量身訂做的考試或測驗

穿著合適的衣服，最能展現體態與韻味。其實考試或測驗也可以像試穿衣服一樣，給最符合考生能力水準的考題，最能反映她／他的真實能力。適性測驗就是要做這樣的事。

■何榮桂

台灣是一個考試名目非常繁多的國家，不論男女老少、各行各業的人，都有很多考試的經驗。

台灣是一個考試名目非常繁多的國家，不論男女老少、各行各業的人，都有很多考試的經驗。有人屢考屢勝，風光順遂；也有人屢考屢敗，聞考色變。考試幾乎影響了一個人的成敗或事業的起伏，可見考試與我們一生息息相關。考試（examination）說得嚴謹一點或學術一點就稱爲測驗（test），測驗這個詞較廣義一些，例如測量態度的工具通稱態度測驗，若說態度考試就比較不恰當，但兩者交換使用也無不可。

考試或測驗的結果要做爲好壞、高低的判斷，因此必須根據比較客觀的原則才能服眾，這個原則就是測量（measurement），也就是根據某種法則，給予人、事、物、甚至現象等一種符號的歷程。測量的法則不難了解，用我們最切身的例子來說，一個胎兒一離開母親的肚子，接生的產科醫生或產婆馬上就可以根據生理特徵（最明顯的應該是生殖器官——男女有別吧！）判斷新生兒的性別，這可能是人生第一次被測量。這個例子告訴我們測量的最基本意義。

同樣地，各式各樣的考試也都是一種測量活動，評分者必須依據評分的標準（有些很客觀，而有些較主觀一些）評定考生的分數。測驗或考試是測量過程中最常見的手段之一，形式繁多。大部分人考試失敗的經驗總比成功的經驗多得多，因此，大部分的人不喜歡考試，甚至害怕考試。但主考官總會找些理由來考試，例如要進行診斷、選才、安置、預測、成績評定、證照認定等，認爲這些理由都正當、合理，而被考的人則會要求公平、合理與準確。

人的一生雖被考無數次，但一般人對測驗的本質並不十分了解。傳統測驗是否公平？合理？或準確？一般人也似乎不太在意。測驗專家卻針對這些問題不斷進行

貓（CAT）總給人神祕感，測驗也是如此。電腦輔助測驗（computer-assissted test）或電腦化適性測驗（computerized adaptive test）的縮寫都是CAT，既然考試或測驗是人生不能避免的事，那就像這隻酷貓，樂觀地面對吧！

大部分人考試失敗的經驗總比成功的經驗多得多，因此，大部分的人不喜歡考試，甚至害怕考試，補習班也因而隨處可見。

研究與探討，因此有多種測驗理論及考試方法的產生。

心理計量專家認爲，只要是存在的事物，不管是具體的或抽象的，都有被測量的可能。人類的特質大多是抽象的，諸如智力、性向、創造力、問題解決能力、情緒、態度等。因爲絕大部分是抽象的狀態或歷程，也因人類有無法數計的特質，測驗專家只好依各種特質的狀態設計各式各樣相對應的測量工具，間接推測各種特質的狀況或歷程，因而有諸如智力測驗、態度量表、創造力測驗、口語表達能力測驗、語文能力測驗等測驗工具的產生。由此不難想像現今的測驗工具種類繁多，難以數計。

常見的測驗方式不外乎口頭回答（口試）、動作展現或操（實）作、筆試等，其中又以筆試占最多數。筆試也就是測驗專家所稱的紙筆式測驗（paper-and-pencil tests），目前常見的紙筆式測驗大都根據傳統或古典的眞實分數（true score）測驗理論。

這種理論的立論就是「眞實分數」（其實我們不知道一個人的眞正能力有多高或多低，因此常常稱爲潛能）等於「實測分數」（實際測量得到的分數）加上「誤差」（每次測量都會有誤差）。一般在做測驗時往往只考一次，但理論上如測量次數夠多，誤差會相互抵銷，最後「實測分數」就會接近於「眞實分數」。

這種理論簡單易懂，但漏洞多，也解決不了一些常見的問題。例如兩位考生得分相同，是否其眞實能力也一樣？其實未必相同，古典測驗理論就無法回答這種問題。

晚近，測驗專家又提出「試題反應理論」（item response theory, IRT）的現代測驗理論。這種以數學及統計學爲基礎的理論，不僅可解決許多古典測驗理論無法回答的問題，也提供了電腦化適性測驗的理論基礎。

