# 第二章　文獻探討

## 第一節 教學與評量

### 一、 教學與評量的關係

教學（instruction）包括教（teaching）與學（learning），亦即為如何教（How），如何學的過程，是師生共同參與而產生交互影響的動態過程。「教學相長」一詞說明了教與學相互助長的密切關係。而美國教授 Yamamoto在其所著「教學中的評量(Evaluation in Teaching)一文中，主張評量應顧及教學活動的七W層面，即「為何而教」(Why)、「誰教」(Who)、「教誰(Whom)、「何時教」(When)、「教什麼」(What)、「如何教」(How)、「何處施教」(Where)，可深刻的說明教與學的關係。

教學與評量密不可分（謝祥宏，段曉林，民 89），教學想要有成效，「評量」為一個重要的驗證方法。教學與評量猶如一種互為鏡像（mirror image）的關係，在評量的鏡像中可以反映出教學目標是否達成，在教學實況中則反映出評量的目標。評量應是連續的過程，讓教學者經由學習者對評量的回應中獲得一些訊息，以作為教學上修正的依據。因此評量可改變教與學，評量可說是教育改革的中心。

教學與評量皆是教學歷程的一部分，評量是達成教學目標的重要工具，透過評量的回饋作用，可以促進教學成效，下面藉由美國教育學者 Kibler（1974）提出的「教學的基本模式」來更進一步說明。

### 二、 教學評量的目的

評量的本質應該是以提升學習者的學習為第一要務。評量的標準要秉持公平、公正的原則，這樣的評量才有意義，也才能真正達到與教學相輔相成的效果。而公平的評量要有三個先決條件（謝祥宏，段曉林，民 89）：(1）課程的真實度（2）多樣性和機會（3）價值和倫理。

評量是運用科學方法和技術（簡茂發，民 85），蒐集有關學生學習行為及其成就的正確資料，再根據教學目標，就學生學習表現情形，予以分析、研究和判斷的一系列的工作。在整個教學歷程中，評量是承接轉合的關鍵部分。

美國教育學者 Kibler（1974）提出「教學的基本模式」，如圖 2-1.1，把教學的基本歷程分為教學目標、學前評估、教學活動、評量等四部份，進而闡述四者之間的交互關係，特別強調評量的回饋作用及積極功能其模式如下：

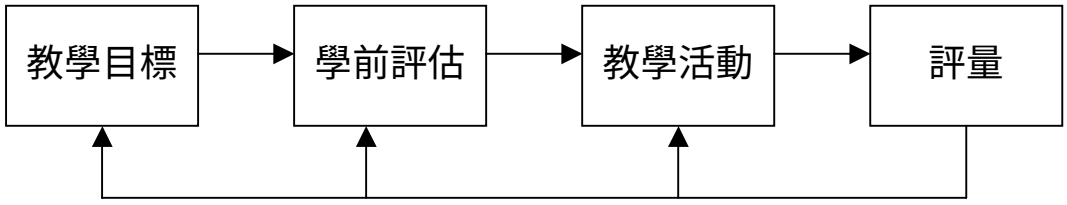


圖 2-1.1 教學基本模式（GMI）

由此基本模式可以了解，評量並非教學基本歷程的最後一站，更不是教學活動的結束，評量應該只是一種方法或工具，而不是最終的目的，評量可作為分析教學得失及評斷學生學習的困難點，以作為實施補救教學和個別輔導的依據，作為下一階段的教學歷程的回饋，以及新的教學活動的開始。

## 三、 教學評量的功能

評量對教師具有下列四項功能（何英奇，民 81）：(1) 了解學生的起點行為－透過預備性評量，教師可以了解學生的起點行為，以作為訂定教學目標的參考（2）建立確實可行的教學目標（3）確定教學目標到達的程度－透過形成性或總結性評量以了解教學目標到達的程度（4）改進教材教法，增進教學效果。

因此教學評量應包括三大部分（簡茂發，民85）：(1) 教師的教學效率的評量

（evaluation of teacher's teaching effectiveness）（2）學生學習成就的評量（evaluation of student's learning achievement）（3）課程的設計與實施的評量（evaluation of curriculum program）。第一部分以教師為評量對象,就教師適性及其教學方法和技術加以評量。其中教學效率的評量包含教學活動設計、教學情境佈置的編選、教法的運用、教學進度的掌握、師生參與活動、作業的規定和批閱訂正、教室管理和常規訓練等項目。第二部分以學生為評量對象,旨在評鑑學生的學習行為和學習結果，宜先了解學生的個別差異,包括身體、智力、性向、人格特質、家庭背景等方面;在學習過程中應注意觀察並記錄學生的學習動機、興趣、態度、方法、習慣、努力情形等；經過一段學習活動以後，再採用科學方法從多方面考查學生的學業成績，分析其優點和缺點，進而診斷其學習困難之所在及原因，據以實施補救教學或個別輔導。第三部分以師生共同參與的課程與教學活動為主，評鑑學校課程計畫與實施之利弊得失，再加以檢討改進，期能有較佳的均衡課程之安排(better-balanced curriculum)，而獲致更為良好的教與學(better teaching & better learning)之效果。亦可以說，教學評量主要的目的在於衡鑑教師教學的效率與學生的學習結果。現在教育學者大多認為成績評量有雙重任務，除了評量學生學習的結果之外，同時也在考驗教師教學的效率，看看它是否符合學生的需求；成績評量應以分析教學得失、診斷學生學習的困難、了解學生學習進展情形為重點，而不要以成績評量之結果，作為判決學生行為而予以獎懲的依據。

一般而言評量在教學過程中具有下列四項主要功能（簡茂發，民85）:(1)了解學生的潛能與學習成就,以判斷其努力程度；(2)了解學生學習的困難，作為補救教學及個別輔導的依據；(3)估量教師教學的效率，作為教師改進教材、教法的參考；(4)獲悉學習進步的情形，可觸發學生學習的動機。

教學評量的目的主要在增加學生學習的效果;是以教育心理學家認為教學評量應

具有五種功能（張景媛，民81）：(1）了解學生的起點行為（2）評定學生學習的結果（3）使教師了解教學的得失（4）診斷學生學習的困難（5）激發學生學習的動機。可以將這些功能的評量類型分類如下：(1）了解學生的起點行為－預備性或安置性評量（2）評定學生學習的結果－形成性或總結性評量（3）使教師了解教學的得失－形成性評量配合診斷性評量（4）診斷學生學習的困難－診斷性評量（5）激發學生學習的動機－預備性評量。

## 四、 教學評量的基本原則

教學的重點是在考查學生的學習成就及衡鑑教師的教學效率，是以在進行教學評量時，必須把握以下十三點原則，方能達其功能（何英奇，民81）：（1）符合原則（2）綜合原則（3）回饋原則（4）歷程與結果兼顧原則（5）個別原則（6）連續原則（7）民主原則（8）合作原則（9）科學原則（10）公平原則（11）保密原則（12）描述原則（13）研究發展原則。

綜合其主要內容如下：（1）評量應根據教育目標進行，避免為升學而做評量。教學前宜先分析教學目標，然後依教學目標進行教學與評量。完整的教育目標兼顧認知、情意、與技能三領域。認知的評量也不應只強調記憶性，而需包合理解、應用、分析、綜合、評鑑等高層目標。（2）教學評量不能只用紙筆測驗，而應依評量的目的選用適當的方法，諸如論文式考試、教師自編測驗、標準化測驗、口試、作業、實作測驗、表演、欣賞、實驗報告、唔談及觀察等方法。（3）在教學過程中宜充分應用安置性、形成性、診斷性、與總結性評量，以提高教學效果。（4）教學評量只是手段，它必須在教學過程中隨時回饋給學生和教師，以作為補救教學的依據。（5）教學評量不只重視月考、期考，平常考也應重視，以便能做立即性回績。此外教師除重視學生答是否正確外，尚需兼顧學生思考的歷程。實驗與技能的評量也不能只重視結果或產品，宜兼顧實

驗與製作的方法及過程。（6）評量需顧及每個學生個別差異。（7）評量為一連續長期的歷程，不但要根據過去狀況，診斷目前的問題原因，更要能預測未來及進行追蹤量評量。（8）評量者應告知受評者評量的方法與標準，受評者有權知道與申訴評量的結果。（9）評量應由教師、學生、家長與有關人員共同合作。學生可以個別和教師訂考查辦法，然後依考查辦法評量。（10）評量應力求：合理性、實證性、客觀性。（11）評量應基於個別差異原則追求「實質的公平」，避免形式上的公平。（12）學生的學業成績應建立隱私權觀念，不宜公開。健全的評量不宜太重視量化，應包括值得描述性評量。（13）實驗與研究脫離不了評量，正確使用評量方法與技術，有助於教育的發展。（14）評量結果的解釋可分為常模參照評量（相對比較）與標準參照評量（絕對比較），二者宜清楚加以區分。例如升學考試屬於常模參照評量，試題應特別重視鑑別力，以便能有效區分考生程度優劣而加以取捨。再如技能檢定考試屬於標準參照評量，這種考試特別重視試題要能涵蓋將來在實際應用時必備的知能及事先要確定該知能的精熟程度，作為及格與否的依據。（15）評量配合電腦使用的發展趨勢，由於電腦的普及，對學校評量有極大的影響，從測驗的實施、計分、建立題庫、試題分析、試題選擇、測驗結果的解釋分析等皆電腦化，達到迅速、正確、簡便的要求，例如在實施精熟學習的班級中，學生的形成性評量結果，利電腦化的S—P表分析(Student-Problem chart Analysis)，可進行診斷教學。（16）診斷性評量的方法與原則，教學評量特別重視回饋原則，尤則當學生學習錯誤或遭遇困難時，極需作深入診斷，以作為補救教學的依據。（17）成績的評定原則評量者宜預先界定評量的項目及標準。

教學在本質上是師生共同參與而交互影響的活動，以學生的「學」為主，以教師的「教」為輔，故評量的重點，除考查學生的學習成就外，也應重視教師的教學效率，是以在進行教學評量時，教育學家簡茂發博士提出，必須把握下列五項「教學評量的基

本原理原則」（簡茂發，民85）：（1）決策原埋：在教學過程中，隨時會遭遇各種教學問題，有必要採取革新措施以增進教與學的效果。決策是評量的理論基礎，為一種有次序且週而復始的連貫性歷程，從各方面蒐集正確可靠的資料，並參照合理而適當的價值標準，以定取捨。（2）回饋原理：評量旨在教學歷程中提供各種必要的「回饋」(feedback)和「引導」(guide)，一方面針對教學上的缺失而檢討改進，另一方面設法突破學習上的障礙，以提高其成就水準。（3）完整原理：評量需要全面性、多元性的綜合資料，並從各個角度和不同觀點加以分析研判，故蒐集的資料愈多、愈齊全，則愈能掌握整體而加以靈活運用。（4）合作原理：評量計畫之擬訂、評量工具之設計與編製、評量之實施及其結果之分析、解釋與應用，均須集結全校師生的力量，始能順利進行，達成教學評量的任務。（5）研究發展原理：評量與研究乃一體之兩面，為求突破目前教學的瓶頸，革新教學措施，必須運用評量的方法和技術、進行教學的實驗研究，使理論與實際相互印證，並開拓教學與評量之研究發展的新途徑。

