自動評分之現況與未來可行性評估

謝名娟，國家教育研究院測驗及評量研究中心

測驗及評量組

壹、議題的重要性

大型評量現受限於經費與成本，多半只能使用紙筆測驗，不僅施測容易、且評分單純。然而，紙筆測驗卻會局限了考試的內容，尤其是實作、技能方面的能力很難評量出來。近年來教育改革浪潮興起，其中師資培育白皮書中更質疑現行教師資格檢定以紙筆測驗方式是否能檢核師資生活用知識的程度、尤其親子天下（2013，5 月29 日）從「被標準答案綁架的老師」來訴說檢定試題題型的僵化，已限制住原有欲評量師資應具備的基本能力。

不可諱言，若要能大規模執行實作評量，自動評分為必要的一種作法。例如，若欲評估老師的口語表達能力，必須花費大量的經費與施測人力。針對每一位師培生進行數分鐘的問答，而後再採用兩位口試委員做為評分者來進行客觀評分，現有近萬名考生，要花費口試委員的時間與施測成本將難以估計。

目前在臺灣自動評分的研究相當有限，但在美國已行之有年，並有測驗公司進行長期的研發。本文將提出在美國自動評分的做法、使用的題型、其相關文獻的信度評估，以做為未來臺灣在大型評量中，執行自動評分的參考。

貳、主要國家（美國）具體作法

在共同核心標準的要求，美國各州都必須持續舉行大規模施測，過去為了節省施測成本，多使用選擇題。然而，在新式評量的呼求聲浪中，單一形式的選擇性題型已不符合各州之需求，各種形式的建構式試題相應而生。建構式試題最大的問題為批改的龐大人力成本，因此近五年來，許多大公司（如ETS, Pearson）都發展自動化評分系統，希望除了能節省閱卷成本之外，還能降低人工閱卷的誤差，提高評分的信度。

目前在美國已針對各種建構式試題寫出自動評分的程式，包括聽、說、讀、寫的各種類型。有短文、口語表達、短題問答與數學的題表與數字題。在真人評分時，其評分信度的高低仰賴試題本身是否敘述清楚、人員妥善訓練。但是在機器的自動評分上而言，則著重在對建構式試題的答案限制。答案限制越多，則評分一致性越高。

表1呈現目前有在使用，且使用機器自動評分的題型。樣本數為實驗的人數，而機器與真人評分分數的相關性有時比兩位真實評分者所得到的相關還要高。

表1

自動評分題型

題型	作者	樣本數	機器 vs 真人	真人 vs 真人

			相關性	相關性
短文（6年級~12年級）	Practice Hall	400	0.89	0.86
短文（4年級~12年級）	MetaMetrics	635	0.91	0.91
摘要彙整	Council for aid to Education	1239	0.88	0.79
口說測驗（成人）	Balogh& et al .（2005）	50	0.97	0.98
口說測驗	Bernstein et al.（2009）	134	0.97	0.99
口說流暢度（一年級~五年級）	Downey et al.（2011）	248	0.98	0.99

此外，自動評分系統在速度上也比真人評分快速的多，就ETS的e-rater報告指出，在GRE的作文評分時，為了節省成本，一份作文試卷要求閱卷者僅能花兩至三分鐘就評閱完畢，而e-rater可以在20秒內，評閱16,000份試卷。

然而，機器畢竟不是人類，仍然有其缺點。The New York Times（2012）報導指出，機器評分最大的問題是無法評斷作者描述事件的真實性。例如若作者寫一次大戰時間發生時間是2015年，這種胡扯性的事實機器評判不出來。　此外，機器評分具有邏輯性，喜歡固定的語句和寫法，若作者把Or 或And 放在開頭通常比較低分，若是文中有however, morever則比較高分。機器評分喜歡長篇大論，不喜歡短小精幹的文章。若了解機器評分邏輯的考生，則也可能輕易用一些沒意義的語句拿高分。

表2則為目前已發展成熟並適用於大型測驗之各種自動評分測驗的題型，包括聽、說、讀、寫四種混雜的類型。其中就寫而言，包括限制性的打字、重組句子到較為開放性的，如寫一封簡短的email等，就口說部分，也有朗讀、重述、句子重組到較為開放性的問題像是重述一篇故事、回答對話等問題等。

