教育研究資訊，1997.3，5(2)，頁 21-30

上限水準的設定與測驗信度關係之研究

林惠芬

本研究的主要目的係探討以連續錯二題、三題、四題和五題爲上限水準以及傳統不設定上限水準的計分方式對於「國民中學新生數學能力測驗」數與式（30題），測量（20題）以及幾何、機率和統計（30題）三個分測驗信度估計值的影響。研究樣本爲國中一年級新生1025位。依全體受試者的作答情形，試題由易而難排序後，以隨機方式將全體受試者分爲10組，每組各100名。在每一分測驗裡，依不同上限水準設定方式算出其內部一致性Alpha信度估計值，並轉換爲Zr值。然後以Dunnett事前比較分析法以及重複量數變異數分析法進行資料分析。研究結果指出在不同上限水準的設定之下，測驗信度估計值有所不同，其中以連續錯二題爲上限水準的信度估計值爲最大，而傳統不設定上限水準的信度估計值爲最小，此外，隨著上限水準設定的放寬，信度估計值逐漸遞減。本研究並提出未來研究之建議及方向。

關鍵詞：信度係數、上限水準、成就測驗、常模參照測驗。

Keywords: reliability, ceiling levels, scoring methods

壹、緒論

一般在評鑑一份測驗適用與否，除了會探討其效度之外，信度也是一項重要的考量要項。依據古典眞分數理論(classical true score theory)，所謂信度係指眞實分數(true score)與實得分數(obtained score)變異數的比值(Allen & Yen, 1979)。而實得分數是眞實分數與誤差之和。

因此，當實得分數不變時，眞實分數愈大，亦即誤差愈小，則信度值會愈大；反之則信度值愈小(Gullisken, 1950; Allen & Yen, 1979)。然而，Pedhazur(1982)曾指出有測量便會有誤差的存在，這是一種慣例而非特例，也因爲如此，一般的測驗使用者均能滿足使用信度不高的評量工具(Hamblton, van der linden, 1982)。不過，由於在教學過程中，對於學習者各項學習能力的瞭解是件相當重要的工作，因此測驗編製者在編製測驗時，莫不以探討如何提高測驗的信度爲重要任務。

一般而言，提高測驗信度的方法有：(1)增加測驗的長度亦即增加測驗試題，以便試題更具代表性；(2)增加受試者在測驗上的變異性，以及(3)調整測驗試題的難度，多採用難易適中的試題（葛樹人，民79；郭生玉，民78；陳英豪、吳裕益，

民79)。

在重測信度、複本信度以及內部一致性信度三種信度估計方法中，以內部一致性信度的求法最能反應出試題的難易程度，以及試題與整個測驗間的關係。從內部一致性信度可以得知，測驗信度與測驗分數的變異量有直接關係存在，當測驗分數的變異量增大時，測驗信度便會因而提高，而測驗分數的變異量是由各試題的變異量(item variance)和試題間的共變量(interitem covariance)所組成。因此欲提高測驗信度時，便須提高各試題的變異量或試題間的共變量(Kuder, Richardson, 1937; Ferguson, 1981; Hambleton & Swaminathan, 1985; Ferguson, Takane, 1989)。

是故，以提高試題變異量而言，大部份的學者均認爲試題難度約爲.50或界於.30~.70之間，最具鑑別度，亦最能提高受試者在測驗的變異性。而以提高試題間的共變量而言，便是在編製測驗時，所使用的試題以具有高相關及高共變量的試題(higher interitem covariance and correlation)爲主(Ferguson, Takane, 1989)。

依據Ferguson(1981)、Ferguson, Takane (1989)，試題間的共變量(r ijSiSj)係等於受試者在某兩個試題均答對的百分比(Pij)減去分別在此兩試題答對百分比的乘積(Pipj)。當受試者的作答反應(item response)與試題難易呈一致關係時，則試題間的共變量會增加，亦即當試題簡單時，受試者答對該題，當試題較難時，受試者答錯該題，此時，試題的共變量會增加，信度值也會因而提高。此外，Horst(1966)、Terwilliger和Lele(1979) 更進一步指出，試題間的共變量爲最大(maximun)是在於當答對較難題目的百分比(Pj)減去與答對較簡單之試題百分比的乘積(PiPj)（在此，第j題較第i題爲不易作答）。因此爲了要得到最大之試題間共變量，試題便須由易而難依次排序，然後再從中找出最難的題目，使其與其它各題目間的共變量爲最大(Suriyawongse ,1987)。從測驗的觀點來看，答對此題目便是受試者的最大表現，則其得分應至此題爲限，否則便會有誤差出現。

