

我國調查研究電腦化測驗的實施情形與展望

陳繼成*

摘要

隨著素養導向評量改革深化，紙筆測驗已逐漸難以充分評量跨文本閱讀與情境判斷等能力，電腦化測驗（CBT）遂成為大型評量之重要方向。本文以「臺灣學生成就長期追蹤評量（TASAL）」為例，期能為同樣面臨數位轉型挑戰之國家考試，提供另一種測驗治理與技術發展之補充視角。

本文首先梳理政策背景與數位轉型面臨之「測驗模式效應（test mode effect, TME）」挑戰；其次分析受測、後臺、閱卷與監考四大介面，如何支撐大規模之標準化施測。在測量技術上，本文說明如何運用多向度隨機係數多項洛基模式（Multidimensional Random Coefficients Multinomial Logit Model, MRCMLM）分析跨域素養能力結構，並透過差異試題功能（differential item functioning, DIF）檢測機制檢視試題公平性；在試務制度方面，則介紹制度化的線上人工閱卷機制。同時，檢視長篇數位測驗中的認知疲勞、設備差異與分數可比性等實務挑戰。

最後，本文立基於補充視角，萃取上述調查研究之實徵經驗，提出大型測驗在擴展數位轉型時可供參酌之治理面向。文中客觀探討了監考防呆、申論題線上評閱與公平性常態監測之實務運作，並點出作答歷程資料（process data）的應用潛力與系統長期維運之重要性。期能透過不同性質測驗間的經驗分享，為國家考試未來持續建構數位選才公信力提供他山之石。

關鍵詞：電腦化測驗、試題反應理論、差異試題功能、線上閱卷、作答歷程資料

* 國家教育研究院測驗及評量研究中心副研究員

The Implementation and Prospects of Computer-Based Testing in Survey Research in Taiwan

Chi-Chen Chen*

Abstract

With the deepening of competency-based assessment reforms, traditional paper-and-pencil tests are no longer sufficient to evaluate abilities such as cross-textual reading and situational judgment. Consequently, computer-based testing (CBT) has become an important approach for large-scale assessments. Taking the "Taiwan Assessment of Student Achievement - Longitudinal Study" (TASAL) as a case study, this paper aims to provide a supplementary perspective on test governance and technological development for national examinations facing similar digital transformation challenges.

First, this paper outlines the policy background and the challenges posed by the “test mode effect” (TME). Next, it describes how the four major interfaces—testing, back-end management, grading, and proctoring—support standardized large-scale administration. In terms of psychometric techniques, the paper explains the application of the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) to disentangle the interactions within the structure of cross-disciplinary competencies, and examines item fairness through the Differential Item Functioning (DIF) detection mechanism. Regarding test administration, the paper introduces an institutionalized online human scoring mechanism. Concurrently, it reviews practical challenges in lengthy digital assessments, such as cognitive fatigue, hardware disparities, and score comparability.

Finally, adopting a supplementary perspective, this paper extracts empirical experiences from the aforementioned survey research and proposes governance dimensions that may serve as references for the expansion of digital transformation in large-scale assessments. It objectively discusses the practical operations of fail-safe proctoring systems designed to minimize administrative errors, online scoring of essay questions, and routine fairness monitoring. It also highlights the potential applications

of process data and the importance of long-term system maintenance and operation. By sharing experiences across different types of assessments, this paper hopes to provide valuable insights for national examinations to continuously build public trust in digital talent selection.

Keywords: computer-based testing (CBT), Item Response Theory (IRT), Differential Item Functioning (DIF), online scoring, process data



* Associate Research Fellow, Research Center for Testing and Assessment, National Academy for Educational Research

壹、前言

一、108 課綱與素養導向評量的制度轉向

108 課綱實施後，我國不同學習階段的課程與評量的核心關懷，已由單純知識再現逐漸轉向能力整合、情境應用與跨領域理解。此一轉變不僅反映於課程目標，也直接重塑評量設計的思維。過去以單一知識點辨識、固定題幹閱讀與短篇材料作答為主的紙筆測驗，固然便於大規模施測與計分，然而在面對長文本閱讀、多媒體資訊理解、圖表判讀、資訊搜尋整合、分階段作答與複雜問題解決等任務時，其表徵力與測量效度已逐漸顯現限制。

素養導向評量強調受試者的知識、技能與態度在真實情境中的統整表現，這意味著試題材料可能更長、情境更複雜、作答歷程更具層次，甚至需結合工具使用與資訊操作能力。在此脈絡下，評量載體本身已不再只是承載題目內容的媒介，而成為影響作答表現、題型設計與測量推論的重要條件。換言之，當教育現場的學習情境已逐步數位化，評量若仍停留在紙本邏輯，便可能無法忠實捕捉學生在真實或擬真情境中的能力表現。

二、電腦化測驗導入之必要性與「測驗模式效應」

電腦化測驗之價值，不僅在於將紙本題目轉移到螢幕上作答，更在於其可重新定義題目呈現、作答支援、試務監控、資料蒐集與結果分析的方式。首先，數位介面可以整合計算機、公式表、標註工具、繪圖模組、分頁式文本呈現與互動題型，使過去不易在紙本中充分呈現的素養任務得以具體化。其次，電腦化施測可提升測驗管理的即時性與標準化程度，包括受測登入、作答時間控制、設備狀態監控、異常事件記錄與施測歷程保存等。再者，數位測驗天然產生的歷程資料，也為後續分析受測者的作答行為、時間配置、答題路徑與認知負荷提供可能。

