評量的蛻變與突破——
從哲學思潮與效度理論思考起

吳毓瑩

國立台北師範學院國教所副教授

台北市政府教育局　　函　　中華民國八十四年三月九日
北市教三字第１０２７３號

正本：本局所屬各公私立國民小學
副本：國立台北師院實小、政大實小、本局第三科，督學室
主旨：爲減輕本市國小學生考試壓力並改進教學評量方式，各校自八十三學年度第二學期起應採用紙筆測驗以外之方式辦理第一次定期考查，請查照。
說明：各校得採用鑑賞、晤談、報告、表演、實作、資料蒐集整理、設計製作、作業、實踐、及其他方式取代紙筆測驗方式，辦理第一次定期考查。

這是一份正式的公函（劃線部分爲作者自加），許多老師也曾爲這件事苦惱過。上學年，就教育改革而言，是熱鬧繽紛的一年。教育改革中，最具體可行、最直接、又成本最小的切入點，就是評量的改革。上述台北市教育局的公文中，明白地要求紙筆測驗頻率的減少。在台北縣的開放教育中，評量方式是其中七項基本認識中的一項（盧美貴，民83）。評量，在我們的社會中，很顯然的，是教育改革的指標與先鋒，是第一個應該開刀的目標，也是最具體可以看見的成效。評量改革來自410的現代教育實驗班、評量教學改進班、開放教育、台北市教育局的公文等，清楚的訴求是：我們不要塡充題、是非題、選擇題，我們不要傳統的紙筆測驗，然而，要的是什麼呢？好像也可以朗朗上口，如上述公文中的「鑑賞、晤談、報告、表演、實作、資料蒐集整理、設計製作、作業、實踐、及其他方式」，可是這些又是什麼呢？一直不很清楚。此點與Cizek(1991)談到評量改革在美國的浪潮時，非常相近，Cizek提醒教育界：

這些定義尚待澄清的另類評量，會比傳統測驗花上更多錢……而且著著急急地投資於剛出爐的改革方案，花下的錢會買到什麼？……舉例來說，Mitchell倡導的另類評量「可以包含所有想像得到的形式」。那麼也就是說教育界正被要求要購買一個看不到前景、且形式模糊的產品。

在我們的社會中，我尚未聽到有人公開發表（published）如此反潮流的聲音

，因爲要倡導另類評量（alternative assessment）都還來不及了，大概沒有人忍心將其打壓。不過，換句話說，在我們只聽到一種聲音（例如，廢除傳統紙筆測驗）傳播於媒體間、還沒聽到相反看法的時候，我們更要小心地自我檢視這個聲音在社會中的意義、在學校中的功能、或者在學術研究中的角色；我們還要知道，這個聲音來自什麼樣的思考，代表的是什麼樣的想法。這是我撰寫這篇文章的出發點。我不至於如Cizek 般的反此潮流，但是，我希望從三個角度來想想另類評量的意義。在本文中，我會先對評量的意義作一釐清，接著，討論另類評量的內涵，然後，再分別呈現不同思潮如何在評量之中找到自己的位置。

測、量、評、鑑的分別

談到評量，似乎須先對評量的意義作一釐清。評量的根源似乎是測驗，而其間關係如何呢？什麼是測？什麼是量？什麼是評？什麼是鑑？我對這一系列評量過程的解釋爲，這四者像是一層層的包裹「測」（測驗）爲最單純的方式，指在特定時間內要求受測者自己完成一系列問題的回答，測驗的形式與內容皆已事先結構清楚，每個受測者的測驗情境也力求一致，由此蒐集得的資料，作爲測驗的結果。「量」（測量）則是對某一特質作數字上的描述，可根據測驗得來的答案，也可根據其他形式的蒐集，如他人觀察，來形成量的資料。「評」（評量）則對於特質作數字上及語言文字上的描述、解釋及詮釋，採行方式的彈性比測或量都要大許多。「鑑」（評鑑）的意義有人認爲與評的意義相當，如Linn and Gronlund（1995, p.5）在他新版的書中，將鑑的意涵納入評之中。我自己將「鑑」的範圍定義得比「評」爲廣，與Airasian（1994, p.6）較爲類似。「鑑」因爲描述的過程已牽涉決策，故雖然一樣在爲資料作解釋，但解釋的根據會比評量時所動用的資料更要廣，也許包括經費的考慮，設備的狀態，人事的配合等。所以「鑑」可採納測、量、評三個方式的結果，再加入現實環境的考慮，對一特質作價值判斷或決策。

舉一個例子，學生接受大大小小的數學測驗（測），按照比例加減乘除後，老師會爲他這一學期的數學能力給一個等第（量），老師或他自己也會爲這一能力及學習過程作一個語言上的解釋，包括能力、動機、及特色（評），若這結果要用來作爲數學資優班學生的篩選，則描述可能還要包括此學生的現況是否可持續、潛力如何、資優班的名額、家長的配合等（鑑）。由於評鑑所考慮的因素較爲複雜，已超過學生與老師之間，但是它又植基於「評」所得的資料，因此我今天討論的內容將以「評」爲主。

另類評量－－它是什麼？

「教學經驗的增加，使得我這一、二年來，方得以從容思考教學與評量的問題；而隨著人生經驗的體會，也讓我有以下的教學概念：知識是取之於世界，故對學生的教學與評量也要回歸於原點，能在世界應用。」（陳素櫻，南勢國小）

