國民教育研究學報，第 8 期：1-24
民 91 年國立嘉義大學國民教育研究所

量化研究的品管：統計考驗力與效果值分析

李 茂 能

嘉義大學教育學院國教所教授

摘　要

本文旨在剖析樣本規劃、臨床或實用顯著性與統計顯著性的密切關係，以提昇量化研究的品質。首先點出與第一類型錯誤、第二類型錯誤、統計考驗力、P 值、及樣本大小有關之迷思、誤用或誤解。其次，探討統計考驗力與效果值分析的意義和功能，同時舉一實例說明如何透過統計考驗力與效果值分析，提昇量化研究的品質。接著，說明研究者進行樣本規劃並非難事，只要透過 Cohen 氏查表法與統計考驗力分析軟體 G*Power，即能快速達成。文末，試著針對論文編審提出量化研究品管的具體方法。

關鍵詞：統計考驗力分析，效果值，量化研究

壹、問題背景

Fisher 氏的『顯著性考驗』(significance testing)與 Neyman-Pearson 氏的假設考驗(hypothesis testing)在紛擾了一、二十年之後，雖然受到嚴厲的挑戰(Carver, 1978; Rozeboom, 1960)，終於 1980 年代末期確定了『統計顯著性考驗』的一致運作模式(Nix & Barnette, 1998)。這種虛無假設的顯著性考驗之探究模式，似已成爲科學研究不可或缺的下決策工具。它是客觀的表徵，也是量化研究的表徵，一直給人的印象是只談第一類型錯誤α，避談第二類型錯誤β與統計考驗力(power)、只關心 P 值是否小於α，不在乎效果值(effect size)是否具有應用價值(Rojewski, 1999; Thompson, 1996 , 1998)。這些似是而非的迷思，導致不少的困惑或誤解：

一、許多研究者，常把虛無假設考驗的統計顯著性視爲研究價值的指標(李茂能，民87; 張德榮，民71)。當研究結果達到預期的統計顯著水準(例如α=.05)，即下結論說研究結果在臨床上或實際應用上，具有顯著效果(Cohen 1977 , 1988; Daniel, 1998; Johnson, 1999; Thompson, 1998)。

二、不少研究者認爲犯第一類型錯誤較嚴重必須控制，犯第二類型錯誤可以不管，而使用了統計考驗力很低的統計考驗，許多相互矛盾的研究結論因而產生(Aron & Aron, 1999/2000; Borenstein, 1994 & 1997; Bezeau & Graves, 2001; Helberg, 1996)。

三、對於實驗效果的考驗，因襲統計量的假設考驗，忽視了區間估計的效用性，而導致臨床應用上二分法(全有效或全無效)的誤導(Borenstein, 1994 , 1997; Bezeau & Graves, 2001)。

四、很多研究者於解釋P值時，常把P值看成效果值的代名詞而有不當之詮釋，導致在解釋p值時，產生很多的誤用與迷思。例如，當P值小於.05，則宣稱達到『顯著水準』，P值等於.01，宣稱達『非常顯著水準』；P值等於.001，則宣稱達到『極顯著水準』(Cohen,1990)；而當P值等於.054時，則解釋爲『接近顯著水準』(Daniel, 1998; Thompson, 1998)。

五、使用大樣本永遠比使用小樣本好(Nix & Barnette, 1998)，以追求統計上的顯著性。

上述這些迷思與誤解易導致統計顯著考驗的誤用，以致於研究結果的誤判。我國量化研究正值高度量產期，雜誌或刊物的主編或研究機構的論文審查者，實應正視這

些令人憂心的議題，以導正量化研究的方向與品質。緣此，本文試圖拋磚引玉，盼更多的研究者能警惕到統計考驗力分析(power analysis)與效果值分析的重要性。

量化研究品質的提昇，除了取樣需具代表性與蒐集資料的工具信度、效度佳之外，端賴統計考驗力分析與效果值分析。抽樣與工具的問題在一般的研究法書籍中已有充分說明，不在此贅言。以下爲使一般讀者更易了解『統計顯著性考驗』與樣本大小、α、β、P 值和效果值間的錯綜複雜關係，先就此做一簡要說明。接著，分析國外對於統計考驗力的研究結果，其次，再分析統計考驗力在樣本規劃上與效果值分析在結果解釋上的重要性，以及如何利用查表或電腦分析軟體 G*POWER(參見圖 3 與圖 4)去規劃樣本大小。此外，並論及統計考驗力分析與效果值分析在論文結論中的應用方法與要點。文末，提出量化研究品質管制的途徑與具體作法。

貳、統計考驗力與 α 、β 之關係

第一類型錯誤α(拒絕虛無假設所犯的錯誤)與第二類型錯誤β(接受虛無假設所犯的錯誤)互爲消長，一般教育與心理研究者均視α爲比較嚴重，需控制在.05 以下，而β則希望控制在.20 以下。統計考驗力是指對立假設爲真時，研究結果可達到統計顯著性的機率(等於 1-β)。因此，設定的α愈小，則統計考驗力也愈小，一般均希望統計考驗力能大於.80，但不要過大(>.95 以上)。因此，統計考驗力與β成反比，而與α成正比。如同時要獲得『低的α與高的 power』，增大樣本是唯一途徑。

參、統計考驗力與樣本大小之決定

統計考驗力與樣本大小具有非常密切的關係。例如，此密切的關係可從下列 t,χ^2 與 F 三種重要統計公式中推知：在 t 考驗中假如分子保持恆定，樣本平均數的抽樣標準誤會因 n 逐漸增大而使 t 值變大，在χ^2考驗中假如分母保持恆定，χ^2值亦會因 n 逐漸增大而使χ^2值變大。同樣的，在 F 考驗中通常分子(組間)的自由度(df_1)並不大，假如分母的自由度逐漸增大，定會使 F 值逐漸變大。由此觀之，當樣本大小趨於無限大時，這些統計量不管其實際之效果值有多小，勢必會達到統計上的任何顯著水準，當研究者使用的樣本很小時，這些統計量不管其實際之效果值有多大，勢必永遠無法達到統

計上的任何顯著水準。

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$F = \frac{\frac{\chi_1^2}{df_1}}{\frac{\chi_2^2}{df_2}}$$