因此，有些測驗編製者便從設定上限水準的方式來提高測驗的信度，Paraskevopolus和Kirk(1969)便曾指出設定基礎水準(basal level)和上限水準可以提高測驗信度。目前許多標準化測驗，例如魏氏智力測驗(WISC-R)，伊利諾心理語言測驗(ITPA, Paraskevopolus, Kirk, 1969)，非語言智力測驗(Test of Non-verbal Intelligence Test, TONI, Brown, Sherbenou, Johnson, 1990)，畢堡德個別成就測驗(Peabody Individual Achievement test, PIAT)……等測驗在計分時，均設有其上限水準(ceiling level)，然而上限水準的設定，因著不同的測驗有不同的設定標準，同時某些測驗，其不同分測驗有不同的上限水準，以國內「新編中華智力量表」（國立台灣範大學特殊教育研究所，民82）爲例，其上限水準設有連續錯2題、3題、4題及5題；伊利諾心理語言測驗(Paraskevopalus, Kirk, 1969)，其上限水準包括沒有特別設定、連續錯2題、3題、6題以及7題中連續錯3題；非語文智力測驗

(Brown, Sherbenou, Johnson, 1990) 則是 5 題內連續錯 3 題爲其上限水準。但這些測驗爲何會如此設定，在使用手册上並無清楚的說明及解釋。

Browing, Salvia 和 Yesseldyske (1979) 和 Suriyawongse (1987) 曾探討上限設定與測驗信度的關係。Browning 、Salvia 和 Yesseldyske(1979) 探討五種計分方式對於畢堡德個別成就測驗 (PIAT) 的數學和閱讀兩份分測驗在平均數、變異數以及內部一致性的影響。其結果顯示在數學分測驗裡，其內部一致性信度值會因著試題的排序，以及設定基礎水準或上限水準而有影響，但是在閱讀分測驗因試題是填充題的題型，受試者較無猜測機會，而對測驗的信度沒有影響，對受試群體得分變異量的影響亦相當有限。

Suriyawongse (1987) 以連續答錯一題、二題、三題、四題、五題以及沒有特別設定的傳統計分方式探討對於 Iowa Test of Basic Skill (ITBS) 信度值的影響，其結果指出，傳統的計分方式均較以連續答錯一題、二題、三題、四題及五題爲上限的計分方式，所估計信度值來的低，而連續答錯 2 題爲上限水準時，其信度值爲最高。但是不同的測驗，不同的受試者，其結果是否仍是如此？Suriyawongse (1987) 指出今後的研究宜繼續再做這方面的探討。目前國內所編製之測驗大都以傳統的方式計分，並未考慮特別設定上限上水準，同時現階段國內很少有研究者探討上限水準的設定對於測驗信度的影響。因此，到底上限水準的設定是否會提高測驗信度？以及那一個上限的設定會較爲恰當？均值得研究。

一、研究目的

本研究係探討有設定上限水準的計分方式與傳統不設定上限水準的計分方式對於測驗信度的影響，以瞭解上限水準的設定和測驗信度間之關係，同時並作爲測驗編製者在編製測驗時是否設定上限水準的參考。具體而言，本研究之目的有二：

㈠探討有設定上限水準的計分方式與傳統不設定上限水準的計分方式，其測驗信度值的差異情形。

㈡探討不同的上限水準計分方式，其測驗信度值的差異情形。

二、待答問題

茲依據上述之研究目的，提出下列待答問題：

㈠有設定上限水準的計分方式與傳統不設定上限水準的計分方式，其測驗信度值是否有統計上顯著差異？

㈡在不同的上限水準設定之下其測驗信度值是否會有統計上顯著差異？

三、名詞詮釋

㈠上限水準：係指在一份選擇題型的測驗中，受試者連續答錯多少題後，後面的試題縱使答對，亦不予計分。本研究所使用之上限水準有四種：連續答錯二題、三題、四題，以及五題。