# 第二節 評量的理論

## 一、評量的意義

在各種教育文獻中，可常看到下列各名詞：測驗(testing)、測量(measurement)、評量(evaluation)、評鑑(assessment)。

(一) 首先，研究者綜合各教育學者的觀點，分析並解釋各名詞如下：

1. 測驗(testing)：(1) 指利用測量工具（instrument）來衡量受試者的身心特質，而測量工具包括測驗、問卷與量表等（何英奇，民 81)。(2) 狹義的觀點認為測驗就是用以測量學生行為樣本(behavior sample)的工具；廣義的觀點則認為測驗是測量行為樣本的系統化程序(systematic procedure)，是對行為樣本所作的一種客觀的和標準化的測量（楊銀興，民 89)。
2. 測量(measurement)：(1) 依特定量尺，對個人的特質，用量化方式加以描述的過程 (何英奇，民 81)。(2) 是對學生的屬性(attribute)、特徵(characteristic)現象(phenomenon)進行觀察，然後依據一定的原則(例如答對多少題)，將所觀察到的資料轉換成數字的一套系統，也就是以數量來描述學生具有某種特質之程度的過程，它不包含質的描述及價值判斷的成份（楊銀興，民 89)。
3. 評量(evaluation)：(1) 所謂評量（國內有時譯為評鑑)，係指依據某項標準，配合其他資料，將測量所得的數量作價值判斷的歷程（何英奇，民 81)。(2) 經由測量而獲致量化（quantitative）的資料，再根據比較可靠的數字性資料，進行精細而深入的分析與研判（簡茂發，民 85)。(3)「評量是獲取資訊，進而行程判斷，並據以做成決定的過程」(TenBrink，1974)。(4) 採用科學的方法與途徑，多方面蒐集通切的事實性資料，再參照合理的評量標準，加以比較分析並做綜合研判。所以評量比測量多了價值判斷的工作，會根據一定的標準，對所蒐集

的事件(event)、物體(object)、或個體(individual)做價值判斷（楊銀興，民89)。

4. 評鑑(assessment)：(1）或譯為評價，視為針對個人問題所作的診斷（何英奇，民 81)。(2)「記述獲取和提供有用的資料以評斷數種方案之效能而作決定的歷程」(D.L.Stufflebeam，1971)(3）包含獲得學生學習訊息的各種方法，例如傳統的紙筆測驗、擴展式反應(extended responses)的題目(如申論題)以及真實情境的表現(如操做實驗)，然後根據所獲得的訊息，對學生學習進步的情形做價值判斷的工作（楊銀興，民 89)。

(二) 其次，針對各教育學者對這四類名詞的異同，綜合分析說明如下：

1. 教育行政與課程專家則慣用「評鑑」一詞，強調「鑑」字的多重涵義，即鑑定、鑑別、鑑賞及視之為一面鏡子而發揮「反映」作用，據以檢討得失，力求革新進步。至於使用「評價」一詞者，從其本質加以闡釋，極力主張它是一種價值判斷的動態過程。對評量、評鑑、評價三個中譯名詞，在涵義上正可互補不足，相得而益彰。綜上所述，評量係採用科學方法與途徑，多方面蒐集適切的事實性資料，再參照合理的衡量標準，加以比較分析與綜合研判的系列過程。吾人宜從整合的觀點，把握評量、評鑑、評價的充分涵義，視之為動態的歷程，而非靜態的實體（簡茂發，民 85）。

2. 測驗只是測量的方式之一。測量除了使用測驗外，尚可使用評定量表、檢核表、問卷、晤談、觀察等工具。評量就是透過測量（如測驗）與非測量（如觀察）等方法，將所蒐集到的資料加以綜合統整，進而進行價值判斷，並據以作成決定的過程（何英奇，民 81)。

3. 評鑑和評量的含義比測量廣且深入，而測驗只是測量的一類，評量和評鑑概念似

乎比較接近。不過以目前正盛行的真切性評量(authentic assessment)、實作評量(performance assessment)而言，強調要學生在各種情境下展現其能力，學者所用的都是「評鑑」，其用意在強調實作時情境的真實性和複雜性，因而評鑑時應採用各種不同的方法(variety of data collection procedures)多方蒐集學生的資料，並據此對學生的學習情形做綜合研判，所以兩者還是加以區分為宜(楊銀興，民 89)。

(三) 最後，從教學評量發展歷史的演進觀點看，測量、評量、評鑑的觀念是依序發展的，從國內學者簡茂發(民 88)探討來說明，其依據「教學評量」的名詞與涵義可分為三個階段：第一階段，屬於美國在 Tyler 的(The eight-year study)時代之前，強調的是「Measurement」(即測量)，強調以量化的方法取得正確可靠的數據。第二階段，學者們將「Measurement」改為「Evaluation」(即評量)學者則認為應該從教育的目標、人格的發展各方面來進行評量，亦即除客觀的數據之外，尚須有一些價值標準來加以衡鑑。第三階段，將「Evaluation」改為「Assessment」，強調評量時應考慮、各種相關的整體情境，從各種可行的途徑，蒐集全面性、多元化的資料，再從各個角度和不同觀點加以比較分析與綜合研判，進行整合性的詮釋，獲致充分的了解。

## 二、教學評量的意義

從國內外學者的觀點，依發表的先後分述說明如下：

1. Stanley 和 Hopkines (簡茂發，民80)：教育歷程分為三大部分，教育目標、學習歷程、成績評量。認為對學生的學習成就評量，必須先以課程標準所訂的教學目標為依據，再配合教學活動，給予適當的測量與評斷。
2. Dececco(簡茂發，民80)：教學歷程分為四大部分，教學目標、起點行為、教學

活動、教學效果評量。先對學生行為狀態加以衡量，再配合教學目標，針對學生需要，提供適合的教學情境，讓學生進行教學活動後所進行的評量工作，再依據評量結果改進教學活動中所用的教材教法。可以說是強調教學效果的評量。

3. 簡茂發(民80)：就教學與評量的關係而言，評量是運用科學的方法和技術，蒐集有關學生學習及其成就的正確資料，再根據教學目標，就學生的學習表現情形，予以分析、研究和評斷的一系列工作。就教學評量的內涵與功能而言，包括對教師教學效率的評量、學生學習成就的評量、課程的設計與實施之評量。

4. 何英奇(民81)：教學評量是依據教學目標，透過測驗、量表、問卷、晤談、觀察等方法與技術，蒐集到完整的量化或質化的資料，採取統整的觀點，對學生的學習結果做價值判斷的歷程。

5. 張春興(民83)：教學評量是指有系統的蒐集學生學習行為的資料，加以分析處理之後，再根據預定之教學目標做價值判斷的歷程。

6. 周文欽(民85)：教學評量是對教師的教學活動與學生的學習活動從事價值判斷。價值判斷的依據則是從教學活動中所蒐集而來的量化資料與質化資料。教學評量的目的，是依據價值判斷後的結果，衡鑑教師的教學效率與學生的學習結果。

7. Airasian(1996)：教學評量是教師將課堂上所蒐集到的種種量的或質的資訊加以選擇、組織並解釋之，以有助於學生做決定或價值判斷的過程，評量可分成「量的描述」與「質的描述」。

8. 楊銀興（民89）：教學評量可視為在教學的活動中，教師採用各種不同的方法，多方面蒐集學生的各種資訊，以獲致量化的及質化的資料，再參照教學的目標，做綜合的價值判斷，以評估學生的學習結果、教師的教學效率以及課程與教材的適切程度。

9. 劉豐銘(民90)：教學評量對教師、學生有回饋與決策的功能。對教師而言，可以了解學生的起點行為、建立確實可行的教學目標、確定教學目標到達的程度以及改進教材教法、增進教學效果；對學生而言，可增進學生瞭解教學目標、激發學生學習動機、協助學生了解自己的能力、性向、潛能與人格質等、診斷學生的學習與適應困難、輔導學生做最佳的升學、就業、生活方式等選擇路。

綜合上述學者們的說法，可知學者們在文字的解釋上或有不同，但看法上差異不大。

## 三、評量的類型

在教學歷程中，教學評量有其適用的時機，及其功能和作用。教學前評量在於了解學生的身心發展在教學前的成熟程度、教學前的基礎和生活經驗背景；教學中評量，藉由隨時觀察和記錄學生在學習各方面的表現，了解學生在學習過程中行為變化的情形，找出學習困難的地方及分析原因，並藉由調整教材教法，給予學習困難的學生必要的學習輔導；教學後評量，在於評鑑學生的學習成就，評定其學習成績及課程內容的精熟程度。

(一) 從教學目標的領域分類：可分為（1）認知領域：指學生心智能力方面，包括知識、理解、應用、分析、綜合、評鑑等六個能力層次。（2）情意領域：指學生的態度、價值、興趣、理想、欣賞等方面。（3）動作技能領域：學生動作技能的學習行為，可分為知覺、心向、模倣、機械反應、複雜反應、適應、創作等七個層次（何英奇，民 81)。

(二) 從教學理論，可將評量依方法分類（簡茂發、李虎雄等著，民 88）為：實地觀察、簡單口頭問答、面談、紙筆測驗、報告、操作評量、專題報告、卷宗評量 (portfolio assesment)（或檔案評量)、進階式評量（Hierarchy assessment)、實作評量

(performance assessment)。茲將評量的方式及其功用綜合如表 2-2.1。

表2-2.1 評量的方式及其功用（簡茂發、李虎雄等著，民88）

| 評量方式 | 特殊作用 |
|---|---|
| 【實地觀察】<br>現場活動實況 | 可看出進行的實況, 只是應儘量避免由觀察所引起的干擾現象 |
| 【簡單口頭問答】<br>現場、事後 | 可針對重點發問, 且可對行為的動機做進一步的了解 |
| 【面談】<br>針對某主題作結構性的深談 | 針對某主題, 經一系列問答得到某人對此一問題的認識、見解等相關的資訊 |
| 【紙筆測驗】<br>可做群測或個別測試 | 對於概念認知及其應用能力有很好的評量功能<br>對於實作及工作流程和態度比較難做有效的評量 |
| 【報告】<br>針對某主題作簡報 | 瞭解撰寫者對整個事件的瞭解程度及對重點的掌控情形 |
| 【操作評量】<br>儀器使用、標本辨識 | 適用於確定的目標 ，非語言文字所能區辨的能力測試 |
| 【專題報告】針對一主題，報告中包括動機、問題闡述、策略、執行、結果、提出解釋、發現新問題等。 | 瞭解撰寫者規劃、執行、研究等各方的能力 |
| 【卷宗】（Portfolio Assessment）<br>（Jongsma K. 1989；King，R. 1995）<br>針對某主題，搜集相關資料作展示 | 可評量出對問題的認識程度, 對變困的重要性認識, 能作有層次有組織的表達能力 |
| 【進階式評量】（Hierarchy Assessment）<br>利用電腦，作滋生式的評測。 | 可對解決問題的動機、理由做詳細的瞭解，便於了解推理的過程 |
| 【實作評量】(MSPAP 1994A)<br>（Performance Assessment）<br>以解決一個小問題為題，一邊處理問題一邊記錄 | 可評測學生概念認知的程度與處理問題能力 |