另類評量、真實評量、與實作評量－－他們是一樣的嗎？

評量，加上另類二字，要傳達的意思是：傳統紙筆測驗以外的方式。它本來

是單純地由是一個形容詞加一個名詞組合而成，就如同我第一次聽到時的感覺一樣，表示別的其他種的評量方法，或者更廣義言，泛指所有評量方式的選擇。然而用得多了，漸漸就變成似乎是一種專有名詞，專門對傳統、大量被採用的紙筆測驗的一種抗議。想當然，它可以是任何傳統紙筆以外的形式，所以它還常常與實作評量（performance assessment）、眞實評量（authentic assessment）混用。Herman, Aschbacher, and Winters （1992）在他們的書中將這三個詞同義使用（use synonymously, p.2），泛指實作評量的各種形式，大凡要求學生用實作的方式而不是選擇的方式來反應者即是。國內有學者將之翻譯爲變通性評量（莊明貞，民84）。

既然另類評量是實作評量的各種形式，實作評量的意義顯然是其核心。Linn and Gronlund（1995, p.238）卻不認爲這三個名詞可以混用，另類評量在他們的定義中，如上所述，是相對於傳統紙筆測驗的其他種評量方式；可是眞實評量就有自己的另外一層意義，指的是評量的工作項目（tasks）可實際應用於眞實世界中，強調評與用之間的相等，例如寫一封道歉信，來表示文字溝通能力而非採行改錯字測驗，也就是將評量與生活結合。至於實作評量的定義，Linn與Gronlund（1995, p.238）則認爲是執行一項工作時展現出的過程與成果，教師即在評量此過程與成果的有效程度。因此Linn等便直言他們較喜歡使用實作評量這一詞，因爲它比另類評量更清楚地描述評量的重點與過程，又不似眞實評量矯情。

怎麼說眞實評量矯情呢？因爲眞實性事實上是一種程度與定義的問題，學校生活再如何眞實，也不過是社會生活的一種模擬，例如分數語言（1/6是1/3的幾分之幾？），幾乎很少在眞實生活中使用。往往教師提供的似眞實情境要學生揣摩並想出解決之道，所以評與用之間，不可能達成眞正眞實的標準。而眞實評量的強力推薦者，Wiggins，爲ERIC寫的一篇文摘中（Wiggins, 1990），採取另一種角度，他將「眞實」定義爲教師希望學生會什麼，就測他什麼，也就是課程目標在培養學生什麼樣的能力，就挑戰他那樣的能力。眞實存在於教與評之間的接近，亦即課程與評量的結合，不同於Linn所認爲的眞實性存在於評量與生活應用之間的接近。

一體三面的意義

其實這三個評量改革的新名詞，各有自己強調的角度，如同一件事物的不同面向，我倒不覺得有如此衝突，但是我也不認爲三者可以交互使用。另類評量形容的是評量改革所著重的方式，突破了以往的傳統紙筆測驗，涵蓋面很廣，強調的是評量方式的彈性，眞實評量從情境的角度著眼，強調評量項目的情境若愈符合眞實生活情況（real life）或是教育目標，則學生愈能活用他所學的知識與能力；表現評量則強調評量的依據乃是學生創作或學習過程的表現。因此我對教育評量改革的描述便是：在儘量合乎眞實的情境中（眞實評量），觀察並紀錄學生完成工作的過程與最後表現（表現評量），這便是一種不同於紙筆測驗評量的方式（另類評量）。

如果說眞實評量的眞實指的是於教與評之間的接近，那麼，我們國中的月考、期考、段考、模擬考就某種程度而言是符合眞實評量的定義。是的，如果教學的目標在幫助學生通過聯考，則符應教學目標的評量就是眞實評量。因此，從這

一觀點而言，教育之沈痾，罪不在評量，而在教育目標。這定義中的眞實評量一點也不矯情，反而中性的反應了現實——什麼樣的教育目標導引出什麼樣的評量。但是如果再往上推，則好像國中老師會告訴我們，罪不在教育目標，而在聯考的存在，然而聯考又是評量、篩選的一種方式，因此在這循環之中，有人認爲要突破此循環的關鍵就從評量的改革做起。

另類評量不是實施了嗎？

那麼就從評量做起吧，評量的定義根據以上所述，在蒐集資料爲學生的學習過程與成果作解釋，聽起來也不太困難，而且台北市教育局不也是公布實施了嗎？可是爲什麼進行起來好像不太像？

台北市教育局於上述第一張公文發出之後約三個月，又發了一份公函（北市教字第26343 號）如下：

> 台北市政府教育局　　中華民國八十四年五月二十九日
> 北市教字第２６３４３號
>
> 主旨：各校應依「國民中小學成績考查辦法」辦理學生成績定期考查，並依教學目標、學科性質、學生能力、及家長配合度妥善規劃評量方式，請查　照。
>
> 說明：各校可依教育專業立場決定學生定期考查之評量方式，不限定第一次定期考查採用非紙筆測驗之評量方式。