因此，樣本大小的規劃需靠統計學者與學科專家共同合作，才能訂出合情合理的估計值。假如樣本太小，易導致統計考驗力下降，錯失發現重要的研究結果；假如樣本過大，易導致統計考驗力過當(>.95)，因而致使沒有實質效益的效果亦達到統計上的顯著水準(Nix & Barnette, 1998)。樣本太小或太大均易導致(1)金錢、時間與資源的不必要浪費，或(2)使受試者遭受不必要的實驗傷害。不過，樣本大小的估計常因實驗性質不同，而有不一樣的考慮因素；例如，在醫療實驗上，受試者常易遭致潛在的危險，研究者最好使用小樣本的多次實驗，而農業實驗上，則因實驗期可能很長(如造林)，爲避免效益嚴重損失，可考慮使用較大的樣本進行實驗(Lenth, 2001)。研究者如欲提昇統計考驗力，尚可考慮其它途徑，如使用變異較小的母群爲研究對象、使用信、效度較佳的工具、使用單側考驗、使用較敏感的母數統計考驗、或增加實驗強度等(Aron & Aron 1999/2000)。

肆、效果值的意義、計算與解釋

缺少效果值，統計考驗力分析是無法進行的。因此，在研究過程中，效果值的估計是不可或缺的。效果值(effect size)可視爲在母群中，H_0與H_1的距離指標，它代表處理效果的大小。它可分爲兩類：(1)自變項與依變項關連指標，例如：Cohen's f^2, r, R^2, η^2，(2)效果值指標，例如：Cohen's d , f, & w, Glass' Δ, 與 Hedges' g。前者係變異導向的效果值係數，後者可爲原始分數的差異分數，亦可轉化爲標準化的差異分數。研究者在資料蒐集前，所預估的效果值，乃是他認爲比此值更大的效果才具有應用價值的臨界值，亦即最起碼的效果值。效果值的設定最好根據過去的研究結果、或根據理論對於效果值的上下限間，做明智的推估、或進行前導性 Pilot 研究再加以

確定。假如這些途徑皆不可行，Cohen (1977, 1988)根據經驗法則所提出的權宜措施，訂定了大、中、小效果值之判定標準(摘述如表 1)亦可作爲社會科學研究者參考之依據，詳細之計算公式參見下節說明。不過，Lenth(2001)建議避免使用這些套裝效果值，因爲它們可能不適用於某些特定的研究領域。一般教育或心理研究者如仍無法估計效果值之大小，可暫定爲中效果，事後再檢討之。一般在 SPSS 或 SAS 的變異數分析報表中，則以 eta-squared(η^2)或 partial η^2 來表示樣本效果值的大小，以ω^2 表示母群效果值，ω^2 永遠比η^2 來的小。另外，Schwarzer (1989)所研發的免費整合分析(Meta-analysis) 軟體，可協助研究者計算與轉換各種效果值，堪稱便利，值得讀者下載參考。

表 1

Cohen 氏大、中、小效果值之判定標準

指標		效果值		
		小	中	大
平均數的 t 考驗	d	.20	.50	.80
相關係數的 t 考驗	r	.10	.30	.50
ANOVA 的 F 考驗	f	.10	.25	.40
MCR 的 F 考驗	f^2	.02	.15	.35
χ^2 考驗	w	.10	.30	.50

伍、非中心性參數與統計考驗力、效果值與樣本大小之關係

非中心性參數(Noncentrality Parameter，簡稱 NP)與統計考驗力、效果值與樣本大小具有密切關係。t 分配的 NP 稱爲 delta(δ)，在 F 分配與χ^2分配中稱爲 lambda(λ)。NP 爲效果值與樣本大小之函數，而統計考驗力又爲 NP、自由度與顯著性臨界值之函數。以下將統計方法中常用的δ與λ的計算公式扼要加以說明，讀者如欲獲得更詳細的內容，請參閱統計考驗力分析軟體 G*POWER 的操作手冊(Erdfelder, Faul, & Buchner,

1996)或 Cohen(1988)所出版的統計考驗力分析專書。

一、雙樣本的t考驗

$$\delta = d \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

二、相關係數的t考驗

$$\delta = \sqrt{\frac{\rho^2}{1 - \rho^2}} \times N$$

三、其它的t考驗

$$\delta = f \times \sqrt{N}$$

這類分析包含重複量數 t 考驗、單一樣本 t 考驗、與 z 考驗；上式中 f 的計算公式，按序分別說明如下：

$$f = \frac{\mu_y}{\sigma_y}, \quad y = x1 - x2$$

$$f = \frac{|\mu - c|}{\sigma}, \quad c : \text{常數}$$

$$f = \frac{|\mu_1 - \mu_2|}{\sigma}$$

四、F考驗(含ANOVA, MCR等)

$$\lambda = f^2 \times N = df_b \times F$$

$$f^2 = \frac{\eta^2}{1 - \eta^2} \qquad \eta^2 = \frac{F \times df_{between}}{F \times df_{between} + df_{within}}$$

五、其他 F 考驗

適合進行多變項分析及重複量數分析。以多變項分析爲例，

$$\lambda = s(h) \times N \times f^2$$

$$f^2 = \frac{V(h)/s(h)}{1 - V(h)/s(h)} = \frac{V(h)}{s(h) - V(h)}$$

上式中 V(h)爲待考驗效果的 Pillai-Bartlett 值，s(h)爲待考驗效果的自變項

數或依變項數的最小值，V(h)/s(h)的值介於 0 到 1 之間，相當於多變項的 R^2 或 η^2。

六、χ^2考驗(含適合度與獨立性考驗)：

$$w = \sqrt{\sum_{i=1}^{m} \frac{(P_{0(i)} - P_{1(i)})^2}{P_{0(i)}}}$$、m 表 cell size

$$\lambda = w^2 \times N$$

上式中 $P_{0(i)}$與 $P_{1(i)}$分別爲虛無假設的 P_0與對立假設的 P_1在細格中的機率。

根據前述的 NP 參數，配合 SPSS 的內建反分配函數 IDF.T、IDF.F 與 IDF.CHISQ 所求得三種分配的顯著臨界值 q，再帶入 SPSS 內建的非中心性分配 NCDF.T、NCDF.F 或 NCDF.CHISQ 函數，即可間接求得各種統計考驗力。以 F 考驗爲例，首先利用 IDF.F 函數計算 F 反分配的臨界值 q，帶入公式：1-NCDF.F(q,df1,df2,NCP)，即可求得統計考驗力。SAS 亦有類似的函數可以運用，例如，使用 SAS 的 FINV 與 ProbF 函數，亦可間接求得 F 考驗的統計考驗力。讀者如有興趣，可以將 SPSS GLM 報表上的 NP 參數，驗證一下是否可以得到報表上所列的事後統計考驗力(Observed Power)。

