

## 《教育與心理測驗》撰文-葉華老師

### 心理與教育測驗的發展&應考準備

#### 壹、測驗理論之發展趨勢

測驗是一門新興的科學，其發展歷程不過百年歷史。測驗主要可分為兩大領域，一是心理測驗；另一則為教育測驗。不過，無論是何種的測驗內容，測驗理論可統整稱為解釋測驗資料間實證關係(empirical relationships)之有系統的理論學說(余民寧，民 80)，測驗理論的發展迄今已邁入不同的新紀元，隨著測驗理論的不斷發展與創新，依提出時間與內容可區分為兩大里程：一為古典測驗理論(classical test theory)，以真實分數模式(true score model) (Gullikson, 1987; Lord & Novick, 1968)為骨幹；另一為現代測驗理論(modern test theory)，是以試題反應理論(item response theory) (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980)為架構。

歐美測驗理論的發展起源於現代心理學的研究，乃需要對人類行為加以測量與觀察，其產生的時代背景可由下列四項研究中窺知：

##### 一、法國智能不足兒童的研究

法國對智能不足兒童的分類與訓練，奠定了智力測驗發展的基礎，比奈(Binet)受教育部之委託，於 1905 年發展出比西智力量表，而成為近代的智力測驗之父。

##### 二、德國實驗心理學的研究

德國因實驗心理學的研究，而奠定了測驗程序標準化的基礎，實驗心理學之父馮德(Wundt)於 1879 年在德國的萊比錫(Leipzig)大學創立了第一個心理實驗室，致力於實驗研究，提供了日後重視測驗程序標準化與精確測量的依據。

##### 三、英國個別差異的研究

英國對於個別差異的研究，提供了測驗資料分析的統計方法，生物學家高爾登(Galton)致力於個別差異的研究，現今吾人所使用之評定量表、問卷方法與自由聯想技術，可說高登是其先驅者。此外，他又創用座標圖表示兩變項之間的相關，之後由其同僚皮爾遜(Pearson)發展成為機差相關法，而成為現代重要的統計分析工具與測量工具編製之基礎。

##### 四、美國測驗運動的推展

心理測驗萌芽於歐洲，但卻於美國發揚光大，影響美國測驗發展的功臣首推卡泰爾(Cattell)，心理測驗(mental test)一詞為其在 1890 年發表論文中首先被使用；此外，桑代克(Thorndike)乃為測驗理論迅速散佈全美的另一關鍵人物，其第一部論及教育評量的教科書，影響了美國的教育學家與心理學家，成為美國推展測驗運動的主要力量。

#### 貳、我國對於測驗的運用

我國對於測驗的應用則始於二千餘年悠久的文官考試制度，遠在漢朝時代(西元前 206 年)，即創用科舉制度，採用口試與筆試方式拔擢優秀人才，科舉考試成為平民才俊躍升官人的一大機會，二千餘年來，文人士子無不寒窗苦讀，以求得天下功名利祿，而此方法不僅為當時政府選賢與能的良好制度，更為我國今日的考試制度奠下深厚基礎。以下先對測驗理論的沿革，簡要介紹兩大測驗理論，在後續主題中，將會再特別針對時下最熱門的測驗理論—「現代測驗理論」作詳細的介紹。

##### 一、古典測驗理論

(一)意義：古典測驗理論又稱傳統測驗理論，或稱真分數理論，因其較易符合一般研究資料，故又稱為弱真分數理論，為與 Lord(1965)所提出的強真分數理論作區別，強真分數理論的測量標準誤比弱真分數理論的測量標準誤更能適應個別差異的需要。

(二)假定：

1. 觀察分數(X)是真實分數(T)與誤差分數(E)之和。亦即  $X=T+E$
2. 觀察分數的期望值等於真實分數，亦即  $E(X)=T$

3. 誤差分數與真實分數之間無相關，亦即  $\rho_{TE} = 0$
4. 不同測驗的誤差分數間為零相關，亦即  $\rho_{E1E2} = 0$
5. 不同測驗的誤差分數與真實分數間為零相關，亦即  $\rho_{E1T2} = 0$