對這份公文（劃線爲作者自加）的詮釋，我聽到的有至少三種版本，有人認爲：

「除了第一次非紙筆外，第二、及第三次也可以非紙筆，但至少第一次要非紙筆，其他兩次沒關係。」

也有人認爲：

「只要有一次非紙筆，不限定在第幾次，第一、第二或第三都可以。」

還有人認爲：

「第一次不一定要非紙筆，至於第二或第三因爲沒規定所以可自由選擇，這表示三次都不做任何規定了，也可以紙筆或非紙筆。」

這件事情，給我很大的啓示。由第一份「強制執行非紙筆測驗」的公函，妥第二份「不限定採用非紙筆測驗」的公函的過程中，我體會到任何改革一定要與時代的潮流同時並進，改革運動必須在社會的脈絡中進行才能成功。評量改革的種子已在很多老師的心中等待時機發芽。不過在現場的老師們需要的大概不是一紙命令，或一個手勢，他們需要一些時間來體會時代的脈動，及思索每一動的意義。老師們（也許是校長）對第二份公文做了配合自己需求的詮釋，表示著推動做法的背後就是想法與感覺。因此接下來我想以三個層面來思考評量觀念背後的想法（或說思潮）是什麼，也許最後我們心中會漸漸浮現出答案來，妥談改革，最先要變化的應該是什麼？

首先我討論建構學習論對評量的影響，因爲評量是學習的一部份；接著從後結構思潮對知識形成的理解思索評量如何看待知識，來連接評量如何期待學生的表現；然後是較實務層面地探討效度理論如何涵納評量的變革。

評量與建構學習

個人建構與社會文化的交互

評量既要討論學習的過程，那麼學習理論的發展，必會影響評量的內容與方式。學習理論近年的變化，逐漸走向學習過程的建構觀點，認爲學習包含兩部份：社會文化觀與個人建構觀（Cobb,1944）。我對這兩個觀點融合起來的詮釋是：「在學習的社群中受到社群激發出的想法與實務的啓發後，對這些想法與實務形成個人的意義，並貢獻回社群之中，與社群產生互動，在互動中，個人又從中創造自己的意義，再帶回社群之中，一方面個人不斷地知道（knowing），另一方面，社群在形成共享的知識（knowledge），因此是一個不停的循環。」知道與知識的想法來自Smith（1995）。

我若將這句話放進教室的現場中，就是學生在教室中與同學、老師一起讀書、解決問題、分享心得，並對這一切有了自己的瞭解與收穫，而於班級的討論互動中，互相分享瞭解，可是每一個人對此瞭解賦予自己的意義。建構觀點所強調的學習過程，已然從教師中心移轉到學生中心，教師對學生而言，是促進他學習時社群中的一部份，而不是他學習的主宰。但是有一點是一直爲大家所忽略的，教師不是學習的中心並不意味教師不重要了，相反地，我要強調教師的地位更形重要，正處於津要關卡之處，是學習者與知識之間的仲介者（Driver, et. al., 1994），來幫助學生對於流通在教室之中的知識形成自己的意義，或如Leont'ev（1981）所言（引自Cobb,1944），教師角色的特徵，即是仲介於學生的個人意義與社會上藉文化建立起來的意義之間。在學生形成及轉化意義的過程中，教師是一個從旁激發、支持、並適時介入的人，然而無法替他們結構整個意義化的歷程。因此在建構教學的教室中，教師布題、帶引討論、仲介於題目與學生的思考之間，而不是在黑板上呈現連串的解題技巧、訣竅，來結構學生的理解及意義化的方向。

由上的詮釋可知，社會文化觀者認爲，知識與了解建構於社會（或說教室）中爲了解決共問題與任務時發生的言談活動，因此個人對知識的意義便逐漸在彼此的對話之中形成，而學習就是較無經驗的社會成員，由有經驗成員引介文化的過程。至於個人建構論者，強調教室互動中個人的詮釋活動，關心的重點在學生認知上的自我結構。是故Cobb（1944）認爲這兩個理論各說了故事的一半，當社會文化論者談學習的文化薰陶過程時，前提假設是學生會主動參與並個人建構自己的學習；相對的，建構論者討論學習者的自我認知結構過程時，也事先假想學生是參與於文化的實際活動中，故二者是互補的。我有一個貼切的比喻，就是這兩個學習的觀點，互相假設對方是自己的背景，而自己則立於背景之前演出，觀眾則各聚集在不同的舞台前，擁護他們欣賞的主角。但是，觀眾們要清晰的了解到，主角的精采表演，是因爲有相稱的背景。

社會文化論對評量的影響－－學習在哪裡發生的？

那麼這兩個互爲背景的學習理論，對評量的影響如何？評量實施應脈絡化於教室之中，以教室中的學習爲開端，便反應了社會文化的學習觀。因爲學習是在教室之中發生，是由學生與老師共同策劃進行的過程，因此，只有他們才知道他

們學習的內容及意義何在。這意義，很可能跨班就不同了，遑論跨校或跨市。

舉一個例子如下：

在一個國小二年甲班的教室中，大家正在學二位數的加法，老師布了一個題目，〝媽媽買了兩盒草莓回家，第一盒有草莓19顆，第二盒有草莓17顆，她一共買了草莓多少顆？〞老師選第一、二、三組的2號生上台呈現他們的做法後，與全班同學一起討論這三種做法的意義，老師同時也在座位之間走動。在黑板上的討論大約要結束時，她於學生座位間看到一種算式，於是她特別將這算式寫在黑板上：

$$\begin{array}{r} 10 \\ +\ 10 \\ \hline 20 \\ +\ 9 \\ \hline 29 \\ +\ 7 \\ \hline 36 \end{array}$$ （第一式）