陸、統計考驗力的分析

Cohen(1962)率先針對 1960 年變態心理學雜誌(The Journal of Abnormal Psychology)，調查刊內各篇論文的統計考驗力，發現小效果值的平均統計考驗力僅爲.18，中效果值的平均統計考驗力僅爲.46，大效果值的平均統計考驗力爲.83，亦即要這些論文的中效果值能達到統計顯著水準的機率還不到一半。統計考驗力分析始受到高度重視。其後，Chase & Chase(1977) 針對 1974 年應用心理學雜誌(Journal of Applied Psychology)，調查 121 篇論文的統計考驗力，發現小效果值的平均統計考驗力僅爲.25，中效果值的平均統計考驗力僅爲.66，大效果值的平均統計考驗力爲.84。統計考驗力偏低的情況並未改善多少。Sedlmeier & Gigerenzer(1989)再次針對 1984 年變態心理學雜誌(The Journal of Abnormal Psychology)調查 54 篇論文的統計考驗力，發現(1)中效果值的平均統計考驗力僅爲.44，(2)僅有兩篇提及統計考驗力分析，(3)54 篇論文中有 7 篇未達統計顯著水準，其中效果值的中位數爲.25。最近，教育與心理學界再度關注到統計考驗力分析的重要性，由表二可看出，不少期刊上之研究或博士論文的統計考驗力要偵測出小效果值的研究結果，其機率大約在.13~.41 之間，實

在過低，而中效果值的統計考驗力介於.53~.81 之間，尚屬偏低。推究原因可能係(1)採用 30 人以上即認爲大樣本的經驗法則，與(2)小效果值的實驗必須付出很大的成本代價：數百以上的樣本(參見表 5)。雖然 Cohen(1990)認爲統計考驗力分析要成爲研究者的例行工作，或要普遍成爲教科書的重要題材，尚須假以時日。可喜的是，近年來，統計考驗力分析已逐漸受到其它醫學(Goodman & Berlin, 1994)、漁業(; Edwards, & Perkins, 1992; Cahalan, 2000)、農林業(Stear, Reid, & Gettingby, 1996；Nemec, 1991)、野生動物(Steidl, Hayes,& Schauber, 1997; Johnson, 1999)、與生態學(Peterman, 1990)等各界應用與探討。反觀國內這方面的探討與研究可說鳳毛麟角，首推張德榮(民 71)對於教育心理研究的統計考驗力分析，發現國內教育心理研究的統計考驗力與 Cohen(1962)早期的研究結果頗爲類似，亦即犯第二類型錯誤的機率蠻嚴重的；其後只有謝季紅、涂金堂(民 87)介紹 t 考驗的統計考驗力與其計算方法，期間再無其它相關之研究報告，國內研究品質是否已顯著提昇，尚待我國學界的正視與努力。

表 2

期刊、博士論文的統計考驗力分析摘要

作者	年代	研究期刊	篇數	小	中	大
Rossi, J. S.	1990	Jr. of Consulting & Clinical Psychology + Jr. of Abnormal Psychology + Jr. of Personality and Social Psychology	221	.17	.57	.83
Daniel, T. D.	1993	Jr. of Research in Music Education (1987-1991)	78	.13	.64	.97
Coblick, G. E.	1998	Nursing Research(1995)	52	.25	.80	.94
Coblick, G. E.	1998	Western Jr. of Nursing Research (1995)	37	.27	.77	.92
Maddock, J. E.	2000	Health Psychology(1997)	共	.34	.74	.92
Maddock, J. E.	2000	Addictive Behavior(1997)	187	.34	.75	.90
Maddock, J. E.	2000	Jr. of Studies on Alcohol(1997)	篇	.41	.81	.92
Tener, M. A.	2000	Jr. of Athletic Training(vol. 32~34)	36	.18	.53	.75
Deng, H.	2000	田納西州 5 個大學教育行政與教育領導博士論文(1996~1998)	80	.34	.79	.94
張德榮	1982	師大、高師、彰師等學術期刊	50	.18	.63	.85

柒、實得效果值的分析在結論中的應用

理論上的效果值(hypothesized effect size)用於樣本規劃，實得效果值(observed effect size)則用於結果解釋。由於P值是效果值與抽樣誤差的函數，當P值小於事前所設定的α時，並無法區分到底是效果值或抽樣誤差所造成的顯著性結果。許多的研究者(Borenstein, 1994 , 1997; Goodman & Berlin, 1994; Nix & Barnette, 1998)主張發現統計的顯著性(statistical significance)之後，應再探討效果值的大小，以確定其臨床或應用的顯著性(clinical or practical significance)。他們認爲建立效果值的信賴區間，可用以評估樣本大小與誤差的大小。Nix & Barnette(1998)認爲不管關連量數或效果值量數均可作爲實用性的指標(measures of practical significance)。因此，醫學文獻上的關注重心逐漸從統計的顯著性考驗轉移到效果值的分析與信賴區間的運用上。Borenstein(1994 , 1997)並以癌症病人接受舊式治療與新式治療爲例，說明效果值信賴區間分析，能提供比統計顯著性的考驗更多且有用資訊(參見圖1與圖2說明)。

一、假如達到統計顯著水準時：

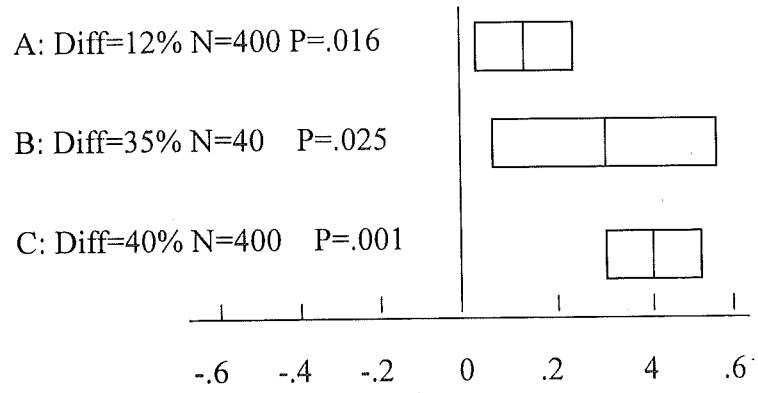

圖1、三個達到顯著水準研究的腫瘤實驗處理效果之信賴區間

在圖1中，橫軸右側表有利於新式療法(0以上)，左側表有利於制式療法(0以下)。由圖1知，第一個研究(每組200人)可以降低腫瘤再生率12%(Diff=12%)，第二個研究(每組20人)可以降低腫瘤再生率35%(Diff=35%)，第三個研究(每組200人)可以降低

