

試題反應理論的介紹

一測驗理論的發展趨勢(一)

✿ 余民寧

考試制度的創設雖然源自中國，綿延數千年後，世界各國爭相採用，以作為建立文官制度的選拔依據，但是中國卻一直沒有針對「考試」這門學問進行比較科學化的量化分析，致使近代的心理計量學 (psychometrics) 卻發展且發揚於外國，西風東漸後，才傳入中國。

心理計量學是一門研究心理測驗 (psychological testing) 與評斷 (assessment) 的科學 (Cohen, Montague, Nathanson, & Swerdlik, 1988, P.26)，是一門包括量化心理學 (quantitative psychology)、個別差異 (individual differences)、和心理測驗理論 (mental test theories) 等研究範圍的學問。比奈——賽門 (Binet-Simon) 的智力測驗，可說是人類有史以來第一個心理測驗，測驗理論便是起源於此，並由此繼續往前發揚光大，成為心理計量學的主要架構。

測驗理論 (test theory) (或全稱叫「心理測驗理論」) 是一種解釋測驗資料間實證關係 (empirical relationships) 的有系統的理論學說，它的發展，迄今已邁入不同的新紀元，測驗理論學者通常把它劃分成二大學派：一為古典測驗理論 (classical test theory) ——主要是以真實分數模式 (true score model) (Gullikson, 1987; Lord & Novick, 1968) 為骨幹；另一為當代測驗理論 (modern test theory) ——主要是以試題反應理論 (Item response theory) (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Hulin, Drasgow, &

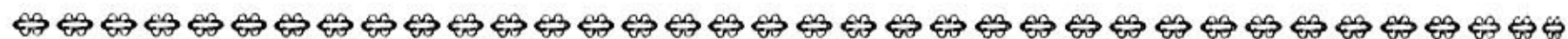
Parsons, 1983; Lord, 1980) 為架構。這兩派理論目前並行流通於測驗學界，但試題反應理論卻有後來居上，逐漸凌駕古典測驗理論之上，甚至進而取而代之之勢。

本文作者擬撰寫一系列文章，介紹試題反應理論的主要理論內涵及其應用，在此之前，我們有必要從歷史的觀點，來回顧與展望測驗理論的發展趨勢，以明瞭測驗理論發展的來龍去脈，這也正是本文的主要目的。

兩派測驗理論之比較

比奈——賽門的第一個心理測驗問世後，正是心理計量學誕生之始，後經諸多學者（如：Cronbach, 1951; Guilford, 1954; Gullikson, 1987; Guttman, 1944; Lord & Novick, 1968; Richardson, 1936; Terman, 1916; Thurstone, 1929; Tucker, 1946) 的研究與闡述，終於歸納形成古典測驗理論等學說。

古典測驗理論的內涵，主要是以真實分數模式（亦即，觀察分數等於真實分數與誤差分數之和，數學公式為 $X=T+E$ ）為理論架構，依據弱勢假設 (weak assumption) 而來，其理論模式的發展已為時甚久，且發展得相當規模，所採用的計算公式簡單明瞭、淺顯易懂，適用於大多數的教育與心理測驗資料，以及社會科學資料的分析，為目前測驗學界使用與流通最廣的理論依據。



然而，除上述各項優點外，古典測驗理論卻有下列諸項先天的缺失 (guion & Ironson, 1983; Wright, 1977)：

(一) 古典測驗理論所採用的指標，諸如：難度 (difficulty)、鑑別度 (discrimination)、和信度 (reliability) 等，都是一種樣本依賴 (sample dependent) 的指標；也就是說，這些指標的獲得會因接受測驗的受試者樣本的不同而不同，因此，同一份試卷很難獲得一致的難度、鑑別度、或信度。

(二) 古典測驗理論以一個相同的測量標準誤 (standard error of measurement)，作為每位受試者的測量誤差指標，這種作法並沒有考慮受試者能力的個別差異，對高、低能力兩極端組的受試者而言，這種指標極為不合理且不準確，致使理論假設的適當性受到懷疑。

(三) 古典測驗理論對於非複本 (nonparallel) 但功能相同的測驗所測得的分數間，無法提供有意義的比較，有意義的比較僅局限於相同測驗的前後測分數或複本測驗分數之間。

(四) 古典測驗理論對信度的假設，是建立在複本 (parallel forms) 測量的概念假設上，但是這種假設往往不存在於實際測驗情境裡。道理很簡單，因為不可能要求每位受試者接受同一份測驗無數次，而仍然假設每次測量間都彼此獨立不相關，況且，每一種測驗並不一定同時都有製作複本，因此複本測量的理論假設是行不通的，從方法學邏輯觀點而言，它的假設也是不合理的、矛盾的。

