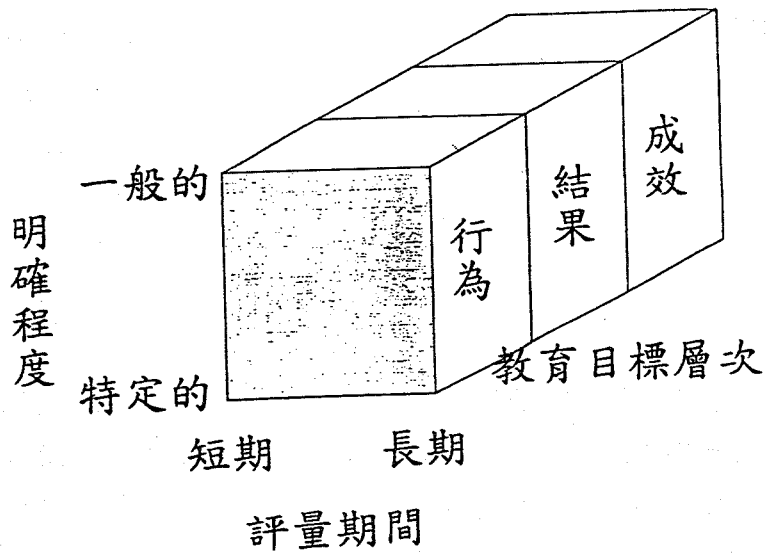
由多元評量的觀念看傳統評量的角色與功能

林世華
國立臺灣師範大學　教育心理與輔導學系

首先從傳統評量的定位來看，多元評量、另類評量，或變通評量，基本上都是現在評量中相當重要的趨勢。有時候甚至爲了對照的需要，經常把傳統評量定位在紙筆式的測驗，甚至更窄化到可以電腦化閱卷的選擇題，也就是說大部分是以選擇題作爲多元評量的一個對照基準，從事科學研究很喜歡用這種方法，因爲本身無法很順利彰顯功能或特色，總是要有一個對照，所以傳統評量在這角色上扮演得相當不錯。今天我會是從一般評量的觀點、一般教學的觀點來看評量。教學基本模式大家都知道它的順序，教學目標先發生，起點行爲的測量再發生，再來是教學活動的發生，最後才是教學評量，所以是目標先發生，評量是根據目標來決定，所以應是教學目標引導且決定評量，這是學理。不幸的事實是教學評量卻嚴重影響了教學活動，甚至教學目標，學理上這是不對的，但它卻發生了，因此在從事課程教學或師資培育上，一定要記住這個事實是不正確的。雖然不對，很多老師是如何學會的，師資培育學校老師沒教，任教老師卻自己自然而然學會了。尤其在國中教書以後老師經常會直接提一個問題：「如果你不告訴我要考什麼，我怎麼知道要教什麼？」我覺得這問題是有邏輯上的問題。因爲若不考的話，豈不是就不用教了嗎？小學就不用教了，因爲沒有考試。問題發生的關鍵點在那裏呢？在 1993 年 Gitomer(1993)的一篇「實作評量與教育測量」的文章中就提到這個問題，我稱它爲多元評量與傳統評量所遭遇的一個點，Gitomer 引用兩個例子(Chase&Simon,1973)：好的閱讀者能較正常、快速的辨認文章中的字；另外一個例子，下棋高手能比生手更能精確的記憶棋子在棋盤中的位子。這兩個例子告訴我們的幾個觀念及他們所發生的幾個問題：文字辨認能力與棋盤記憶的能力或技能；另外兩個可能是我們較關心的，閱讀能力及棋奕能力，前面兩個事實上是當他下棋下得很好或閱讀能力很強時，所必然發生的特徵，叫 emergent properties（未料特徵），作者的意思是說，用評量的觀點來看，若測量學生的文字辨認與棋盤記憶的能力，我們就會把這兩個視爲閱讀能力及棋奕能力的代理測量(proxy measures)。因此訓練文字辨認及棋盤記憶的能力並不意味著你的閱讀或棋奕能力是能同時被培養出來的。但這也未必，因爲代理測量與能力培養的完整領域之間的關係目前並不明確。各位若類推這個關係，回過頭來看國內大型的考試，其實今天我們發生的問題是在大型考試或聯招裡頭，事實上是大量使用代理測量的概念，但因爲這些代理測量必須對外公告，事實上對應代理測量的能力的練習便一再發生，相對的失掉國中教育真正要培養的能力，所以評量本身是無過的，是後來的使用發生