（三）依評量的功能分類：

1. 何英奇（民 81）：將評量的種類及其功能，整理為表 2-2.2。

2. 簡茂發（民 85）：教學評量可分為「形成性評量」(formative evaluation)和「總結性評量」(summative)。沙克利芬(M.scriven)在 1967 年所發表的「評量方法論」(The Methodology of Evaluation )一文中創用「形成性評量」一詞。前者係在教學過程中，就教師的教學情形與學生的學習表現加以觀察和記錄，通常採用評定量表為工具，進行非正式的評量，後者係在教學活動之末或結束之後，以定期

考試或測驗的方式，考查教師的教學成果與學生的學習成就，通常採用標準化學科測驗及教師自編課堂測驗為工具，進行正式的評量。

表2-2.2 評量的種類及其功能（何英奇，民81）

| 評量種類 | 使用階段 | 評量的功能 |
|---|---|---|
| 安置性評量 | 教學前 | 測量學習前所需的知能及決定學習前已達課程目標的程度。 |
| 形成性評量 | 教學過程中 | 提供學生和教師有關學習進步的回饋，藉以激勵學生並做為補救教學的依據。 |
| 診斷性評量 | 教學過程中，學生的學習有困難時 | 診斷學生學習困難的原因。 |
| 總結性評量 | 教學結束時 | 用來評定等第與證明已熟練教材 |

（四） 依評量的結果解釋分類：

1. 何英奇（民 81）：可分為（1）標準參照評量：指評量的結果，不與團體作比較，而與既定的標準比較，評判是否達到熟練的程度。（2）常模參照評量：指評量的結果，和團體相比較，以評斷個人在團體中的相對位置。

2. 簡茂發（民 85）：分為（1）常模參照評量：係以同年級或其他條件相若的一群學生在某項成績上的分配情形，據以比較分析學生之間相對的優劣。（2）標準參照評量：則以事前認定的絕對性標準為評斷的依據，考驗個別學生的知能是否已達要求的程度，從而判定其成績的及格或不及格。

（五） 其他：教育學者林寶山（民 83），將教學評量分為六大類，如表 2-2.3。

表2-2.3 教學評量的種類及其功能（林寶山，民83）

| 評量種類 | 評量的功能 |
|---|---|
| 安置性評量 | 決定學生在學習前所具備的技能 |
| 診斷性評量 | 鑑定造成學習困難的原因 |
| 形成性評量 | 提供回饋訊息以增強學習 |
| 總結性評量 | 決定學生在期末的學習成就 |
| 標準參照的評量 | 描述學生在某一教學領域中的學習任務 |
| 常模參照的評量 | 描述學生的表現在某些團體中的相對位置 |

## 四、評量的類型的比較

(一) 依功能分類類型（安置、診斷、形成性、總結性評量）比較：

1. 耿筱會（民 88）：對於不同的性質的評量，會有不同階段的評量目的，由建構主義的觀點來看評量，教學目標強調學生如何學習(How to Learn)？透過一連串的學習過程學生可以建構概念、學習知識，利用表 2-2.4 說明不同性質評量間的差異性。

表2-2.4 不同性質評量之差異的比較表(耿筱會,1999)

| | 形成性評量 | 總結性評量 | 診斷性評量 |
|---|---|---|---|
| 評量目的 | 引導學習者學習的方向 | 評量學習者的學習成就 | 診斷學習者的學習情況，引導學習者的自我建構 |
| 評量時機 | 教學過程中 | 教學結束時 | 教學前、教學過程及結束時均可， 依情況而定 |
| 評量模式 | 非正式觀察<br>隨堂測驗<br>小組討論<br>實驗操作<br>家庭作業<br>課堂表現<br>學習活動單 | 正式的筆紙測驗<br>學習總結成果報告 | 個別唔談或施予診斷性評量 |
| 評量功能 | 改進教學 | 評定學習成就 | 改進教學及了解學生建構的過程 |

2．簡茂發（民85）：將評量的類型及其功能與差異比較，整理為表2-2.5。

表2-2.5 診斷性、形成性、總結性評量之比較（簡茂發，民85）

| 比較項目 | 診斷性評量 | 形成性評量 | 總結性評量 |
|---|---|---|---|
| 功 能 | 決定學生的成熟度、預備狀態、起點行為、與學習有關的特質，予以分組安置：診斷學習困難的原因。 | 提供學生進步的回饋資料，指出教學單元結構上的缺陷，以便實施補救教學。 | 在某一教學單元、課程或學期之末，就學生們的學習成就進行評量，決定其成績的等第、及格與否。 |
| 時間 | 教學之初或學習困難之時。 | 教學進行之中。 | 教學之末。 |
| 評量重點 | 認知、情意、技能方面的行為、身心及環境因素。 | 認知方面的行為表現。 | 一般以認知行為為主，但有些科目也涉及技能、情意方面的學習結果。 |
| 工具種類 | 學前的測驗、標準化成就測驗、診斷測驗、教師自編測驗、觀察和檢核表。 | 為教學需要而特別設計的評量工具：評定量表、作業及其共同訂正口頭考問、實際演示、問題研討。 | 期末或教學單元結束時的考試 |
| 行為目標行為樣本之選擇 | 必備的地點行為： 單元目標的要項和教學有關的學生身心特質及環境因素所涉及的行為。 | 教學單元層次結構中所有相關的行為項目：教材細目、學習的動機、態度、方法、習慣等。 | 依教學目標和教材內容的相對重要性而擇定評量項目，使其有適當的比例分配。 |
| 項目難度 | 為診斷必備的知能及基本技能，大部分試題是簡易的，通常P在0.65以上。 | 隨實際情形的需要而定，未能事先決定。 | 大部分試題的P在0.30至0.70之間，容易、很難的試題也有一些。 |
| 計分 | 常模參照和標準參照。 | 標準參照。 | 通常是常模參照，但有時也可能是標準參照。 |
| 通報分數的方式 | 把各方面知能程虔以側面圖呈現之。 | 把個人在教學單元層次結構中各項結目及格與否的組別呈現出來，以便了解個別情形。 | 很據行為目標列出各項分數和總分。 |

3. 何英奇（民 81）：（1）將評量的類型中各種功能、特徵的異同，整理為表 2-2.6。

（2）將教學歷程中的評量之使用時機及其相互關聯性，做成圖 2-2.1。

表2-2.6 診斷性、形成性、總結性評量之比較（何英奇，民81）

| 測驗的類型 | 測驗的功能 | 取樣的考慮 | 題目的特徵 |
|---|---|---|---|
| 安置測驗 | 測量學習前所需之知識和能力<br>決定學習之前，已達到課程目標之程度 | 包括各種學習前所需之起點行為<br>選擇能代表課程目標之樣本 | 通常題目之難度低,是效標參照的<br>題目難度之分散，範圍較廣,是常模參照的 |
| 形成性測驗 | 為學生及教師提供學習進步情形之回饋 | 如果可能則包括所有單元目標或最必要的目標 | 題目配合單元目標的難度,是效標參照的 |
| 診斷性測驗 | 診斷學習困難之原因 | 以共同的學習錯誤為依據 | 通常,試題是容易的,用來指出特殊學習錯誤之原因 |
| 總結性測驗 | 在教學結束時，用來評定等第或證明已熟練教材 | 選擇一種代表課程目標的樣本 | 通常題目難度之分散範圍廣,是常模參照的 |

本表資料採自 P.W.Airasian and G.F.Madaus，Functional Types of Student Evaluation. Measurement and Evaluation in Guidance,4(1972),221-223.。

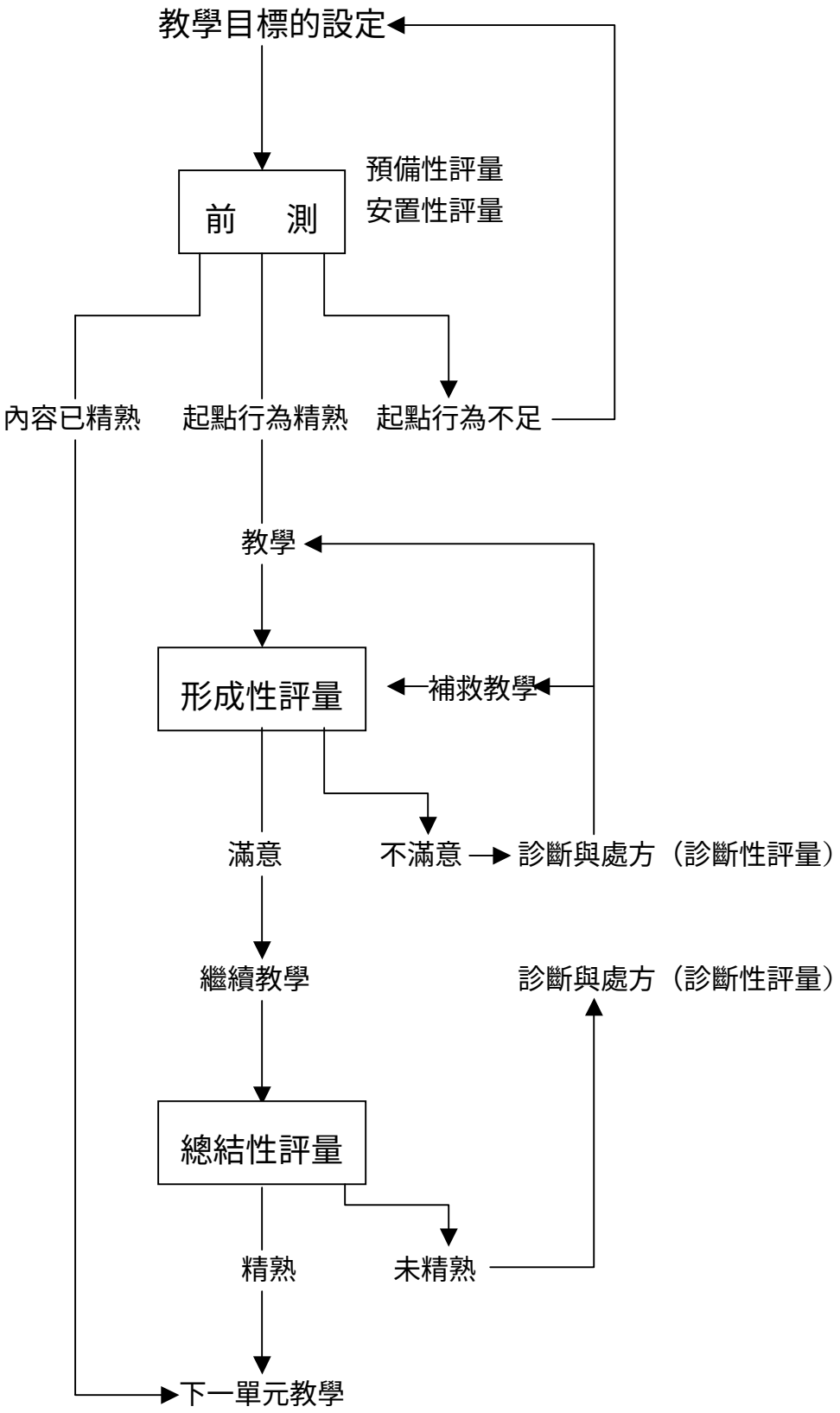


圖 2-2.1　教學過程中的評量（取自 Brown，1981）

(二) 依評量的結果解釋類型（常模參照、標準參照評量）比較：

1.簡茂發（民 85）：將評量依結果解釋的不同，比較整理為表 2-2.7。

表2-2.7 常模參照評量與標準參照評量之比較

| 比較項目 | 常模參照評量 | 標準參照評量 |
|---|---|---|
| 主要目的 | 學習成就的相互比較。 | 特定精熱性的考驗。 |
| 評量內容 | 涵蓋廣泛的成就領域。 | 針對界定的學習項目。 |
| 量尺定準點 | 中間,事後決定。 | 兩端,事前決定。 |
| 參照點性質 | 實際的、相對的。 | 理想的、絕對的。 |
| 評量功能 | 鑑別：:比較團體成員之間的差異情形，找出最有潛能者。 | 檢定:找出超乎某一特定的能力水準以上者。 |
| 數據性質 | 分數的變異性愈大愈好。 | 注重各題反應與效標之間的關聯性 |
| 結果表示法 | 百分等級、標準分數。 | 及格或不及格(滿意或不滿意)。 |
| 記分制 | 常態等第制。 | 傳統百分制。 |
| 主要用途 | 安置：分班編組。 | 診斷：補救教學。 |