(五) 古典測驗理論忽視受試者的試題反應組型 (item response pattern)，認為原始得分相同的受試者，其能力必定一樣；其實不然，即使原始得分相同的受試者，其反應組型亦不見得會完全一致，因此，其能力估計值應該會有所不同。

一般說來，為了克服古典測驗理論的缺失，才有當代測驗理論的誕生。當代測驗理論的內涵，主要是以試題反應理論為理論架構，依據強勢假設 (strong assumptions) 而來，其理論的發展為時稍晚，理論模式也不斷的在發展當中，所採用的計算公式複雜深奧，艱澀難懂，為一立論與假設均合理與嚴謹的學說，所適用的測驗資料種類雖屬有限，但深受測驗學者的青睞，已有逐漸凌駕古典測驗理論之上，甚至進而取而代之之勢。

當代測驗理論是為改進古典測驗理論的缺失而來，它具有下列幾項特點，這些特點正是古典測驗理論所無法具備的 (Hambleton, 1989; Hambleton & Cook, 1977; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980)：

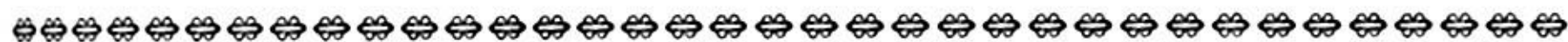
(一) 當代測驗理論所採用的試題參數 (item parameters) (如：難度、鑑別度、猜測度等)，是一種不受樣本影響 (sample-free) 的指標；也就是說，這些參數的獲得，不會因為所選出接受測驗的受試者樣本的不同而不同。

(二) 當代測驗理論能夠針對每位受試者，提供個別差異的測量誤差指標，而非單一相同的測量標準誤，因此能夠精確推估受試者的能力估計值。

(三) 當代測驗理論可經由適用的同質性試題組成的分測驗，測量估計出受試者個人的能力，不受測驗的影響 (test-free)，並且對於不同受試者間的分數，亦可進行有意義的比較。

(四) 當代測驗理論提出以試題訊息量 (item information) 及試卷訊息量 (test information) 的概念，來作為評定某個試題或整份試卷的測量準確性，倒有取代古典測驗理論的「信度」，作為評定試卷內部一致性指標之勢。

(五) 當代測驗理論同時考慮受試者的反



應組型與試題參數等特性，因此在估計個人能力時，除了能夠提供一個較精確的估計值外，對於原始得分相同的受試者，也往往給予不同的能力估計值。

(六)當代測驗理論所採用的適合度考驗值(statistic of goodness-of-fit)，可以提供考驗模式與資料間之適合度、受試者的反應是否為非尋常(unusual)等參考指標。

綜合上述，當代測驗理論似乎是絕對優於古典測驗理論，但是事實上，當代測驗理論被採用於解決真實測驗資料者，比起古典測驗理論廣泛地被應用的情形而言，尚屬少數，微不足道。其主要原因有下列諸點：

(一)當代測驗理論係建立在理論假設嚴謹的數理統計學機率模式上，是一種複雜深奧、艱澀難懂的測驗理論，這對於在數學方面訓練有限的教育與心理學界學者而言，莫非是一大挑戰。閱讀有關此理論之數學方面的研究報告與專書，已頗感困難，實在更難以深入將之發揚光大。

(二)多數當代測驗理論學者都是出身自數學界或曾是數學主修者，或至少在數理統計學上訓練有素者，他們偏愛對理論模式的探討，遠勝於對實際應用的推廣工作。

(三)過去，電腦科技的進步有限，沒有電腦套裝軟體程式的即時配合，當代測驗理論中對模式參數的估計，難以用手算或小型計算機順利進行，因此，在應用上更受限制。

(四)有些古典測驗理論的擁護者，對當代測驗理論的研究與發展，所能獲致之成效與應用性深表懷疑。為了證明與解釋疑惑，當代測驗理論學派的支持者，便更朝理論模式的量化技術方面探討，致使當代測驗理論的發展愈趨數學化、數量化、與電腦化。

(五)礙於嚴苛的基本假設，當代測驗理

論所能適用的教育與心理測驗資料有限，並且需要大樣本的配合，因此使得它的應用性大打折扣，未獲一般測驗使用者的全力擁戴。

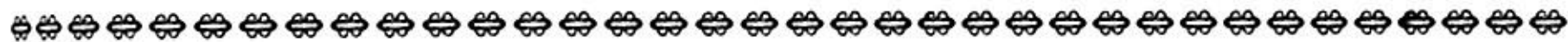
由上述兩派測驗理論的比較可知，古典測驗理論雖然不夠嚴謹，但理論淺顯易懂，便於在實際測驗情境（尤其是小規模資料）實施；當代測驗理論雖然嚴謹，但理論艱深難懂，僅適用於大樣本測驗資料的分析。所以，這兩派測驗理論各有所長，在應用上也各有其限制，我們僅能靜觀測驗理論的發展，逐步歸納出其未來的發展趨勢。