現代測驗理論不僅以數學模式（考生大可不必了解這些複雜的數學或統計學模式，相信這些測驗專家不會騙人也不會唬

現代測驗理論不僅以數學模式來校準每一個試題的難度、鑑別度、可能被猜測的程度等，更可以用視覺化的圖形來表示每個試題的難度、鑑別度與猜測度的相對位置，以及每個試題用來測試考生後所可反映的訊息量。

人的）來校準每一個試題的難度、鑑別度（可以把不同能力區分出來的指標）、可能被猜測的程度等，更可以用視覺化的圖形來表示每個試題的難度、鑑別度與猜測度的相對位置（測驗專家稱爲試題特徵曲線），以及每個試題用來測試考生後所可反映的訊息量（稱爲試題訊息曲線）。

對測驗專家或主考官來說，選取難度適中、鑑別度高、猜測度低和訊息量豐富的試題組成一份考卷或測驗卷，最能考出考生的眞實能力。這些看似簡單，但做起來並不容易，這也就是爲什麼有那麼多測驗或考試專家一直很用心地從事測驗或考試的研究吧!

這裡要談一個有趣的問題，就是在前面提到的得分相同的考生能力是否也一樣呢？筆者曾實際分析某考試機構某一科目的考試成績，其中有8名考生在40個選擇題中都答對了39題，但答錯的1題都不一樣，例如有考生答錯第13題、有的答錯第15題等。如果答對得1分，依古典測驗理論的記分方式，這8名考生都得39分。但筆者利用試題反應理論中貝氏（Bayes）能力估計法的原理，分析這8名考生的考科能力，結果發現這8名考生的能力其實都不同。

這個例子告訴我們，現代測驗理論對考生眞實能力的估計，比古典測驗理論精確多了! 現代測驗理論不僅有這些優點，更重要的是提供了實施適性測驗的理論基礎。

資訊科技的發展與應用，大大地影響了考試的方式，電腦輔助測驗（computer-assisted tests, CAT）也應運而生。目前已有很多著名的測驗用電腦來實施，如托福、GRE等。電腦輔助測驗乍聽好像很有現代感，其實測驗電腦化並非新的東西，在大部分的我們還沒出生前，測驗已經開始電腦化了。1934年，美國哥倫比亞大學教授班・伍德（Benjamin Wood）就與IBM的工程師合作，開始發展電腦閱卷機。他們的創意隨即被一位高中科學教師實現了，類似目前使用的電腦計分卡在那個時代就已誕生。

創造力也可用紙筆測驗加以測量，威廉斯是這項測驗的編製者。

性向（aptitude）是「認知能力」，非「性」格的傾「向」。

這項發明使班・伍得教授贏得教育改革者的美譽，而電腦閱卷機也立即對測驗產生影響。例如減輕了人工閱卷的勞力負擔，且允許測驗題本與答案卷（卡）分開印製或處理，換句話說，測驗題本可重複使用。此外，因爲使用電腦閱卷，刺激了大規模測驗的實施，也增加了選擇題記分的可靠性。

運用電腦於測驗中雖然很早就實現了，但往後測驗電腦化的發展大抵與電腦的發展平行進行。1970年代的電腦，體積龐大，人機介面以文字模式爲主，導致電腦的工作大都隱藏於幕後，用於處理繁重且重複的事物。1980年代後，因個人電腦的出現及大量使用圖形介面，促使電腦走到幕前，儼然成爲每個人的工作伙伴。

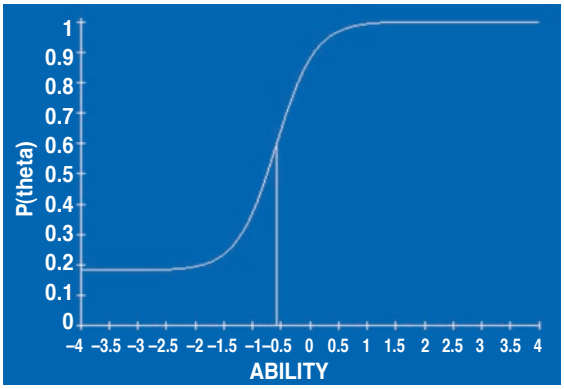
電腦應用於測驗的發展也大致如此。1970年代，測驗僅用電腦閱卷、記分及處理測驗報告。到了1980年代，才開始使用電腦來實施測驗。換句話說，電腦與測驗的結合與應用，到了個人電腦出現後才逐漸普及。