(三)缺點：

1. 抽樣變動大：古典測驗理論使用的項目分析法(又稱為試題分析, item analysis)，所得之項目統計量數受樣本的抽樣變動影響甚大。換言之，試題的各項指數受到當次受試者的表現而決定，幾乎永遠都無法獲得一個恆久的試題指標值。因此，古典測驗理論第一個受到批評的就是試題指數訊息難確定。關於測驗或問卷題項的指數可分為三類，分別為認知測驗中的難度指數、鑑別度指數，以及針對調查問卷的信度指數等。
  - (1)項目難度指數：所謂難度(difficulty)，指的是測驗中各個题目的難易程度，亦即是某個群體通過(或答對)某個測驗题目的百分比，當獲得的通過率指數愈高時，表示試題愈容易或難度較低。通常難度分析都是用在能力測驗、成就測驗或有標準答案的測驗上。
  - (2)項目鑑別度指數：所謂鑑別度(discrimination)指各個測驗题目能夠測量到所預測量之特質的程度，或指各個測驗题目反應情形與測驗總分的一致性程度，當鑑別度指數愈高時，表示該試題能夠更精確的測量出所欲測量的特質，而當樣本較具同質性者，鑑別度值愈小；反之，愈大。
  - (3)信度指數：古典測驗理論所指的信度估計方法，乃是受樣本對測驗理論分數的變動性的影響。
2. 能力難比較：依據古典測驗理論，受試者間能力的比較只能在相同測驗或平行複本測驗中進行。然而平行複本測驗的編製並不容易，因此，如何精確的估計受試者的能力表現，實為一大問題。
3. 複本難實施：信度的界定是古典測驗理論的假設。但是，事實上，平行複本測驗非常難編製，且又因受試者會因遺忘、焦慮程度、習得新知識與動機的改變而影響測驗的結果，使得不緊受試者的能力表現難以估計，也使得測驗的信度不易達到穩定。
4. 缺乏預測力：無法預測受試者在一個新測驗時的表現，即因上述幾項的缺失可知，測驗試題的難度、鑑別度受抽樣變動影響大，以及問卷的編製不易、信度不穩定等因素，使得測驗的效果大大降低，因此，預測效果即大打折扣。
5. 等測量標準誤：古典測驗理論假定所有受試者的測量標準誤都是一樣的。亦可由以上的論述瞭解到這是不可能存在的完美測驗結果，受試者可能會在各次不同的問卷，或在各次的複本測驗表現中產生許多無法控制的誤差。因此，等測量標準誤的假設是幾乎不可能存在於真實測驗之中。

綜合以上針對古典測驗理論的批評，遂有現代測驗理論的誕生，現代測驗理論又稱為試題反應理論，即是所謂的IRT(Item Response Theory)。因應教育改革運動，國內升學考試制度亦有重大變革，在考試制度方面，即揚棄傳統的古典測驗理論計分方式，改以試題反應理論進行試題的分析，以及學生能力表現的估計，目的即是在於藉由試題各項參數的設定，以提供更公平的評分方式，且對學生能力更精確的評定。

以下先簡單介紹現代測驗理論的發展與相關學者的貢獻成果，未來的系列文章中，將會探討新式測驗對於全球考試制度的影響與貢獻。

## 二現代測驗理論

- (一)發展：1916年比奈(Binet)和西蒙(Simon)首創以圖面表示兩變項間的關係，如：年齡與答對機率，即誕生了今日的項目特徵曲線(item characteristic curve, 簡稱ICC)。ICC是現代測驗理論的重要觀念之一，比西智力量表即是第一個適性測驗。適性測驗(tailoring testing)濫觴於比西量表。意義是指針對受試者先前經驗選取受試者能力的題目進行施測，作答立即給分，以決定下一個題目，再評分再決定下一題的測驗方式，直到預定的題數完成為止，也就是在題庫中選取符合受試者能力的難度題目進行施測。現代則因應電腦的誕生，利用電腦化適性測驗(computerized adaptive test, 簡稱CAT)來進行施測，目前最為世人