她問同學這樣寫對不對，小朋友起起落落，有說對有說不對，有一個女生很大聲的說：「可以說對也可以說不對。」老師請她上台解釋。她說：「對的是答案，都是36。錯的是算式，她要在第一步的時候把題目的意思就表達出來，然後畫圈圈代表先不要算，用箭頭拿下來等一下算。」她在黑皮上做了修改：

$$\begin{array}{r} 19 \\ +\ 17 \\ \hline 20 \\ +\ 16 \\ \hline 36 \end{array}$$ （第二式）

老師稱許她的想法，全班同學似乎也很滿意，這時剛好下課。下課後，我看看隔壁乙班，也正好上完數學，黑板上還留著類似第一式的算式，顯然上課時是作爲一種方法來呈現，我找了一個時間問乙班老師對這樣算式的處理爲何。她說：「我們班今天第一次上二位數加減，所以對於那樣的直式寫法（指第一式），我不特別提醒對或錯，因爲它是一個很容易表現計算過程的方式，我覺得會幫助他們的瞭解，不過大概下一堂課我會適時提出吧。」

接著，有一天，我到二年丙班的教室去看他們上數學課，也問了丙班的老師對這樣直式算法的感想，丙班老師的回答是，「對於這樣紀錄歷程的算法（指第一式）可否被接受，我沒有任何特別的偏好，不過等到二位數的加減法都結束後，我會適時將這個提出來，讓小朋友漸漸轉化去熟悉較有組織的表達法（指第二式）。」

由以上例子，我們可以看到，各個教室對於第一式的解法在學習歷程的意義，有不同的看法與處理。這就是教師專業自主的呼聲中，老師期待自己扮演的角色。如果我們尊重如此不同的詮釋，我們應該也可以理解課堂中的活動，會延續到評量的多種解釋，在教室與教室之間，會產生不同的評量內容，與同一個答案的不同評價。因此評量的教室意義位於全校共同意義的核心，更是超越了跨校統

一的考卷與答案。統一考試的進行，我認爲是抹煞了老師對課程內容詮釋的活潑性，它應該據有的位階最多到達基本能力的評量。評量回歸到教師自主、教室情境，才有可能與課程結合，此想法多稱爲教室評量（classroom assessment）（Tindal and Marston, 1990; Airasian, 1944）。這是社會文化論對評量的影響，尊重學習發生的情境。

建構論對評量的影響——是誰在學習？

在建構論的影響下，評量走向個人學習的意義，討論一個人如何詮釋他自己的學習歷程。卷宗評量就是一個很好的方法（參見吳毓瑩，民84）。卷宗評量將學生的表現包括一次次的考試、每次的作業、筆記、作品等，依學生的方式或全班共同的方法，放進個人卷宗裡，並加入學生自評、老師評語、同儕互評、家長感想等，以展現個人學習的歷程及意義。我讀到一個數學老師進行卷宗評量的感想（Knight, 1992）。他說：

「有一天，我發給學生一人一個卷宗，然後在黑板上寫下卷宗兩個字。我問他們，〝什麼東西可以放進卷宗裡，什麼東西可以表達你們在數學的學習與努力，哪些活動你覺得最有意義？〞我寫下他們的建議：有筆記、個人預算計劃報告、彩卷計劃報告。最好的考試、最差的考試、每週問題、和家庭作業等。然後我要他們挑選其中的五項，來代表自己的能力與努力。」

這個數學老師第一次進行卷宗評量，他讓學生想想這一學期來自己努力了多少，學到了什麼，並對自己作一個反省與評論。這老師說：

「他們的反省告訴了我他們自己是怎樣的人，他們如何學習數學。我們一起在卷宗之中學習。」

要組成一個卷宗，學生必須選擇適當項目放進卷宗裡，然後描述選擇的理由並評析成果，這些反省皆反映了學生自己的觀點，有誰能比自己更瞭解自己的動機、能力、與感受？因此何不讓他們表達自己的狀況？學習並不是只有認知的工作，情意的接納與啓發，遠比認知還要重要，這才是認知發展的基礎。羅吉斯說：「我們比我們的聰明才智還要有智慧。」（We are wiser than our intellects.）（Rogers, 1980；引自李安德，民77）這句話我尚未讀到原文的上下文，但仍深有所感。

瞭解社會文化論所強調教室在評量中的意義，乃建構論所強調個人意義化過程在學習中的重要之後，我們要問的是，難道以前進行的如是非、選擇、應用、填充、改錯字、連連看、閱讀測驗等，不能達成如上的目標嗎？在教室中，老師仍可以進行全班皆有意義的考試，不和隔壁班一模一樣，有他選題的特色，符合社會文化論的精神，不過倒也不必大費周章讀那些不容易評分的申論題，他可以舉行都是各式紙筆測驗的教室評量。或者，他也可以讓學生選擇自認爲最有代表性的考卷，放入卷宗中，形成從頭到尾都是大考小考測驗卷的卷宗，仍然可以呈現學習歷程，似乎抓到了精神，不過是以非常傳統的紙筆測驗方式行之而已。

然而，評量的精神，除了教室中文化的意義，除了兒童學習的建構過程外，還有一個我認爲是關鍵的思潮，在隱隱主導著評量與學習理論。這思潮可以回答爲什麼上述都是考卷的卷宗，或者發生在教室之中的知識測驗題，仍不太像具有眞實或實作評量的精神。