然而，將傳統紙筆測驗轉移至電腦化平臺並非單純的技術升級，而是涵蓋測驗理論、試務治理、資訊安全、閱卷流程、公平性檢核與使用者經驗設計的複雜「制度工程」。其中，測量學界最關注的核心效度挑戰即為「測驗模式效應（test mode effect, TME）」。儘管過去有文獻指出紙筆與電腦化測驗的成績並無顯著差異（Wang et al., 2007）；但面對現今強調長文本與多媒體整合的素養導向測驗，決策者必須審慎評估：數位介面的操作門檻是否會引入干擾測量的「構念無關變異」（construct-irrelevant variance）？考生在螢幕與紙本作答的分數是否具備實質

「可比性」(comparability)？尤其對於大型評量與高風險測驗 (high-stakes test) 而言，任何一項系統設計或模式轉換上的疏漏，都可能嚴重影響考生權益、測量品質與制度公信力。因此，國內既有大型電腦化測驗的實徵經驗，或可提供未來大型高風險測驗在穩健推動數位轉型時之參考，協助降低制度試行的成本與風險。

三、TASAL 作為我國調查研究電腦化測驗的重要案例

「臺灣學生成就長期追蹤評量 (TASAL)」兼具大規模抽樣、跨年施測與政策用途，須在高標準化與多元評量間取得平衡。該計畫之施測範圍極具廣度，涵蓋全國各類公私立高中職、不同學群科別（如普通型、技術型與綜合型高中）以及多元的學生背景。此種跨越多元學制與廣泛施測對象的特性，不僅能全面檢視評量系統在不同學校運作環境下的穩健度，更特別能反映電腦化測驗在大規模真實教育場域中的可行性與制度限制。其在系統建置、線上閱卷、模式效應處理與測量公平性上的實徵經驗，對未來國家考試推動數位轉型深具前導與借鏡價值。本文之核心目的，並非單純介紹某一評量計畫的執行成果，而是透過其系統建置與施測經驗，探究我國調查研究電腦化測驗如何在技術與制度層面逐步成熟，並在承認兩者測驗性質與目的不同之條件下，將此類實徵經驗轉化為國家考試未來擴展電腦化測驗時，可供參酌之補充視角與他山之石。

四、本文架構與論述主軸

本文以下分為四部分。第二部分以 TASAL 為例，說明其電腦化測驗之實施情形，聚焦於系統架構、四大介面設計與試務標準化流程。第三部分分析實施過程中的技術創新與實務挑戰，包括多向度測量、公平性檢核與線上閱卷機制等問題。第四部分則將上述經驗轉譯為國家考試之政策啟示，討論監考、評閱、公平性監測、作答歷程分析與人工智慧導入的發展方向。第五部分提出結語。

貳、我國調查研究電腦化測驗之實施情形—TASAL 系統與試務運作

一、調查研究導向之電腦化測驗需求與系統架構

大型調查研究（如 TASAL）中的電腦化測驗，與一般學校自辦線上測驗具備本質上的差異。前者不僅須處理抽樣代表性與跨群體可比性，更需在全國不同

縣市、學校類型與設備條件下，維持高度的施測一致性。此外，為因應素養導向題型中常見的長文本、多來源資訊整合與多步驟情境任務，數位測驗工具不能僅是「將紙本掃描上線」，而必須與題型需求協同發展，支援受測者在閱讀、運算與標記間反覆切換的作答歷程。

面對「穩定施測」與「支援素養作答」的雙重需求，大型評量的系統建置已非單一作答頁面所能應付，必須發展為一套多角色、多流程且前後臺高度整合的測驗平臺。以 TASAL 為例，其系統架構具體發展出「受測、後臺、閱卷與監考」四大核心介面。此四者共同構成了從命題、組卷、施測、監控到評閱的完整工作鏈，使電腦化測驗不再只是一場單次性的數位施測活動，而是一套「可穩定、可監測、可追溯」的完整評量治理系統。

二、四大介面之建置與功能設計

（一）受測介面：以作答支持提升素養題型的可測量性

受測介面的設計，首先必須回應素養題型的實際需求。在數學、科學或跨領域應用題中，受測者可能需要進行運算、查閱公式、繪圖或以圖像方式呈現想法。為此，系統整合了計算機、公式表及單字表等答題輔助工具。針對數學與繪圖題型，系統更內建方程式模組及繪圖模組，具體提供鉛筆、橡皮擦、三角板、圓規等功能。在紙筆測驗轉換至電腦化的過渡期，提供與傳統作答習慣相近的基礎輔助工具，正是為了防止滑鼠與鍵盤操作門檻成為干擾成績的元凶。若系統未能提供適切的輔助工具，測驗分數將混入考生的數位操作熟練度，進而產生嚴重的構念無關變異。換言之，提供此類基礎防干擾的輔助功能，是為了排除數位介面效應，確保測驗能真實還原學生的學科素養，維持測量的公平性。

（二）後臺介面：題庫建置與組卷治理的中樞

大型評量無法僅靠人工方式管理龐大題庫，後臺介面的價值在於支撐試題生命週期管理與組卷治理。系統內建進階搜尋功能，研究人員可透過特定的資料庫語法快速提取題目或題組，例如：使用「#」搜尋標籤、「@」搜尋標題、「\$」搜尋試題編號。更重要的是，系統具備「試題版本控制」功能，不僅區分命題、統計與研究人員的帳號權限，更完整保留每道試題歷次編修的歷史紀錄與參數變化供隨時查閱。

在試務與組卷治理方面，後臺系統支援「受試者大批匯入與建置」，並可直

接由資料庫追蹤受試者所有的歷史答題紀錄；同時亦提供「線上組卷與施測安排」模組，讓試務人員能精準控制各卷別的派送邏輯。此設計使題庫不再僅是靜態的儲存空間，而是具備高度擴充性與管理量能的測驗治理平臺，對於未來國家考試推動穩定的電腦化題庫與組卷制度尤具參考價值。