表2

Versant Pro 中使用的自動評分類型

測驗類型		測驗描述
寫	打字	使用60秒，逐字擅打短文。評估打字速度與正確性。
寫	完成句子	使用25秒，進行克漏字填空。
寫	語意重組	先閱讀螢幕上的短文30秒衝，而後再使用90秒鐘寫出內容大要，須越詳盡越好。
寫	寫email	根據主題與情境，來回覆email，撰寫9分鐘。
讀、說	朗讀	使用30秒鐘，將文章逐字逐句唸出來，儘量完整且清楚的閱讀。
聽、說	重述	將剛剛所呈現的句子重述一遍。
聽、說	短答	根據題目所問的內容，提出簡短的回答。
讀、寫	句子重組	將名詞重組成有意義句子
讀、說	說故事	根據所讀的故事，重新用自己的話說出來。
聽、說	對話	回答的對話內容。
讀、說	短文理解	根據短文的內容，回答三個相關問題。

參、我國現況概述

目前有許多研究計畫與學位論文（李垚曒，2011；游雅婷，2006；戴予恆，2013），試探自動評分在寫作測驗上的應用。

在大型施測的應用上，目前有師大心測中心和交大合作，研發國中基測中文寫作自動評分系統（Automatic Chinese Essay Scoring，簡稱ACES），預試了689名和1,390名學生，透過兩個作文題目進行電腦和真人的評分，其評分一致性高達了9成，然而，電腦評分對於寫作創意部分，則有較大辨識誤差。由於基測屬於高風險的測驗，還需多加測試。

另外，師大科教中心為了能夠讓開放性問題的題型可以落實於教學評量，針對開放性問題或短文作答研發自動評分。（計畫網站http://www.dorise.info/main/iAssessment/002-plans.html）。其使用演算法有兩種：其一是根據一個或數個文本特徵進行統計分析，採取的處理方式大多是針對文本進行一些表面的斷字或斷詞。其二則是完整的自然語言處理分析（Full Natural-Language Processing），根據文本的語言學結構，來判斷語意及語法，其演算法則會依據語言的種類，以及語言學的分析模型而不同。目前仍處於實驗階段。

就目前臺灣自動評分發展現況，多聚焦在寫作測驗上，但在口語方面的相關發展較少，可能是語音要轉化成文字上的技術層面較為困難，但相關研究值得未來深入探究。

肆、對我國的啟示與建議

一、機器可輔助評分，非代替人工評分

現今越來越多的測驗，講求走出紙筆與選擇題的方式，使用更多元的方式來測驗學生在各個面向的多元能力。然而，在大型評量上，由於實作評量涉及人工評分，執行上很難推動。此外，凡測驗必有誤差，尤其人工評分，受限於疲倦、偏見等影響，若是一份測驗僅使用一位評分者評分，其誤差甚大。若能採用自動評分的方式，一個分數採用人工評分，另一個採機器評分，而有歧異的時候，再使用另一位人工進行評分的方式，除了能節省經費，也較能保證客觀性。

二、須長期、持續性的經費支持研究

要能建置一個可用的自動評分系統，需要投入長期的時間於研究人力。評分的準確性，和系統資料庫蒐集的語料庫大小有關，當語料庫的資料越豐富，系統判別會較為準確。雖然臺灣與大陸有現成免費的中文語料庫，但仍須加以測試精進。

三、優先試行低風險性的考試

在實務執行方面，可優先考量大型且低風險的測驗，例如國教院所建置的TASA，等系統較為穩定成熟，才逐步推廣至中、高風險的考試，如教檢、基測或是會考。另外，系統建置容易，但維護才是大工程，為了避免成為蚊子系統，長期、持續性的投資與維護具有必要性。

參考文獻

李垚暾（2011）。**使用模糊理論於中文寫作自動評分之新方法**（未出版之碩士論文）。高雄應用科技大學資訊工程學系，高雄。

游雅婷（2006）。自動評分系統與雙語索引系統對臺灣大學生寫作之影響（未出版之碩士論文）。清華大學外國語文學系，新竹。

親子天下（2013）。被標準答案綁架的老師。取自 http://www.cw.com.tw/article/article.action?id=5049435

戴予恆（2013）。**具有自動評分功能的類別圖考試系統**（未出版碩士論文）。交通大學多媒體工程研究所，新竹。

The New York Times（2012）. Facing a Robo-Grader? Just Keep Obfuscating Mellifluously.取自 http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html