㈡信度：係指測驗分數一致性的程度。在常模參照測驗裡，信度包括重測信度、複本信度和內部一致性信度三種。本研究所指之信度係指Cronbach α 內部一致性信度值。

貳、研究結果

㈠有設定上限的計分方式與傳統不設定上限水準的計分方式，其測驗信度估計值上的差異考驗情形。

表 4-1 爲以 Dunnett 事前比較分別分析以連續錯二題、三題、四題和五題和傳統不設定上限水準的計分方式，其測驗信度估計值在數與式、測量以及幾何、機率和統計三個分測驗的差異考驗結果。

由表 4-1 的資料顯示在三個分測驗裡以連續錯二題、三題、四題及五題爲上限水準的計分方式，其測驗信度估計值 Zr 的平均數均大於傳統不設定上限水準的計分方式，且其差異達統計顯著水準 (P<.001)。

表 4-1　不同上限設定法與傳統計分法在各分測驗信度估計值 Zr 之 Dunnett 事前比較分析摘要表

	數與式			測量			幾何、機率與統計		
	平均數	標準差	t-值	平均數	標準差	t-值	平均數	標準差	t-值
錯二題	1.626	0.051	45.536**	1.728	0.045	38.170**	1.627	0.045	38.161**
傳統	0.977	0.058		1.159	0.028		0.991	0.047	
錯三題	1.568	0.051	39.646**	1.582	0.029	28.376**	1.546	0.064	33.301**
傳統	0.977	0.058		1.159	0.028		0.991	0.047	
錯四題	1.497	0.050	33.675**	1.446	0.049	19.253**	1.479	0.050	33.675**
傳統	0.977	0.058		1.159	0.028		0.977	0.058	
錯五題	1.372	0.05	26.498**	1.35	0.057	12.813**	1.347	0.055	21.361**
傳統	0.977	0.058		1.159	0.028		0.991	0.047	

** P< .001

㈡以連續錯二題、三題、四題和五題爲上限水準的計分方式，其測驗信度值間的差異考驗

表 4-2 爲以連續錯二題、三題、四題和五題爲上限水準的計分方式在「國民中

學新生數學能力測驗」三個分測驗的變異數分析結果。

表 4-2　上限水準設定法在「國民中學新生數學能力測驗」信度估計值 Zr 之變異數分析摘要表

分測驗	變異來源	離均差平方和	自由度	均方	F 值
數與式	上限設定法	0.37	3	0.12	150.27*
	樣本人數	0.07	9	0.01	
	殘差	0.02	27	0.0007407	
測　量	上限設定法	0.81	3	0.27	221.75*
	樣本人數	0.04	9	0.0044444	
	殘差	0.03	27	0.0011111	
幾何、機率與統計	上限設定法	0.44	3	0.15	93.31*
	樣本人數	0.07	9	0.01	
	殘差	0.04	27	0.0014814	

* P<.001

由表 4-2 的資料顯示，在三個分測驗裡，各上限水準設定法的主要效果均達統計顯著水準(P< .001)。本研究乃繼續以薛費多重比較進行事後比較考驗，以進一步瞭解各上限水準設定法所估計之測驗信度平均值的差異情形。

由表 4-3 的資料顯示，在三個分測驗裡，各上限水準設定法所估計之測驗信度平均值 Zr 的差異彼此間均達統計顯著水準(P<. 05)，且其中均以連續錯二題爲上限水準所估計的測驗信度值爲最大，然後隨著錯的題數的增加，其測驗信度平均值 Zr 逐漸遞減。

表 4-3　上限水準設定法在「國民中學新生數學能力測驗」信度估計值 Zr 之事後比較分析表

分測驗	上限設定法	連續錯	2 題	3 題	4 題	5 題
數與式		平均數	1.626	1.568	1.479	1.372
	連續錯 2 題	1.626	—	0.058*	0.147*	0.254*
	3 題	1.568		—	0.089*	0.196*
	4 題	1.479			—	0.107*
	5 題	1.372				—
測　量		平均數	1.728	1.582	1.446	1.350
	連續錯 2 題	1.728	—	0.146*	0.282*	0.378*
	3 題	1.582		—	0.136*	0.232*
	4 題	1.446			—	0.096*
	5 題	1.350				—
幾何、機率與統計		平均數	1.627	1.546	1.453	1.347
	連續錯 2 題	1.627	—	0.081*	0.174*	0.280*
	3 題	1.546		—	0.093*	0.199*
	4 題	1.453			—	0.106*
	5 題	1.347				—