（三）閱卷介面：兼顧效率、品質與一致性的線上人工評閱

對包含建構反應題的測驗而言，線上閱卷介面的品質直接影響評分信度與整體作業時程。為提升效率，系統支援「鍵盤快速閱卷」，閱卷人員可利用數字鍵盤（0-9）對應計分，並透過 Q、A、Z 等字母鍵對應不同子題選項（如 Q 為 1A、A 為 1B），同時支援 Enter 鍵跳下一題、Tab 鍵切換分數、Backspace 鍵回上一題。此外，系統具備「漏改警告」防呆機制，提醒未批閱試題並自動歸類至「待閱設定」區，且再次登入時會自動記憶前次進度。更重要的是，評分一致性被內嵌於系統規則中，系統會自動計算兩位委員在評閱共同分派到之該批考生作答時的一致性。當該批作答的整體評分一致率未達 80%時，系統即會將該兩位委員評分不一致的試卷，自動派送進入第三閱（三閱討論）；若該批整體一致率達 80%以上，則針對少數不一致之作答，逕以一閱（主閱）分數為最終成績。此機制形成了一套制度化且能精準控管不同評分者組合間信度（inter-rater reliability）的爭議處理流程。

（四）監考介面：即時監控與異常處理的數位化支撐

電腦化測驗的試場管理，與紙筆測驗相比有不同的風險結構，監考介面之設計實質上是試務治理能力的展現。監試人員登入後可看到數位化的座位表，即時監控受測者的登入及作答狀況，並顯示座號與姓名等資訊。系統特別建置「呼叫燈號」功能，受測者若發現個人資料有誤，按下「資料有誤，呼叫老師」後，監考端該受測者的座號燈即會閃爍並跳出警訊，使其能在不擾動全場秩序的前提下獲得協助。系統亦提供預設為下拉式表單的「特殊狀況紀錄」，將異常事件標準化並自動存檔，供後續資料清理、成績判讀與制度檢討使用。

三、組卷策略與試務標準化流程

大型調查研究與國家考試在擴展電腦化測驗時，皆須處理龐大的線上題庫建置、測量誤差控制與跨卷別等化等共通挑戰。透過嚴謹的組卷與試務設計，除確保調查結果的公信力外，其處理大規模數位施測的量尺連結經驗，亦可作為其

他大型測驗在題庫治理上之補充視角。

（一）標準化的電腦化試務推動期程

相較於紙筆測驗，電腦化測驗的試務準備工作必須投入更長的前置期程。TASAL 建立了標準化的工作期程：在正式施測前，即密集舉辦說明會。特別值得注意的是，TASAL 不僅針對學校聯絡人辦理試務說明，更要求「資訊教師」必須參與，確保其充分瞭解硬體檢測細節與電腦化監考流程。同時，試務中心會統一寄送包含「帳號密碼卡」與聽力或影音題必備之「耳機」等實體調查資料及物資至各校。此種將「資訊技術人員」納入試務培訓之標準化前置作業，正可作為國家考試未來擴展電腦化考場時，建構專屬技術支援 SOP 之重要借鏡。

（二）資料安全與機密性保存機制

在數位評量中，資安為制度公信力之核心。TASAL 制定了嚴格的規範：最關鍵的電腦化測驗電子檔案，不僅儲存於調查平臺，更同步下載至「雲端加密伺服器」進行異地備份保存。未來釋出供學者進行二次分析之電子資料，亦全面落實去識別化作業。此種結合雲端加密與實體控管的資安防護網，亦可供國家考試全面數位化時，建構雲端異地備援架構與資安治理參考。

參、電腦化測驗實施過程面臨的創新與挑戰

一、評量模式與技術的應用

（一）多向度隨機係數多項洛基模式（MRCMLM）拆解跨域素養之交互影響

素養導向試題的一項核心特徵，在於其作答表現往往不由單一能力決定。例如，數學或科學素養題中所需的不僅是學科概念，還包含閱讀理解與情境轉譯能力。若分析模型僅以單一潛在特質處理所有題目，便可能混淆學科能力與閱讀負荷的影響，進而降低測量詮釋的精確性。為解決此問題，在電腦化測驗的資料分析上，可導入「多向度隨機係數多項洛基模式（MRCMLM；Adams et al., 1997）」此一進階的測量模式（measurement model）。

考量素養導向試題所涉及的核心素養極多，實務上難以窮盡。本文特以「閱讀素養」作為多向度分析之實徵範例，主因在於近年長文本題型增加的趨勢下，「閱讀負荷是否成為干擾學科測量的元凶」屢屢成為媒體與輿論熱議的焦點。透過此一多向度 IRT 模式，我們得以將試題作答中糾結的「學科素養」與「閱讀素

養」加以區辨。實徵資料顯示（陳繼成，2024），數學與閱讀的相關係數為 .68，科學與閱讀為 .92，社會與閱讀為 .89。此結果不僅以客觀數據解答了外界對於長文本干擾的疑慮，證實了素養導向試題確實需要一定程度的閱讀素養來進行評估，更展現了測量分析能精準避免將受測者在長文本中的表現完全歸因於學科能力高低，為精進試題設計提供了科學依據。

（二）同時估計、可能值與抽樣加權之量尺化技術

在釐清多向度構念後，大型教育評量之核心挑戰在於如何穩健估計受測者之潛在特質（latent traits），並妥善處理測量過程中的不確定性。為確保不同題本間之試題難度與受測者能力具備跨卷別之可比性，TASAL 採用了「同時估計（concurrent estimation）」技術，透過共同題之參數校準，將各卷別之試題參數與能力估計值置於共同量尺（common scale）上進行運算（Hanson & Béguin, 2002）。

