降與系統性偏高之現象。

綜觀四構念之分析結果可見，生成式 AI 之系統性偏差與構念特性有關：結構性面向（P1、P2）與進階認知面向（P3）之平均差距絕對值均小於 0.30，未呈現方向性偏差；唯獨高階認知面向（P4）之平均差距達+0.73，明顯偏離其他構念之趨勢，顯示高估幅度與構念抽象程度呈正相關。

伍、研究結論

本研究目的為檢驗生成式 AI 在 STEM 跨領域問題解決流程圖評量中的評分一致性，依據研究問題進行分析，結論如下：

一、結構性面向之一致性

在結構性評分面向（問題解決步驟呈現程度與程序邏輯連貫性）上，生成式 AI 與人工評分達中度一致水準。

二、構念抽象程度與一致性變化

在進階認知面向（策略檢視與調整表現）上，生成式 AI 仍維持中度一致水準，與結構性面向表現接近。然而，在高階認知面向（跨學科整合品質）上，一致性明顯下降，完全一致率亦大幅降低。

三、系統性偏差

在結構性與進階認知面向中，生成式 AI 之平均差距均接近零，未呈現穩定方向之系統性偏差。然而，在高階認知面向中，生成式 AI 評分平均高於人工評分，顯示存在明確之正向系統性偏差。

陸、研究討論與建議

一、研究討論

（一）結構性評分之一致性與應用潛力

本研究發現當評分構念具備明確規準與可觀察特徵時，生成式 AI 能有效掌握評分依據，並維持與人工評分相近之判斷趨勢。此一結果與過去在文本評量與圖像評量研究中皆指出一致，當任務涉及具體且結構清晰之特徵時，AI 分系統較易達成穩定表現（Zhai et al., 2020; Lee & Zhai, 2025）。本研究進一步將此發現延伸至 STEM 流程圖評量情境，顯示生成式 AI 在結構性評分任務中具備實務應用潛力。

（二）構念抽象程度與評分一致性

本研究進一步發現，生成式 AI 之評分一致性會隨構念抽象程度提升而下降。此結果直

接呼應 Zhai 等人（2020）所提出之構念複雜度框架，即評量任務愈涉及高階認知歷程（如整合、推論與評價），自動評分系統之準確性愈受挑戰。相較於結構性與程序性判斷，跨學科整合需同時考量多領域知識之連結與應用，其判斷標準具有高度情境依賴性與詮釋空間，因而增加 AI 評分之困難度。透過案例分析可進一步釐清此現象：AI 對表面詞彙之誤判與對因果連結之識別不足，即為構念抽象程度提升時，AI 判讀精細度下降之具體表現。本研究結果顯示構念抽象程度不僅影響文本評分，同樣對流程圖等視覺化表徵評分一致性產生影響。

（三）系統性偏差與評分效度邊界

本研究發現生成式 AI 在高階認知面向，呈現明確之系統性偏差。從評量效度觀點而言，此發現具有重要意義。跨學科整合能力被視為 STEM 教育中的關鍵構念（Roehrig et al., 2021），因此即使在評分上具有較高難度，仍具有納入評量架構之必要性。然而，Messick（1987）所提出之效度觀點強調，評量結果應能合理反映受試者之實際表現與能力。因此，當評分工具在特定構念上出現系統性偏差時，其解釋與使用即需審慎考量。本研究結果顯示，生成式 AI 在跨學科整合構念上的評分仍存在正向偏誤，顯示其尚未能穩定反映學生實際表現。基於此，在實務應用上，教師應避免單獨依賴 AI 評分結果，而應採取人機協作之評量模式，以提升評分之準確性與公平性。

二、研究建議

基於本研究結果可知，生成式 AI 應採取分層與分工之應用策略，以兼顧評量效率與評量效度。在「評量操作層面」，可由生成式 AI 負責結構性與程序性特徵之初步判讀與大量作品篩選，以降低教師在重複性判斷上的負擔；而涉及高階判斷與跨學科整合之構念，則應由教師進行專業裁量與最終決定，以形成有效之人機協作評量模式。在「評量設計層面」，應依構念抽象程度審慎界定生成式 AI 之使用範圍。對於結構性與進階認知構念，AI 評分結果可作為輔助依據；然而，在涉及高階整合品質之構念時，則不宜直接採用 AI 評分結果，以避免因構念複雜度而產生評分誤差。在「教師專業發展層面」，應強化教師對生成式 AI 評分機制與潛在偏誤之理解。特別是在高階認知構念評量中，教師需具備辨識 AI 系統性偏差之能力，並能進行必要之判讀與修正，以確保評量結果之準確性與公平性。

三、研究限制與未來方向

（一）本研究目的聚焦於評分一致性指標，尚未探討生成式 AI 評語回饋對學生學習歷程之影響。未來可結合形成性回饋研究設計，檢驗生成式 AI 回饋是否能促進學生跨學科整合能力之發展。