評量與後結構思潮

評量的範圍，雖然不是以認知爲主，但確定的，他無法避免要評量學生的知識。因而，知識是什麼？評到的是知識嗎？知識經由什麼表現，可讓評量抓到？甚或，什麼才是知識？都是很基本卻又很重要的問題。對知識的定義，很顯然的，會影響評量知識的方式。

結構主義說：知識怎麼來的？

我們先來看看，我們所熟悉的傳統紙筆測驗，背後對知識的定義是什麼？那是植基於結構觀點的系統。如果很難想像，那麼我們來讀一讀一個大陸學者一段話：

「人們所認識的社會現象是雜亂無章的，要達到有秩序的認識，就要掌握現象的結構。那麼，又怎樣掌握、認識現象的結構呢？表層結構是現象的外部連繫，通過人們的感覺就可以認識，而深層結構就是現象的內部連繫，不能通過經驗的概念去獲得它，只有通過理論模式才能認識它……而一但根據具體材料把模式建立了起來，它就脫離了這些材料而獨立地起作用，這也就是說，它（模式）可以是知識的來源而不必求助於現實。」（徐崇溫，民77，頁22）

這一段話，如果還是覺得太理論了，那麼我們來想想，皮亞傑對兒童發展的階段論。皮亞傑認爲，兒童主動建構自己的智力發展，這一點，是前面所述學習建構論的起源，不過建構的過程如何呢？根據石偉平等合譯Gibson（1984）的結構主義與教育一書（民84，頁55）中所說，皮亞傑認爲所有兒童都必須經過一系列共同的智力發展階段，在任一階段的表象上，會出現各種不同的行爲模式，但在這些模式深層存在著共同的結構，可以解釋兒童外顯的行爲。這就是結構主義論。

結構思潮影響了教育對學習的看法，在學校生活中，可見到清清楚楚的痕跡，從上課、下課、午休、打掃、到放學。課程的編排，必有一套很規則的流程與發展方向，深信如此的結構，可以適合全部的小朋友。而評量的日期、範圍也是跟著既定的結構進行，全校統一的考試內容，一樣的試題結構，國語、數學、自然、社會、生倫考卷，閉起眼睛一想，就是那樣的形式。如果還記得以前常舉辦的教師評量設計競賽，其中的第一件基本大事，就是Bloom的雙向細目表。Bloom（1956）對於人的認知複雜層次，有一個我們耳熟能詳的結構，回憶、理解、應用、綜合、評估，雖然也許我們曾經質疑過認知複雜的順序是這樣的嗎？不過，給一個清楚堅定的階層，就是結構思潮的產物。換句話說，以一個紮實的結構，將學習及評量模式化。

爲了尋找這模式，結構主義主張從集體、社會的主體出發，「認爲結構是世界上萬事萬物存在的方式，就是說，他們認爲，世界是由關係所組成，事物不過是這些關係的支撐點。」（徐重溫，民84，頁35）。因此，一個人的能力，必須在它所屬的結構中找到很清楚的定位，例如常模中的百分位數，例如全班的名次。而名次所依據的分數，也是根據一套清楚的公式計算出來，例如是非題每一題2分，簡答題5分，應用題10分，所以不易評分的論述因爲沒有具體貢獻，就自然被忽略了。評量系統是時代的產物，傳統的紙筆測驗，在結構思潮的時勢下，貢

獻仍不可抹滅，具有它當時（或說現在）存在的意義。

從結構到後結構－－知識又可以怎麼來呢？

那麼，在此刻，我們好像已感受到時代浪潮頂端的方向在調整，關心評量的人，已嗅得出評量轉化的內在潛力。對結構主義的再省思，轉化出的後結構主義，力圖〝恢復主觀性、歷史活動、和實踐的問題〞（徐崇溫，民77），就是一個前瞻的方向。因此，知識的後結構觀點，相對於結構觀點的對錯分明，已走向「知識就是人們在他們的言談話語（discourse）之間所創造表達出來的東西。」（Delandshere and Petrosky, 1994）。像這樣對知識概念的轉變，從以前的絕對方式，到現在歡迎各式對話的開放空間，也影響了我們評量知識的方法。而什麼方法可以抓得到知識產生自話語的精神呢？顯然，評量者必須創造「可以導引出話語」的評量環境。所以傳統紙筆測驗爲什麼開始受到質疑，因爲是非題，選擇題，填充題，甚至應用題，要求答案的提供，卻不容許論述的彈性，已經不能反映我們對知識產生方式的理解。後結構觀點試圖打開封閉的結構系統，例如標準答案與計分方式，要求主觀異質的詮釋，來豐富結構的層次。漸漸的，實作評量與卷宗評量便開始受到重視，而他們所強調的，就是闡述、自評、討論、與實踐。

我國唐代以降的科舉制度，其實一直就是服膺知識於論述中創造出來的理念，我們所熟悉的蘇東坡，在應試時，對於時政及社會現象的觀察與討論，鞭辟入理，以致於閱卷的歐陽修以爲是自己的學生曾鞏，爲了避嫌而將之降格爲第二名，這是一個實作評量的例子，套具現代用語，此種題目，稱爲〝弱結構〞（ill-structured）題目（Delandshere and Petrosky, 1994; Linn and Gronlund, 1995, p. 240）。學生可以選擇對自己有意義的方法來完成工作。因爲每一個人的結構、方法、內容各有特色，自己就是這評量工作項目的中心，故而此評量方式也完成了人本教育所主張的〝兒童爲中心〞的教育理念。