2.何英奇（民81）：將評量依結果解釋的不同，比較整理為表2-2.8。

## 五、總結

評量的目的在於了解實況（謝祥宏，段曉林，民 89），但在不同的情境下，該採用適當的評量模式和內容來評量，簡單的說，評量應包括學生做功課時，最有價值、有意義的事項。一個具有能由觀測中獲得全相的、真實的反映出實況功能的評量，可稱為「全真評量」（或稱真確評量(authentic assessment)），可含括晤談、小組解題、歷程檔案、概念圖、實作評量、開放性問題等。可說是各種「評量的方式」努力追求的標竿。評量的目的亦在於改進教學，而教學的目的在促進有意義的學習，要診斷學習是否發生，就必須有具體的指標供教學與評量之用，即「評量的方式」；美國新評量的指標與進行的方式－「科學實作評量」，其具體的做法可作為國內進行評量指標之參考（邱美虹，湯偉君，民89）。

表2-2.8 常模參照評量與標準參照評量之比較

| | 常模參照評量 | 標準參照評量 |
|---|---|---|
| 主要的用途 | 綜合性的測量 | 精熟性的測量 |
| 主要的重點 | 測量成就的個別差異、鑑別 | 敘述學生能做的工作、檢定 |
| 結果的解釋 | 和別人的成就表現比較 | 和具體明確的成就領域比較 |
| 涵蓋的內容 | 涵蓋廣大的成就領域 | 集中在有限的學習內容 |
| 測驗計畫的性質 | 使用雙向細目表 | 使用詳細的教材領域細目表 |
| 題目選擇的方法 | 選擇最能區分個別差異的題目（分數的變異性最大）。刪除容易的試題。 | 包含所有能適當敘述表現的題目，不試圖改變題目的難度，或刪除容易題目，來提高分數的變異性 |
| 量尺定準點 | 中間，事後決定 | 兩端，事前決定 |
| 參照點性質 | 實際的、相對的 | 理想的、絕對的 |
| 數據性質 | 分類的變異性愈大愈好（亦即鑑別力愈高愈好） | 注重各題反應與效標間之關聯性 |
| 結果表示法 | 百分等級、標準分數 | 及格或不及格（滿意或不滿意） |
| 成就的標準 | 依據在團體中的相對地位來決定成就水準（如在20人中第5等級） | 依據絕對的標準來決定成就的水準（如能界定90%的專有術語代表學習精熟） |
| 主要用途 | 安置性評量（分班編組）、總結性評畫、升學、考試 | 預備性、形成性與診斷性評量（補救教學）、技能檢定 |

一般學校的段考、期考可算是一種總結性評量；大學入學考試則為一種升學考試，屬於一種「甄選用」的總結性評量。總結性評量的目的，在評定學習成就的高低及其在團體中所佔的相對位置，所以「題目難易分配」的講求及「評量技術」的應用就非常重要。在評定學習成就的高低方面，就大學考試的統計資料顯示：欲鑑別考生程度，測驗題的設計應以「中等偏易」時最易達成目標。在評定其在團體中所佔的相對位置時，評量技術中的「常模參照評量」與「標準參照評量」也是需要了解的。

就目前二階段的大學入學考試：「指定科目考試」即偏向「標準參照評量」的應用；「學科能力測驗」則偏向「常模參照評量」。但物理考科為一門思考性的學科，在入學考試中命題常偏難，題目難度較高常被定位為「投資報酬率低」的學科，整體分數偏低（例如，91 年指定科目考試，高標為 30 分，均標為 17 分，低標為 5 分），除扼殺學生的學習興趣外，也造成在選才的比重上吃虧很多，與加重計分原意背道而馳，是否在「指定科目考試」中採用「標準參照評量」則值得研究。

本研究即希望藉由相關文獻的分析與蒐集，探討高中物理實驗評量的方式，並分析紙筆測驗題型的不同，研究測驗對『評量學生的學習成就與能力指標的檢定』的優缺點，藉由測驗的結果與檢定方法的討論，提供高中教師處理教學評量時的參考。

# 第三節 評量的技術

## 一、評量應具備的條件

測驗的品質良好與否，可經由質的分析與量的分析。教育測驗可提供有關實際教育決定所需的資料。測驗學者們認為良好的測驗應該具有下列幾個主要的特徵（盧欽銘、范德鑫，民 81）：(1）能提供「正確決定的資料」；(2）所蒐集到的資料具有一致性或穩定性；(3）所需的資料容易蒐集。

一個良好的測驗應具備的品質（陳英豪、吳裕益，民 87）：(1）適切性－問題類型，適切於學習結果（2）平衡性－試題比率與細目表一致（3）有效性－試題編製、評分及作答時間要經濟（4）客觀性－試題要明晰，若為該領域專家應得滿分（5）特殊性－要正確解答試題一定要具備該領域的特殊知識（6）適當的難度（7）良好的鑑別度（8）信度－測驗的結果要有一致性或可靠性（9）效度－所編製或選用的工具要能真正測得所預測量之特質或學習結果（10）公平性－對每一位受試者有良好和相等的機會（11）非速度性－試題數量要配合作答時間。

測驗亦須具備測量某一特質前後能相互一致的特性，且能提供正確決定的資料，以及容易蒐集的特性。心理測驗學者郭玉生（民 76）認為這特質為：(1）信度（reliablity）－結果的一致性（2）效度（validity）－判斷的適切性（3）實用性（employability）－收集資料的容易性。另外 Kline（1986）認為良好的心理測驗應為等距量尺，應具有鑑別力。

## 二、評量的方法

不同的評量方法所適用之評量工具亦不同。就教育目標來說，可分為認知、情意與技能三種評量；就解釋方法來看，可有常模參照評量與標準參照評量；就教學歷程來分，可為安置性(或預備性)、形成性、診斷性與總結性評量；以評量工具之編製來說，

可有標準化測驗與教師自編測驗；以成績評定（何英奇，民81）來分，可大概分為相對比較、絕對比較與自我比較。

（一） 美國教育學者瓦特金斯(R.K.Watkins)將學校成績評量方法歸納為下列九種：（1）教師評判（2）口頭述誦（3）論文口試（4）標準化客觀測驗（5）教師自編測驗（6）學生作品評定（7）操作的評定（8）非正式記述的評量（9）機械紀錄。

（二） 日本學者艉田叡一則將教學評量方法分為七種：（1）標準化測驗（2）教師編製測驗（3）問卷法（4）口頭問答法（5）觀察紀錄法（6）報告（Report）法（7）學生作品評定法。

日本教育評量學者橋本重治，把教師自編測驗與觀察技術分為七大類：（1）論文式測驗法（2）客觀式測驗法：又分為單純回憶法、填充法、序列法、改正法、是非法、選擇法、配合法、選擇填充法（3）問題情境測驗法（4）問卷法（5）猜是誰技術（guess who technique）（6）觀察評定法：可分為檢核表法、評定量表法（7）其他教育工學的技術。

（三）橋本並進一步把教師自編測驗與觀察技術使用在不同教育目標上的適當性作詳細整理，如表 2-3.1。

## 三、評量工具的特徵

一個良好的測驗所需的評量工具，需具備信度與效度；而測驗的特徵依測驗的類型：安置、形成性、診斷、總結性測驗，因類型不同其功能與目的也不同，其對教學目標所要測量的學習結果亦不同。

1．測驗所需的信度與效度（何英奇，民81）：

（1） 信度：指一個測驗所測到之特質的穩定性和可靠性。其類別有：

(A) 重測信度－一個測驗所得的特質，隨著時間保持穩定的程度。

(B) 複本信度－一個測驗之兩個複本相等的程度。

(C) 內部一致性信度－有折半信度、庫李信度、α係數等，用以了解一個測驗內容取樣是否適當或內容是否同質性。

(D) 評分者信度－不同評分者所評閱的分數間一致性的程度。

表 2-3.1 教師自編測驗的測量目標與評量技術

| 測量目標 | | 適當的評量技術 |
|---|---|---|
| 理　　解 | | 論文式測驗,客觀測驗(特別是選擇法、配合法、選擇配合法、填充法) |
| 知　　識 | | 論文式測驗與各種客觀測驗,如單純回憶法、選擇法、配合法、選擇配合法、是非法、序列法、改正法等 |
| 技能 | 讀、寫、算 | 各種客觀測驗 |
| | 會話、討論等的社會技能 | 檢核表、猜是誰技術 |
| | 圖表、圖書、機械器具等的使用技能 | 檢核表、評定量表、各種客觀測驗 |
| | 構圖、描畫、工作、裁縫、實驗、運動的技能 | 檢核表、評定量表、各種客觀測驗 |
| 思考力 | | 問題情境測驗〈論文式和客觀的測驗〉 |
| 圖畫、工作等的作品<br>音樂、體操等的表現 | | 檢核表、評定量表 |
| 態度 | 價值觀、意見 | 問卷、論文式測驗 |
| 興趣 | 興趣 | 問卷、猜是誰技術 |
| 習慣 | 學習習慣,健康習慣 | 檢核表、問卷、猜是誰技術 |
| 鑑賞 | | 檢核表、評定法、問卷 |

(取自橋本 ,1981)

(2) 效度：指一個評量工具能夠測量到它所想要測量之特質的程度。其類別有：

(A) 內容效度－指測驗內容的代表性或取樣的適切性。

(B) 效標關聯效度－指測驗分數與一些外在效標間之相關的程度。包括（a）

同時效度：估計個人在效標方面的目前實際表現的程度。（b）預測效度：指測驗分數預測個人在效標方面未來表現的程度。

（C） 構念效度－指測驗能夠測量到理論上的構念或特質的程度。

不同評量工具所要求的信度與效度並不同。例如在信度方面，智力與性向測驗特重『重測信度』，即測驗的穩定程度；而情意與技能目標之觀察評定則特別重視『評分者信度』要可靠。在效度方面，成就測驗特別重視『內容效度』，即試題有無代表性，本研究即特別重視內容效度；而性向測驗則特重『預測效度』。

2．測驗所要測量的目標：

一個測驗所測量的學習結果，應能忠實反映教學目標，所以測驗編製時首要確定為測量的教學目標。依據學者布魯姆等人所著『教育目標的分類』一書（Taxonomy of Education Objectivies）（見 Bloom et al.，1956；Krathwohl et al.，1964；Harrow，1972）。教育目標分為三大領域：認知、情意、動作技能。而學科知識方面的學習測量屬『認知領域』。

