古典測驗理論

余民寧　教授

摘自「教育測驗與評量：成就測驗與教學評量」一書（2002，台北，心理）

雖然根據歷史學家（DuBois, 1970）的描述，早在西元一千多年前科舉時代的中國，即有能力測驗（即科舉考試制度）的雛型產生。但是，對「測驗」這門學問進行科學化的量化研究者，卻始於歐美各國，西風東漸之後，才又傳入中國。

西元 1905 年，Binet-Simon 在法國所發展的智力測驗，可以說是人類第一個客觀的心理測驗，也是測驗理論的真正濫觴。至此，這門專研心理測驗與評量（psychological testing and assessment），內含：量化心理學（quantitative psychology）、個別差異（individual differences）、和心理測驗理論（mental test theory）等研究範圍的科學，即稱為「心理計量學」（psychometrics）（或又譯成：「心理測驗學」），正式確立。心理計量學的誕生，乃心理學者企圖將心理學發展成為一門「量化的理性科學」（quantitative rational science）的結果，到目前為止，它雖然已邁入不同的新紀元，但成長與茁壯的腳步，卻未曾停止過。

談到測驗理論的發展，很多人喜歡以某某學派來作為區分，雖然這種分法不見得正確，但為了討論方便起見，我們亦可以一本著作或一位人物，作為某個學派理論的開始或代表。如此一來，我們大概可以將測驗理論粗分為下列兩派：

1.**古典測驗理論**（classical test theory，簡稱 **CTT**）：代表人物和作品分別為 **H. Gulliksen** 的「**Theory of mental test**」（1950）。

2.**試題反應理論**（item response theory，簡稱 **IRT**）：代表人物和作品分別為 **F. Lord** 的「**Applications of item response theory to practical testing problems**」（1980）。

底下，僅先就古典測驗理論的重要內涵做個扼要的評述，下一節再敘述試題反應理論。

「古典測驗理論」是最早的測驗理論，至今，它仍然是最實用的測驗理論，許多通用的測驗仍然是根據傳統方法來編製，並且建立起測驗資料間的實證關係。古典測驗理論也叫「古典信度理論」（classical reliability theory），因為，它的主要目的是在估計某個測驗實得分數（observed score）的信度；亦即，它企圖估計實得分數與真實分數（true score）間的關聯程度。因此，有時候它又稱作「真實分數理論」（true score theory），因為它的理論來源都是建立在以「真實分數模式」（true score model）為名的數學模式基礎上。

當某位受試者接受一份測驗的施測後，他（或她）在該測驗上的得分（即「**實得分數**」），即代表在某些特定的情境下，他（或她）在這些試題樣本上的**能力**（ability）。當然，有許多因素會影響受試者在測驗上的表現。即使在內容範圍相同但試題樣本不同的條件下，或在不同的時間、主測者、與施測情境條件下，受試者的表現也都有可能會不一樣。因此，如果

我們在所有可能的施測情境下、在所有可能的不同時間範圍內、或儘可能使用不同試題樣本，來針對同一位受試者進行同樣的測驗多次（理論上是無窮多次），則我們可以獲得許多有關該受試者的實得分數。這些實得分數的平均數（又稱為期望值（expected value）），即代表該受試者能力的不偏估計值（unbiased estimate)，該估計值即被定義為「**真實分數**」。因此，所謂的「真實分數模式」，即是指一種直線關係的數學模式（linear model），用來表示任何可以觀察到、測量到的實得分數（又簡稱為**觀察值**或**測量值**）皆由下列兩個部份所構成的一種數學涵數關係，這兩個部份分別是：一為觀察不到，但代表研究者真正想要去測量的潛在特質（latent trait）部份，叫作「**真實分數**」；另一為觀察不到，且不代表潛在特質，卻是研究者想要極力去避免或設法降低的部份，叫作「**誤差分數**」(error score)。這兩個部份合併構成任何一個真實的測量值（即實得分數），並且彼此之間具有及延伸出多種基本假設，能符合這些基本假設的測量問題，即為真實分數模式所探討的範疇。