腫瘤再生率 40%。三個研究(A,B,C)的 P 值均低於α=.05 顯著水準，其相對應的信賴區間亦都不包含零。如僅依統計顯著性考驗結果，每一個研究的效果，不管制式或新式療法均能顯著降低癌症再發率，其結論是相同的。但如以效果值的信賴區間來看，第一個研究 A 的處理效果相對來說是較小的(.12)，第二個研究 B 的處理效果很不錯(.35)，但變異很大(.04~.59)，而第三個研究 C 不僅處理效果最大(.40)且變異很小，最可靠(.31~.48)。可見效果值的信賴區間比統計顯著性考驗，能提供更多有用的資訊。因爲醫生與病人都希望明確知道新舊式腫瘤醫療的效果是多大，或效果不穩定(如第二個研究)的程度。此外，研究者如光以 P 值大小去論斷哪一研究較顯著(例如，A 研究比 B 研究效果顯著)，這顯然是不正確的說法。筆者認爲信賴區間的大小亦與效果值的變異量具有密切關係，在研究結果的解釋時，不僅要注意效果值的平均值，而且也要注意效果值的離散情形(如查看標準差的大小)，才能作出正確的解釋。

二、未達到統計顯著水準時：

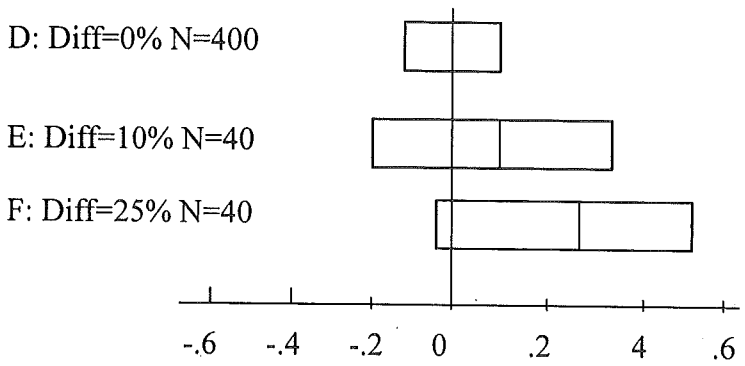

圖 2、三個未達顯著水準研究的腫瘤實驗處理效果之信賴區間

由圖 2 的三個信賴區間(D,E,F)均包含 0，三個研究的處理效果均未達到統計上的顯著水準。從 D 研究之信賴區間知，兩組的組間平均差異等於零(Diff=0%)，且其相對處理效果差異不大是非常明確的，至於 E 與 F 研究的信賴區間甚廣，可能有利於新式療法，亦可能不利於新式療法，但 F 的研究不僅平均處理效果較大，且較有利於新式療法。因此，新式療法在臨床上似乎更具有應用價值。

由此觀之，統計顯著水準並不能保證臨床上或實際應用上一定具有顯著效果，因爲虛無假設考驗只關心 p 值是否小於α值，完全不涉及相關效果值的大小與其可能的範

圍，難怪有些學者就建議揚棄假設考驗，改用提供資訊更多的信賴區間考驗(Borenstein, 1994 , 1997; Gardner & Altman, 1989)。例如，同樣是達到統計上.05顯著水準的同性質醫療實驗，效果值大且其離散量小的實驗應是最佳醫療選擇。同樣地，同性質的幾個醫療實驗如均未能達到統計上.05顯著水準，如能了解效果值的分佈與離散量，對於研究結果的解釋與應用常有意外發現。例如，有些研究結果雖沒達到統計上的顯著水準，但卻有應用價值，這只能從效果值的大小與分佈離散量去做主觀性判定(Borenstein, 1994 , 1997)。例如，假如研究者認爲處理效果差異達 20%以上才具有應用價值，那麼 A 研究雖已達統計上之顯著水準，但卻低於最低效果值 20%，實質上 A 研究並無實用價值；而 F 研究雖未達統計上之顯著水準，但卻高於最低效果值 20%，實質上 F 研究仍具實用價值。其實，所有的研究多少都涉及主觀的認知與判斷(Huberty & Morris, 1988; Rojewski, 1999)，量化研究亦無法完全價值中立。

綜合以上圖 1 與圖 2 之資料分析，可歸納出如表 3 與表 4 的結論：一個研究達到統計的顯著水準可能係因樣本過大所致，統計的顯著性不代表臨床上一定具有應用價值(如 A 研究)，而統計上未達顯著水準，可能係統計考驗力太低、或效果太小，但不代表沒有臨床上的應用價值(如 F 研究)。因此，研究者如遇上述兩種情況，可另選樣本進行交互驗證研究(cross-validation study)。不過，醫學上的應用價值，除了醫療效果的大小外，可能尚須考慮其他因素(如醫療成本、醫療時間與醫療副作用等)，不能單靠統計的顯著效果下決策。在教育或心理的研究應用上，亦何嘗不是如此。至此，讀者當不難看出虛無假設檢定只能提供我們有關實得差異是否爲『機遇』所造成的資訊，反而信賴區間可以用來評估抽樣誤差、效果值的離散情形、工具可靠度、與樣本大小的合適性。因此，效果值的信賴區間分析似乎是未來統計顯著性考驗的後續充分條件。2001 年第五版 APA 寫作手冊已強制要求研究者在提供 p 值時，一併要提供相關之效果值與區間估計值(APA, 2001)，即是明證。

表 3

統計顯著性與臨床顯著性之關係

		統計顯著性	
		達	未達
臨床具有	是	C	
	不確定	B	E/F
顯著效果	不是	A	D

表 4

統計顯著性、樣本大小與研究結論的關係

		樣本大小	
		小	大
統計顯著性	Yes	重要結果 (C)	視效果值而定 (A)
	No	視效果值而定 (B/E/F)	H_0 可能爲真 (D)