在認知領域中智育的學習結果可分為兩大類（陳英豪、吳裕益，民 87）：（1）知識（2）心智的能力與技巧。而布魯姆的認知目標分類為六個層次：（1）知識（2）理解（3）應用（4）分析（5）綜合（6）評鑑。

上述六個層次，茲解釋如後（何英奇，民 81）：（1）知識—能記憶名詞、事實、規則、原理等（2）理解—能把握所學過的知識或概念的意義（3）應用—將所學到的概念、方法、原理原則等應用到新的情境（4）分析—將所學到的概念或原則，分析為各個構成部分，或找出各部分之間的相互關係（5）綜合—將所學到的片斷概念、知識、原理原則等綜合成新的整體（6）評鑑—能依據內部證據或外在效標作價值判斷。

測驗編製時，必須依學習結果測量目標的不同，擬定所要測量認知層次的重要領域，編製一份測驗『雙向細目表』（two-way specification table）。

雙向細目表可說是測驗編製的藍圖，根據雙向細目表的計畫來編擬試題，使測驗能充分連結『教學』與『評量』，完全達到測量學習結果的教學目標。

# 第四節 實驗的評量

## 一、科學實驗的教學目標

由文獻中（許榮富、趙金祁，民 76），整理出幾個『實驗教學的目標』：

(一) Shulman & Tamir(1973)將實驗室的教學目標歸類為五類：

(1) 技能：操作性的(manipulative)、探究性的(inquiry)、探討的(investigative)、組織的（organizational）及溝通（communicative）等技能。

(2) 概念：如假說、理論模型、分類（Taxonomy category）等。

(3) 認知能力：批判思考、解決問題、應用、分析、綜合、評鑑、決策、創造力等。

(4) 理解科學的本質：科學的領域、科學家及科學家的工作、科學方法的多樣性、各種學科中的科學及技術的相關。

(5) 態度：如：好奇心、興趣、冒險、客觀性、精確性、成就感、負責、一致性、合作及喜愛科學等。

(二) Klopfer(1971)認為實驗教學的目標為：

(1) 知識與理解。(2) 操作技能。(3) 科學探究技能：(a) 觀察與測量 (b) 解釋數據 (c) 辨認問題 (d) 尋求解決問題的方法。(4) 鑑賞科學方法。(5) 科學態度與興趣。(6) 應用科學知識及方法。

Klopfer 進一步指出：對於中等學校而言，應強調 1、4、5 三項。對於應用科學科系之學生則應強調 1、4、6 三項；對科學及工程學系之學生則要著重在 1、2、3、5 四項。

(三) Pella (1961) 認為實驗室教學目標為：

(1) 瞭解科學課程的內容。(2) 發展科學態度。(3) 學習科學的方法。(4) 發展良好的社會態度。(5) 刺激對科學的興趣。(6) 學習如何應用科學原理。(7) 發

展科學知識成長及發展鑑賞的能力。

（四）SASSP 公報（1953）則認為實驗室教學目標為：

（1）發展批判思考及解決問題的技能。（2）學習科學觀察。（3）對學生啟蒙、想像力及互助合作能力的發展。（4）洞悉科學家的工作及實驗室在人類文明中所扮演的角色。（5）增進基本原理、概念、科學事實的瞭解。（6）增加技能的熟習：如紀錄、組織及分析資料、讀取數據、操作儀器等。（7）發展對科學原理及過程的態度。

（五）Sund & Trowbridge(1967)認為實驗室的技能分為五項：

（1）可學習而獲得的能力：（a）聽講（b）觀察（c）蒐集資料（d）探討（e）探究的能力（f）蒐集數據（g）研究。

（2）組織的能力：（a）紀錄（b）比較（c）對比（d）分類（e）組織（f）描述（g）回顧。

（3）創造的能力：（a）事先計畫（b）設計新問題、方法設計或系統（c）發明（d）綜合。

（4）操作技能：（a）使用儀器（b）照顧儀器（c）示範（d）實驗（e）維護（f）建構（g）定刻度。

（5）溝通能力：（a）發問問題（b）討論（c）解釋（d）報告（e）寫作（f）評論（g）繪圖（h）教導同學。

（六）Bringman(1969)認為實驗室的技能為：

學生必須要有三種技能：（1）問問題（2）做觀察（3）組織觀察。他列了六個探究的一般因素：（1）形成問題（2）形成假設（3）設計研究（4）執行探討計畫（5）解釋數據或研究發現（6）綜合探討而得的知識。

（七）Nedelsky(1965)在＂Science Teaching and Testing＂中認為實驗室的教學目標

有三個層次：L1、L2、L3。

(1) L1 是指有關實驗室的知識，可分為三部分 (a) L1.1：儀器或材料的知識 (b) L1.2：實驗室步驟的知識（c）L1.3：數據及數據之類化之間的關係的知識。

(2) L2 是指實驗的理解（Laboratory understanding），可分為兩個部分（a）L2.1：理解測量的過程（b）L2.2：理解實驗，即理解自然現象與理論（或書本）間的關係。

（a）L2.1，再分為二個次子層（Ⅰ）L2.11：儀器部分（Ⅱ）L2.12：測量部分。

（b）L2.2，再分為三個次子層（Ⅰ）L2.21：實驗設計（Ⅱ）L2.22：實驗過程（III）L2.23：實驗數據的解釋。

(3) L3 是指從觀察或實驗中學習的能力，可分為四部分：（a）L3.1：實驗探究的能力（b）L3.2：實驗技能的能力（c）L3.3：學術思考的能力（d）L3.4：在實驗室中想像的能力。

（八）Ganiel & Hofstein(1981)歸納實驗室的成就為五個成分：

(1) 組合實驗是儀器及其他操作技能 (2) 觀察及測量 (3) 設計實驗步驟 (4) 處理數據 (5) 下結論及批判思考。

## 二、實驗評量的理論

確認了教學目標後，如何瞭解教學後的學習成果即評量教學成果，就相當重要。以下整理幾個有關實驗評量的理論：

（一）物理實驗的評量（簡茂發、李虎雄等著，民 88），可有：實地觀察、簡單口頭問答、面談、紙筆測驗、報告、操作評量、專題報告、卷宗評量（portfolio assessment）(或檔案評量)、進階式評量（Hierarchy assessment)、實作評量（performance assessment)。

（二）實作評量（簡茂發、李虎雄等著，民 88），係一種針對自然科學的學習，相當適切的、也頗接近「全真評量」的評量方式；這種評量是用來測驗學生，應用知識和技能於真實情境中的能力，企圖藉由學生處於真實情境中主動解決問題，以呈現學生的真實能力。以科學教育目標來看藉由實作評量能了解學生高層次思考能力，這種評量才是不偏廢的全方位評量。

（三）實驗評量（Laboratory Examination）、操作評量（Manipulating skill testing、Practicing skill test）（簡茂發、李虎雄等著，民 88），經常運用於自然科學的學習中，其內容比較注重在執行能力以及資料的整理、分析、歸納研判方面，尤其操作評量其著重點更集中在執行、研判方面，它是實作評量的局部，在特別強調觀測某一能力時適用。

（四）廣義的實驗（楊文金、許榮富，民 76），就是完整的實驗，它應包括『做』實驗前的認識問題、辨認變因、形成假說、設計實驗等心智過程，及實驗進行中之觀察、測驗、紀錄等實驗操作技能，還有實驗後之資料處理、下結論、推衍等能力。

（五）實驗室教學成果之評鑑（許榮富、趙金祁，民 76）可分為四種：（1）紙筆測驗（2）實作測驗（3）評量表（checklist or rating scale）（4）書寫報告評量。

## 三、實驗的評量技術

由上述對評量理論的瞭解，接著探討評量的技術。

實驗評量的實施可分為三個步驟進行：實驗前、實驗中、實驗後。本研究教學評量的範圍為屬總結性或成就性評量，即實驗後的評量，測驗的目標為『學生的科學能力與物理實驗技能』，其評量的方法與技術，至少包含紙筆測驗、寫實驗報告、實作評量等。

### （一）評量的方法

科學實驗較易使用的評量方法：（1）實地觀察（2）紙筆測驗（3）操作評量（4）

專題報告（5）卷宗評量（6）實作評量。

科學實驗較難採用的評量方法：(1）進階式評量（2）簡單口頭問答（3）面談（4）報告。

Lunetta(1981)對下列四種評量提出評論（許榮富、趙金祁，民 76）：

(1）寫報告：無法提供學生在操作儀器、觀察、組織或執行探討活動等技能的直接訊息。

(2）紙筆測驗：這種測驗方式對實驗活動之執行階段的評量是不完全的。如果全部用紙筆測驗做為評鑑基礎，則學生可以根本不會操作或執行實驗。而依然可有好的表現。

(3）實作測驗：這種評量方法雖然可以測量學生實際執行實驗的能力，但一般說來，實作測驗只能評量低層次的操作及觀察技能而已。

(4）觀察評量（observational assessment）：這種評量方式在七十年代末期被發展出來，它先要有依完整的評量標準（assess criterion）及評量表（checklist），以長時間的觀察來評定學生之實驗活動的各種目標的成果。

總結，觀察評量、實作評量雖較能測出學生的學習成就與技能，唯認定標準較難，技術複雜執行困難，常為教師所忽略，『大型的入學考試』亦未採用，國內的入學考試則僅採用『紙筆測驗』。在尚未出現替代的評量方法前，雖然採用紙筆測驗，有其評量上的盲點，但仍可就現行的方式上尋求改善，以期達到評量的目標。

## (二）評量準則與評量工具

1．許榮富、趙金祁（民 76）於國科會的研究報告中，提出幾個實驗的評量標準與評量工具：

(1）CITO（1982）發展出一個稱為『客觀的學生物理實驗室連續評量』工具。此評

量工具以實作做為測驗的主體配合評量標準（assessment criterion）實施。

(2) Krusalk（1955、1958）分別設計適用於大學階段的物理實驗測驗工具，1954 的實作測驗，1958 則為複選的紙筆測驗。

(3) Lunetta (1978) 發展一套專門評鑑科學實驗手冊之評鑑工具 LAI (The laboratory Structure and Task Analysis Inventory)，並曾對 PSSC、HPP、BCCS 等教材進行分析。1981 年則發展一套實驗的觀察評量標準。

(4) Shymansky(1979)設計一套名為 SLIC（Science Laboratory Interaction Categories）的評量工具。此評量工具是專用於教室環境（Classroom environment）之評鑑而設計。

(5) Walberg（1968）發展學習環境的評量工具 LEI（Learning Environment Inventory)。

(6) Rentoul & Fraser（1978）發展 ICEQ（Individualize Classroom environment Questionaire)。

(7) Maynes(1986)提出以變因作為評量學生在形成假設、設計實驗、資料處理及下結論等過程技能之階層性（hierachical）的選擇格式測驗的評量準則。該評量準則掌握了以變因的認識能力，來評鑑學生的能力層次。這種評量準則在選擇格式的測驗中，只要在選擇題之選項上滿足其標準，則除了該標準的正確性外，是一種客觀的評量方法。因為在選擇格式的測驗中，選項及題幹都是高度結構的，對學生而言，所面臨的情境是固定的，而且該標準已然假設學生的能力層次的分佈範圍及反應型態，而學生在固定的選項中，選擇最適當的選項，基本上這是一種辨認或比較的能力。