根據古典測驗理論的假設，受試者所具有的某種潛在特質，無法單由一次測驗的實得分數來表示，它必須由受試者在無數次測驗上所得的實得分數，以其平均數來表示，該數值即是受試者的潛在特質之不偏估計值，即是前述的「真實分數」；真實分數的存在並不受測量次數的影響，它代表長期測量結果「不變」的部份。而實際上，單獨一次測量所得的實得分數，總會與真實分數間產生一段差距，這段差距即稱作「隨機誤差分數」(random error score)，或簡稱為「誤差」(error)；誤差分數深受測量工具之精確度的影響很大，它代表某次測量結果「可變」的部份。若以數學公式來表示，這兩種分數與實得分數間的關係可以表示如下：

$$\chi = \mathbf{t} + \mathbf{e}$$

其中，χ 代表實得分數， **t** 代表真實分數， **e** 代表誤差分數。

古典測驗理論即是建立在上述這種真實分數模式及其假設的基礎上，針對測驗資料間的實證關係，進行有系統解釋的一門學問。

壹、真實分數理論的基本假設及其結論

真實分數模式的成立，必須滿足一些基本假設，這些基本假設就是真實分數理論所賴以建立的基礎。

真實分數理論的基本假設，可以歸納成下列七項：

1. $\chi = t + e$ (即實得分數等於真實分數與誤差分數之和)；
2. $\mathrm{E}(\chi) = t$ (即實得分數的期望值等於真實分數)；
3. $\rho_{te} = 0$ (即真實分數與誤差分數之間呈零相關)；
4. $\rho_{e_1e_2} = 0$ (即不同測驗的誤差分數間呈零相關)；

5. $\rho_{e_1 t_2} = 0$ (即不同測驗的誤差分數與真實分數間呈零相關)；

6. 假設有兩個測驗，其實得分數分別為χ和χ'，並且滿足上述 1 到 5 的假設，且對每一群體考生而言，亦滿足 $t = t'$和$\sigma^2_e = \sigma^2_{e'}$ 等條件，則這兩個測驗便稱作「複本測驗」 (parallel tests)；

7. 假設有兩個測驗，其實得分數分別為χ和χ'，並且滿足上述 1 到 5 的假設，且對每一群體考生而言，亦滿足 $t_1 = t_2 + c_{12}$，其中 c_{12} 為一常數，則這兩個測驗稱作「本質上τ相等測驗」(essentially τ-equivalent tests)。

根據上述七個基本假設的數學公式所示可知，古典測驗理論對測量問題所持的觀點，可以做如下的詮釋：

1.假設具有潛在特質存在。

從第一個假設可知，測量必須要有對象，此對象即是我們所假定的潛在特質（亦即是 t 所代表者），它是看不見的東西，但我們必須先假設它的存在，如此才值得我們去測量它，若不先假設它是存在的，則我們的任何測量行為都將失卻目標，變得盲目無效。

2.多次測量的推論結果。

既然上述所假設的潛在特質是看不見的，因此，我們就無法直接進行測量它。我們僅能從數學觀點去假設它與我們從外觀測量得到的數據間具有某種數學關係（通常都假設成直線關係），為了釐清這種關係，通常需要使用多次的測量數據，再透過統計學的估算（如：求期望值），才能估計出這種潛在特質的量到底是多少，並且推論出它與外觀測量得到的數據間具有什麼關係。

3.單獨一次的測量必有誤差存在。

既然潛在特質是經由多次測量才推論得到，因此，單獨一次的測量結果，除了測量到所要測量的潛在特質外，也必定同時測量到誤差成份。但是，在經過多次的測量後，我們由上述說明所推論出來的結果將愈來愈接近真正的潛在特質，因此，這麼多次測量值所含的誤差分數也就可以彼此抵銷。這項結論也就是上述第一和第二個假設合併起來的推理結果。

4.假設潛在特質與誤差之間是獨立的。

第四個假設把測量問題單純化，僅假設潛在特質與誤差之間是獨立的。由於有這項假設存在，在測量時，我們不必考慮其他可能干擾測量結果的來源，僅將潛在特質以外的干擾，統統歸類到所謂的測量誤差（measurement errors），不再進一步細部分析，如此，可以把測量結果的推論問題單純化。附帶一提的是，這項假設亦延伸出第四和第五個假設；但是，這種把測量問題單純化的假設，卻是造成古典測驗理論飽受批評的地方。

5.複本測驗的嚴格假設。

古典測驗理論對測量結果的解釋和比較，是建立在複本測驗的嚴格假設上。換句話說，從第六和第七個假設可知，唯有滿足複本測驗之嚴格假設的兩個測量結果間，才可以直接進行比較大小和解釋優劣；若非滿足此假設，則任何兩次測量結果間的解釋和比較，均是無意義的。