更進一步，國際大型教育評量（如 PISA 或 TIMSS）普遍採用量尺化與加權技術以精準推論母體並控制測量誤差。具體作法為：首先，自受測者能力值之後驗分配（posterior distribution）中抽取多個「可能值（plausible values, PVs）」以表徵其能力分布的機率特徵（Braun & von Davier, 2017）。其次，在進行跨背景變項或群體表現之次級分析（secondary analysis）時，將這些可能值結合「母體估計權重（sampling weights）」進行加權運算，據以估算出各群體表現之加權平均數與估計標準誤（SEs）（Arıkan et al., 2020）。最後，再透過線性轉換將結果轉換為平均數 500、標準差 100 之標準分數。

此種將「能力估計誤差」與「抽樣代表性」同時納入模型之大型分析策略，能有效避免因測量誤差而高估群體間的實質差異；儘管國家考試性質不同，實務上無須進行抽樣推論與加權，但此類運用共同題（common item）進行同時估計與量尺連結的技術，對於未來國家考試若欲建立跨年度、跨卷別的電腦化測驗題庫與常模，仍具備方法學與實務參酌價值。

（三）確立測量不變性：試題公平性檢核與 DIF 監測機制

高風險或大規模測驗若欲具備制度正當性，必須客觀證成其對不同背景群體（如性別、城鄉）具備「測量不變性（measurement invariance）」。在測量實務上，找出「差異試題功能（differential item functioning, DIF）」—即分屬不同群體但能力相同的受試者，因題目本身的偏誤而導致答對機率不相等—是把關試題

公平性的核心機制。然而，電腦化情境與素養導向測驗因涵蓋較多構念無關因素（如閱讀負荷、數位介面操作），若測驗中潛藏較多具偏差風險的試題，極易在 DIF 分析時造成互相干擾與誤判。

為此，TASAL 於公平性檢核的實務突破在於建立一套能排除干擾變項的直接檢驗機制，具體作法乃是採用「多重指標多重原因模式（MIMIC）」結合「先定錨後檢核策略（DFTD）」進行 DIF 分析（Shih & Wang, 2009）。實徵結果顯示，在性別差異檢核上，數學科未發現任何 DIF 試題，提供了該科目具備優良試題品質與公平性的直接證據；而在科學科的分析中，更深刻揭示了「閱讀素養」作為干擾變項的影響：若未納入閱讀素養考量，系統原先會誤判一道具性別 DIF 的試題（對女性有利）；但當將閱讀素養納入多向度模式共同分析並控制後，該題的 DIF 現象即不復存在（未達顯著水準）。

此一實徵結果對國家考試的宏觀啟示在於：維護考試公平性不應僅憑命題委員的主觀審查，或過度仰賴放榜後的試題疑義申訴處理。誠如實務界所深切關懷的，國家考試對城鄉、地區、性別等「外顯變項」的公平性要求極高；然而許多外顯的群體表現差異，往往根源於試題中隱含的「潛在干擾變項（如數位操作門檻、閱讀負荷）」。尤其在推動電腦化測驗與情境化試題時，國家考試可建立常態化、系統化的 DIF 檢測機制。透過將「閱讀負荷」與「數位操作」等潛在干擾變項納入統計模型中進行客觀檢驗，方能從命題源頭精準防堵偏誤，達成「從潛在因素解決外顯不公」的治本效果。這不僅能大幅降低數位轉型過程對不同背景群體（如城鄉設備落差）的干擾，更能以堅實的科學數據回應外界對高風險測驗的質疑，從而大幅提升國家選才制度的公信力與專業度。

（四）數位實作任務與作答歷程資料（process data）之校標效度驗證

傳統紙筆測驗僅能記錄受測者的最終作答結果（對或錯），難以窺探其解題歷程中的策略運用與動態調整。然而，素養導向評量（如：規劃執行與創新應變）高度仰賴學生在多變情境中的資源評估、步驟規劃與決策調整；此種評量需求已超越了僅提供靜態基礎輔助工具的層次，進而發展出具備動態互動操作特性（如調整變項、動態實驗）的多步驟數位實作任務，充分展現 CBT 的高階評量價值。

以陳繼成（2025）的研究計畫中導入之數位化「河內塔（Tower of Hanoi）」測驗為例，透過電腦化施測系統，研究者不僅能記錄受測者最終是否成功解題，更能自動在後臺擷取極具價值的「作答歷程資料」，包括受測者在三至五層任務

中的「錯誤步數」與「花費時間」。實徵分析結果顯示（陳繼成，2025），受測者的「規劃執行與創新應變」素養得分，與其在任務中的錯誤步數及花費時間呈現穩定的顯著負相關（ r 介於 $-.20$ 至 $-.29$ ， $p < .05$ ）。此一實證結果不僅為評量工具的校標關聯效度提供了堅實支持，更具體展示了電腦化測驗如何突破紙本限制，將「評斷測驗結果」昇華為「捕捉思考歷程」。

二、建構標準化之線上人工閱卷與品質監控機制

如前所述，電腦化測驗雖能提供高效率的線上閱卷介面，但在大型測驗中，包含建構反應題（如非選擇題或實作題）的評分信度，歷來是測驗實務的一大挑戰。TASAL 在推動線上閱卷時，並非追求技術上的特立獨行，而是務實地將「閱卷標準作業程序（Standard Operating Procedure, SOP）」與數位系統深度結合，透過平臺防呆與流程檢核，確保大規模評分的客觀性與穩定性。其標準化作業主要涵蓋以下兩大階段：