2. 許榮富、趙金祁 (民 76)，發展一套 SLSI (Science Laboratory Structure Inventory)

科學實驗評量量表，並針對 (1) 形成假說 (hypothesising) (2) 設計實驗 (designing an experimental investigation) (3) 資料處理 (data processing) (4) 下結論 (concluding) 等四項科學實驗過程統整技能，發展出選擇格式、問答格式、實作格式之示範、設計、引導實驗等五種評量模式工具。

3. 王澄霞、洪志明（民 81）於研究報告中，提出實驗技能的評量工具應有的標準與信度：(1) 評分者間信度－評分者由使用實作觀察量表，觀看學生之實驗技能錄影帶，並用討論的方式，建立一致的評分標準，而確立評分者間的信度。(2) 用 S-P 表 (student-problem score table) 分析－由前測與後側的 S-P 表分析結果，如果整個 S 曲線持續地向右移動，顯示評量工具對學習進步的一慣性和均勻性，則此評量工具有信度。(3) 兩對等形式 (two equivalent forms) 測驗結果比較－由使用兩對等形式的實驗技能紙筆測驗，其所得的成績的一致性表示出評量工具的信度。

根據以上的研究，如 Maynes(1986)選擇格式測驗的評量準則，許榮富、趙金祁 (民 76) 發展的 SLSI 科學實驗評量量表暨紙筆測驗準則，建立一份優良的物理實驗客觀式的紙筆測驗試題相當可行。

是以本研究僅針對『紙筆測驗』進行探討，分析近年來『入學考試』中的物理實驗試題，大多採用『選擇題型』與『問答題型』的測驗試題。本研究希望藉由『題型的探討』，比較分析各種題型的優缺點，以期提供教育單位及教師教學上參考。

## 四、評量的選用：

本研究專注於『物理實驗的成就評量』，在國內『入學考試』仍維持紙筆測驗的情況下，是以本研究僅探討『實驗評量中的紙筆測驗』。

測驗的選用，必須考量測驗的題型，不同的試題類型有不同的測驗功能及優缺點，皆是編製測驗時所需的評估。

（一）紙筆測驗中的題型分析

試題的類型，可分為兩大類（余民寧，民 86）：(1) 選擇型試題（selection-type items）(或慣稱為客觀測驗（objective test)）：如選擇題、是非題、配合題、填充題、解釋性習題（interpretive exercise items）(2) 補充型試題（supply-type items）（或慣稱為論文測驗（essay test)）：如簡答題、限制性反應題（restricted response essay questions)、申論題。

也有將簡答題歸類為客觀式測驗者（陳英豪、吳裕益，民 87)。廣義上，填充題、解釋性習題、簡答題等皆可說為限制性反應題。無論如何分類，客觀式測驗、限制性反應題大多僅能測量『知識』、『理解』、『應用』、『分析』等層次的學習結果，而『擴展反應式問題』（extended-response questions）的論文式試題，則能評量較高層次如『綜合』、『評鑑』等的學習結果。

依據布魯姆認知領域中（陳英豪、吳裕益，民 87)，六個層次的教學目標分類中較複雜的綜合、評鑑層次，難以客觀式的測驗測量，論文式的測驗較適合測量此類高層次的學習結果；而客觀式的試題最適用於『知識』層次的測量。

而客觀式與論文式測驗的優缺點比較（郭玉生，民 81)，如表 2-4.1。

綜合上述各家的分析，研究者將『入學考試中的物理考科命題題型』的優缺點，整理得表 2-4.2。

（二）大型考試的評量選用

國外常見的大型考試：TOEFL（托福)、GRE、SAT 等測驗常採用選擇題型的紙筆測驗，在合乎測驗的評量準則下，因為在選擇格式的測驗中，選項及題幹都是具有高度結構的，並加以嚴格的測驗信度與效度考驗，則基本上皆能達成測驗學生能力的目的。

表 2-4.1 客觀測驗和論文測驗比較

| | 客觀測驗 | 論文式測驗 |
|---|---|---|
| 測量的能力 | 適合測量『知識』、『理解』、『應用』和『分析』層次的學習結果；但不適合測量『綜合』和『評鑑』兩種學習結果。 | 不適合測量『知識』層次的學習結果；適合測量『理解』、『應用』和『分析』等層次的學習結果；最適合測量『綜合』和『評鑑』兩種學習結果。 |
| 內容的取樣 | 採用大量之試題，涵蓋的範圍較大，內容的取樣有代表性。 | 使用相當少的試題，涵蓋的範圍小，內容的取樣較沒有代表性。 |
| 試題之準備 | 準備優良的試題很難且費時。 | 準備優良的試題也很困難，但比客觀試題容易。 |
| 記分 | 客觀、簡單並且相當可靠。 | 主觀、困難且較不可靠。 |
| 影響分數的因素 | 閱讀能力和猜測。 | 寫作能力和吹噓的功夫。 |
| 對學習的影響 | 促進學生記憶、解釋和分析他人的觀念。 | 促進學生認識、統整和表達自己的觀念。 |

（採自 Gronlund，1982，p.73）

表 2-4.2 物理科入學考試題型的優缺點比較表

| | 優點 | 限制 |
|---|---|---|
| 選擇題 | 1. 計分容易，受猜測的影響較是非題低，信度較高。<br>2. 選項及題幹具有高度的結構，增加測驗信度與效度考驗。<br>3. 可廣泛測量各層次的認知能力目標。<br>4. 可以透過精心設計的誘答選項，提供有價值的診斷訊息。 | 1. 編製出具有良好誘答力的誘答選項不容易，當命題技巧不良時，容易編製出僅測量低層次認知能力（如記憶）的試題。<br>2. 無法測量問題解決、組織、與表達能力。<br>3. 容易引發猜題與作弊行為，間接增加能力估計的困難。<br>4. 較不適合測量『綜合』和『評鑑』兩種學習結果。 |
| 填充與簡答題 | 1. 適合測量『知識』、『理解』、『應用』和『分析』層次的學習結果。<br>2. 命題較簡便、容易。<br>3. 不易引發猜答行為。<br>4. 可以用來測量數學或自然科學之問題解決能力。 | 1. 需採人工計分，計分不客觀。<br>2. 作答及閱卷相當費時。<br>3. 無法測量複雜的學習結果。<br>4. 較不適合測量『綜合』和『評鑑』兩種學習結果。 |
| 問答題 | 1. 適合測量較高層次的學習結果，如綜合、評鑑。<br>2. 可以測量學生的組織、統整、歸納、問題解決、表達觀點的能力。<br>3. 命題較簡便、容易。<br>4. 不易引發猜答行為。 | 1. 試題取樣不具代表性，無法涵蓋教材內容的全部，造成測驗內容效度降低。<br>2. 評分不容易，易受評分者主觀影響。<br>3. 評分易受寫作能力與作答技巧影響。<br>4. 作答及閱卷相當費時。 |

反觀國內大學入學考試：未有專業研究建立各科測驗的評量準則，亦未有專業機構對測驗的信度與效度執行考核，考量試務龐雜且規模龐大，並因時效、環境上的限制，近年來皆採用紙筆測驗方式。如「學科能力測驗」、「指定科目考試」，其中除國文、英文的作文屬性接近簡短的『論文式測驗』外，其餘各科的題型無論是選擇式測驗題、計算、問答、說明題等皆屬『客觀式測驗』或『限制性反應題』。

「學科能力測驗」的考科為國文、英文、數學、自然、社會等五科，目標在評量考生是否具備接受大學教育所應具備的「一般性知識與技能」，是故大多採用『客觀式測驗』。而「指定科目考試」的考科為國文、英文、數學、物理、化學、生物、歷史、地理等科目，由大學各學系指定考試科目，目標在評量考生具有就讀該學系，所需要的專精或特殊的知識與技能，是否採用『論文式測驗』值得探討研究。

分析近年來大學多元入學的取材方法，可以將各種評量歸類為（1）申請入學：有採用客觀式的『學科能力測驗』的測驗成績，加上報告、卷宗評量、專題報告（大多以小論文方式呈現）等方式。（2）推薦甄選：也採用客觀式的『學科能力測驗』的測驗成績，加上採用客觀式的『自訂紙筆測驗』、簡單口頭問答、面談（以面試方式呈現）、或報告與專題報告（以小論文方式呈現）等方式。（3）聯合分發：採用客觀式的『指定科目考試』的測驗成績。

總結各種評量類型與測驗的功能，評量的選用在『大學入學的選才』方式上，尤其是『申請』與『甄選』入學其評量方式，尚可考量：卷宗評量、進階式評量、實地觀察與操作評量、實作評量等，值得研究。

（三）國內『指定科目考試』的題型與測驗目標分析

本研究的測驗為『成就測驗』，相當強調測驗的測量目標，而研究標的為『高中物理實驗評量』，比對『大學入學考試』的題型，研究者統計 78 至 92 年這 15 年來

『入學考』實驗題的測量目標如表 2-4.3。由表知，高中物理實驗在『入學考』的測驗目標上停留於評量『知識』、『理解』與『分析』的層次上。至於『應用』層次則付之闕如，而『綜合』與『評鑑』則因可能題型上沒有優勢、命題設計的難度高、考試時間需較長等因素，亦未見命題出現。也許命題者，認為此三種層次的考試，可以在物理理論命題上測量即可。

依據陳英豪、吳裕益合著『測驗與評量』（民 87）一書中所提：『大多數的客觀式測驗僅適用於『知識』層次，但選擇題例外，可用來測量複雜的學習結果，其中包含『理解』、『應用』和『分析』層次。』此說明，若以往年國內『入學考試的物理實驗測驗』的測量目標層次為準，選擇式的測驗試題測量學習結果的層次涵蓋『簡答題』或說是『問答與說明題』的層次和功能。這也是本研究所要探討的問題。

因此，針對『入學考試的物理實驗測驗』，本研究鎖定『知識』、『理解』、『分析』這三種層次的評量設計測驗卷。

表 2-4.3 中的測量目標代號 1、2、3、4、5、6 分別代表實驗測量目標：（1）實驗裝置與儀器（2）實驗步驟與方法（3）實驗目的與原理說明（4）實驗注意事項與原因（5）實驗數據處理與分析（6）實驗結果解釋與問題討論。而依布魯姆的認知目標六層次知，可大致將代號 1 及 2 歸類為『知識』層次，而代號 3 及 4 歸類為『理解』層次，代號 5 及 6 項歸類為『分析』層次。