由表 3 與表 4 知，統計上之顯著性考驗和統計考驗力分析是決策的必要條件，而效果值的大小是決策的充分條件。

捌、統計考驗力分析的類別

統計考驗力的分析可以分爲三大類別：

一、事前統計考驗力(A Prior Power)分析：統計考驗力分析在研究計畫階段就已進行，旨在規劃適當樣本大小，以獲得適當的統計考驗力。

二、事後統計考驗力(Post-Hoc Power)分析：用在實驗後資料分析時，旨在了解實得統計考驗力(observed power)，以正確解釋研究發現。

三、折衷式統計考驗力(Compromise Power)分析：旨在規劃出能同時兼顧低的 α與高

的Power的樣本大小(同時控制α與β)，即要滿足β/α=q的要求。

部份學者(Gerard, Smith, & Weerakkody, 1998; Hoenig & Heise, 2001; Lenth, 2001; Lewis, 2000; Thomas, 1997)堅持反對使用事後統計考驗力分析，認爲進行事後統計考驗力分析，會導致下列弊端：第一、當發現研究結果未達統計上的顯著水準時，趨使研究者一味追求統計上的顯著去增加樣本，而忽視科學上實質的需求。第二、進行實得統計考驗力(Observed power)的分析，計算時只使用了樣本實得效果值與實得變異量而非學理依據的效果值，並沒提供比 P 值更多的資訊。第三、當發現研究結果未達統計上的顯著水準，且又具有很高的統計考驗力時，一方面虛無假設似乎獲得支持，另一方面因爲統計考驗力高而 P 值也愈小，而出現不利於虛無假設爲真的矛盾現象。因此，應儘量避免使用事後統計考驗力分析。如欲使用，需使用實際之樣本大小、α與理論上的效果值而非實得效果值去估計事後統計考驗力。Lewis(2000)則認爲解決之道是改用信賴區間方式去詮釋研究結果，而非使用事後統計考驗力分析。

至於折衷式統計考驗力分析，筆者認爲是最理想的樣本規劃，尤其當您能正確判斷犯α與β兩種錯誤的相對嚴重性大小時。不過，一般在教育研究上要判斷α與β錯誤的相對嚴重性大小非常不易，應用上通常以 1:4 比率(β/α=.20/.05) 處理。因此，用於樣本規劃的事前統計考驗力的分析是最有效與最無爭議的用途。

玖、事前統計考驗力分析

事前統計考驗力分析旨在樣本規劃。Cohen (1992)有鑑於統計考驗力的分析常大費周章，異常繁複，爲讓讀者能快速查出 Power=.80 時所需之樣本人數，提供如表 5 之樣本規劃簡表供查用。研究者只要先確定(1)所需用到的統計量，(2)α水準與(3)效果值大小，即可查到雙尾檢定時所需的樣本總人數(表 5 中統計別 2, 4, 6, 8)或各組的樣本人數(表五中統計別 1, 3, 5, 7)。例如，假如研究者希望α設定爲.05，統計考驗力爲.80，效果值爲中效果(d=.50)，而所需的統計量爲兩個獨立樣本的 t 考驗，利用表 5 的第一個統計量：Mean diff，往右即可查出所需的各組人數爲 64 人，兩組共需 128 人。再如研究者希望進行 3x4 的列聯表關聯性考驗，α設定爲.05，統計考驗力爲.80，效果值爲中效果(w=.30)，利用表 5 的第六個統計量：χ^2，往右即可查出自由度爲 6(2x3)時，所需的總樣本人數爲 151 人。值得一提，由表 5 之樣本人數估計表知，傳統上認爲每組超過 30 人之樣本，即認爲大樣本的思維，除非具有大效果，在大部分

的情境下是不正確的。

表 5 係『雙側考驗』用的統計考驗力分析，讀者如欲進行『單側考驗』，或其它統計考驗力(如.85 或.90)之下的樣本人數，可仿製表 5，利用下節介紹的 G*POWER 軟體進行樣本之規劃，亦相當便利。

表 5
效果值大、中、小的樣本大小估計與各統計量數效果值的設定表 (Power=.80)

	α						效果值		
	.01			.05					
統計別	Sm	Med	Lg	Sm	Med	Lg	Sm	Med	Lg
1. Mean dif	586	95	38	393	64	26	.20	.50	.80
2. Sig *r*	1,163	125	41	783	85	28	.10	.30	.50
3. *r* dif	2,339	263	96	1,573	177	66	.10	.30	.50
4. *P* = .5	1,165	127	44	783	85	30	.05	.15	.25
5. *P* dif	584	93	36	392	63	25	.20	.50	.80
6. χ^2							.10	.30	.50
1*df*	1,168	130	38	785	87	26			
2*df*	1,388	154	56	964	107	39			
3*df*	1,546	172	62	1,090	121	44			
4*df*	1,675	186	67	1,194	133	48			
5*df*	1,787	199	71	1,293	143	51			
6*df*	1,887	210	75	1,362	151	54			
7. ANOVA							.10	.25	.40
2*g*[a]	586	95	38	393	64	26			
3*g*[a]	464	76	30	322	52	21			
4*g*[a]	388	63	25	274	45	18			
5*g*[a]	336	55	22	240	39	16			
6*g*[a]	299	49	20	215	35	14			
7*g*[a]	271	44	18	195	32	13			
8. Mult *R*							.02	.15	.35
2*k*[b]	698	97	45	481	67	30			
3*k*[b]	780	108	50	547	76	34			
4*k*[b]	841	118	55	599	84	38			
5*k*[b]	901	126	59	645	91	42			
6*k*[b]	953	134	63	686	97	45			
7*k*[b]	998	141	66	726	102	48			
8*k*[b]	1,039	147	69	757	107	50			

註：Lg, Med, & Sm分別表大、中、小效果值，涉及2組以上的比較時，表中之N係指各組在雙側檢定時所需的人數。
[a] 表組別數。 [b] 表獨立變項數。
修正自Cohen(1992). A power primer. Psychological Bulletin, 112(1), 155-159.

