AAS

**ABSTRACT:**12313

**Keywords:** 45646

主體架構參考paper：

https://doi.org/10.1016/j.compedu.2025.105385

https://doi.org/10.1080/02602938.2021.2001431

https://doi.org/10.3390/math12213416

# Introduction

隨著Artificial Intelligence(AI) 於教育領域中扮演日益關鍵的角色，技術導向的學習模式不僅重塑學習方法，更成為促進知識建構與創新應用的核心策略(Lan & Zhou, 2025)。為了在變動快速且自主性高度要求的環境中獲致良好學習成效，學生需展現高度的自我調節學習(Self-Regulated Learning, SRL)。例如有效的時間與資源管理、個別學習目標的設定，以及對策略與進度之持續性監控與評估(Chang et al., 2023)。自我評估(Self-Assessmen) 成為SRL中不可或缺的一環。學生能否正確地判斷自己學習進度與表現，會直接影響後續學習策略的調整與資源分配。若學生對自身表現的判斷偏差過大，可能導致錯誤的學習決策，使學習成效大打折扣。

自我評估準確性(Self-Assessment Accuracy, SAA) 被描述為學生對自身學習表現的主觀判斷與實際表現之間的一致程度(Wang et al., 2025)。作為SRL的重要組成，SAA會直接影響學生是否能根據真實的學習狀況調整策略與目標。然而Panadero et al.,(2016)指出：多數學生傾向高估自身表現，導致評估結果偏差，阻礙有效的學習調節。正因如此，越來越多研究主張應採取具體且系統性的干預措施，以協助學生提升其自我評估的準確性(Lan & Zhou, 2025)。在眾多干預措施中，提供具針對性的Feedback為一項被廣泛證實有效的干預措施，有助於促進學生對任務要求與表現標準的理解，提高其自我評估的準確性(Braumann et al., 2024)。表現較差的學生在SSA上通常落後於同儕，更需要依賴教師或AI系統提供的Feedback。這些學生透過具體與即時的Feedback，可以獲得學習方向上的明確指引，有效修正錯誤認知並調整策略，從而促進其學習動力與元認知發展(Liu et al., 2025)。

對於教師而言，提供Feedback是一項極具挑戰性的任務，尤其是針對複雜的寫作任務等涵蓋極高的時間與認知成本(Liebenow et al., 2025)。隨著Large Language Model(LLM) 的迅速發展，AI展現強化教學回饋的潛力，不僅能為大量學生提供即時、個人化的意見，也有助於減輕教師負擔並提升回饋覆蓋率。例如Meyer et al.,(2024)證明LLM生成的Feedback對學習成果具有正向影響，像是寫作表現、動機與情緒狀態的提升。然而LLM所生成的Feedback是否能有效支持學生的SSA仍未獲得一致性證據。雖然LLM可以提供快速且一致的Feedback。但Feedback在準確性、適切性與認知支持層面，與教師仍存在顯著差異(Liebenow et al., 2025)。LLM所生成的Feedback可能缺乏情境敏感性，對於策略引導與錯誤識別的深度仍有待提升，進而影響其對 SAA 的潛在成效。Lew et al.,(2010)則提到，提供明確績效標準或具結構性的反饋對學生的SAA提升尤為關鍵，尤其是針對初始表現較差之學生，績效導向的Feedback能協助其重構自我理解與策略認知。

因此，本研究旨在探討LLM生成的Feedback是否能如同教師給予Feedback般有效促進學生的SAA。透過隨機對照實驗設計檢視學生在接收LLM後SAA的變化趨勢與學習績效之關聯。進一步，為了探究LLM之Feedback對於不同學生群體的影響，研究亦納入學生初始表現與初始SAA表現作為調節變項，旨在驗證研究核心假設：對於在自我監控與自我評估能力上相對薄弱的學生而言，結構化的外部回饋能發揮補償性作用，進而促進其學習調整與策略修正。本研究具體問題如下：

* RQ1：LLM所生成的Feedback是否能提高學生的 SAA？
* RQ2：Feedback是否根據學生之初始表現和初始SAA表現產生不同的影響？
* RQ2a：與初始表現較高的學生相比，表現較差的學生是否可以通過LLM所生成的Feedback更好地提高他們的 SAA？
* RQ2b：與初始SAA表現較高的學生相比，初始SAA表現較差的學生是否可以通過LLM所生成的Feedback更好地提高他們的 SAA？