由於有了試題反應理論爲依據，配合現代電腦的高速運算能力，有別於傳統測

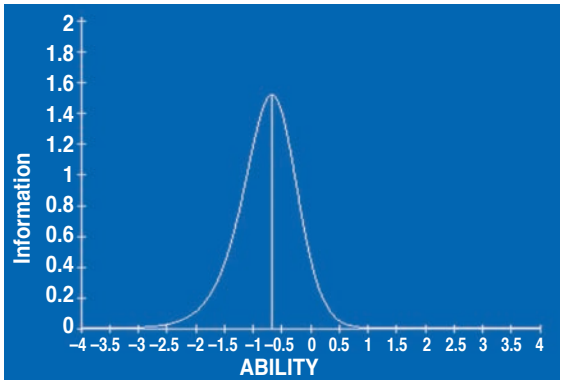
資訊科技的發展與應用，大大地影響了考試的方式，
電腦輔助測驗也應運而生。目前已有很多著名的測驗
用電腦來實施，如托福、GRE等。
電腦輔助測驗乍聽好像很有現代感，
其實測驗電腦化並非新的東西，
在大部分的我們還沒出生前，測驗已經開始電腦化了。

驗電腦化的適性測驗（adaptive test）的實施就成爲可能。適性測驗就是量身訂做（tailored）的測驗，早期稱量身訂做的測驗（tailored test），實比目前用適性測驗（adaptive test）傳神多了。

量身訂做的測驗，意思就是給考生做符合她／他能力或特質水準的題目，這樣最能反映她／他的能力或特質。對考生來說，太難或太簡單的題目都不適合，也就是說，題目太難或太簡單，即使考生答或猜了這些題目，也無法從這些答題反應獲知考生的能力或特質。或是說，這些題目並無法回饋考生的訊息給主試者，這樣的測量就失去意義。



試題反應理論可依每個試題的參數（難度、鑑別度及猜測度）繪製試題特徵曲線，曲線愈靠右表示愈難，愈陡則鑑別度愈高，截距愈小則猜測度愈低。



試題反應理論也可依每個試題的參數（難度、鑑別度及猜測度）繪製試題訊息曲線，曲線所圍的面積愈大表示測驗訊息愈豐富。

適性測驗既然稱爲量身訂做的測驗，在這裡就以試穿衣服爲例，說明適性測驗的施測原理。假設某一款式的衣服有5種尺寸，1代表最小，5代表最大。當選購者對尺寸大小或售貨員對選購者的身材都一無所知的情況下，最好是先讓選購者試穿中等尺寸的衣服，因爲身材中等的人最多，誤差的機率最小。在這個例子中，最好先試穿3號衣服，而在實施適性測驗時，3號衣服就相當於初始題目。

我們都有類似的經驗，在這麼多不同尺寸的衣服裡，第1次就選到合身的機會並不大，不合身當然要再挑選。如果太小，下一件當然要挑大一點的；反之，則選小一點的。問題是如何再從大一點的2件或小一點的2件中有效率地挑一件就合身，這在適性測驗裡就涉及選題策略（item select strategies）了。選題策略是適性測驗過程中很重要的工作，而設計有效率的選題策略更是適性測驗專家的研究重點。

「我喜歡做的事」是國內著名的職業興趣測驗，也曾發展為電腦化測驗。

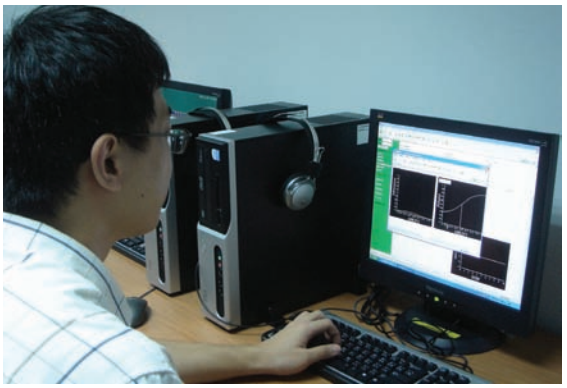
就像試穿衣服一樣，適性測驗的實施過程，就是要從一組題目或題庫中找尋符合考生能力的題目給她／他作答，「找的方法」就是適性測驗的選題策略，也就是判斷某題目是否符合某考生能力的方法。

一般而言，在一組題目或一個題庫裡，不容易一下子就找到符合某一考生能力的題目，通常要經過好幾題的測試，才能漸漸找到。爲了測試某題是否符合考生的能力，必須每答一題，就要重新估計一次考生答了這個題目後的能力，這是適性測驗的一大特色。而每一階段（即每答一題）的能力估計，也是適性測驗施測過程中非常重要的工作。