舉個例子如何？

從建構論的興起到後結構主義對結構主義的再省思，我們可以在評量之中，看到這兩個思潮影響的痕跡。我們逐漸重視兒童在學習過程中所形成的解釋與意義，強調他們對自己的評析與對學習脈絡的掌握，而同時在心中所形成的想法、理解與感受，亦須藉語言的作用、透過話語來形成知識，因此我們會看到像這樣的題目：

「媽媽收到了一盒蘋果，每個看起來大小差不多，媽媽把每個蘋果都切成六等分放在盤子裡，弟弟每次拿一塊來吃，總共拿了9次，請問弟弟總共吃了幾個蘋果？甲說9個，乙說9/ 6個，丙說1又3/6，丁說1又1/2個，你認爲哪一個人的說法比較合乎題意，爲什麼？」（鄔瑞香與林文生提供，東園國小四年級下學期分數概念評量）

或是這樣的題目：

「爸爸去超市買水果。橘子一個8元，他買了5個。蘋果一個10元，他也買了5個？回家以後，哥哥知道這件事，他認爲是8×5＋10×5=90，共90元；

弟弟覺得應該是

8×5＋10×5=40＋10×5=50×5=250，共250。請問爸爸應該付多少錢？你認

爲哥哥和弟弟誰的講法比較有理？爲什麼？」（鄔瑞香提供，東園國小四年級上學期四則運算數學評量）

也可能是這樣的題目：

「各位同學，現在實驗桌上有四種溶液，有酸性也有鹼性，而且他們的強度也不同。請你利用這些溶液調配出讓廣用試紙不變色的溶液……在實驗進行中，你只需將你做的過程重點式的摘錄下來，等到實驗結束後，會有充分的時間讓你慢慢整理你的實驗記錄。」（桂怡芬提供，力行國小六年級上學期自然評量）

從以上題目，我們可以知道，評量對學生的要求不僅要有能力計算，有能力做實驗，還要能理解過程與步驟的意義，同時藉著話語的呈現，來整理自己體會到的學習，形成自我結構出來的知識。

如果我以上的討論能稍微釐清〝評量爲什麼會朝這個方向發展？〞的疑問，下一個我想討論的問題是〝這樣評量的方式能說服大眾嗎？〞

評量與效度理論的延展

「我們的社會是文憑主義的社會、競爭的社會、考試的社會，您說我們應該如何來教育我們的孩子？……社會的價值觀改變了多少？有多少人眞能從潛意識中徹底拔除文憑主義？社會制度改變了多少？是否可以廢除考試制度？這就是我們孩子要面對的社會！！」（鄭鼎耀，銘傳國小）

評量改革有效嗎？什麼是有效？

面對這樣的質疑，我們會自問也會互相問：「這個——這個新的評量方式有效嗎？」要回答這個問題，我想先討論什麼是有效。如果有人問：「魏氏智力測驗有效嗎？」其實這是一個不能回答的問題，如果魏氏智力測驗測量情緒，則顯然它是很無效的，所以在評量領域中，工具本身不具備有效度或是信度，因爲我們甚至於可以將兒童生活適應量表拿去測驗國語能力而造成無效的結果。

因此，有效性是從人們在評量上的反應中產生而出的，然而評量的反應又無法替自己說話，是靜態的產物，因此評量的有效性我們只能從對反應的解釋上去尋找（Messick, 1989）。是故，解釋反應的人就非常重要，他有可能因爲不明瞭評量設計的目的與所要指涉的能力，而將有效性完全毀滅。一般所熟知的效度概念，來自很強的證據基礎，評量的設計者與解釋者要能根據反應的質與量來詮釋評量背後的構念，這是所謂到建構效度，要有憑有據、忠實反映、並且價值中立，然後評量的工作也似乎到此結束。至於評量結果的詮釋對個人的影響是什麼？可以做什麼樣的應用？會導致什麼樣的社會風氣？常常不在效度的考慮之列。作爲評量的專業工作者，如此狹窄的效度概念，會讓我覺得自己缺乏應該有的社會良心與責任。

效度概念是不是有所突破？

舉一個例子，來自一篇教師評量學生等第研究（Brookhart, 1993）的聯想：如果有一個學生能力表現丁等，可是很認眞、學習動機很強，若老師認爲等第必須忠實反應他學習到的知識與能力，依建構效度證據取向的解釋，會給他丁等的成績。而另一位老師，爲了鼓勵他的兢兢業業態度、又顧及丁等成績可能會打擊他的動機，而將其成績往上調高一等，考慮到等第對學生學習的影響。如此的考

慮，有其意義存在，可是，是不是會造成效度降低呢？Messick於1989提出一個效度概念的漸進矩陣（progressive matrix），認爲效度的概念雖然以概念爲基礎，但這僅止於詮釋的證據面而已，效度應該涵蓋測驗結果的詮釋與使用兩方面，而不僅以證據爲基礎，還應擴展到影響後果的層面，因此這個2×2的矩陣包含兩個向度：詮釋及使用爲一個向度（inpterptretation and use），證據基礎及後果基礎（evidential basis and consequential basis）爲另一個向度，二者交互後便形成四種情況，這四種情況就是效度應該涵蓋的層面。Messick更於1992時，將效度由建構概念一層層推演到社會影響層面的企圖闡釋得更清楚：

表一：效度漸進矩陣的層面

	測驗解釋	測驗使用
證據基礎	建構效度	建構效度＋適切性／使用性
後果基礎	建構效度＋價值意涵	建構效度＋適切性／使用性＋價值意涵＋社會後果

資料來源：Messick, S.(1992). Validity of test interpretation and use. In M. C. Alkin(Ed.), Encyclepedia of educational research, (pp. 1487-1495). New York: Macmillan.