# Literature review

## Self-assessment accuracy

Self-assessment涵蓋不同的技術與形式，協助學生自我監控學習歷程並判斷學習表現，進而促進學習調整與成效提升(Yan & Brown, 2017)。Self-assessment accuracy(SAA) 又被稱為calibration accuracy(Hacker & Bol, 2019)或metacognitive monitoring accuracy(de Bruin & van Merriënboer, 2017)，描述學生自我評估表與自身實際表現結果相對應的程度。SAA的研究受SRL所影響，被視為元認知歷程中的關鍵組成。若學生能準確評估自身學習狀態，更能設定合理目標、精準監測進展，並針對學習策略做出有效調整(Rickey et al., 2025)。而自我評估活動本身亦已被證實能提升學生的反思能力與自我監控意識，是培養SRL技能的重要歷程之一(Andrade, 2019)。Ernst et al.,(2025)則指出高準確度的SAA能促進學生對自身學習過程的現實理解，有助於減少過度自信或錯誤判斷所導致的策略失誤，提升學習效能與動機。具備較高 SAA 的學生，能更現實地理解自身學習狀況，做出策略性調整，提升整體學習成效。這種能力使學生得以識別其學習中的知識盲區，並針對性地投入更多資源進行補強。Thiede et al.,(2010)在文本理解領域的研究指出，SAA較高的學生更能有效辨識需要重新學習的內容，並因此表現出更佳的學習成果。凸顯SAA不僅是一項元認知指標，更是實現有效學習的必要條件。Feedback

學生在Self-assessment出現不準確的情形已被研究多次提到(León et al., 2023; Panadero et al., 2016)。也突顯出有效介入措施學生SAA的重要性。此現象可以通過檢索利用理論(Cue utilization theory)解釋(Kakaria et al., 2024)。依據Cue utilization theory：學生會使用某些線索作為自我評估的依據。例如與自己相關的表現資訊。而這些線索的品質並不相同。Koriat (1997)為此區分出可預測實際表現的具診斷性線索(Diagnostic cues)與法預測實際表現的非診斷性線索(nondiagnostic cues)。例如學生會根據自身閱讀速度(歸類於Nondiagnostic cues) 來判斷對自身對文章的理解程度，從而做出錯誤的評估。而當學生收到教師給予的文本修訂Feedback時(歸類於Diagnostic cues)，這些高品質的Feedback能幫助學生進行更準確的自我評估。學生需要被提供有效的診斷性線索，才能提升SAA(Winstone et al., 2017)。這也說明了Feedback在學生SAA上傳遞Diagnostic cues的重要功能(Butler & Winne, 1995; Panadero et al., 2016)。Gutierrez de Blume(2022)在統合分析中指出，Feedback有助於提升學生自我評估的準確性，具中等程度效果。

於此同時，關於學生水準與Feedback之間交互作用的研究仍相對稀少。過往研究多忽視一項重要變項：SAA 本身亦可能成為回饋效應的調節因子。若學生初始SAA較低，代表其在校準與監控上存在明顯困難，此時所接收的結構化回饋可作為校準參照點，有助於辨識學生學習落差並進行策略性修正(Ernst et al., 2025)。這亦呼應了診斷性回饋(diagnostic feedback)在支持高風險學習者自我監控與調節中的核心作用。探討學生的初始表現與SAA水平如何共同調節回饋效應，不僅有助於釐清回饋機制運作之精細路徑，亦對高等教育中個別化教學介入設計具實質啟發意義。

## feedback literacy與self-assessment之間的相互作用

學生的反饋素養(feedback literacy) 是決定學生是否能夠進行有意義的自我評估並從中受益的重要因素。雖然本文不以全面重新詮釋Feedback literacy為目標，但我們仍希望將其視為一項與自我評估實踐密切相關的能力來進行討論。Feedback literacy的範疇包括處理情緒與理解回饋作為一種互動歷程(Molloy et al., 2020)。其中主動尋求、處理與採取行動這些行為層面，特別適用於自我評估中的回饋歷程。Carless and Boud (2018)將Feedback literacy描述為理解資訊並利用它來增強工作或學習策略所需的理解、能力與傾向。而後續研究進一步強調課程設計應融入反饋素養的三個核心面向，分別為主動尋求相關資訊、有效處理回饋內容，以及根據回饋採取調整行動(Malecka et al., 2020)。Nicol (2020)則提出內部反饋(Internal Feedback)的概念，主張學生在回饋歷程中應發展自我生成的認知評估機制，以深化回饋的理解與應用。本研究聚焦這些行為層面，試圖建立Feedback literacy與AAS之間的連結，並進一步探討兩者交互作用所帶來的教學意涵。