拾、事後統計考驗力分析

事後統計考驗力分析，需使用實際之樣本大小、α與理論上的效果值而非實得效果值去估計事後統計考驗力，才有助於研究結論的正確解釋(Thomas, 1997)。Cohen(1988)主張統計考驗力至少要有.80 以上，但也不要過高，以致過當而使研究結果沒有實質效益。Dilullo(1997)針對數學教育研究期刊上在 1976-1995 年中 81 個統計考驗進行統計考驗力分析，發現有一半未達統計上顯著性水準的研究，統計考驗力未達.50，有 40%結果達顯著性水準的研究，其統計考驗力高達.95。他們建議研究者當研究結果未達統計上顯著性水準時，其統計考驗力未達.50 者(過低)，或研究結果達顯著性水準時，而其統計考驗力高達.95 者(過高)，解釋結果時要特別小心，或應進一步探究之，勿遽以論斷。前者可能會使具有實驗效果的實驗獲得肯定的機會渺茫，而喪失應有的效益，後者則可能使小效果或沒有應用價值的實驗達統計上的顯著性，而徒然浪費人力、時間與金錢在後續的應用或研究中，甚至使受試者遭致無謂傷害。研究者，如能再配合 Borenstein(1994 , 1997)醫學模擬實驗結果的效果值分析方法，以效果值大小與其信賴區間去作全盤考慮，以去除量化研究報告只推論不描述的偏差。

因此，當 p 值小於預定的α值時，如果該研究使用了很大樣本，其研究結論事實上是不可靠的，因爲當樣本趨於很大時，任何微小的效果均可達到統計上的顯著水準。又如當 p 值大於預定的α值時，如果該研究使用了較小樣本，其研究結論事實上也是無法確定的，因爲統計考驗力太低所致(Aron & Aron, 1999/2000; Borenstein, 1994 , 1997)。Dilullo(1997)分析從 1976 年到 1995 年間的數學焦慮與表現之論文，亦發現研究結論不一致，他們之間存在著極大差異，主因即出自於有些論文統計考驗力太低所致。

另外，有些研究者誤解了實得統計考驗力，常宣稱『本研究不僅達到統計上之顯著水準，而且統計考驗力亦很高』或宣稱『本研究未達到統計上之顯著水準，因統計考驗力亦較低所致』(Lenth, 2001)。其實，一個研究就是因統計考驗力高才會達到統計上之顯著水準，因統計考驗力低才不能達到統計上之顯著水準。因此，在進行事後統計考驗力分析時，必須以理論上的效果值而非實得效果值去估計事後統計考驗力，才更具實質意義。

拾壹、G*Power 統計考驗力分析軟體簡介與應用

統計考驗力分析的軟體已不少，其中以 SPSS 的 SamplePower(提供事前統計考驗力分析)及 SPSS GLM(提供事後統計考驗力分析)、SAS8.02 所提供的 Sample Size 副程式(在 Statistics 之下)、UnifyPow Macro 程式(O'Brien, 1998)、與 G*Power(Erdfelder, Faul, & Buchner, 1996)最爲大家所樂用，尤其 G*Power 係免費共享軟體，其下載網址爲:

http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/index.html

G*POWER 可同時分析三種不同類別的統計考驗力分析，值得在此推介使用。由圖 3 使用者界面知，使用時要先在視窗的最右側 Analysis 欄與 Test is 欄中確定您要進行哪種考驗力分析(事前統計考驗力分析與單側檢定爲內定值)，再先點選功能表單上的『Tests』，選擇待考驗的統計量(含 t 考驗、F 考驗與χ^2考驗等 7 種)，接著填入效果值(Effect size)大小、Alpha 大小與所需的 Power，再按視窗中的『Calculate』，電腦即會將您所規劃的樣本大小 顯示在視窗的中央，而整個計算參數、過程與結果都會顯示於視窗的下方 Protocol 中。圖 3 視窗中央的數據即是 t 考驗所需的樣本人數規劃(Total sample size=102)。

統計考驗力、效果值的分析與樣本大小、α、和β具有密切的函數關係。讀者亦可應用 G*POWER 探求他們之間的關係。例如，過去最常見研究者利用統計考驗力、效果值與α、β的大小，去估計所需的樣本大小(選擇 A Prior Analysis 或 Compromise Analysis)。其次，研究者亦可利用樣本大小、效果值與α、β的大小，去估計統計考驗力(選擇 Post Hoc Analysis)。同樣地，研究者亦可利用統計考驗力、樣本大小與α、β的大小，計算效果值大小(按『Calc Effectsize』)，以進行研究結果的 Meta-analysis，有興趣的讀者可參讀整合分析與應用一書(應立志、鍾燕宜，民 89)。

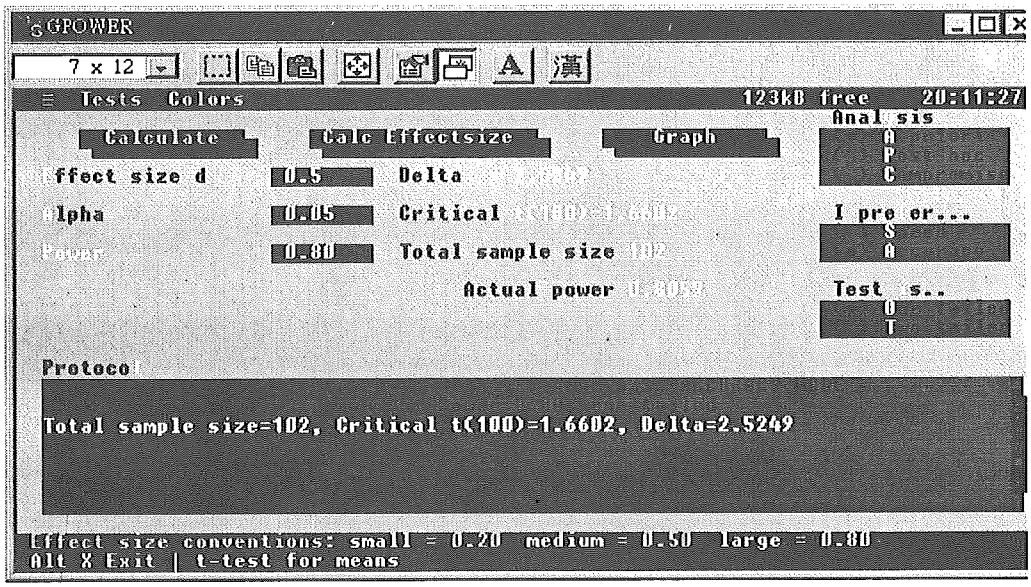

圖 3、G*Power 的使用者界面

讀者打開功能表單上的『Tests』時，可發現 G*POWER 提供了 t、F 與χ^2考驗的統計考驗力分析，簡介如下：

一、 t考驗的統計考驗力分析

1. 獨立雙樣本的 t 考驗
2. 相關係數 t 考驗的統計考驗力分析
3. 其它 t 考驗的統計考驗力分析

本類的統計考驗力分析包含重複量數 t 考驗、單一樣本 t 考驗、與 z 考驗。讀者如欲進行 z 考驗時，需將 df 設定在 32000 以上，以逼近於 z 分配。另外，在本類考驗時必須明確告訴電腦 N 與 df。因此，G*POWER 軟體無法進行事前統計考驗力分析，只能重複的調整 N 與 df 進行事後統計考驗力分析，一直到滿意的統計考驗力爲止，以間接獲得所需的樣本大小。