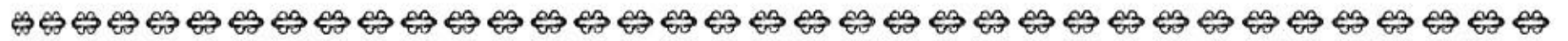
測驗理論的發展趨勢

自從 Lord(1980)發表第一本以「試題反應理論」為名的專書後，當代測驗理論正式以試題反應理論為其中心架構；在此之前，試題反應理論有個別稱：「潛在特質理論」(latent trait theory)，由於潛在特質理論一詞還包括「因素分析」(factor analysis)、「多元度量法」(multidimensional scaling)、與「潛在結構分析」(latent structure analysis)等，涵蓋面甚廣，無法精確反應出受試者在試題上的反應狀況，因此，自 Lord 發表專書後，試題反應理論於是正式正名，且宣告誕生。所以自 1980 年後，測驗學者逐漸以試題反應理論為當代測驗理論的代表。

試題反應理論雖然自 1980 年才正式正名成立，然而在 30 和 40 年代，試題反應理論便已有初步的理論架構。其中，Tucker(1946)便是第一位使用「試題特徵曲線」(item characteristic curve, 簡稱 ICC)一詞的心理計量學家，這一名詞也逐漸成為試題反應理論的中心概念。

茲將對試題反應理論發展有實際貢獻的代表性作者及著作，條列簡述於表一，由表一的



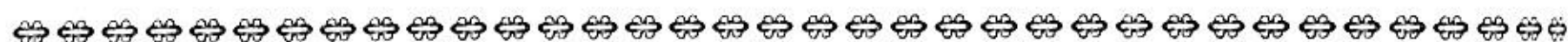


	試題反應理論及其應用的進階專輯。
Wainer & Messick(1983)	編輯而成的論文集，以表揚 Lord 一生對試題反應理論的貢獻，並兼論該理論的應用與未來。
Weiss(1983)	編輯而成的論文集，專談試題反應理論的應用與未來，並介紹它在電腦化適性測驗上的應用。
Hulin, Drasgow, & Parsons (1983)	為一本試題反應理論的教科書，增加對「適合度測量」概念的說明與應用。
Hambleton(1983)	編輯而成的論文集，專談試題反應理論的模式與應用。
Embretson(1985)	編輯而成的論文集，專談試題反應理論的未來發展。
Baker(1985)	為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 等人 (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，適用於非數學主修的初學者閱讀。

其實，隨著近年來人類在電腦科技上的突飛猛進，各種適用於試題反應理論的電腦軟體程式（如：目前最常用，也最有名的程式 BILOG 和 LOGIST 等）相繼誕生與再版修訂，已使得美國很多研究機構、地方政府機關、和私人團體，都率先採用試題反應理論作為他們編製測驗、施測、計分、解釋、與提供諮詢服務的依據。

此外，由表一所示的發展趨勢可見，當代

測驗理論的發展趨勢不外朝兩個方向同步進行——理論的發展愈趨數學化與理論的應用愈依賴電腦。相信在可預期的將來，測驗理論的使用者必須同時具備數學與電腦方面的良好訓練，方能對試題反應理論的瞭解與應用駕輕就熟，而測驗理論在愈趨專業化、專家化後，也唯有在專家或專家指導下方能推廣應用試題反應理論，不過照目前的發展趨勢來看，試題反應理論要取代古典測驗理論是指日可待的事。



參考書目

- Baker, F.b.(1985). The basics of item response theory. Portsmouth, NH: Heinemann.
- Crocker, L. & algina, J.(1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston. Guion, R.M. & Ironson, G.H. (1983).
- Cohen, R.J., Montague, P. Nathanson, L. S. & Swerdlik, M.E (1988). Psychological testing: An introduction to tests. and measurement, Mountain view, CA: Mayfield.
- Gulliks on, H.(1987). Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates. (Originally published in 1950 by New York: John Wiley & Sons)
- Guion, R.M. & Ironson, G.H.(1983) Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31, 54-87.
- Hamblet on, R.K.(1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.) Educational Measurement (3rd ed, pp.147-200). New York: Macmillan.
- Hamblet on, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff.
- Hamblet on, R.K. Swaminathan, H. & Rogers, H.J. (1991). Fundamentals of item response theory. Newbury Park, CA:SAGE.
- Hulin, C.L. Drasgow, F. & Parsons, C.K. (1983). Item response theory: Application to psychological measurement. Homewood, IL:Dow Jones-Irwin.
- Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Suen, HK. (1990). Principles of test theories. Hillsdale, NJ:Lawrence Erlbaum Associates.

(作者：國立政治大學教授兼附小校長)