（一）落實閱卷前置作業與規準共識

依據 TASAL 之實務經驗，系統上線前的規準確認為評分信度之基石。在正式閱卷工作啟動前，由命題專員召開「評分規準討論會議」，邀請學科專家共同檢視並確認每道試題的評分規準。隨後進入「閱卷訓練」階段，要求每位閱卷者依據分派之試題先行試閱 20 筆作答反應。若負責同一試題之一閱與二閱人員評分結果出現分歧，即刻由專員進行解說與再訓練，直到所有評閱人員對規準的理解完全同步，方可進入正式閱卷。須特別註明的是，此種透過一、二閱試評比對以凝聚共識的訓練模式，即是專為平行兩閱機制所設計之重要配套。然而，即便在國家考試常見的單一閱卷情境中，此種正式閱卷前的規準討論與試閱確認（如由單一閱卷者與召集人或標準卷進行對接），同樣是防堵閱卷初期標準漂移、維持評分者內一致性的關鍵 SOP。

（二）內嵌於系統的一致性控管與三閱機制

在傳統紙本或早期的人工評閱中，評分一致性的檢查往往仰賴人工抽驗或事後統計。TASAL 則將品質監控邏輯直接寫入系統規則中：在正式閱卷期間，每道試題皆由兩位閱卷者獨立進行一閱與二閱。當系統偵測到同一作答反應經一、二閱後，其評分一致性未達 80% 時，系統即自動將該卷派送至「第三閱（三閱討論）」，並以三閱分數為最終成績；若一致性達 80% 以上，則以一閱（主閱）

分數為準。這種將雙閱機制與爭議處理流程「內嵌」於平臺的標準化作法，免除了傳統人工挑卷的耗時與錯置風險。透過系統自動把關，將閱卷管理從單純的「效率追求」落實為嚴密的「品質監控」，務實且穩健地提升了評量結果的公信力。

值得進一步指出的是，雖然 TASAL 採行平行兩閱機制，但此類數位化介面與監控功能，對於實務上常因應龐大試務量而採行單一閱卷的大型測驗（如多數國家考試）而言，同樣深具價值。透過系統內嵌的評閱規準、防漏改機制，以及評閱者個人給分趨勢（如偏嚴或偏寬）之可視化圖表，能具體協助單一閱卷委員在長時間評閱中維持高度的評分者內一致性（intra-rater reliability），避免評分標準隨疲勞或時間發生漂移。

雖然大型教育評量與國家考試在風險上的程度不同，但兩者在建構反應題評分上面臨的核心議題相當接近，包括效率、保密、評分公平與品質追蹤。因此，若國家考試未來推動建構反應題線上評閱，教育評量所發展出的快捷鍵評閱、漏改警告、待閱標記與第三閱機制，皆可作為制度設計的重要參考。

三、電腦化測驗的實務挑戰

（一）長篇數位測驗中的認知疲勞與介面設計優勢

電腦化測驗雖擴展了題型可能性，但也引出新的作答負荷問題。誠如實務所見，長篇文章造成的認知疲勞並非電腦化測驗所獨有，傳統紙筆測驗同樣面臨此一挑戰。然而，兩者在閱讀歷程上有著本質上的差異：紙筆測驗多屬單一方向的「線性閱讀」，而電腦化測驗則常涉及需頻繁檢索、跳躍與多媒體互動的「非線性閱讀」。

實務觀察顯示，學生面對長篇數位測驗時，容易出現注意力下降、作答耐力減弱與測驗後段試題表現下滑等認知疲勞現象。這種疲勞不僅來自題目本身的難度，也可能來自長時間閱讀螢幕、持續滾動資訊、頻繁視覺切換與操作壓力。然而，若介面設計得宜，電腦化測驗反而能轉化為處理長文本的絕對優勢。例如，透過將「文本」與「試題選項」分置於螢幕的左右兩側，讓受測者能參照著看，便可大幅減少傳統紙筆測驗中頻繁翻頁所造成的「記憶斷層」與訊息負荷。

值得注意的是，無論是紙筆或數位載體，認知疲勞若未妥善處理，將直接影響測量效度。因為後段題目分數下降，未必代表能力較低，而可能是作答資源耗竭的結果。若此一現象在特定群體中更為明顯，甚至可能進一步衍生群體比較偏

誤。

（二）透過多卷別題序調動因應認知疲勞之實務操作

面對不同科目或素養的數位測驗帶來的認知疲勞，TASAL 在實務上採取的一項重要應對策略為「題序調動」。在施測設計中，研究團隊透過不同的測驗卷別，刻意調整了試題的出現順序，將較具趣味性、情境感或視覺吸引力較高的題目（例如：網路梗圖題或生活情境題）安排於部分題本的前段。此作法之目的在於透過較高的作答投入感（engagement），協助受測者順利進入狀態，減緩一開始即面臨高密度文本所造成的心理負荷。TASAL 實施此一策略的核心邏輯，在於將題序調動視為一種事前的方法學控制，旨在從系統設計端直接排除疲勞效應的干擾，而非事後去測量疲勞的衰退幅度。初步的實徵統計亦顯示，受測者在前段題與後段題間的答題正確率與花費時間並無顯著落差。此一結果恰好印證了，此種透過多題本設計來分散疲勞風險、避免所有考生都在測驗同一個階段面對高負荷試題的實務操作，已成功抵銷了長篇測驗可能帶來的認知疲勞，成為大型數位評量中維持測量穩定性的重要防線。

（三）學校硬體設備與網路穩定度落差

我國各地學校在設備更新速度、終端數量、教室網路環境與資訊支援人力上，仍存在明顯差異。當大型調查研究進入真實校園場域，這些差異會直接轉化為施測風險。部分學校可能面臨裝置老舊、解析度不足、鍵盤滑鼠狀況不佳、瀏覽器版本不一或網路不穩等問題。對紙筆測驗而言，這些因素幾乎不構成風險；但對電腦化測驗而言，卻可能影響受測者作答流暢度與試務秩序。