從前面例子的說明，可以理解適性測驗的原理其實很簡單，先給考生一個難易適中的考題，如果她／他答對了，表示這個題目對這考生來講並不難，則下一個考題會難一點；反過來說，如果答錯了，下一個考題就應該容易一些。如此反覆進行

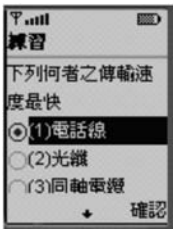
對考生來說，太難或太簡單的題目都不適合，
也就是說，題目太難或太簡單，即使考生答或猜了這些題目，
也無法從這些答題反應獲知考生的能力或特質。
或是說，這些題目並無法回饋考生的訊息給主試者，
這樣的測量就失去意義。

測試，每測試一題，測量誤差就會逐步遞減，一直進行到誤差小到可以容忍或接受的程度時，考試就可結束了。而考生在答最後一題時，能力估計的落點會停留在一個大約從－3到＋3的能力量尺上。



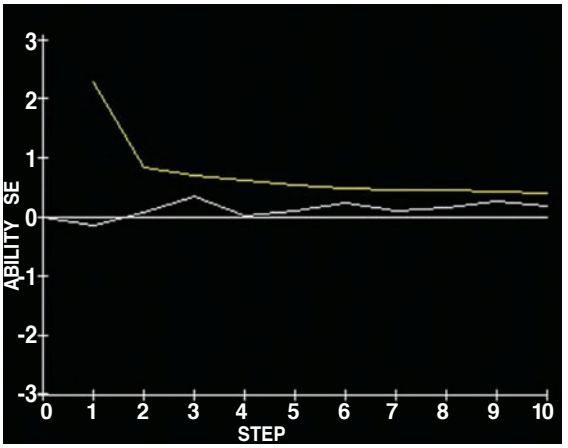
用電腦做測驗，情境標準化，較準確又較經濟。

用 PDA 也可進行測驗



用手機也可進行測驗

因爲人類的眞實能力或特質無從直接得知，理論上這把量尺應從負無限大到正無限大，但實際測量時，大概99.9％的考生的能力估計值都會落在－3～＋3的範圍內。若某考生的能力估計值落在量尺上0的位置，表示這位考生的能力中等；如果落在2.9的位置上，就表示這位考生的能力高人一等。由於考生的能力不同，因此電腦從題庫挑選出來給不同考生作答的考題及最後答題的題數也未必一樣，但並不影響能力估計的準確性，這也就是適性測驗與傳統測驗不一樣的地方。



這個圖告訴我們某一考生經過 CAT 測試10題後，這位考生的能力估計值是 0.17（中等），誤差是0.41（可以接受）。

另一個與傳統測驗不一樣的地方，是適性測驗每次選題都儘量挑選最接近考生能力的題目，且其結束是以誤差的極小值爲依據。不像傳統測驗以固定時間考試，且要求全數作答，不管考生會不會回答。因此，適性測驗所用的時間比較經濟，作答的題數也比傳統測驗少很多，這也是適性測驗的一大特色。

適性測驗的道理雖然簡單，但製作過程卻相當繁瑣（這應該是測驗專家的事）。在實施過程中，每階段的選題與能力估計都涉及複雜的計算，如果沒有運算快速的電腦來輔助，實施起來就很困難。這也就是何以適性測驗需要電腦化的理由，也因而稱爲電腦化適性測驗（computerized adaptive tests, CAT）。目前電腦硬體的功能都很強，軟體也十分親和，發展電腦化適性測驗已經是容易的事，甚至以行動載具如平板電腦、PDA、手機等來實施測驗，都不是困難的事。

測驗理論的發展伴隨著資訊科技的進步，可以預見未來大部分紙筆測驗會被電腦化測驗取代，而適性測驗也可能逐漸取代傳統測驗。這不僅是理論與技術的進步而已，我們對人類內在潛藏的能力或特質也能更深入了解！ □

何榮桂
台灣師範大學資訊教育研究所

深度閱讀資料
1.何榮桂（民89），量身訂製的測驗－適性測驗，測驗與輔導，157，3288－3293。
2.許擇基、劉長萱（民78），試題作答理論簡介，行為科學社，台北。

適性測驗每次選題都儘量挑選最接近考生能力的題目，且其結束是以誤差的極小值為依據。不像傳統測驗以固定時間考試，且要求全數作答，不管考生會不會回答。因此，適性測驗所用的時間比較經濟，作答的題數也比傳統測驗少很多。