參、結論與建議

一、研究發現

經由研究資料分析的結果，本研究的發現如下：

1. 以連續錯二題、三題、四題、五題爲上限水準的計分方式與傳統不設定上限水準的計分方式，其所估計的測驗信度平均值有差異，且其差異達統計顯著水準。
2. 在有設定上限水準的計分方式裡，以連續錯二題所估計之測驗信度值最大，然後測驗信度值隨著錯的題數愈多，其值愈低，且彼此之間的差異達統計顯著水準。

二、討論

由本研究之研究結果及發現，現就有關事項討論如下：

㈠設定上限水準對測驗信度值的影響

依據本研究的研究結果，在「國民中學新生數學能力測驗」的三個分測驗裡，在不同的上限水準設定之下，其信度估計值均不同，且也與傳統不設定上限水準的測驗信度估計值不同。Browning, Salvia 和 Yesseldyske (1979) 曾指出測驗的信度值會受到是否有設定上限水準，以及上限水準設定的不同而有所影響，本研究的研究結果再次證實 Browning 等人的論點。因此，爲提高測驗的信度，標準化的常模參照測驗，是宜考慮設定上限水準。

㈡何種上限水準的設定，其測驗信度值是最大？

依據本研究的研究結果，連續錯二題爲上限水準，其測驗信度估計值爲最大，且隨著上限水準的提高，其信度估計值逐漸減小，同時，以傳統不設定上限水準的信度值爲最小。在 Suriyawongse (1987) 的研究裡亦是有相同的結果。此兩個研究在不同樣本，不同試題，但研究設計相同的情形下，其結果是相同，似乎傳達著下列的訊息：

1. 在衆多的計分方法中，以連續錯二題爲上限水準是爲較恰當的計分方式。
2. 從誤差的觀點來看，當試題由易而難排序時，錯一題也許是粗心，看錯答案或寫錯答案；連續錯二題爲上限水準，則可能代表著受試者的能力是至此爲止；而以連續錯三題或三題以上才停止計分，則有可能包含如猜測之類的誤差，此種誤差的存在便會降低測驗信度的估計值。因此，爲提高測驗的信度，是宜設定上限水準，以減少測驗誤差的存在。而在設定上限水準時，以連續錯二題爲上限水準似較爲恰當，這是因爲 Suriyawongse (1987) 和本研究的結果均有相同的發現。

二、結論

由上述發現，本研究之結論爲：

測驗信度的估計值隨著不同上限水準的設定而有所不同，其中以連續錯二題的上限水準的計分方法，其測驗信度估計值爲最高，然後隨著錯的題數的增多，而測驗信度估計值逐漸遞減，同時以傳統不設定上限水準的計分方法，其測驗信度值爲最小。因此，上限水準的設定對於測驗信度是有影響，而且在各上限水準設定中，以連續錯二題爲計分方式的測驗信度值爲最高。

三、建議

茲就本研究所得的研究結果以及研究者所做之結論，提出有關在編製和使用測

驗以及未來研究上的一些建議。

㈠在測驗編製和使用方面

1. 高信度估計值是一份優良測驗必備的條件之一，也是每位測驗編製者在編製測驗時，期望達到的目的之一。由本研究的結果得知，設定不同的上限水準會影響測驗信度值，是故，在建立信度的過程中，測驗編製者宜就不同的上限水準的設定對該份測驗的影響進行探討及分析，以期透過設定適當的上限水準來提高測驗的信度估計值。同時在有關的使用手冊上亦應註明其上限水準設定的原由，俾利使用者瞭解其中的道理。

2. 測驗使用者在選用有關測驗工具時，可就該份測驗是否設有上限水準做爲選用的條件，因爲，由本研究結果得知，設有上限水準的測驗，其信度估計值較高，而信度值高的測驗，其測量標準誤會較小，對於受試者學習能力的瞭解會較正確。此外，設有上限水準的測驗，在施測及計分上亦較節省時間，是故，不論就測驗的品質或施測的方便性來考量，測驗使用者是應考慮使用設有上限水準的測驗。