（四）設備與環境差異對公平性的潛在影響

設備差異不只是一項行政問題，也可能成為測量公平性議題。若學生因設備反應遲滯、畫面顯示不佳或網路中斷而增加額外負荷，則其分數便可能混入非目標構念因素。尤其在城鄉差異明顯的情況下，技術條件差距甚至可能與社經背景交疊，形成制度上的複合不利。

因此，電腦化測驗的公平性，不應僅理解為題目內容是否偏向特定群體，也應包括作答條件是否在實質上等值。此外，在 TASAL 的實徵經驗中，初步雖未接獲現場教師回報顯著的網路延遲或設備異常問題；但為進一步精準掌握數位公平性，亦規畫透過系統日誌（log）進行深度檢核，釐清是否有因潛在的硬體

等級或網路延遲而導致的作答離群值 (outliers) 現象，以作為未來優化跨校施測環境與數據清理的科學依據。

誠如實務現況所見，國家考試目前皆於設備與網路環境一致之「認證合格電腦試場」中進行，已有效避免此類問題，這正是維持高風險測驗公平性的關鍵基石。然而，TASAL 深入全國各類型學校的實徵經驗反向突顯了一項重要啟示：未來國家考試若因應數位轉型的全面擴張，而必須大量擴增委外考場（如租借更多不同層級學校之電腦教室）時，對設備規格與網路環境一致性的嚴格把關標準絕對不可妥協，因為高風險測驗更不能容許作答環境的微小差異轉化為制度性的不公平。

四、測驗模式效應與分數可比性之效度挑戰

在將傳統紙筆測驗轉移至電腦化平臺時，測量學上最核心的效度挑戰即為「測驗模式效應」。決策者與學者勢必關注：考生在紙本與電腦上的作答表現是否具備可比性？數位介面的操作是否引入了干擾測量的「構念無關變異」？TASAL 的實務經驗顯示，為確保測量效度與跨期分數的一致性，不能僅是「將紙本掃描上網」，而必須在系統與試題設計上進行調適。例如，為降低數位工具操作門檻對數學或實作題型造成的干擾，系統必須內建與紙筆作答邏輯相近的方程式模組及小鍵盤等輔助工具，以確保測驗真正測量的是學科素養，而非數位操作能力。此外，為精準掌握轉換過程中的能力變化，TASAL 在組卷設計上特別保留了來自前一年度的共同題。透過這些共同題在 IRT 模式中的參數校準與同時估計，得以將不同測驗情境下的受測者能力穩健地連結至共同量尺上，以此科學方法回應測驗模式轉換間的可比性挑戰。

肆、調查研究經驗對國家考試之啟示與展望

一、教育評量數位轉型經驗之參考價值

國家考試與大型教育評量在目的上雖有差異，前者偏重資格篩選與任用正當性，後者偏重整體監測與政策分析；但兩者皆屬大規模、高標準化且高度重視公信力之測驗制度。因此，教育評量在電腦化測驗中所累積的治理經驗，或可作為另一種實務對話視角。

尤其在國家考試推動數位轉型的過程中，政策制定者最需要的往往不是抽象原則，而是已在大規模施測中被驗證可行的制度做法，包括如何控管異常、如

何評閱建構反應題、如何監測試題偏差、如何處理考生疲勞與如何面對設備差異。就此而言，TASAL 類型計畫可視為一種「低一階風險、但高一階複雜度」的試驗經驗，應有值得大型測驗借鏡之處。

二、對國家考試試場管理之啟示

（一）建立監考防呆與異常處理的可視化系統

國家考試若採電腦化施測，監考工作勢必從紙本點名、口頭提醒與人工巡場，轉向結合數位平臺的試場治理。教育評量中的監考介面經驗顯示，即時座位表、登入狀態顯示、呼叫燈號與異常紀錄，可有效降低監試人員資訊落差與處理遲滯。對國家考試而言，建議未來建置具備下列功能之監考支援系統：

1. 即時顯示考生登入、作答中、暫停、中斷與完成狀態。
2. 異常事件標記與處理紀錄自動留痕。
3. 呼叫機制分級化，例如設備問題、身體不適、離座需求等分流處理。
4. 與考場主控臺連動，於必要時快速啟動備援設備或延長處理程序。

此類機制的核心價值，在於減少人為疏失對高風險測驗造成的不確定性，並提升事後申訴處理的可驗證性。

（二）將試務規則轉化為系統規則

高風險考試最大的治理挑戰之一，在於「同一規則是否被一致執行」。若試務規則僅存在於手冊與監試訓練中，便可能因考場差異而出現不同解釋。相對而言，若能將關鍵規則內嵌於系統，例如逾時自動鎖卷、異常中斷自動記錄、未完成程序無法進入下一步，則制度執行可更一致。這正是電腦化測驗與紙筆測驗在治理能力上的本質差異。

（三）確立因應高法律風險之系統故障救濟程序

除了試場的常態監控外，國家考試與大型調查研究（如 TASAL）在試務治理上最大的本質差異，在於面對「系統與設備故障」時的法律風險與容錯率。相較於 TASAL 作為政策調查，當面臨單一考生的設備當機或網路中斷時，其行政處理多以「更換備用機、重新登入續答」或極端情況下的「註記為遺漏值」為主，對整體抽樣推論的影響相對有限。