儘管feedback literacy與self-assessment本質上是不同的概念，但兩者在自我調節和終身學習等實現教育的核心目標上都發揮著重要作用(Boud, 1999; Winstone & Carless, 2019)。為了更好地理解這兩個概念並有效應用於實務中，應採取一種整合性的觀點，並審視兩者之間的互動關係。在self-assessment的過程中，存在著許多發展feedback literacy的機會。具備良好feedback literacy的學生，也更有可能進行更有意義的self-assessment。其他長期能力一樣，回饋素養的發展是**一個持續且漸進的過程。**透過適當的設計，self-assessment不僅能讓學生主動尋求外部回饋，能透過將自身表現與各種參照資訊進行比較來產生內在回饋。這種重複性的實踐有助於提升學生的feedback literacy。

self-assessment如何影響feedback literacy上，過去並沒有相關研究在此琢磨過；但feedback literacy在self-assessment中的促進角色，可以從以下兩個層面進行談討。第一，self-assessment不單純只有「自我(self)」，在self-assessment中，「他人(others)」也存在同樣關鍵。雖然self-assessment主要是一種內省行動，其不僅有關於 ”self”。除了”self”之外，”others”也在自我評估過程中存在至關重要的作用(Yan & Brown, 2017)。為促進以學習為中心的Feedback，學生須具備主動覓得並整合來自具適切性的人際或環境資源的能力，使其自我評估歷程更具反思性與調節性。self-assessment歷程需要學生主動從教師、同儕、家長等相關環境與人員尋求feedback(Boud, 1999)。具備feedback literacy的學生，不僅更傾向主動尋求回饋，也更了解專業性、可信度與人際關係等因素會如何影響回饋歷程與品質，因此他們更可能主動尋找對自我評估真正有幫助的回饋資訊(Malecka et al., 2020)。第二，self-assessment中會產生內在回饋(Internal Feedback)。學生會將自己的作品與某些參照標準進行比較，從而產生Internal Feedback。這些Internal Feedback能支援自我評估歷程中的多個面向，例如確立自我評估的標準、辨識優勢與弱點，以及調整當前的學習策略。具備feedback literacy的學生較能產生具備學習導向的高品質內在回饋，因此更有潛力有效運用自我評估的結果來促進自身的學習與改進(Yan, 2020)。

尋求回饋是指學生主動獲取與自己作品相關的資訊，以促進改進與學習。Feedback literacy中的核心行為要素，他將學生的學習與外部資訊連結起來。尋求回饋可以分為兩種策略：**主動詢問（inquiry）與觀察監測（monitoring）。主動詢問（inquiry）意旨**學生直接向他人請求對自己進展或已察覺問題的意見；**觀察監測（monitoring）**則是指學生從環境中獲取資訊，例如比較自己過去的表現與他人表現、參考範例、評量標準，或查閱其他相關資源(Ashford & Cummings, 1983; Joughin et al., 2021; Leenknecht et al., 2019)。然而，僅有外部的回饋資訊並不必然能帶來學習成效，除非學生能透過處理與應用這些資訊，進一步產生內部反饋(Internal Feedback)，才能真正促進學習成長。因此，產生內在回饋(Internal Feedback)是Feedback literacy的另一個核心行為要素。內部反饋(Internal Feedback) 關乎學生自行發展洞見、從Feedback中建構意義，並將其應用於未來的行動調整。根據Nicol (2020)的定義，內在回饋是學生透過將當前表現與某些參照資訊進行比較後所產生的新知識。為了對學習產生實質影響，所有回饋資訊最終都必須內化。不論回饋的來源為何，學生都必須透過自身的認知系統加以詮釋與篩選後才能真正運用。