了問題。另外，國內有些特別現象，尤其在聯考一再發生一些不應發生的事情，聯考有規定：所有試題的用字用語不能超出課本之範圍，英文科是不可以用 fly 這個字，因爲它沒有出現在課本，所以出題用 fly 這個字便是超過範圍。事實上這是在執行整個入學考試計畫的一個很大的弊端。代理測量的公佈使大家便於訓練代理測量，變成大家在訓練文字辨認、棋盤記憶之能力，而不是培養閱讀及棋奕能力。



圖一　教育脈絡中多元評量的面向 (摘自 Gitomer, 1993)

接著，我也仿效陳教授所說的定位問題，在多元評量的觀點中，事實上是把評量的觀念擴大而不是多元評量觀念出現之後排擠另外一些存在相當有歷史的評量方式，所以我借用 Smith 在 1976 年提出的一個模式，這模式把評量從三個面向來看，這三個面向分別從評量期間、明確程度及教育目標層次。所謂教育目標層次是指：它有多接近教育目標的程度，此一模式是否好，我不知道，不過它至少符合我的需要，所以它不是公認的定位，是我接受的定位。當中所說評量期間的長短，若是傳統的評量，紙筆測驗或選擇題通常時間較短，若是實作評量是比較長的時間；若從明確程度來看，是要看評量所提供之訊息，明確程度有多大，假定用寫作能力觀念來看，若只是檢測他寫作中的主詞、是檢測他寫作中的主詞、動詞，是否能配合來使用之能力，這項測量的訊息是非常明確的，但若是看寫作中作品產

生的溝通的訊息是否有效，這就不是很明確，比較一般性。剛剛陳教授所提的實作評量中，大部分之目標較一般或完整一點，若從教育目標層次來看，其實是看它與真正教育目標有多接近，分別是分成行爲、結果、成效的三個層次來看，假若從讀及寫這兩個能力來看，行爲的層次很可能會去測量他的編碼之速度，因爲它與讀寫有密切關係，它是一種行爲，其實有點類似代理測量的概念；若從結果層次來看，就是看閱讀的水準有多高；若從成效觀點來看就不單純了，從教育的目標來看，要評的是到底達到哪個層次，例如：達到高中以上的水準，有多少人通過高中的程度，有點類似教育指標的觀念，我們國家須有一些教育指標來了解國人能力的品質如何，是屬於這個層次。若這樣看，傳統評量的定位就會比較清楚，它就會落在此一模式左下角的位置，多元評量大多數可能會在右側，上面一點，較偏向結果，跟傳統評量的位子不太一樣。基本上這是我對它的定位。

接下來，從多元評量的觀點比較回來，主要從評量所用問題的性質來看，剛剛陳教授也評論到傳統評量所面對的問題，通常比較結構、比較明確的，通常也是比較沒有條件的，另外一個就是它所涉及到的知識通常是非常有限的。多元評量的問題性質，通常是比較非結構，問題的解釋有條件限制，另外涉及到比較大的問題是多元評量的解答是會廣泛涉及到各種知識，特定的主題的知識領域，這是一般我們在看多元評量及傳統評量的比較時，從問題的性質上來看。給各位一個例子，這個例子是改自 GMAT，這是管理學科研究所的入學考試出現這樣的一個題目「文件影印一頁的價格是 4 元，一份有 x 頁的文件，若要影印 x 份，請問多少錢？」各位會覺得困惑的是，這與管理有何關係，這涉及到它在管理上性向測量時所用之代理測量觀念是什麼？多元評量的問題會是如下：「影印部門想要更新影印設備，以提昇營運效率，請問應做些什麼？」這兩個問題以本質來看，後者較接近管理的觀念，前者則比較間接。在比較多元評量與傳統評量時，從問題的性質來看這兩個問題其實是非常不一樣。不一樣就可能代表它可能在用途目的上不見得一樣。

最後，從傳統評量裡一個較小的問題，其實也是較被忽略的問題，剛陳教授也提到師大測驗統計這門課之教學，或是評量測驗之教學，事實上比較強調的部份一直是測驗的技術，以測驗爲單位的技術，始終稍微被忽略的是命題這部份，也就是如何把試題寫出來。因爲事實上把試題寫出來後，就較接近多元評量的觀念，所以特別把這部份挑出來看看傳統評量在目前發展的情況及其所作的努力，傳統評量對試題的意義，基本上是認爲任何一個評量上所用到的試題，不一定只是選擇題，它是具有刺激及引導的作用，能引導學生作答的一種測量單位，用於引起學生的反應，而基本上是假設這反應是建立在學生的心理建構，而我們對此心理建構是感興趣，比如說，測這個知識、測這個能力或測他的動機，涉