表 2-4.3　75 至 92 年大學入學考試物理科實驗題測量目標與層次

| 年度 | 實驗單元名稱 | 題型 | 測量目標代號 | 測量目標層次 | 備註 |
|---|---|---|---|---|---|
| 78 | 光的干涉與繞射 | 單選 | 1 | 知識 | |
| | 力的合成與分解 | 多重選擇 | 2、3、4 | 知識、理解 | |
| 79 | 共鳴空氣柱 | 說明題 | 2、3 | 知識、理解 | |
| | 水平地磁強度 | 說明題 | 1、2、3 | 知識、理解 | |
| 80 | 針孔照相機 | 多重選擇 | 1、2、3、4 | 知識、理解 | |
| | 等位線與電場 | 問答題 | 1 | 知識 | |
| | 電流磁效應 | 問答題 | 5、6 | 分析 | |
| 81 | 牛頓第二運動定律 | 問答題 | 5、6 | 分析 | |
| 82 | 電流磁效應 | 問答題 | 1、2、3 | 知識、理解 | |
| 83 | 惠司同電橋 | 單選 | 2、3 | 知識、理解 | |
| | 非彈性碰撞 | 問答題 | 1、2、3 | 知識、理解 | |
| | 光的干涉與繞射 | 問答題 | 1、2 | 知識 | |
| 84 | 冰之熔化熱的測定 | 單選 | 2、3 | 知識、理解 | |
| | 波以耳定律 | 問答題 | 1、2、3 | 知識、理解 | |
| | 共鳴空氣柱 | 問答題 | 1、2、3 | 知識、理解 | |
| 85 | 直線運動定律 | 問答題 | 1、2、3、4 | 知識、理解 | |
| | 日光燈 | 問答題 | 1、2、3 | 知識、理解 | |
| 86 | 力學能的轉換 | 多重選擇 | 1、2、3 | 知識、理解 | |
| | 冰的熔化熱測定 | 問答題 | 2、3、4 | 知識、理解 | |
| | 水波槽實驗 | 問答題 | 1、2、3 | 知識、理解 | |
| 87 | 雙狹縫干涉實驗 | 多重選擇 | 1、2、3、6 | 知識、理解、分析 | |
| | 拋體運動 | 問答題 | 1、2、3、5、6 | 知識、理解、分析 | |
| 88 | 共鳴空氣柱 | 多重選擇 | 1、2、3、4、6 | 知識、理解、分析 | |
| | 等位線與電場 | 問答題 | 1、2、3、4 | 知識、理解 | |
| 89 | 惠司同電橋 | 多重選擇 | 1、2、3、4、6 | 知識、理解、分析 | |
| | 金屬的比熱 | 問答題 | 2、3、4 | 知識、理解 | |
| 90 | 非彈性碰撞 | 單選 | 3 | 理解 | |
| | 日光燈實驗 | 問答題 | 1、2、3 | 知識、理解 | |
| | 電流磁效應 | 問答題 | 2、3 | 知識、理解 | |
| 91 | 自由落體數據處理 | 單選 | 5 | 分析 | |
| | 二維空間碰撞 | 多重選擇 | 1、2、3、4、6 | 知識、理解、分析 | |
| 92 | 金屬的比熱 | 多重選擇 | 3、4 | 理解 | |
| | 波以耳定律 | 多重選擇 | 2、3、4、5、6 | 知識、理解、分析 | |

# 第五節 測驗的編製

一份優良的成就測驗編製有五大步驟（余民寧，民 86）：（1）準備測驗編製計畫（2）編擬測驗試題（3）試題與測驗的審查（4）試題與測驗的分析（5）測驗的編輯。

依據本研究的目的，擬定研究計畫後，根據測驗編製的理論與本文獻探討中的理論，研究者試著依步驟（1）確立測驗的目的與目標（2）設計雙向細目表與選定測驗的題型（3）試題的編定與審查（4）題目的分析與信度效度研究。藉由完成一份測驗的初步編製，同時指出本研究所採用的測驗編製理論與方法。

## 一、確立測驗的目的和目標

一份成功的測驗，必定依測驗的目的與所欲測量的目標而設定。本研究的測驗目的為測量教學後學生的學習成就，依測驗的類型知其特徵為『總結性測驗』，測驗的功能與目的為測量『學生的科學知識與能力及對課程內容的精熟程度』。教學目標為『認知領域』中『知識方面』的學習。

分析 78 至 92 年這 15 年來『入學考試物理實驗題』的測量目標如表 2-4.3。由表知，測驗的目標層次大都是『知識』、『理解』方面，少數為『分析』的層次。本研究主要目的探討入學考試測驗題型的優缺點，強調比對的對象為入學考試，是以本研究也鎖定這三種層次的評量設計測驗。

## 二、設計雙向細目表與選定測驗的題型

本研究標的為物理實驗課程內容測驗，為增加測驗內容的效度，必須含括所有欲測驗的實驗內容。分析近年來的『入學考試物理實驗題』題型，大多以單選、多重選、問答題與說明題的題型出現，如表 2-4.3。而其評量認知領域層次為『知識』、『理解』、『分析』等層次的學習結果。

本研究依上述條件與本章前述的文獻理論，發展編製測驗所需的雙向細目表，同時

依研究者的教學經驗與教材內容，參酌相關題庫及大學聯考命題資料以及測驗題型，據以編製一份與『入學測驗』相當且評量層次接近的物理實驗成就測驗雙向細目表，如表2-5.1。表中的A、B卷主要為分析多重選擇與問答題的題型優劣所設計，測驗概念數指的是每一個實驗題的概念子題數目，詳細設計原理請參考第三章說明。

表2-5.1 物理實驗成就測驗雙向細目表

| 卷別 | 實驗單元名稱 | 題型 | 學習結果測量目標代號 | | | | | | 總計測驗概念數 |
|---|---|---|---|---|---|---|---|---|---|
| | | | 知識層次 | | 理解層次 | | 分析層次 | | |
| | | | *1* | *2* | *3* | *4* | *5* | *6* | |
| AB共同卷 | 自由落體運動 | 單選 | | 1 | 2 | 1 | | 1 | 5 |
| | 二維空間碰撞 | 單選 | 1 | 1 | 1 | 1 | | 1 | 5 |
| | 牛頓第二運動定律 | 多選 | | 1 | 1 | 1 | | 2 | 5 |
| | 波以耳定律 | 多選 | | 1 | 2 | | 1 | 1 | 5 |
| | 斜面上運動 | 問答 | 1 | | 1 | 1 | 1 | 1 | 5 |
| | 水波干涉與繞射 | 問答 | 2 | | 3 | | | | 5 |
| A或B卷 | 金屬的比熱 | A多選<br>B問答 | | 1 | 1 | 1 | | 2 | 5 |
| | 靜力平衡 | A多選<br>B問答 | | 1 | 3 | 1 | | | 5 |
| | 量熱器的水當量 | A問答<br>B多選 | | 1 | 2 | 1 | | 1 | 5 |
| | 共鳴空氣柱 | A問答<br>B多選 | 1 | 1 | 1 | 1 | | 1 | 5 |
| 總計測驗概念數 | | | 5 | 8 | 17 | 8 | 2 | 10 | 50 |

※表中的測量目標代號1、2、3、4、5、6分別代表實驗測量目標：（1）實驗裝置與儀器（2）實驗步驟與方法（3）實驗目的與原理說明（4）實驗注意事項與原因（5）實驗數據處理與分析（6）實驗結果解釋與問題討論。

## 三、試題的編定與審查

一般而言試題的編擬，首先需充分瞭解各類題型的優缺點（參考表2-4.2）與命題

要點，考量測驗的目的與目標並了解課程內容與學生特性後，編擬出一份初步的題庫，供預試測驗或經過專家審查後再修正完成。

歸納國內外學者專家（余民寧，民 86；陳英豪、吳裕益，民 87；張景媛，民 81；盧欽銘、范德鑫，民 81；簡茂發，民 85；郭玉生，民 81；陳李綢，民 81；Kennth D.Hopkins，1990）的見解，本研究試擬一份『命題原則暨自我審查要點』，以供命題時參考，避免命題的疏失並做為試題的自我審查使用：（1）試題取材需廣泛且均勻，並具教材內容的代表性（2）試題的敘述應力求簡單扼要、題意明確（3）題目宜有公認或相對較佳的答案，且明確不具爭議性（4）各個試題盡量彼此獨立，互不牽涉，並避免含有暗示答案的線索（5）試題應重視重要概念或原理原則之瞭解與應用，避免偏重瑣碎的知識記憶（6）命題不可超過共同教材的範圍（7）選擇題的誘答選項必須可行、且吸引未具備相關知識學生（8）問題的結構，必須能把正確反應的曖昧降到最低，且避免定義不清（9）試題是否與測量目標一致（10）試題的難度是否與入學考試一致。

根據上述的命題原則，設計了測驗 A、B 兩卷，經過研究者學校周老師鴻案做初步的審查，並在沈教授青嵩的指導下，完成整份試卷的編輯，如附錄四。

## 四、題目的分析與信度效度研究

本研究由於時間、人力的限制等因素，無法進行測驗的預試，是以探討測驗的可研究性，主要為測驗後試題的可信度與效度準備。以下探討一些試題分析的理論，並同時說明本測驗欲採用的方法。

### （一）試題分析的功用

1. 余民寧(民 86)，認為有（1）作為改進學生學習的參考（2）作為實施補就教學的依據（3）作為修改課程建議的憑據（4）增進教師編製測驗的經驗（5）增進測驗題庫運用的效能。

2．陳英豪、吳裕益（民 87）認為有（1）有助於測驗結果的討論（2）可作為實施補救教學的依據（3）可作為改進班級教學的依據（4）可增進編製測驗的技能（5）可以提高測驗的信度和效度。

（二）試題的難度與鑑別度、選項分析

本研究採用『常模參照測驗』的試題分析步驟（余民寧，民 86）（1）將學生依測驗成績高低依序排列，並分為高中低分組（2）計算高低分組學生在每一個試題的答對率（3）計算每一個試題的難度指數（4）計算每個試題的鑑別度指數（5）對每個試題的選項進行『正答』與『誘答』分析。

上述試題分析法的理論相當多，歸納國內外學者專家（余民寧，民 86；陳英豪、吳裕益，民 87；簡茂發，民 85；郭玉生，民 81；陳李綢，民 81；黃政傑，民 85；Kennth D.Hopkins，1990）的見解，概略介紹於後。

(1) 高低分組法：高分組為全體學生總分前 25%或前 33%的學生群，低分組為全體學生總分後 25%或後 33%的學生群，其餘歸類為中分組。由於『大考中心』對聯考的試題分析較常採用前後 33%的分組方法，且由數學統計觀點知採用前後 33%的分組方法較為嚴僅，故本研究採用此分組法。

(2) 答對率（number correct ratio）：即計算答對每個試題的人數佔總人數的比值。本研究分為全體答對率（PR）、高分組（前 33%）答對率（PH）、低分組（後 33%）答對率（PL）。

(3) 難度指數（difficulty）（P）：表示法有二

（a）以答對率表示：試題的難度一般均以答對率來表示。也有測驗專家以高分組答對率（PH）與低分組答對率（PL）相加除以 2 來表示。如此，P 值越高，表示題目越容易。而難度指數以 0.30 至 0.70 為理想，越接近 0.5，區分高低分組的能

力越高。

(b) 以等距量尺表示：由於答對率所建立的難度指數，只能用來表示試題難易的相對位置，無法指出各難度間差異大小的數學涵義。假設所有試題所測量的特質均呈常態分配，可根據常態曲線概率表，將試題的難度轉換成具有相等單位的等距量尺（interval scale)，如此就可比較各試題的難度數值。

本研究採用美國教育測驗服務社（Educational Testing Service，簡稱ETS）的等距量尺來表示試題的難度指數，為將難度指數轉化為正值之等距量尺分數。其轉換公式為：△＝13+4Z；其中，△（delta）代表試題難度指數，Z 為標準化常態分配量尺上的標準分數，13 代表轉換公式的平均難度，4 為轉換公式的標準差。△值愈大，表示難度愈高。而經由 PH，PL 值可以透過『范式試題分析表』(Fan‘s item analysis table）查得相關的 P、△值及 r（試題反應與效標的二列相關係數)。