根據上述的詮釋，從真實分數理論的基本假設可以推導出下列十八項結論，這些結論正是古典測驗理論的研究主題所賴以推理及演繹的依據：

1. $E(e)=0$（即誤差分數的期望值為零)；
2. $E(e,t)=\rho_{et}=0$（即誤差分數與真實分數之期望值為零)；
3. $\sigma^2_x=\sigma^2_t+\sigma^2_e$ (即實得分數的變異數等於真實分數的變異數與誤差分數的變異數之和)；
4. $\rho^2_{xt}=\sigma^2_t/\sigma^2_x$ (即實得分數與真實分數間之相關係數的平方等於真實分數之變異數和實得分數之變異數的比值)；
5. $\rho^2_{xt}=1-\sigma^2_e/\sigma^2_x$ (即實得分數與真實分數間之相關係數的平方等於1減去誤差分數之變異數和實得分數之變異數的比值)；
6. $\sigma^2_x=\sigma^2_{x'}$ (即複本測驗的實得分數之變異數相同)；
7. $\rho_{xy}=\rho_{x'y}$ (即複本測驗分數與另一變項分數間的相關係數相同)；
8. $\rho_{xx'}=\sigma^2_t/\sigma^2_x=\sigma^2_{t'}/\sigma^2_{x'}$ (即複本測驗分數間的相關係數等於其中一種測驗之真實分數變異數和實得分數變異數的比值)；
9. $\rho_{xx'}=1-\sigma^2_e/\sigma^2_x$ (即複本測驗分數間的相關係數等於1減去誤差分數之變異數和實得分數之變異數的比值)；
10. $\rho_{xx'}=1-\rho^2_{xe}$ (即複本測驗分數間的相關係數等於1減去實得分數與誤差分數間之相關係數的平方)；
11. $\rho^2_{xt}=\rho_{xx'}$ (即實得分數與真實分數間之相關係數的平方等於複本測驗分數間的相關係數)；
12. $\sigma^2_t=\sigma_{xx'}$ (即真實分數的變異數等於複本測驗的實得分數間之共變數)；
13. $\sigma^2_e=\sigma^2_x(1-\rho_{xx'})$ (即誤差分數的變異數等於實得分數的變異數乘以1減去複本測驗間之相關係數)；
14. $\rho_{t_xt_y}=\dfrac{\rho_{xy}}{\sqrt{\rho_{xx'}\rho_{yy'}}}$ (即　任兩個測驗的真實分數間之相關係數等於該二測驗的實得分數間之相關係數除以該二測驗之複本測驗相關係數的相乘積之開根號)；

15. $\sigma^2_{t_X} = N^2\sigma^2_{t_Y}$ (即如果X為N個複本測驗分數Y之和，則X的真實分數之變異數等於N平方倍之Y的真實分數之變異數)；

16. $\sigma^2_{e_X} = N^2\sigma^2_{e_Y}$ (即如果X為N個複本測驗分數Y之和，則X的誤差分數之變異數等於N平方倍之Y的誤差分數之變異數)；

17. $\rho_{xx'} = \frac{N\rho_{YY'}}{1+(N-1)\rho_{YY'}}$ (即如果X為N個複本測驗分數Y之和，則此為 Spearman-Brown 的折半信度公式)；

18. 如果 $\rho_{YY'} \neq 0$，則 $\lim_{n\to\infty}\rho_{xx'} = 1$ (即X和Y的定義同結論 15，如果 $\rho_{YY'}$ 不等於 0， 則 $\rho_{XX'}$ 的極限為 1)。

整個古典測驗理論便是以前述七項基本假設，和推導出的十八項結論為基礎，企圖去估計測驗內（或測驗間）實得分數與真實分數間的關聯強度，這些關聯強度亦即是該理論所要估計的各種可能信度係數，故古典測驗理論又有「古典信度理論」之稱。

除了信度估計之外，古典測驗理論也還探討其他有關聯的話題，例如：效度（validity）、測驗編製（test construction）、常模（norm）、測驗等化（test equating）、測驗偏差（test bias）、試題分析（item analysis）、精熟測驗（mastery testing）、適性測驗（adaptive testing）、題庫建立（item banking）、及其在社會科學研究上的應用課題等；這些課題都是根據它的基本假設和推論延伸而來，並且散見於專書、會議論文、和下列各種重要學術期刊：