及到測量量化之功能，是一種數量詮釋之機制，學生是被刺激物，即試題所引導，學生是在一個被動的情形下，學生的行爲反應是用來詮釋心理的建構。其次是要報告傳統評量對試題的一個評量基準，當然目的必須是吻合，評量試題的目的是要清楚界定；另外必須要把評量試題之測量誤差將之最小化，這是因爲測量本身必須要在減低誤差的可能性，這也是與多元評量在觀點上稍微有點出入的一點；另外評量試題題型必須適合於評量目標，選擇題是繼續在發展，而不是因爲多元評量出來後而停止，但不可否認，評量尤其是傳統，它的整個研究基礎，雖然已使用將近 100 年，但其研究基礎到現在還沒有一個共識性的結果出來，還是非常經驗性，跟著教學一直在走。因此它所出現的各式各樣題型，傳統的選擇題題型是有一個題幹，即題目，有四個選項，其中一個是正確的，另外三個是誘答選項，所謂誘答選項是具有誘惑力，學生會去選它，剛剛陳教授所講的是事實，那三個誘答選項其實是選擇題中最難出的，研究很早就告訴我們，任何人可以寫出一個誘答就很了不起，若有兩個似真誘答有誘答功能，那已經是非常不容易，所以要寫出三個，總是要放進一個白癡選項在裡頭，所以各位一定很容易了解，爲什麼以上皆非、以上皆是很容易出現，就是因爲逗不出第三個了。所以確實這是個很大的困難。另外傳統題型主要有三種題型，分別是：在題幹上出現一個問題的題型，或試題幹上出現一個未完成語句題型，或是最佳答案的選擇，即每個選目都是答案，其中一個是最佳的，這是傳統的選擇題，但傳統的選擇題之外有很多變形。事實上這些變形也是因應它所受到之限制，比如說，選擇題受到最大的一個挑戰是：它總是測到一些零零碎碎的知識，有時勉強可以測到零碎的技能，這是一不爭的事實，所以選擇題是否有空間來作高層次思考的問題，答案是正向的，傳統評量也是在這個部份，甚至在 90 年以後他們還大量研究這些東西，所以包括題型在內，比如說增列選項的概念。事實上選擇題另外一個被批評的是：它可以用猜的，所以從事數學教育的人都很聰明，你這樣來，我就這樣對付你，他出了 10 個選項，我說四個都寫不出來，還寫 10 個，第一個是 0.5，第二個是 1，第三個是 1.5，第四個是 2，你要我寫 100 個我都寫的出來，但問題是答案會這麼湊巧嗎？他不這樣問，他說你解出答案最接近上面哪一個，爲什麼要這樣做，在做傳統評量的人是非常巧思的，因爲數學題經常考試考生會倒過來做，從選目裡回過來，所以你考的不是他解題的能力，而是驗算的能力，他看看老師出題能力好不好，出的好不好，所以老師就還以這種東西。重要的題型其實是選擇題目前的發展，其中一個是多重是非題，還有一個是多重選擇題，在大專聯考出現過，還有一種是脈絡依賴題型，實際上是類似傳統的題組，這個部分在選擇題中是很大的突破，尤其大約是在 90 年後。在 90 年代以前，傳統心理計量觀點認爲題組它彼此是內部相依的，題目之間會互相依

賴，即若這組答對他會得高分，若答錯他就會全軍覆沒，分數會比較容易兩極化的現象。但在 90 年以後，他們把計分的技術改變，他們使用題束的觀念，事實上在 90 年代以後，心理計量的觀點和學科專家的題組觀念又開始結合在一起，因爲傳統評量學者對於題組的觀念抱以非常高的期許在高層次思考的評量，傳統評量學者均認爲透過題組的使用應可以設計出相當理想的評量工具，以測量高層次思考的能力。

參考文獻

1.Gitomer, D. H., (1993). Performance assessment and educational measurement. In: R. E. Bennett & W. C. Ward (Eds.) *Construction versus choice in cognitive measurement*. Hillsdale, New Jersey.

2.Haladyna, T. M. (1999). *Developing and validating multiple-choice test items*. Mahwah, New Jersey.

棋盤策略題

設計者：陳昭地

在一張方格紙中，放了一堆的棋子，每一顆棋子放在一個小方格中，並依下列規則拿掉棋子：

每次移動跨越一小格，將一顆棋子橫向或直向跨越相鄰且有棋子的一個小方格，而進入下一個沒有棋子的小方格，即這個方格必須是沒有棋子，否則不被允許，隨即把被跨越過的棋子拿掉。

如果在方格紙之中央附近分別擺設如下形式的棋子，哪些形式經過若干次取走棋子後，到最後會僅剩下一顆棋子在棋盤上？【把辦得到的圈出來，並，在(13)~(16)中，就能辦到的，舉一例，寫出你的步驟】

(1)有何心得？

(2)若不能夠，請說明理由。

（取材自：國立臺灣師範大學科學教育中心舉辦之臺北地區國中學生創意競賽題目）