（二）由於本研究設計僅採用單一模型與固定提示設計，未來可進行跨模型比較或提示策略實驗，以檢驗不同模型架構與提示設計對評分一致性之影響。

（三）本研究樣本蒐集聚焦於國中學生，未來可擴大樣本來源與教育情境，檢驗不同年齡層或不同學科情境下生成式 AI 評分表現是否呈現相同之構念層級差異。

（四）本研究結果發現高階整合面向出現系統性高估現象，未來研究可進一步分析模型高估之具體模式，例如是否受到文本長度、關鍵詞密度或結構複雜度之影響，以釐清偏差來源。

參考文獻

Al Zubaer, A., Granitzer, M., Geschwind, S., Graf Lambsdorff, J., & Voss, D. (2025). GPT-4 shows comparable performance to human examiners in ranking open-text answers. *Scientific Reports*, *15*(35045). https://doi.org/10.1038/s41598-025-21572-8

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® V.2. *ETS Research Report Series, 2004*(2), i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Caraeni, A., Scarlatos, A., & Lan, A. (2024). Evaluating GPT-4 at grading handwritten solutions in math exams. *arXiv*. https://doi.org/10.48550/arXiv.2411.05231

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. https://doi.org/10.1037/h0026256

Cañas, A.J., Novak, J.D. (2014). *Concept Mapping Using CmapTools to Enhance Meaningful Learning*. In: Okada, A., Buckingham Shum, S., Sherborne, T. (eds) Knowledge Cartography. Advanced Information and Knowledge Processing. Springer, London. https://doi.org/10.1007/978-1-4471-6470-8_2

Chinofunga, M., Chigeza, P., & Taylor, S. (2025). How can procedural flowcharts support the development of mathematics problem-solving skills?. *Mathematics Education Research Journal*, *37*, 85-123. https://doi.org/10.1007/s13394-024-00483-3

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

George, A., Ranjha, S., & Kulkarni, A. (2021). Enhanced problem solving through redefined 8D step completion criteria. *Quality Engineering*, *33*, 695-711. https://doi.org/10.1080/08982112.2021.1969665

Heimberg, M., & Bernhard, W. (2025). *GPT-4 as an automatic grading system for open-ended questions in computer science education.* IEEE Integrated STEM Education Conference. https://doi.org/10.1109/ISEC64801.2025.11147305

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G.L., 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*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kim, Y. (2025). Automated essay scoring with GPT-4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment With Human Scores. *TESOL Quarterly*, *59*(S1), S318-S329. https://doi.org/10.1002/tesq.3405

Lee, G., & Zhai, X. (2025). NERIF: GPT-4V for automatic scoring of drawn models. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-025-10262-9

Liu, Y., Lu, X., & Qi, H. (2025). Comparing GPT-based approaches in automated writing evaluation. *Assessing Writing*, *66*, 100961. https://doi.org/10.1016/j.asw.2025.100961

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

Messick, S. (1987), Validity. *ETS Research Report Series*, *1987*(2), i-208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

OECD (2019), *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en.

OECD (2023), *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing. https://doi.org/10.1787/c74f03de-en.

OpenAI. (2023). GPT-4 technical report (arXiv:2303.08774). *arXiv*. https://arxiv.org/abs/2303.08774

Roehrig, G., Dare, E., Ring-Whalen, E., & Wieselmann, J. (2021). Understanding coherence and integration in integrated STEM curriculum. *International Journal of STEM Education*, *8*, 1-21. https://doi.org/10.1186/s40594-020-00259-8.

Ray, P. P. (2024). Integrating AI in radiology: Insights from GPT-generated reports and multimodal LLM performance on European Board of Radiology examinations. *Japanese Journal of Radiology*, *42,* 1083-1084. https://doi.org/10.1007/s11604-024-01576-6

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, *33*(6), 569-600. https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

UNESCO. (2021). *AI and education: Guidance for policy-makers.* UNESCO. https://doi.org/10.54675/PCSP7350

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. Journal of Research in Science Teaching, 57(9), 1430-1459. https://doi.org/10.1002/tea.21658

A Study of Scoring Agreement of Generative Artificial Intelligence in the Assessment of STEM Interdisciplinary Flowcharts

Yu-Jen Sie*

National Chung Hsing University
Institutional Research Center
Postdoctoral Researchert

Abstract

As large language models have increasingly demonstrated multimodal comprehension and semantic reasoning capabilities, the question of whether generative AI can reliably support teachers in scoring students' visual artifacts has become an important issue in educational assessment research. This study aimed to examine the scoring agreement between generative artificial intelligence (AI) and human raters in the assessment of STEM interdisciplinary flowcharts. Using 81 STEM interdisciplinary flowcharts produced by junior high school students as the research sample, this study compared the level of agreement between generative AI and human scoring. The results indicated that generative AI achieved moderate scoring agreement in structural and intermediate cognitive dimensions. However, lower agreement was observed in the higher-order cognitive dimension involving conceptual integration and cross-disciplinary reasoning. These findings suggest that, under clearly defined scoring rubrics and structured procedural criteria, generative AI may serve as a supplementary assessment tool rather than a full replacement for human judgment. The outcomes of this study provide empirical evidence for future STEM interdisciplinary assessment design and AI-assisted scoring applications.

Keywords: generative artificial intelligence, STEM education, flowchart assessment

* **Corresponding author: Yu-Jen Sie**, E-mail: umi.sie168@gmail.com
Manuscript received: Mar. 3, 2026; Modified: May 22, 2026; Accepted: Jun. 8, 2026
DOI:10.6249/SE.202606_77(2).0013