然而，國家考試攸關人民服公職與執業權利，具備極高的法律風險與利害關係，任何設備卡頓或系統中斷皆可能損及考生權益並引發後續之行政爭訟。因此，國家考試的數位試場管理不能僅停留在「異常記錄」，更必須預先建立嚴謹且法制化的「系統故障救濟程序」。例如：明訂設備當機中斷時的「作答時間補償計算基準」、因全面性網路或電力癱瘓而啟動的「紙本備用卷備援機制」，以及若遭逢不可抗力導致成績減損時的「重測與等化計分補救規範」。唯有將這些救濟程序予以明文化，並與線上監考系統所留存的異常斷點紀錄（log）緊密結合以作為客觀舉證基礎，方能確保國家考試在數位轉型過程中的程序正義與法律正當性。

三、對建構反應題線上評閱制度之啟示

（一）建構反應題評閱數位化的發展契機

我國國家考試長期仰賴建構反應題選才，以評估應試者之論證、分析與專業表達能力。然在閱卷量大、時程緊湊與公平要求高的情況下，傳統紙本評閱方式已逐漸面臨限制。如前文實務經驗所示，電腦化閱卷能具體回應上述挑戰：針對「閱卷量大與時程緊湊」的痛點，數位平臺透過自動分群派卷、鍵盤快速評閱及即時進度監控，大幅提升了卷務流轉效率，免除傳統人工搬卷與核算的耗時；針對「公平要求高」的挑戰，系統則透過漏改警告防呆機制、內嵌評閱規準，以及未達一致性自動啟動三閱等制度化設計，有效排除人為疏漏與標準漂移。將數位技術結合卷務流轉與品質監控，對國家考試具高度啟發性。

（二）線上評閱平臺的制度設計原則

未來國家考試建構反應題線上評閱，可參考以下原則：

1. 雙閱與第三閱制度化。當評分差距超出容許範圍時，系統自動送交第三閱，而非由人工事後挑卷。
2. 漏改防呆機制。未填寫必要評分欄位不得送出，以避免評分遺漏。
3. 評閱規準同步化。評分指引、示例答案與爭議案例可於平臺中即時呈現與更新。
4. 進度與一致性監控。無論測驗係採行平行兩閱或單一閱卷，平臺皆可即時呈現各評閱者之進度，以供即時掌握評分者內一致性（intra-rater reliability）。
5. 匿名化與資訊隔離。考生識別資訊應完全遮蔽，以維持評分公正。

（三）從效率管理走向品質治理

國家考試若導入線上評閱，目的不應僅是加快放榜流程，而應進一步建立以資料為基礎的評分品質治理模式。亦即，不只知道「評得快不快」，更要知道「評得穩不穩、準不準、公不公平」。如前文所提及，透過線上評閱系統內建的「防漏改防呆機制」、「評分者偏嚴/偏寬數據之可視化監控」，以及「評分歧異自動派送第三閱」的標準化流程，皆能以客觀的系統數據，輔助召集人即時校正單一或多位評閱者之偏差，有效解決傳統紙本閱卷難以即時掌握評分一致性與公平性的實務痛點。這種轉向，正是大型教育評量在長期實務中累積的重要制度啟示。

四、對紙筆測驗轉型過渡期之測驗公平性與分數可比性啟示

任何大型測驗在推動數位轉型時，若欲將具備長遠「紙筆傳承」的傳統考科轉移至線上，考生與社會大眾往往會對「紙本換成電腦考，分數會不會不準或不公平？」抱持疑慮。面對從紙筆轉成電腦化測驗所造成的測驗不公平現象與模式效應（TME），測驗主辦機構不應僅仰賴行政層面的宣導，而應建立嚴謹的科學驗證與檢核機制。

首先，針對測驗載體改變可能帶來的干擾，國家考試未來不宜僅依命題委員專業判斷試題是否公平，亦應建立常態化、模型化的 DIF 檢核流程。尤其針對具有長文本、情境判斷、圖表整合或高度數位操作要求的題目，更應檢視其是否因「介面轉換」而對不同性別、年齡、地區或教育背景考生產生額外的不利效果。

其次，在擴展數位化的過渡期中，無可避免地必須直接面對 TME 的檢驗。在推動傳統重點科目數位化時，若可行，建議可在全面上線前，先進行小規模的「紙筆與電腦化測驗之可比性（comparability）研究」；在正式轉型初期，題庫組卷亦可策略性地放入曾在紙筆測驗中施測過且具備穩定參數的「共同題」。透過統計等化技術，精準評估並校正「測驗載體改變」對考生成績所造成的影響。唯有以客觀的信效度指標與等化數據，證明轉換過程中的 CBT 考試分數與過去紙筆考試具有實質等值性，方能有效消弭外界對測驗模式轉換的疑慮，確保測量品質的絕對公信力。

五、因應測驗目的與數位情境之題序取捨：兼顧「信心維護」與「疲勞控管」

高風險的國家考試與作為國家政策調查的 TASAL，在測驗目的上本有根本

差異。國家考試攸關個人職涯與公平篩選，傳統試題編排多依循「先易後難」的政策原則，旨在穩定考生情緒，避免一開始即打擊作答信心；而 TASAL 作為長期追蹤調查，更關注如何在高密度的素養題本中，維持群體受測者的作答動機。因此，兩者在題序設計上本就會因目的不同而有所取捨。

然而，當未來國家考試逐步轉向電腦化測驗，並不可避免地導入長篇素養題型時，考生在螢幕前面臨的挑戰將不再僅是單純的「信心受挫」，更包含了嚴重的「認知疲勞」。TASAL 透過多卷別調動題序來分散疲勞風險的實務經驗，確實證實了作答負荷控管對維持測量效度的重要性。然而，誠如考選實務所見，國家考試囿於試後公布試題之慣例及社會對公平性之絕對要求，同類科應考人均須作答同一套試題；且現行國家考試 CBT 系統之亂題亂序功能主要係發揮防窺與維護試場秩序之效用。若全面啟動亂序，傳統先易後難的排序原則勢必被打破。