所知的電腦化適性測驗即是全球的托福考試，可精確估計受試者的能力之外，更可避免考試作弊行為的發生。表一，將對百年來現代測驗理論的誕生與茁壯列表作逐一的介紹，讓世人瞭解投身於測驗理論學者的諸項貢獻。

表一 試題反應理論之相關代表性文獻資料

作者(年代)	代表作及其貢獻
Binet& Simon(1916)	首創以圖形方式表示兩變項之間的關係，如年齡與答對機率。兩位亦是首先應用 ICC 的心理學者。
Richard(1936)	導出現代測驗理論理論參數與古典測驗項目指標間的關係，為 IRT 參數估計的最早方法。
Lawley(1943, 1944)	提出新的參數估計方法，且對未來的 Lord 影響甚深。
1945	電腦誕生。
Tucker(1946)	第一位提出試題特徵曲線(ICC)概念的學者。
Gulliksen(1950)	提出真實分數(true score)模式概念，為古典測驗理論之濫觴。
Lazarsfeld(1950)	專攻態度測量，可能為最早使用「潛在特質(latent trait)」一詞的學者。
Cronbach(1951)	提出 $\alpha$ 信度係數概念。
Lord(1952)	第一位導出兩個參數常態肩形模式(two parameter normal ogive model) $P_i(\theta) = \int_{-\infty}^{a_i(\theta-b)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ 的參數估計公式( )，為第一位從事試題反應理論應用性的學者。其著作被視為是 IRT 的起源。
Birnbaum(1957, 1958)	提出對數型模式(logistic model)的統計基礎。
Rasch(1960)	試題反應理論中 Rasch 模式的創始者。
Wright(1967)	在美國教育測驗服務中心(ETS)的測驗研討會上，演講 Rasch Model 編製而大受矚目，成為 1970 年代 Rasch Model 在美國測驗發展史上的催生者與領導者。
Lord & Novick(1968)	其著作為介紹古典測驗與當代測驗理論模式的經典作品，引發學者對「潛在特質」概念的重視與研究，並對現代測驗理論研究具有啟發作用。
Wright & Panchapakesan (1969)	美國地區第一篇介紹 Rasch 模式的參數估計法，並發展有名的 BICAL 電腦程式，此程式為 Rasch Model 應用時最重要的電腦程式。
Samejima(1969)	發表一系列作品描述新的試題反應模式及其應用，其中包含處理多分法與連續性資料的模式，甚至擴展到多向度的試題反應模式，為一艱澀難懂的重要著作。
Bock(1972)	提供許多估計模式參數的重要新概念，如：估計連續性類別資料的概念。
Cronbach , Glaser , Nanda & Rajartnam(1972)	提出推論力理論。
Andersen(1973)	歐洲地區談論測驗模式的重要著作。
Lord(1974)	發展新參數估計法。
Fisher(1974)	提出線性對數型模式(linear logistic model)。
Bashaw, Lord, Marco, Rentz, Urry & Wright (1977)	在教育測量季刊(Journal of Educational Measurement)第四季出版一冊專門探討試題反應理論的專輯。
Baker(1977)	第一篇評論試題反應模式參數估計法的文獻探討。