Messick當初提出這概念時，應是基於測驗評量對社會影響應該負起的責任。此時，另類評量或眞實評量尚未蔚爲風潮，而從90年代以降1，新的評量方式與精神的突破，造成以往較狹隘的效度概念，不能發揮他作爲評量效度把關的角色，研究者及倡導者漸漸地也拓展了自己效度的概念，以Messick的矩陣作爲效度的理論基礎，著重評量的後果影響面，一般將此稱之爲後果效度（consequensial basis of validity）（吳毓瑩與桂怡芬，民84；Linn, 1993; Messick 1994; Moss, 1994），意含除了構念效度的證據考量外，評量對學生個人、老師教學、及班級文化、甚或社會習慣的影響，都是評量效度的範圍。

評量效度如何面對質疑？

對於實作評量的期望與效度的原則，Linn(1991)曾爲文討論了8項效度的準則，其中第一項就是評量的影響後果。也就是評量的實施，很明顯地，都會影響學生對時間的安排、對科目的選擇、及自我的看法等，這些其實也都包含在我們教育的目標當中，因此一個有效的評量，不僅能呈現所欲反映的構念，且能產生符合教育方向的影響。他特別提醒大家，「我們不能假設一個比較眞實的評量就一定會引發教室中有建設性的學習活動」。Messick(1994)對於脈絡中的評量，也有他的擔心，他說：「能夠引發某一個學生投入且興致勃勃的評量特色，（在同一個教室中）也許讓另一個學生混淆不清且覺得很陌生，而使得他的表現有偏差及扭曲。」在Messick列舉的建構效度的六個面向中，後果面向不可或缺的也是其中一項。

評量背負著許多教育改革的期望，而評量在教室中的影響，也不可能一一照著我們期望的方向去發展。因此，對於評量效度的後果面向的研究與探討，也是我對評量研究未來發展的期望。效度有了這層考慮，評量的改革在教育中的影響，才有立論的基礎。因爲對表現的詮釋與評析，主要憑藉證據支持，而其傳達的又是社會價值（或說教室價值），效度概念便存在於證據基礎與後果基礎二者的互相影響及演變（interplay）之中（Messick, 1994）。如果我們能將這一層考

慮，放進我們的思考架構中，成爲我們已經熟知的效度概念的延展，則「評量在教室中如何影響教師與學生」的探討，便已涵蓋在效度的範疇之中，因此，一個完整的評量，才會有機會再從效度的基礎回到社會文化觀點的教室脈絡之中，與個人建構觀點的學習過程之中。有了這層思考架構，我們再來讀本節開始處鄭老師的擔心，便不會將之視爲評量改革的反對浪潮，反而他提醒我們，評量必然不然跳脫社會文化的脈絡中，但是它可以如何影響之，這是評量的效度討論必須嚴肅待之的問題。

結語：從情境、仲介、參與、學習、到影響

最後，我要呈現一個老師的教室經驗，並試圖將這個經驗放入這篇文章的脈絡之中，來總結我對評量近時在社會中的蛻變與突破，背後的思考。

我的教室經驗

上三甲自然「水的蒸散」這個單元時，我請學生帶厚度相似的手帕到學校來，請他們在上課時，把手帕弄溼，然後在2*2=4種情況中－－（陽光下或風扇前）*（攤開來或折疊起來）－－觀察討論水蒸散速度的快慢。在等待的過程中，我發現有一位女生拿衛生紙出來，一直想把他折疊起來的手帕沾乾，我走過去向他解釋我們沒有在比賽誰的手帕乾得快，我們只是在看在哪一種情況下水的蒸散最快，同時我也向全班再度說明實驗所要觀察的內容，並以她的例子說明該注意的事項。上完課後，我馬上決定把這個事件，放入這一班的評量之中，因爲再也沒有更活生生的題材，可以讓他們討論及瞭解實驗的意義。（陳素櫻，南勢國小）

首先，我們看到陳老師在實驗課中，佈置了一個情境，這裡我們可以知道，雖然老師不是學習的主宰，但是她是關鍵角色，她仲介知識於學生的學習情境中。學生與老師在教室裡，共同享有此情境，然也各自形成個人的意義。其中有一個女生對於實驗做了自己的詮釋（用衛生紙沾乾手帕），我們看到了以社會文化論爲背景的個人建構意義化過程。而陳老師不僅布題、引介，還參與學生的學習，她說：「我發現有一位女生……」此時陳老師「走過去向她解釋……」，傳達社會上對實驗的共識，陳老師沒有另外描述學生的反應，可能她聽懂了點一點頭，也可能提出她的理由與老師據理力爭。陳老師接著將這一事件帶入全班共同關心的情境中，將兩個人的互動擴展到全班的互動，這時，我們又看到了以個人建構論爲背景的社會文化互動。而這一事件因爲當下發生，故而對這班特別有意義。陳老師將這事件作爲評量的內容之一，將評量與課程、教學結合在一起，形成他們自己教室中的評量。這一「活生生的題材」、有意義的學習，是統一考試所無法取代與發現的，陳老師希望學生能「討論及瞭解……」，說明了後結構中所談的知識創造於話語言談當中。這個評量效度，在結果解釋的證據基礎上，討論的是陳老師如何將結果詮釋印證學生的實驗概念與操作；而在結果解釋的後果基礎上，可分爲兩方面來談：其一是在陳老師對學生表現的詮釋中，包含了她的價值觀，認爲實驗中自變項完整的重要，及干擾自變項的嚴重，並將此價值觀藉由評量傳達給學生；另一方面，陳老師使用這詮釋（隱含了她的價值觀）來描述學生時，給這學生的影響有很多可能方向，也許他們的實驗態度將因此很有趣，也許那個女生從此不敢看陳老師，或者她發現讀書與實驗原來更謹慎、也許他們更