## 段落C

形成性評量（formative assessment）旨在持續調整教學內容，以符合學生的需求(Filsecker & Kerres, 2012)。儘管有部分學者提出爭議(Bennett, 2010)。formative assessment仍被認為是促進有效學習最具影響力的因素之一。其中formative feedback便是formative assessment的核心要素之一。formative feedback被描述為傳達給學習者的資訊，旨在調整其思維或行為，以提升學習成效(Shute, 2008)。根據Hattie and Timperley(2007)的研究，有效回饋需要涵蓋三個問題：Feed Up、Feed Back與Feed Forward。Feed Up 涉及明確陳述學生的學習目標、建立清楚的方向與目標感。Feed Back 聚焦於學生當下的表現，並在脈絡中指出錯誤，幫助學生理解自己的學習狀況並辨識具體可改善的區域。Feed Forward 則提供未來行動的指引。這三個問題又可以應用在四個層次上：task level、process level、self-regulation level與self level(Hattie & Timperley, 2007)。

除了內容層面的因素外，文字上語氣、語言清晰度、prompt的適切使用與文字長度也對回饋的有效性具有關鍵影響。回饋的語氣應具有鼓勵性，同時避免過度正向(Kluger & DeNisi, 1996)。這樣的做法能建立支持性的學習環境，並確保回饋具有建設性(Brookhart, 2017)。平衡的語氣能激勵學生，同時保留對其成長至關重要的批判性意見。過於攏統的回饋會形成學生理解上的阻礙。因此，回饋應具體、直接且易於理解(Ossenberg et al., 2019)。在文字長度方面，回饋應力求簡潔明確(Kulhavy et al., 1985)，但Van der Kleij et al., (2015) 在其統合分析中發現，**相較於過度簡化的回饋，內容詳盡的回饋可提升學習成效。**

隨著科技，基於自動化、電腦的回饋也逐漸在教育研究領域當中被證實具有實質效益。其中一大優勢便是能即時產生回饋。對於學業表現較弱的學生而言，**即時回饋已被證實具有良好成效(Shute, 2007)。**

Azevedo and Bernard(1995) 在其關於電腦教學中回饋效果的統合分析中指出，即時回饋對所有學生的效果為 0.80，延遲回饋則為 0.35。此外，使用自動化回饋還有較少偏見的認知優勢(補充文獻)。然而一對一輔導被公認為是非常有效的學習方式(補充文獻)，而電腦回饋系統通常難以模擬人類輔導所具備的適應性。

與傳統的電腦回饋系統不同的是，基於大型語言模型（LLM）的系統能提供更高程度的互動性，從而實現更具適應性與參與感的學習體驗。Nguyen et al., (2023)在數學教育研究中評估 GPT-3.5 在處理學生對小數概念錯誤的回饋效能，並發現其在提供恰當回饋方面整體表現良好。Seßler et al., (2023) 在**寫作教學**方面使用 GPT-3.5 生成即時回饋，證明其在透過建設性評論支持學生寫作發展上的實用性。Gabbay and Cohen(2024) 在程式設計教育領域中提到，發現 GPT-3.5 與 GPT-4 雖能有效偵測 MOOC 課程中的程式作業錯誤，但在提供精確或可行的回饋建議方面仍有不足。Estévez-Ayres et al., (2024)則提到，LLM 在評估涉及並行錯誤（concurrency errors）的練習題時能力有限，凸顯其理解複雜程式概念的挑戰。此外，Koutcheme et al., (2024)觀察到，GPT-4 在初學者程式課程中傾向給予過度正面的回饋，可能忽略一些應被關注的重要問題。

在較早的文獻綜合描述中，Zhai et al., (2020) 檢視了基於AI的科學評量研究，聚焦於技術實作、效度與教學特性三個面向，發現大多數研究偏重於效度，較少處理技術與教學面向的議題。而近兩年已有一些初步嘗試探索語言模型在科學教育評量中的應用。例如Wu et al., (2023) 應用預訓練 BERT 模型搭配zero-shot prompting對學生書寫作答進行評分，展現 LLM 在評分任務中的可行性。Latif & Zhai(2024) 比較微調過的 BERT 模型與 GPT-3.5 在自動評分學生答案的表現，結果發現 GPT-3.5 在評分準確度上顯著優於 BERT。Guo et al., (2024) 則發展一套multi-agent system，用以針對科學教育題項生成回饋。然而，該研究存在幾項重要限制：第一，該系統並未將系統所產生的回饋與真實教師的回饋進行比較，因此其與實際課堂教學的契合度尚未驗證。第二，研究評估面向僅關注過度讚美（over-praise）與過度推論（over-inference），忽略了明確性（clarity）與具體性（specificity）等其他關鍵教學品質指標。第三，缺乏針對多元課堂情境的實證驗證，導致其在教育場域中的普遍適用性仍受質疑。