1. **Annual Review of Psychology**
2. **Applied Psychological Measurement**
3. **The British Journal of Mathematical and Statistical Psychology**（早期刊名：**The British Journal of Statistical Psychology**）
4. **Educational Measurement : Issues and Practice**
5. **Educational and Psychological Measurement**
6. **Journal of Educational Measurement**
7. **Journal of Educational Statistics**
8. **Psychometrika**
9. 中華心理學刊
10. 測驗年刊
11. 測驗與輔導
12. 輔導月刊
13. 國內各大學相關學報及教育領域學術期刊

貳、古典測驗理論的優缺點

古典測驗理論的理論架構，主要是以真實分數模式為主，其理論模式的發展已為時甚久，且頗具規模，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情境，以及社會科學研究資料的分析，為目前心理計量學界應用與流通最廣的一種測驗理論。

然而，若從當代測驗理論（以「試題反應理論」為代表）的觀點來看，古典測驗理論除了具備上述各項優點外，卻含有下列諸項缺失：

1.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，會因為接受測驗的受試者樣本的不同而不同，因此，針對不同潛在特質的樣本，同一份測驗很難獲得一致的難度、鑑別度、或信度等指標。

2.古典測驗理論以一個共同的測量標準誤（standard error of measurement），作為每位受試者的潛在特質估計值的測量誤差指標；這種作法完全沒有考慮受試者反應的個別差異，對於具有高、低兩極端潛在特質的受試者而言，這種指標極為不合理且不精確，致使古典測驗理論模式的適當性受到懷疑。

3.古典測驗理論對於非複本（nonparallel），但功能相同的測驗所獲得之量數間，無法提供有意義的比較；有意義的比較僅侷限在相同測驗的前後測量之量數或複本測驗分數之間而已。

4.古典測驗理論對信度的假設，是建立在複本（parallel forms）測量概念的假設上；但是這種假設在實際的測驗情境裡，往往是不合理或不存在的。因為，在實際的測驗情境下，施測者不可能要求每位受試者在接受同一份測驗無數次後，而仍然保持每次反應結果都彼此獨立、互相不影響；況且，每一種測驗並不一定在編製測驗之時就同時製作複本。因此，複本測量的理論假設是行不通的，不論是從實際層面或方法學邏輯的觀點來看，它的假設既不切實際、又不合理、並且也是矛盾的。

5.古典測驗理論忽視受試者作答的試題反應組型（item response pattern）所代表的意義，對於在原始得分上相同的受試者或正確反應總和相同的試題，即看成是潛在特質（如：能力）或試題參數（如：難度）的估計值相同。這種觀點其實是不正確的，因為，總分相同的受試者或總和相同的試題，其試題反應組型不見得會完全一致，因此，試題反應組型所顯示的意義也不會相同，所估算出的潛在特質和試題參數估計值，應該也會不一樣。

由於古典測驗理論有上述諸項缺失，學者們為彌補這個理論上的缺失，乃轉向尋求理論與方法均較嚴謹的當代測驗理論，於是才會有日後的「試題反應理論」誕生。不過，由於古典測驗理論所採用的數學方法較為簡單易行，廣被中小學教師及一般大眾所能接受，在當今

實務應用方面，古典測驗理論的重要性仍佔有一席之地。

附錄　古典測驗理論的重要參考專書

Allen, M. J., & Yen, W. M. (1979). ***Introduction to measurement theory***. Monterey, CA: Brooks/Cole.

Anastasi, A. (1988). ***Psychological testing*** (6th ed.). New York: Macmillan.

Berk, R. A. (Ed.) (1980). ***Criterion referenced measurement: The state of the art***. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.) (1982). ***Handbook of methods for detecting test bias***. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.) (1984). ***A guide to criterion referenced test construction***. Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (Ed.) (1986). ***Performance assessment: Methods and applications***. Baltimore, MD: Johns Hopkins University Press.

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). ***Psychological testing: An introduction to tests and measurement***. Mountain View, CA: Mayfield.

Crocker, L., & Algina, J. (1986). ***Introduction to classical and modern test theory***. New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1990). ***Essentials of psychological testing*** (5th ed.). New York: Harper & Row.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). ***The dependability of behavioral measures: Theory of generalizability for scores and profiles***. New York: John Wiley & Sons.

Dick, W., & Hagerty, N. (1971). ***Topics in measurement: Reliability and validity***. New York: McGraw-Hill.

DuBois, P. H. (1970). ***A history of psychological testing***. Boston, MA: Allyn & Bacon.