在此實務限制下，國家考試雖無法直接套用 TASAL 的多題本等化設計，但 TASAL 關於長文本易引發認知疲勞的實徵發現，仍可作為數位試卷建置的重要參酌。建議未來國家考試在推動電腦化測驗時，或可於命題或組卷階段導入「試前評估文本認知負荷」之機制，並對高負荷文本進行嚴格把關。具體的控管策略，除了「限制單一套試題中高認知負荷題型的總量與比例」外；針對數位系統的亂題亂序功能，未來若技術允許，亦可評估採行「分區（層）亂序」機制（例如：測驗最前段的基礎題區塊採小範圍亂序以穩定作答信心，中後段高認知負荷區塊再進行大範圍亂序以達防弊之效）。此種在絕對公平與防窺前提下，將教育評量之科學概念轉化為具體試務規範的務實微調，將是未來優化考生經驗、維持測量效度的關鍵配套。

六、對設備治理與基礎建設的啟示

（一）考場條件與設備標準化之重要性

數位轉型有賴於穩定的考場條件。為降低設備與網路條件對考生作答之干擾，未來在考場建置上，或可持續朝向統一設備規格、完善網路備援及現場技術支援等標準化方向邁進。

（二）確立連續性之系統維運治理

在當前的公共治理中，數位系統需長期維運已是基本共識。電腦化測驗系統所依賴的軟硬體環境具備高度變動性，實務經驗顯示，系統維護若有停滯，後續

為因應新環境而重啟維運的修復成本往往更高。因此，大型電腦化測驗平臺若能視為關鍵基礎設施，並在預算規劃上朝向常態化、連續性的維運經費編列，將更有助於確保高利害測驗所需的穩定性與安全性。

七、作答歷程資料應用之未來展望

電腦化測驗最大的新增資產之一，在於自動生成的作答歷程資料。如同 TASAL 在實作任務中精準捕捉考生的「錯誤步數」、「停留時間」與「策略修正軌跡」，這些資料徹底改變了測驗分數的意義。對國家考試而言，未來的選才將不再僅限於評估考生「是否選出正確答案」，更有機會應用歷程資料來建立更細緻的品質監控與風險偵測機制。例如，透過分析頁面停留時間與反覆切換路徑，可協助判別考生是否落入猜測行為、出現認知疲勞，或及早預警異常作答模式。此外，若國考未來擬引入「情境決策模擬題」（例如模擬公務員在危機事件中的處置流程），歷程資料更能成為評量其決策應變能力的重要科學依據。

伍、結語

在我國素養導向課程改革深化之際，電腦化測驗已不再只是單純的技術選項，而是評量制度能否有效回應新型能力需求、維持高風險測驗公信力的重要關鍵。以 TASAL 為代表的國內大型調查研究實務顯示，電腦化測驗的成功推動，必須同時建構題庫治理、線上閱卷、監考支援、試題公平性檢核與歷程資料分析等完整制度配套。

回顧本文探討，TASAL 在四大介面的建置，體現了數位測驗從命題到評分的全流程治理能力；其在多向度測量模型、MIMIC 與定錨策略的 DIF 檢核實踐，展現了電腦化測驗在測量精緻化與公平性保障上的方法學進展；而「河內塔」等實作任務所擷取的作答歷程資料，更突破了傳統紙筆測驗僅能評斷對錯的限制，為捕捉考生思考與策略軌跡提供了實徵依據。同時，TASAL 對認知疲勞的題序調動策略，以及透過同時估計的等化技術之應用，亦深刻揭示了不可忽視的效度挑戰與現場現實。

對國家考試而言，這些經驗最重要的啟示在於：電腦化測驗的推動必須兼顧技術可行、制度正當與考生權益保障。未來若欲穩健推進數位化，可從監考防呆、建構反應題線上評閱、公平性檢核常態化、認知疲勞防範及設備維運治理等面向建立制度基礎。再往前一步，則應審慎導入人工智慧與歷程資料分析，使電腦化測驗不只是更快、更方便，而是更能提升評量品質與公共信任。

總言之，大型教育評量所累積的電腦化測驗經驗，已為我國未來高風險考試的數位轉型提供了具體路徑。如何將這些經驗轉化為可持續、可治理、可被社會信任的國家考試制度，將是下一階段教育評量與考選治理共同面對的重要課題。

參考文獻

陳繼成 (2024)。學科素養與閱讀素養之爭：TASAL 素養導向試題分析計畫。

國家教育研究院研究計畫成果報告 (編號：NAER-2022-017-C-2-1-C8-01)。

陳繼成 (2025)。臺灣高中學生核心素養評量計畫 (二) —臺灣高中學生規劃執行與創新應變之表現評估。國家教育研究院研究計畫期中報告 (編號：NAER-2024-024-C-1-1-C1-02)。

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23. <https://doi.org/10.1177/0146621697211001>

Arıkan, S., Özer, F., Şeker, V., & Ertaş, G. (2020). The importance of sample weights and plausible values in large-scale assessments. *Journal of Measurement and Evaluation in Education and Psychology*, 11(1), 43-60. <https://doi.org/10.21031/epod.602765>

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: Psychometric and statistical considerations. *Large-scale Assessments in Education*, 5(1), 17. <https://doi.org/10.1186/s40536-017-0050-x>

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26(1), 3-24. <https://doi.org/10.1177/0146621602026001001>

Shih, C. L., & Wang, W.-C.. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, 33(3), 184-199. <https://doi.org/10.1177/0146621608321758>

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement*, 67(2), 219-238. <https://doi.org/10.1177/0013164406288166>