Wright & Stone(1979)	第一本描述各種 Rasch 模式理論及其應用的專書--「最佳測驗設計(Best Test Design)」。
Lord(1980)	出版「試題反應理論在測驗的應用」(Applications of Item Response Theory to Practical Testing Problems)，此書為第一本以試題反應理論命名的專書，介紹 IRT 的發展與三參數模式的應用，是現代測驗理論發展的里程碑。
Weiss(1980)	第一本論電腦化適性測驗的論文集，專談試題反應理論的實際應用課題。
Andersen(1980)	對測量模式參數估計法有貢獻的方法學專論。
Bock & Aitkin(1981)	提出邊緣的最大近似值估計法——EM 估計程序，對參數估計法的改進貢獻不少。
Masters(1982)	第一位發表部份知識計分模式，對改進 Likert 式評定量表的計分與次序反應資料的計分貢獻不小。
Wright & Masters(1982)	闡述 Rasch 模式的各種模式成員，證明皆與部份計分模式相通，對 Likert 式評定量表與次序反應資料的計分方式改進不少。
Mislevy & Bock(1982)	發表另一有名的電腦程式：BILOG，可進行 EM、JEM 估計。
Lord et al.(1982)	修改完成第二版的 LOGIST 電腦程式，在使用上更加便捷。並於應用心理測量(Applied Psychological Measurement)期刊第四季，出版一冊專門探討試題反應理論及其應用的進階專輯。
Wainer & Messick(1983)	編輯而成的論文集，以表揚 Lord 一生對試題反應理論的貢獻，並兼論該理論的應用與未來。
Weiss(1983)	編輯而成的論文集，專談試題反應理論的應用與未來，並介紹它在電腦化適性測驗上的應用。
Hambleton(1983)	編輯而成的論文集，專談試題反應理論的模式與應用。
Hulin, Drasgow, & Parsons(1983)	為一本試題反應理論的教科書，增加對「適合度測量」概念的說明與應用。
Embretson(1985)	編輯而成的論文集，專談試題反應理論的未來發展。
Baker(1985)	為一本導論性的試題反應理論教科書，專為沒有數學訓練基礎的讀者而作，並附有 CAI 的電腦教學磁片。
Hambleton & Swaminathan(1985)	為一本進階的試題反應理論教科書。
Crocker & Algina(1986)	談論與比較古典與當代測驗理論的導論性教科書。
Wainer & Braun(1988)	專談有關效度方面的論文集，也談試題反應理論在效度上的應用。
Linn(1989)	負責主編第三版的「教育測量」(Educational Measurement)，其中增加一章專門介紹並評論試題反應理論。
Freedle(1990)	專談人工智慧及其在當代測驗理論上應用之論文集。
Suen(1990)	介紹各種測驗理論方面的教科書。
Wainer et al.(1990)	專談電腦化適性測驗方面的入門書，也談試題反應理論在電腦化適性測驗上的應用。
Hambleton, Swaminathan, & Rogers(1991)	試題反應理論方面的入門書，解說淺顯易懂，適用於非數學主修的初學者閱讀。
1991	網路的誕生，此時的網路資訊僅靠文字傳輸。
1994	WWW(World Wide Web)誕生，使得資訊格式的傳輸突破限制，可傳輸文字、圖片與動態影音等檔案。網路化測驗的理念開始被實踐。

(整理自 Hambleton & Swaminathan, 1985; 余民寧, 民 80a; 王寶壙, 民 84)

由以上介紹可知測驗理論的發展，對於時代有其重要里程碑的意義性。因此，在研讀教育與心理測驗這門課程時，除了針對古典測驗理論應用的熟讀之外，建議準備研究所考試的同學們也不可忽略試題反應理論的重要性。畢竟，順應時代潮流趨勢，更加上國內最重要的學生能力表現測驗(國中基本學力測驗、大學學科能力測驗)已遵循此理論進行。因此，建議同學們不僅要對古典測驗理論的內容熟悉之外，更加需要熟讀現代測驗理論內容。不過，現在測驗理論偏重數理公式的理解與運用，也相對地讓許多同學感到卻步，在此筆者亦要給予同學信心建議，考試的重點在於觀念的理解與應用，而非是理論公式的推導與計算。處於電腦化的 E 時代，許多繁瑣的數學公式運算都已交由電腦處理，無須研究人員或測驗使用者再進行艱深數學公式數字的運算。除了兩大測驗理論是相關考試的重點之外，學生在學校的各項學習表現，以及其他新型態的學校評量方式是另一類重要考題，如：卷宗評量、檔案評量、動態評量、多元智慧評量、以及其他非紙筆測驗等等。關於學校評量的論述，亦將在未來文章中作比較介紹。