Ebel, R. L., & Frisbie, D. A. (1991). ***Essentials of educational measurement*** (5th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Fan, C. T. (1952). ***Item analysis table***. Princeton, NJ: Educational Testing Service.

Gronlund, N. E. (1993). ***How to make achievement tests and assessments*** (5th ed.). Boston: Allyn & Bacon.

Gronlund, N. E., & Linn, R. L. (1990). ***Measurement and evaluation in teaching*** (6th ed.). New York: Macmillan.

Gulliksen, H. (1987). ***Theory of mental test***. Hillsdale, NJ: Lawrence Erlbaum Associates. (Originally published in 1950 by New York: John Wiley & Sons)

Haladyna, T. M. (1994). ***Developing and validating multiple-choice test items***. Hillsdale, NJ:

Lawrence Erlbaum Associates.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). ***Educational and psychological measurement and evaluation*** (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Jensen, A. R. (1980). ***Bias in mental testing***. New York: The Free Press.

Kaplan, R. M., & Saccuzzo, D. P. (1993). ***Psychological testing: Principles, applications, and issues*** (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kryspin, W. J., & Feldhusen, J. T. (1974). ***Developing classroom tests***. Minneapolis, Minn: Burgess.

Kubiszyn, T., & Borich, G. (1987). ***Educational testing and measurement: Classroom application and practice*** (2nd ed.). Glenview, IL: Scott, Foresman & Company.

Lindquist, E. F. (Ed.) (1951). ***Educational measurement***. Washington, DC: American Council on Education.

Linn, R. L. (Ed.) (1989). ***Educational measurement*** (3rd ed.). Washington, DC: American Council on Education.

Linn, R. L., & Gronlund, N. E. (2000). ***Measurement and assessment in teaching*** (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Lord, F. M., & Novick, M. R. (1968). ***Statistical theories of mental test scores***. Reading, MA: Addison-Wesley.

Mehrens, W. A., & Lehmann, I. J. (1991). ***Measurement and evaluation in education and psychology*** (4th ed.). New York: Holt, Rinehart & Winston.

Nitko, A. J. (1983). ***Educational tests and measurement***. New York: Harcourt Brace Jovanovich.

Nitko, A. J. (2001). ***Educational assessment of students (3rd ed.)***. Upper Saddle River, NJ: Prentice-Hall.

Noll, V. H., Scannell, D. P., & Craig, R. C. (1979). ***Introduction to educational measurement*** (4th ed.). Boston, MA: Houghton Mifflin.

Oosterhof, A. (2001). ***Classroom applications of educational measurement***（3rd ed.）. Upper Saddle River, NJ: Prentice-Hall.

Osterlind, S. J. (1998). ***Constructing test items: Multiple-choice, constructed-response, performance, and other formats***（2nd ed.）. Boston: Kluwer Academic Publishers.

Ory, J. C., & Ryan, K. E. (1993). ***Tips for improving testing and grading***. Newbury Park, CA: Sage.

Payne, D. A. (1992). ***Measuring and evaluating educational outcomes***. New York: Macmillian.

Popham, W. J. (1978). ***Criterion-referenced measurement***. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1990). ***Modern educational measurement: A practitioner's perspective*** (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1999). ***Classroom assessment: What teachers need to know***（2nd ed.）. Boston: Allyn & Bacon.

Priestly, M. (1982). ***Performance assessment in education and training: Alternative techniques***. Englewood Cliffs, NJ: Educational Technology Publications.

Sax, G. (1989). ***Principles of educational and psychological measurement and evaluation*** (3rd ed.). Belmont, CA: Wadsworth.

Stiggins, R. J. (1994). ***Student-centered classroom assessment***. New York: Macmillan.

Stiggins, R. J., & Conklin, N. F. (1992). ***In teacher's hands***. Albany, NY: State University of New York Press.

Suen, H. K. (1990). ***Principles of test theories***. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (1991). ***Measurement and evaluation in psychology and education*** (5th ed.). New York: Macmillan.

Tindal, G. A., & Marston, D. B. (1990). ***Classroom-based assessment***. Columbus, OH: Charles E. Merrill.

Wainer, H., & Braun, H. I. (Eds.) (1988). ***Test validity***. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wiersma, W., & Jurs, S. G. (1990). ***Educational measurement and testing*** (2nd ed.). Boston: Allyn & Bacon.

Worthen, B. R., Borg, W. R., & White, K. R. (1993). ***Measurement and evaluation in the schools***. New York: Longman.