二、 F考驗的統計考驗力分析

1. 變異數分析(ANOVA)

讀者在執行單因子 F 考驗時，必須先點選『Hypothesis』欄下『Global』，而在執行多因子 F 考驗與共變數分析時，必須點選『Hypothesis』欄下『Special』，才能正確獲得分析結果。在 G*POWER 中，目前只能進行固定效果模式分析，這是一大限制。

2. 複相關與迴歸分析(MCR)

3. 其它F考驗

適合進行多變項分析及重複量數分析的統計考驗力分析，但 G*POWER 軟體無法進行事前統計考驗力分析，只能重複的調整 N 與 df 進行事後統計考驗力分析，一直到滿意的統計考驗力爲止，以間接獲得所需的樣本大小。

三、χ^2考驗的統計考驗力分析

用以進行適合度與獨立性考驗的統計考驗力分析。

另外，G*Power 亦可按『Graph』，以提供如圖 4 的 Power 與樣本大小之函數圖，研究者可利用此圖迅速找到理想的樣本人數。讀者如欲知道更詳細的操作方法，可從網站上下載 G*Power 的操作手冊。

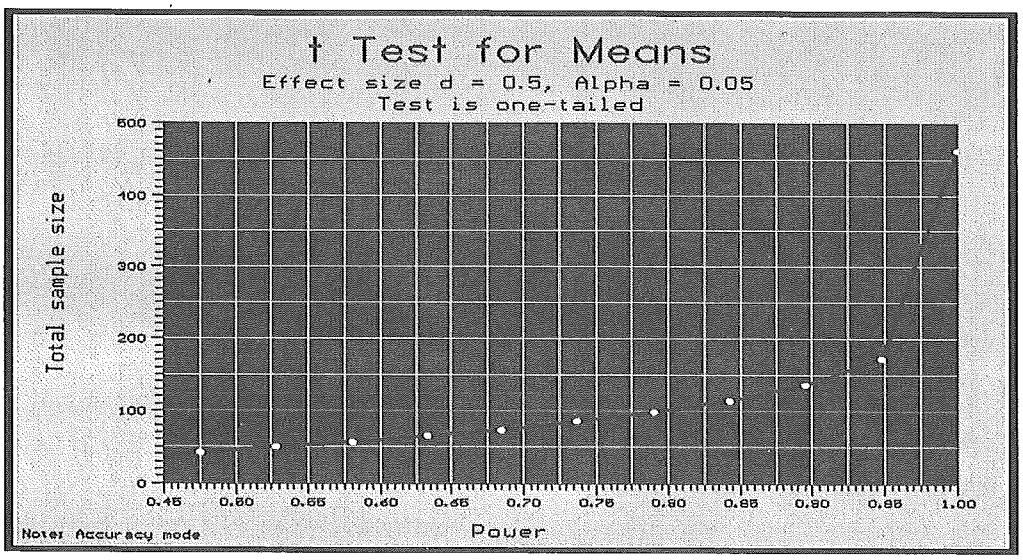

圖 4、G*Power 的圖形視窗

拾貳、量化研究品管的可行途徑

量化研究的品管從早期的樣本抽樣與工具設計即應開始，在量化研究的過程中，講求的是客觀性與推論性，研究者如能抽取適當的代表性樣本，其推論性與複製性必更高；如能講究測量工具的信、效度以減少測量誤差(Helberg, 1996; Rojewski, 1999)，其內在效度必高。最後，研究者對於研究結果之詮釋時亦應縝密且與過去之

研究脈絡相結合，如仍無法定論，需再進行複製研究。如此，量化研究的內、外在效度必能獲得品質保證，所累積的知識才能日臻完善。

爲了糾正過去量化研究的迷思，Huck & Cormier(1996)建議研究者採行下列步驟，進行統計顯著性之考驗，應是量化研究品管的良方(李茂能，民 87)：

一、根據待答問題擬出對立假設與虛無假設，
二、決定適當之顯著水準，
三、估計最低效果值，
四、估計統計考驗力，
五、規劃樣本大小，
六、使用良好工具蒐集資料，
七、選擇適當的統計考驗，
八、比較統計考驗值與臨界值，裁決統計上是否拒絕虛無假設，
九、再根據效果值(效益)大小、效果值離散情形、成本等相關因素，判定決策上之應用價值。

此外，因爲虛無假設檢定，只能提供我們有關實得差異是否純爲『機遇』所造成的訊息，筆者建議將『最低效果值』設定爲虛無假設檢定的目標，統計顯著性考驗將更具應用價值。

鑑於統計考驗力與效果值分析的必要性，建議國科會、教育部、各級學校論文的審查者或期刊的主編，應要求研究者於研究計畫或論文的方法論一節中，規劃樣本大小、統計考驗力分析，於結果分析一節中，除了報告 P 值外，尚須報告實得統計考驗力與效果值分析(Cohen, 1990；Thompson, 1998)，以進行量化研究品質的最後把關。美國心理協會第五版 APA(2001)寫作手冊，已建議作者提供統計考驗力分析，並強制要求報告效果估計值大小，值得國內研究者及期刊主編借鏡，當有助於提高國內研究成果在 SCI 或 SSCI 引用的比率。

參考書目

李茂能(民 87)。統計顯著性考驗的再省思。*教育研究資訊，6(3)*，103-115。
張德榮(民 71)。教育心理研究的統計考驗力分析。*輔導學報，5*，119-137。
應立志、鍾燕宜(民 87)。*整合分析方法與應用*。台北：華泰。

謝季紅、涂金堂(民 87)。t 考驗的統計考驗力之研究。*教育學刊，14*，93-113。

Aron, A., & Aron E. N. (2000). *心理與教育統計學(黃瓊蓉編譯)*. 台北：學富。(原著出版日期：1999)

American Psychological Association. (2001). *Publication Manual of the American Psychological Association(5th ed.).* Washington, DC: Author.