## 參、觀摩試題

以下列舉幾題一般普遍出現的試題，提供初次準備這門課程學生的參考。針對教育與心理測驗的試題，有可能以選擇題(包含複選題)的方式呈現，或結合教育統計學、教育研究法或教育心理學等課程以申論題方式出題。因此，建議學生不僅需對此課程熟悉，亦需將其他課程中所學的知識與本課程知識作結合運用。

(一)有關教學與評量的敘述，下列何者錯誤？

1. 教學基本歷程分為教學目標、學前評估、教學活動與評量。
2. 教學目標包括認知、情意、技能三方面。
3. 教學前需先瞭解學生身心發展的程度，即所謂起點行為。
4. 就評量的時機而言，可分為形成性評量與過渡性評量。

【解析】4；

就教學評量的時機而言，形成性評量在學習中實施；而總結性評量則在學習結束後實施。

(二)國中基本學力測驗各科計分所採用的是：1. 百分數。2. 平均分數。3. Z 分數。4. 量尺分數。

【解析】4；

關於國中基本學力測驗將在後續文章中介紹。

(三)如果有一個題目，全班 40 人中有 30 人答對，請問此一題目的難度多少？1. .25。2. .50。3. .75。

【解析】3；

$$P = \frac{R}{N} = \frac{30}{40} = .75$$

(P：難度指數；R：答對人數；N：受試總人數)

(四)如果有一個題目，全班 40 人中均沒有人答對，請問此一題目的鑑別度多少？1. 0。2. .50。3. 1。

【解析】1；

完全沒有人答對，表示該試題無法區別出優劣成績表現，故鑑別度值為 0。鑑別度公式為：

$$D = P_H - P_L$$

( $P_H$  為高分組的難度指數； $P_L$  為低分組的難度指數)

鑑別度的評鑑標準：

鑑別度指數	試題評鑑
.40	極佳
.30~.39	良好
.10~.29	尚可
.01~.09	差
負數	試題有誤或題意不明

(資料來源：Hopkins, 1998, P. 260)

(五)在其他條件不變下，信度愈大，則測量標準誤：1. 愈大。2. 愈小。3. 不變。4. 不一定。

【解析】2；

參考測量標準誤(SEM)的公式： $SEM = SD\sqrt{1-r_{xx}}$

(六)主要以文字敘述的方式描述與記錄兒童的行為，是屬於以下何種觀察方法？1. 檢核表。2. 等級量表。3. 軼事記錄。4. 評量表。

【解析】3；

其他選項都是是數字形式紀錄與表示。

(七)教師有系統的蒐集兒童的學習作品，與有關兒童發展成長的相關資料，以瞭解幼兒之發展歷程，是屬於以下何種評量方式？1. 常模參照測驗。2. 效標參照測驗。3. 學習檔案評量。4. 檢核表。

【解析】3；

學習檔案乃是動態、長期的評量方式。

(八)一份教師特質量表，如果受試者分數越高的，後來在擔任教師工作的表現也越好，表示該量表具有何種效度？1. 內容效度。2. 表面效度。3. 同時效度。4. 預測效度

【解析】4；

對於未來表現預測的正確性。

(九)某成就測驗提供下列心理計量訊息：等值穩定係數 0.80、 $\alpha$  係數 0.85、折半信度 0.90、評分信度 0.70，則內容取樣之誤差來源的比率為何？1. 0.05。2. 0.10。3. 0.15。4. 0.20。

【解析】2；

等值穩定係數(複本信度)之誤差變異數

= 時間取樣誤差變異數 + 內容取樣誤差變異數 =  $1 - 0.80 = 0.20$   $\alpha$  係數之誤差變異數

= 內容取樣誤差變異數 + 試題的同質性誤差變異數 =  $1 - 0.85 = 0.15$

折半信度之誤差變異數 = 內容取樣誤差變異數 =  $1 - 0.90 = 0.10$

評分者信度誤差變異數 =  $1 - 0.70 = 0.30$ 。∴根據以上討論可得到內容取樣之誤差變異數為：0.10

(時間取樣誤差變異數 =  $0.20 - 0.10 = 0.10$ ；試題的同質性誤差變異數 =  $0.15 - 0.10 = 0.05$ )

參考文獻：略