# 系統架構

本節介紹針對AAS所開發之LLM系統。真實且準確的文章評分需要一個能夠全面評估寫作品質各個組成部分的強大框架。因此，推薦的方法旨在自動評估不同體裁和主題的論文。

## Dataset Analysis

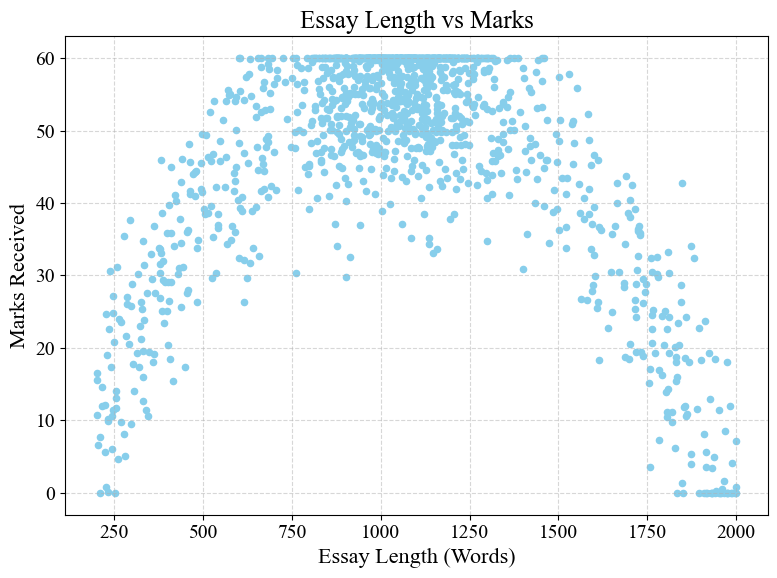
研究所使用之資料集來自校內通識教育中心國文課程學生於 2024 年 1 月至 2025 年 7 月間所撰寫之寫作文本，共計蒐集 7,158 篇文章。這些文章涵蓋議論文和敘述性論文，能有效反映一般大學生在自然語境下的中文寫作表現。本研究依據文本主題與任務性質，將資料集劃分為8個群集，各類型文章分別對語義結構、論證邏輯與敘事技巧有不同要求，有助於訓練語言模型進行回饋生成時，學習辨別與應對不同文本結構與內容特性。文章長度與文章字數分別如figureXX與如figureXX所表示。模型需具備足夠的語境處理能力，以理解篇章上下文並生成具針對性的語言回饋。

|  |  |
| --- | --- |
| *Figure 1*: Distribution of essay lengths | *Figure 2*: Distribution of essay words |

為了檢視語言特徵對得分的影響，研究亦計算了各篇章的**詞彙豐富度（Type-Token Ratio, TTR）。TTR是衡量一段文本中詞彙多樣性（lexical diversity）的基本指標(Richards, 1987)，反映文本中不同詞的數量與總詞數的比例。其中Word types指的是**不重複的詞數，而Word tokens指的是所有詞出現的總數。

**結果如FigureXX所表示，**整體樣本的 TTR mean為0.34，SD為0.06。顯示大多數學生在詞彙使用上具有中等程度的多樣性。TTR 與得分之間雖非完全線性相關，但得分較高之論文往往伴隨較高詞彙變化度，顯示語彙多樣性可能與文本說服力與內容發展程度相關。

*Figure 3*: Average sentence length vs. marks



# Method

## Study sample

本研究的參與者為來自台灣南部某大學工程學院的 64 位大一學生(N=64)，資料收集時間為2024年4月至7月。該實驗課程屬於校內XX課程的一部分。為確保學生在既有學業背景與基礎能力上的同質性，研究採用分層隨機分派法，將學生分為實驗組（EG，n = 32）與控制組（CG，n = 32）。在整體樣本中，男生佔56.75%，平均年齡為 16.01 歲；女生占 52.75%，平均年齡為 16.01 歲。

