

柯氏性格量表效度概化之統合分析

林世華

台灣師範大學教育心理與輔導系

黃寶園

台灣師範大學教育心理與輔導系博士班

本研究之目的乃在透過統合分析法(meta-analysis)，對「柯氏性格量表」過去34年來所累積的研究資料，做客觀、計量性的分析，以瞭解其效度概化的程度。經電腦程式Meta-HC分析後顯示：當以「柯氏性格量表」作為人格上的區辨工具時，以在心理症病人的效度最高，其次依序為特殊組群、不同心理症病人、犯人；而當施測於生理症病人與男女之間時，其區辨力稍嫌不足。綜觀本研究的結果顯示：「柯氏性格量表」的效度概化程度頗高，區辨力頗為理想，因此是一個值得進一步開發、研究以及推廣使用的人格測驗。

關鍵詞：效度概化、統合分析、柯氏性格量表

緒論

一、研究動機

「柯氏性格量表」(Ko's Mental Health Questionnaire，簡稱KMHQ)是國內第一個臨床上自編的大型本土化心理診斷工具，也是目前國內臨床心理學和精神醫學界有關性格評估較常使用的工具(呂傳欽、闕清模、卓良珍、藍忠亮，民81)，且被評為國內應用最廣、效果最好的心理與精神疾病診斷系統(方小萍，民84)。因此其所代表的意義價值非凡，然而其效度為何？是否能達到人格測驗該具有的區辨能力？這相信是一個值得大家去深思探究的問題。

因為「柯氏性格量表」編製得早(民國53年編製完成)，因此所累積的施測報告相當多，其豐富的累積資料正適合使用統合分析法(meta-analysis)來探究其效度概化的程度，以提供測驗編製者及測驗使用者能更深入瞭解「柯氏性格量表」

真正的功能，因此激起了研究者的研究動機，從事本項研究工作。

二、研究目的

基於上述動機，本研究主要的目的有下列二項：
(一)廣泛的收集「柯氏性格量表」應用研究，以了解目前「柯氏性格量表」之使用情形及其效果。
(二)使用統合分析法，為「柯氏性格量表」先前之累積資料做一統合性的分析，進而了解「柯氏性格量表」效度概化的程度。

三、研究問題

依據本研究的目的，提出下列研究問題：
(一)心理症病人與正常人在「柯氏性格量表」結果的表現是否有所不同？其差異程度為何？解釋能力多少？
(二)不同性質的心理症病人在「柯氏性格量表」結果

的表現是否有所不同？其差異程度為何？解釋能力多少？

(三)生理症病人與一般人在「柯氏性格量表」結果的表現是否有所不同？其差異程度為何？解釋能力多少？

(四)犯人與一般人在「柯氏性格量表」結果的表現是否有所不同？其差異程度為何？解釋能力多少？

(五)特殊組群與一般人在「柯氏性格量表」結果的表現是否有所不同？其差異程度為何？解釋能力多少？

(六)男女不同受試者在「柯氏性格量表」結果的表現是否有所不同？其差異程度為何？解釋能力多少？

四、研究假設

依據研究問題，本研究提出下列研究假設：

- (一)「柯氏性格量表」在心理症病人與正常人之間，其施測結果之平均數有顯著不同。
- (二)「柯氏性格量表」在不同心理症病人之間，其施測結果之平均數有顯著不同。
- (三)「柯氏性格量表」在生理症病人與一般人之間，其施測結果之平均數有顯著不同。
- (四)「柯氏性格量表」在犯人與一般人之間，其施測結果之平均數有顯著不同。
- (五)「柯氏性格量表」在特殊組群與一般人間，其施測結果之平均數有顯著不同。
- (六)「柯氏性格量表」在男女之間，其施測結果之平均數有顯著不同。

五、名詞釋義

(一)柯氏性格量表：

「柯氏性格量表」是柯永河根據其心理衛生臨床經驗及明尼蘇達多相人格測驗(Minnesota Multiphasic Personality Inventory, 簡稱 MMPI)之內容於民國 53 年編製完成，並於民國 83 年進行大規模的修訂工作。編製之初的「柯氏性格量表」有