Bezeau, S, & Graves, R. (2001). Statistical power and effect size of clinical neuropsychology research. *Journal of Clinical & Experimental Neuropsychology, 23*(3), 399-406.

Borenstein, M. (1994). The case for confidence intervals in controlled clinical trials, *Controlled Clinical Trials 15,* 411-428.

Borenstein, M. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma, & Immunology, 78,* 5-15.

Cahalan, J. (2000). Application of hypothesis testing and power analysis in the Puget Sound crab fishery: Closure decisions with confidence. *Journal of Shellfish Research, 19*(1), 619.

Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review, 48,* 378-399.

Chase, L. J., & Chase, R. B. (1977). A statistical power analysis of applied psychological research. *Journal of Applied Psychology, 61*(2), 234-237.

Coblick, G. E. (1998). *Statistical power analysis of nursing research.* Unpublished doctoral dissertation, the Auburn University.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences(rev. ed.).* New York: Academic Press.

Cohen, J. (1990). Things I have learned so far. *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences(2nd ed.).* Hillside, NJ: Erlbaum.

Daniel, T. D. (1993). *A Statistical Power Analysis of the Qualitative Techniques Used in the Journal Of Research in Music Education, 1987 through 1991.* Unpublished doctoral dissertation, the Auburn University.

Daniel, T. D. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in Schools, 5*(2), 23-32.

Deng, H. (2000). *Statistical Power Analysis of Dissertations Completed by Students Majoring in Educational Leadership at Tennessee Universities.* Unpublished doctoral dissertation, the East Tennessee State University.

Dilullo, L. K. (1997). *A Post Hoc Power Analysis of Inferential Research Examining the Relationship between Mathematics Anxiety and Mathematics Performance.* Unpublished doctoral dissertation, the Auburn University.

Edwards, E. F., & Perkins, P.C. (1992). Power to detect linear trend in dolphin abundance: estimates from tuna-vessel observer data, 1975-89. *U.S. National Marine Fisheries Service Fishery Bulletin, 90*(3), 625-631.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavioral Research Methods, Instruments, & Computers, 28,* 1-11.

Gardner, M. J., & Altman, D. G. (1989). *Statistics with Confidence--Confidence Intervals and Statistical Guidelines.* London: BMJ.

Gerard, P. D., Smith, D. R., & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management, 62*(2), 801-807.

Goodman, S. N.& Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121,* 201-206.

Helberg, C. (1996). *Pitfalls of Data Analysis.* ERIC/AE Digest.(ERIC Document Reproduction Service No. ED 410 231.

Hoenig, J. M., & Heise, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations in data analysis. *The American Statistician, 55,* 19-24.

Huberty, C. J., & Morris, J. D. (1988). A single contrast test procedure. *Educational and Psychological Measurement, 48,* 567-578.

Huck, S. W., & Cormier, W. H. (1996). *Reading Statistics and research*(2nd ed.). New York; Happer Collins College publishers.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*(3), 763-772.

Kaufman, A. S. (1998). Introduction to the special issue on statistical significance testing. *Research in the School, 5*(2), 1.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *American Statistics, 55*(3), 187-193.

Lewis, R. J. (2000). *Power Analysis and Sample Size Determination: Concepts and Software Tools.* Paper presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine, San Francisco, CA.

Maddock, J. E. (2000). Statistical power and effect size in the field of health psychology. *Dissertation Abstracts International, 60(9-B)*,Apr 2000, 44939. (University Microfilms No. 2000-95006-499.)

Nemec, A. (1991). *Power Analysis Handbook for the Design and Analysis of Forest Trials.* Victoria : Ministry of Forests.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3-14.

O'Brien, R. G. (1998). A tour of UnifyPow: A SAS module/macro for sample size analysis. *Proceedings of the 23rd Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc.,* 1346-1355.

Peterma, R. M. (1990). The importance of reporting statistical power: the forest decline and acidic deposition example. *Ecology, 71,* 2024-2027.

Rojewski, J. W. (1999). Five things greater than statistics in quantitative educational research. *Journal of Vocational Research, 24*(1), 63-74.

Rossi, J. S. (1990). Statistical power of psychological research. *Journal of Consulting & Clinical Psychology, 58(5),* 646-656.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416-428.

Schwarzer, R. (1989). *Statistics Software for Meta-Analysis*. Retrieved October 20, 2001 from the World Wide Web: http://www.yorku.ca/faculty/academic/schwarze/meta_3.htm

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Stear, M. J., Reid, S. W. J. & Gettingby, G. (1996). The likelihood of detecting differences between groups of sheep following deliberate infection with Ostertagia circumcinta. *International Journal for Parasitology, 26*(6), 657-660.

Steidl, R. J., Hayes, J. P.,& Schauber, E. (1997). Statistical power analysis in wildlife research. *Journal of Wildlife Management, 61,* 270-279.

Tener, M. A. (2000). *A Post Hoc Statistical Power Analysis and Survey of the Research Published in the Journal of Athletic Training.* Unpublished doctoral dissertation, the Middle Tennessee State University.

Thomas, L. (1997). Retrospective power analysis. *Conservation Biology, 11,* 276-280.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998). *Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas.* Paper presented at the annual meeting of American Educational Research Association, San Diego.

Quality Control for Quantitative Studies: Power and Effect Size Analysis

Mao-neng Fred Li

Abstract

This paper addresses the close interlock between sample size determination, statistical significance, and practical/clinical significance. Several misconceptions, misuses, and misinterpretations related to Type-I error, Type-II error, power, p-values, and sample size are highlighted. Then, those factors such as α, β, & sample size that may affect power are specified and defined. It is argued that the quality of quantitative studies can only be assured through power and effect size analysis. For a scientific conclusion of statistical significance and practical/clinical significance, power analysis should be done in the planning stage of study and minimum practical effect size should be considered with statistical significance testing. Next, computing power for any specific study may not be a difficult task via Cohen's power table and a computer program called G*Power. Finally, recommendations regarding editorial policies and reviewing practices for the integration of power analysis and effect size into the study plan or the test result are pinpointed.

Keywords: Power analysis, effect size, quantitative studies