㈡未來研究方面

1. 本研究係以1025位國中一年級新生在「國民中學新生數學能力測驗」的作答情形爲分析重點；不同的受試者，不同的測驗，其結果是否仍是如此？宜再進一步探討。

2. 本研究係探討以連續錯二題、三題、四題和五題爲上限水準的計分方式對測驗信度的影響。然而，有許多標準化的測驗裡，除了上限水準的設定外，亦有基礎水準(basal level)的設定，同時在上限水準的設定裡，有的是以多少題裡錯幾題爲上限(例如：5題內錯3題)。在這些考量因素之下，其對信度的影響又是如何？值得進一步研究。

3. 本研究係以1025位受試者的作答情形計算每一試題的難易度值，然後再將全體受試者分爲10組進行分析探討。其結果是以連續錯二題的信度值爲最高；但是若是先將受試者分組後，再分別以各組受試者作答情形來分析，則在考慮各組的差異之下，其結果是否仍然相同？值得再進一步研究。

參考書目

林清山（民81），**心理與教育統計**。台北：東華書局。

吳裕益（民82），**國小高年級學術性向測驗編製第一年報告**。台北：八十二年度心理與教育測驗學術研討會。

周台傑、范金玉（民76），**國民小學數學能力發展測驗指導手冊**。彰化：國立台灣教育學院特殊教育研究所。

周台傑、巫春貴（民81），**國中新生數學能力測驗指導手冊**。彰化：精華印刷企業社。

周台傑（民81），**國民小學國語文成就測驗指導手冊**。彰化：精華印刷企業社。

徐享良、曾秀錦（民82），**國民中學新生自然科學能力測驗指導手冊**。彰化：精華印刷企業社。

郭生玉（民78），**心理與教育測驗**。台北：精華出版社。

陳英豪、吳裕益（民79），**測驗與評量**。高雄：復文書局。

國立彰化師範大學特殊教育學系（民83），**國民中學語文智力測驗指導手冊**。彰化。

葛樹人（民79），**心理測驗學**。台北：桂冠書局。

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

Brown, L., Sherbenou, R. J., Johnson, S. K. (1990). ***Test of Nonverbal Intelligence. Second edition: A language-free measure of cognitive ability.*** Austin, TX: Pro-ED..

Browning, R., Salvia, J., & Yesseldyske, J. E. (1979). Technical characteristics of the Peabody Individual Achievement Test as a function of item arrangement and basal and ceiling rules. ***Psychology in the School,*** 16, 4-7.

Cohen, J. (1988). ***Statistical power analysis for the behavioral sciences.*** (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates Perb.

Ferguson, G. A. (1981). ***Statistical analysis in psychology and education*** (5th ed.). New York: McGraw -Hill.

Ferguson, G. A., Takane, Y. S. (1989). ***Statistical analysis in psychology and education.*** (6th ed.). New York: McGraw-Hill Book Co..

Gulliksen, H. (1950). ***Theory of mental tests.*** New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1985). ***Item response theory: Principles and applications.*** Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K., & Van der Linder, W. J. (1982). Advances in item response theory and application: An introduction. ***Applied psychological measurement, 6,*** 373-378.

Horst, P. (1966). ***Psychological measurement and prediction.*** Belmont: Wadsworth Publishing.

Kirk, R. E. (1982). ***Experimental design: Procedures for the behavioral sciences.*** Monterey: Books/Cole Publishing.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimate of test reliability. ***Psychometrika, 2,*** 151-160.

McNemar, Q. (1962). ***Psychological statistics.*** (3rd ed.). New York: Wiley.

Paraskevopoulous, J. N., Kirk, S. A. (1969). ***The development and psychometric characteristics of the revised Illinois Test of Psycholinguistic Abilities.*** Urbana: Univesity of Illinois Press.

Pedhazur, E. J. (1982). ***Multiple regression in behavioral research*** (2nd ed.). New York: Holt, Rine-hart and Winston.

Salvia, J. & Yesseldyske, J. E. (1988). ***Assessment in special and remedial education.*** (4th ed) . Boston: Houghton Mifflin.

SPSS Inc. (1992). ***SPSS/PC + base system user's guide version 5.0.*** Chicago: Spss Inc.

Suriyawongse, S. (1987). ***A comparison the effects of different sizes of ceiling rules on the estimates of reliability of a mathematics achievement test,*** Doctoral dissertation, North Texas State University.

Terwilliger, J. S. & Lele, K. (1979). Some relationships among internal consistency, reproducibility, and homogeneity. ***Journal of Educational Measurement*** , **16**, 101-108.

林惠芬，國立彰化師範大學特教系教授