清楚地學得了自變項與依變項的關係、也許他們因此發現實驗很有趣，也許那個女生從此不敢看陳老師，或者她發現讀書與實驗原來不太一樣等等。這些後果就是我們一直於建構效度的證據面外，忽略考慮的角度，可是卻又是影響學生學習與成長最劇的面向。

我深信一件事情、一個措施的技術與方法，固然非常重要，然而若我們不相信這件事，不愛這個措施，又怎可能盡力從各個方向來瞭解它做好它。我提出「評量改革爲什麼要進行？」這一個問題，問自己也問大家，來明瞭她在這社會環境之下，受到什麼樣的哲學思潮的影響，及她未來的趨勢是什麼，而且很重要的，我們要瞭解她不僅是被利用來作爲教育改革的工具而已。

感謝

感謝很多人與我一起讀這篇文章的初稿，甚至慷慨地願意提供他們的經驗與珠璣短語來豐富此文的內涵。他們是江淑萍、吳美華、宗美英、桂怡芬、陳素櫻、鄭鼎耀、及劉淑雯。因爲有了他們皺著眉頭咬文嚼字的讀，才讓此文現在可被讀。

參考文獻

李安德（民77）。邁向心理學的新典範——從羅吉斯談起。**輔導月刊**，24(2, 3),9-13。

吳毓瑩（民84）。開放教室中開放的評量：從學習單與檢核表的省思談卷宗評量。載於國立台北師範學院主編，**開放社會中的教學**（頁93-100）。台北：國立台北師範學院。

吳毓瑩、桂怡芬（民84）。形成性評量效度的驗證及教師的角色——以自然科平時評量爲例。**師資培育理論與實際學術研究會**，台灣師大。

徐崇溫（民77）。**結構主義與後結構主義**。台北：谷風。

莊明貞（民84）。一個新的評量取向——變通性評量在國小開放教室的實施。載於國立台北師範學院主編，**開放社會中的教學**（頁77-92）。台北：國立台北師範學院。

結構主義與教育（民84）。（石偉平、王斌華等譯）。台北：五南。（原著出版年：1984年）

盧美貴。（民83）開放教育的本土化實踐。載於鄧運林主編，現代開放教育（頁33-40）。高雄：復文。

Airasian, P. W. (1994). *Classroom assessment* (2nd. ed.). New York, McGraw-Hill.

Bookhart, S. M. (1993). Teachers grading practices: meaning and valuse, *Journal of Educational Measurement*, 30(2). 123-142.

CizeK, G. J. (1991). Innovation or enervation? performance assessment in perspective. *Phi Delta Kappan*, May, 695-699.

Cobb, P. (1994). Where is the mind? constructivist and sociocultural Perspectives on mathematical development. *Educational Researcher*, 23(7), 13-20.

Delandshere, G., and Petrosky A. R. (1994). Capturing teachers knowledge:

performance assessment. *Educational Researcher*, 23(5), 11-18.
Herman, J. L., Aschbacher, P. R., and Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria. VA: Association for Supervision and Curriculum Development.
Knight, P. (1992). How I use portfolios in mathematics. *Educational Leadership*, May, 71-72.
Leont'ev, A. N. (1981). The problem of activity in psychology. In J. V. Wertsch (Ed.), *The Concept of activity in Soviet psychology*. Armonk, NY: Sharpe.
Linn, R. L.(1995). *Measurement and assessment in teaching* (7th. ed.). New Jersey: Prentice-Hall.
Linn, R. L., Baker, E. L. and Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
Messick, S. (1992). Validity of test interpretation and use. In M. C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., Vol. 4, pp.1487-1495). New York: Macmillan.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. New York: Macmillan.
Moss, P. A. (1994). Can there be validity without reliablity? *Educational Researcher*, 23(2), 5-12.
Smith, E. (1995). Where is the mind? Knowing and knowledge in Cobb's constructivist and sociocultural respectives. *Educational Researcher*, 24(6), 23-24.
Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Macmillan College.
Tindal, G. A., and Marston, D. B. (1990). *Classroom-based assessment: evaluating instrucational outcomes*. Columbus, OH: Merril.
Wiggins, G. (1990). The case for authentic assessment. *ERIC Digest*. Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

註釋

【1】我用眞實評量（authentic assessment）作爲關鍵字查1982-191 ERIC資料庫，在1989以前，沒有發現任何文章直接討論之，而於1989當年有兩篇。1990以降，至9/1995，則有324篇。