(4) 鑑別度指數（discrimination index）(D)：分為兩類

(a) 內部一致性（internal consistency）的鑑別度指標：探討個別試題得分和整個測驗總分間的相關。指標公式為高分組答對率（PH）與低分組答對率（PL）差，即 D＝PH-PL。D 值介於-1.00 至 1.00 之間，試題的鑑別指數愈高（即 D 值愈高）代表愈具有鑑別作用。一般測驗專家的看法：鑑別指數以 0.40 以上為理想，0.30 至 0.39 之間為良好，0.10 至 0.29 間尚可，而 0.10 以下則為劣等鑑別度試題，若鑑別度為負值，則需淘汰。

(b) 外在效度（external validity）的鑑別度指標：探討受試者在每一個試題的反應與在效標上的表現之相關情形。大多採用相關係數法，如點二系列相關（point-biserial correlation)、二系列相關（biserial correlation）等，

由於本研究非屬『效標參照測驗』且未有外在效標，不予討論。

(5) 選項分析：每個選項所附的可能答案中，正確的答案稱為『正答』，其他稱為『誘答』(distractors)。而所有可能答案稱為選項。為瞭解試題的有效性，將考生的作答情況統計成選項分析表（如表 3-4.3)。

可以藉由試題分析的指標，做為判斷試題的優劣與診斷學生反應的依據。一般判斷原則為：(a) 至少有一位低分組學生選擇任何一個不正確選項 (b) 選擇不正確選項的低分組學生人數應該比高分組的學生人數還多。

(三) 信度的分析理論與研究

本研究的測驗，比對『入學考試』的選材成就測驗，測驗分析則採用常模參照測驗的信度分析方法。信度分析的理論相當多，整理國內外學者專家（余民寧，民 86；陳英豪、吳裕益，民 87；簡茂發，民 85；郭玉生，民 81；黃政傑，民 85；何英奇，民 81；王文科，民 91；Kennth D.Hopkins，1990）的見解，概略介紹如後。

(1) 信度的意義：

依據古典測驗理論 (余民寧，民86)，信度的定義為真實分數的變異數 (variance) 佔實得分數的總變異數的百分比，即信度＝真實分數的變異數/實得分數的變異數。

信度的涵義指經由多次測量所得結果的一致性或穩定性，或估計測量誤差有多少，以反映出真實量數（true measure）程度的一種指標（Gulliksen，1950/1987)。當測驗分數中測量誤差所佔的比率降低時，則真實量數所佔的比率就相對提高，如此信度就高。

(2) 估計信度的方法：

一般信度的估計方法有四類（余民寧，民86；陳英豪、吳裕益，民87；何英奇，民81；王文科，民91）(a) 重測信度（test-retest reliability） (b) 複本信度

（parallel-forms或equivalent-forms reliability）(c) 內部一致性信度（internal consistency reliability)－分為折半信度、庫李信度、α係數(d)評分者信度(scorer reliability）－分為評分者間相關係數、評分者內同質性信度係數。其名詞解釋，於本章第三節中（三、評量工具的特徵）說明不再贅述。

本研究中，由於時間、人力的限制，未實施預試、重測與複本施測，所以信度的研究限於內部一致性信度與評分者信度，分別採用折半信度（split-half reliability）、交互評分者信度（interscorer reliability）之相關係數。綜合幾種信度係數的類型、目標、使用程序、統計量數與誤差來源，如表2-5.2。

表 2-5.2 信度係數的統計程序、方法與原理

| 種類 | 涵義與目標 | 使用程序 | 統計量數 | 主要<br>誤差來源 |
|---|---|---|---|---|
| 重測信度 | 確定測驗的穩定度。 | 同一測驗隔一段時間先後對同一團體，施測兩次求相關係數 | 相關係數 | 時間抽樣 |
| 複本信度 | 求得某測驗的兩種形式確實相等 | 將同一測驗的兩種形式連續施測同一團體 | 相關係數 | 時間抽樣與內容抽樣 |
| 內部一致性信度－折半信度 | 決定同一測驗內題目測量同一特徵的程度 | 僅施測一次；各半分別計分；計算兩半之間的相關；可用史布校正公式或Guttman 公式。 | 相關係數 | 內容抽樣 |
| 內部一致性信度－α係數 | 試題間的同質性或反應一致性的程度之關連指標 | 僅施測一次；使用 Cronbach 所發明的α係數公式。 | 相關係數 | 內容抽樣與內容異質 |
| 交互評分者信度 | 決定結果的客觀程度；即不問評分者為誰，對同一測驗評定結果將會相同。 | 施測一次；交由兩人評分；計算兩項分數的相關，即 Pearson 積差相關係數。 | 相關係數 | 評分者誤差 |

※ 本表改編自余民寧（民 86）、王文科（民 91）。

通常 Cronbach α係數所估計的信度係數比折半信度所估計的值低，即α係數是所有內部一致性信度估計的下限，所以α係數數值高時，即表示真正的信度值更高。當使用折半信度之史布（Spearman-Brown）公式時，其基本假定為兩半測驗分

數的變異數相等，若變異數不相等，則需使用 Guttman 公式（請參考第一章第四節）。

(3) 影響信度的因素：

測驗的信度低，將使測驗失去意義，是以瞭解影響測驗信度的因素，並進而降低其影響或進行補救，皆是重要課題。

學者余民寧（民 86）認為較常見的影響測驗信度的因素為（a）試題數的多寡：試題數越多，信度會增加（b）樣本能力分配：能力分配的變異數越大，相關係數越高，信度係數越大（c）試題難易程度：難易適中的測驗試題，得分分配越趨近於常態分配，變異程度達最大，信度係數會較大（d）評分的客觀性：計分方式愈主觀者，由於評分者誤差愈大，信度係數愈低。測驗的信度值要高，以選用客觀式的測驗為佳（e）信度的估計方法：選用測驗及解釋其信度資料時，應該考量該測驗所採用的信度估計方法、信度的適用情境、試題間的關連性、及測量誤差的可能來源等因素，方不致於造成濫用或誤用測驗的情勢發生。

本研究考量學生的作答時間，並參考『入學考試』的難度，以致試題的概念數雖達 50 題（參考表 2-5.1），但總完整實驗題數僅為 10 大題，以致影響信度係數。但在樣本能力分配及評分的客觀性上，皆能順利達成，請參考第三章。

(四) 效度的分析理論與研究

整理國內外學者專家（余民寧，民 86；陳英豪、吳裕益，民 87；簡茂發，民 85；郭玉生，民 79；何英奇，民 81；王文科，民 91；Kennth D.Hopkins，1990）的觀點，概略介紹如後。

(1) 效度的意義（余民寧，民 86）：

效度是指測驗分數的有效程度，亦即是測驗能夠提供適切資料以做決策的程度。也就是測驗分數能夠代表它所要測量之潛在特質的程度或測驗能夠達到其編製

目的的程度。

因此，測驗分數必然與所要測量之潛在特質間具有某種程度的關係（即共同變異部分），故根據統計學理論，定義效度為某個測驗和其他測驗（通常指的是外在效標）所共同分享的變異數部分佔該測驗總變異數的比率。即效度＝（共同因素的變異數）/（總變異數的比值）。

(2) 信度與效度的關係：

從前述的信度與效度的定義知（余民寧，民 86）：信度＝效度+獨特性（獨特性即獨特變異數（specific variance）/總變異數）。可知，效度包含於信度之內，信度所涵蓋的範圍比效度所涵蓋的範圍大。

學者簡茂發（民 67）認為：信度低，效度一定低，但信度高，效度不一定高；效度高，信度一定高，但效度低，信度不一定低。學者郭玉生（民 79）則認為效度需要測驗的一致性與正確性；但信度僅需要測驗的一致性即可。

(3) 效度的類型與考驗方法：

效度的種類很多，使用最廣泛的是美國心理學會（American Psychological Association，APA，1974）所採用的分類法：內容效度（content validity），效標關聯效度（criterion-related validity），建構效度（construct validity）三類。學者陳英豪、吳裕益（民 87）將其意義和考驗方法整理為表 2-5.3。

茲將其意義與內涵詳述於後（余民寧，民 86）。

(a) 內容效度：一般而言，測驗試題若能涵蓋所要的教學目標和教材內容，並且是根據雙向細目表來命題，且具有足夠的代表性試題，即能夠確立該測驗內容效度的適當性。因此，教材目標與教學內容即是確立內容效度的兩種重要因素。

表 2-5.3 效度的意義和考驗的方法

| 類型 | 意義 | 考驗方法 |
|---|---|---|
| 內容效度 | 測驗內容能否充分代表其所欲測量的行為領域。 | 比較測驗的材料和所預測量的教學目標及教材內容是否一致。 |
| 效標關聯效度 | 測驗成績對目前及未來某一行為表現（由其他適當的工具測量而得）預測力的高低。 | 求測驗分數與其他測驗成績之相關。其他測驗成績如在同時測量則為同時效度；如在往後測量則為預測效度。 |
| 建構效度 | 測驗成績能以心理學的屬性來加以解釋的程度。 | 建立理論架構，以解釋個體在測驗上的表現；根據理論架構推演出各種假設；收集資料考驗假設是否成立。 |

※ 本表取自陳英豪、吳裕益（民 87）

建立內容效度的方法，可分為：（1）邏輯的分析方法（2）實證的分析方法。以上兩種方法，需邀請專家針對測驗的雙向細目表，做一個檢測，由於相當困難，本研究僅由兩位高中教師針對雙向細目表做初步的教材內容與教學目標檢測。

(b) 效標關聯效度：是指以實證方法研究測驗分數與外在效標間關連性的一種指標。所謂的外在效標，即是指測驗所要預測的某些行為或表現標準。分為兩類：同時效度、預測效度。

本研究利用 A、B 卷同時測量類似效標關聯效度之同時效度。

(c) 建構效度：根據心裡學或社會學的理論建構，對測驗分數的意義所做的分析和解釋，即為建構效度。建構效度的建立是根據理論建構而來，因此，理論所假設的各種原理原則和學說，都必須經過驗證，才能確立建構效度是否成立。

其驗證方法有（Ⅰ）內部一致性分析法：以測驗本身的總分為內在效標，分析個別試題與總分間的相關（Ⅱ）外在效標分析法（III）因素分析法（IV）多特質－多方法分析（multitrait－multimethod aproach）。

而內部一致性分析法，可以採用(a)相關分析法（b）團體對照法：依據學生的測驗總分高低，分成高低兩組，然後比較這兩組學生在每個試題上答對的百分比值。經過統計考驗後，若有顯著差異，表示試題具有較高的內部一致性。

本研究並無外在效標，是以僅採用內部一致性的效度分析法，包含相關分析與團體對照法，來考驗試題的效度，並驗證測驗的可研究性。

（4）影響效度的因素：

影響效度的因素（余民寧，民 86）（1）測驗編製過程是否得當（2）施測程序與情境是否良好（3）受試者的身心反應因素（4）外在效標品質好壞－測驗分數的信度與外在效標的信度愈低，效標關連效度值也愈低（5）樣本能力的變異程度－受試者能力分配的變異程度愈大（即異質性愈高），相關係數值便愈大。

本研究在測驗的編製過程相當審慎且遵守本節所述的步驟，命題時勤於檢查雙向細目表與測量目標，樣本的選擇亦能符合所需，以儘量降低影響效度的因素，提高測驗的可研究性。