## Study design and procedure

整體實驗透過電腦以線上問卷的形式進行。實驗開始前會有一個簡短說明。說明結束後，兩組學生皆開始進行一篇繁體中文的論說文（argumentative essay）。時間上限為50分鐘。論說文題目為以下兩者之一，無論是哪一種題目，學生都需要針對題目表達個人意見，並加以說明與論證。題目一：「你是否同意以下說法？一個人的成功來自選擇而非天賦。請舉出具體理由與例子來支持你的觀點」。題目二：「是否同意以下說法？AI終將取代大多數人類的工作。請舉出具體理由與例子來支持你的觀點」。

在完成第一版初稿後，所有學生須進行一次自我評估前側，以評估自己的寫作表現。隨後，實驗組需要藉由本研究開發之LLM系統生成Feedback，並根據Feedback進行修訂。LLM系統所提供的指示如下：「請根據系統所提供之Feedback(如Table XX所表示) 修訂您的文章，盡可能地進行完善修正，請花足夠的時間進行修訂。」而控制組雖然同樣有機會修訂自己的文章，但LLM系統並不會給予任何具體的回饋內容。相反的，控制組所看到的是一段標準化提示語：「請再次閱讀您的文章，並盡可能進行修訂。請花足夠的時間進行修正。」這段提示的目的僅是為了讓控制組學生在沒有實際回饋的情況下，也能參與修訂過程，以保持兩組在修訂程序上的一致性。兩組的修訂的過程中，皆可以隨時查看原始任務說明與自身的初稿，並直接進行修改。學生修訂時間上限同樣為50分鐘。兩組之間的唯一差異在於：實驗組有實際的回饋表與修訂指引，而控制組僅有一般性指令。

*Table 1*: Feedback example generated by LLM系統

|  |  |  |
| --- | --- | --- |
| **面向** | **改進提示** | **具體改進建議** |
| structure | 文章缺乏清楚的開頭與結尾，使得整體架構不明，讀者較難掌握內容邏輯。建議在段落之間加入銜接語句，以提升段落之間的連貫性。 | 補上引言與結論段落。內容安排上應將相關觀點整理於同一段落中，並使用轉折或因果類語句（如：「然而」、「因此」、「另一方面」）協助引導閱讀。 |
| content | 文章中的論點支持力道不足，缺乏明確且具體的例子來佐證主張，且立場表達不夠清晰，內容顯得籠統。 | 增加具體事例來強化論點說服力，並清楚表達文章立場。避免使用模糊或概括性語句，讓讀者更容易理解作者的觀點。 |
| language | 文章中存在多處拼字或文法錯誤，詞彙使用較為單一。部分句子過長，建議進行斷句與重構，以提升可讀性。 | 仔細檢查並修正拼寫與語法錯誤，嘗試使用更豐富的字彙。將冗長句子拆解成較簡潔的語句，並加入如「然而」、「此外」、「儘管如此」等過渡語來強化語意連貫。 |

在完成文章修訂後，兩組皆需要再次進行自我評估後測，評估自己修訂後的寫作表現。實驗結束時，研究團隊亦蒐集了學生的基本背景資料，包括年齡、性別與大學國文總成績。整個實驗耗費三節課(共150分鐘)完成。

(補一張實驗流程圖)

## Measures

學生在完成自我評估前側與後測候，會根據Likert五點量表(0 = 非常差，6 = 非常好) 對自己的寫作表現進行自我評估，題目為：「請評估你剛才完成的這篇文章，你認為這篇文章的品質如何」。研究以自我評估分數與LLM系統給予分數之間的絕對差距來衡量SAA。其公式參考自Schraw(2009)公式(公式1)，其中N為學生樣本數，為第*i*個學生的自我評估分數； 為第*i*個學生的LLM系統分數。

同時，本研究也評估學生的bias，即自我評估與實際表現之間的簡單差值(正值表示高估，負值表示低估)。依據Schraw(2009)公式(公式3)，納入bias是為了了解學生自我評估準確性的變化情形。透過分析學生的bias分數，我們希望探討學生 SAA 的進步是否可解釋為對自身表現之高估或低估的減少。其中N為學生樣本數，為學生的自評分數減去LLM系統分數。然而，bias分數並不等於絕對準確性(Absolute Accuracy Index)，故在解讀結果時可能具誤導性(Schraw, 2009)。

# Results

## Descriptive statistics