288 題，分為十一個分量尺。而民國 83 年修訂後的「柯氏性格量表 (1996)」其題數增為 300 題，量尺數則增加至 38 個，唯修訂後的「柯氏性格量表 (1996)」仍包括修訂前的十一個分量尺。因此，本研究所指之「柯氏性格量表」乃為修定前、後所共有之前十一個分量尺。

(二)統合分析：

統合分析是一種累積研究成果的計量性方法 (quantitative method)，且能提供效度概化證據的一套統計過程 (Auustas, 1997)，針對此，它統合某個研究領域內之現有研究的結果，給予再分析，以尋求一般性結論並試圖歸納出一個客觀結果的技術。也就是說，它是用來將許多相同現象的不同研究提供系統的合併與統合 (范德鑫, 民 82)。也就是說，統合分析乃嚴格評比論文品質，以統計方法結合數據來解決衝突論點的一種程序。而本研究所使用的統合分析方法依所蒐集的資料，分別採用 Hedges & Olkin (1985)、Hunter & Schmidt (1990)、Rosenthal (1991) 等三種統合分析技術。

(三)效度概化 (validity generalization)：

效度概化是由 Schmidt, Hunter & Urry (1976) 根據統合分析概念所提出來的。其理念在於將先前有關的效度研究做再處理的工作，進一步建立「效標-預期」以求得真正效度的一種方法 (王順正, 民 84)。而本研究所指的效度概化是指「柯氏性格量表」對某一群體施測所得的效度證據是否與其他群體所得的效度證據相同；也就是說，「柯氏性格量表」的施測結果效度對各母群的可推論性大小為何？在各特殊情境、特殊群體的施測結果解釋的效度是否一致？

(四)心理症病人：

本研究所指的心理症病人係指經精神科醫師診斷後確定患有心理症的人，如精神神經病、邊緣性

精神病、分裂性精神病、精神病、躁症、鬱症、精神官能症、焦慮性精神官能症、內因性精神病、生理性精神病等等。

(五)生理症病人：

本研究所指的生理症病人係指經醫師診斷患有某方面的生理疾病者。如癌症患者、關節炎患者、性功能障礙者、性無能者、眼斜視患者以及消化性

潰瘍患者等等。

(六)特殊組群：

本研究所指的特殊組群係指受試者本身在某一方面較為特殊而與一般人有明顯區隔者，如僑生、情緒不穩定者、童子軍、飛行員、警察、有自殺傾向者、工作表現較優者、女性月經週期不順者等等。

文獻探討

本節的目的在探討有關「柯氏性格量表」的相關研究及「統合分析法」的使用技術。首先探究「柯氏性格量表」的發展與演進；其次探究「柯氏性格量表」發展的方向；測驗的效度概化；統合分析的概念；最後則為三種統合分析方法的比較。

一、「柯氏性格量表」的發展與演進

(一)「柯氏性格量表」的編製：

「柯氏性格量表」是為多方瞭解個人與團體心理健康與不健康而發展出來的量表。它是柯永河根據其心理衛生的臨床經驗並參酌明尼蘇達多相人格測驗之內容編製而成的是非題問卷式心理測驗（柯永河，民 66）。

(二)「柯氏性格量表」的使用對象：

據柯永河（民 66）表示，因為「柯氏性格量表」是一種文字測驗，而且在它的一些子題內容中有較為抽象及較為特殊的描述，因此柯氏建議此量表較適用於教育程度為國中一年級以上，且其智商在 90 以上的人。但研究者從實際的文獻搜尋中曾發現有研究者將此量表施測於小學五六年級學生（吳裕益，民 71），關於這種情形研究者曾親自請教過柯永河先生：據柯氏表示他不曾對小學生施用此量表，因此施測後的效度狀況並不清楚，但柯

氏建議若要使用在小學生身上，則施測過程需要彈性些，如施測時間需較長，或是對某些小學生不懂的題意加以說明……等等。

(三)「柯氏性格量表」的使用狀況：

依據過往「柯氏性格量表」的施測報告來看，使用「柯氏性格量表」最多的單位為學校的心理輔導中心，其目的主要乃在探究學生心理健康的狀態，做為學生學校生活適應程度的指標，並以為輔導的依據；其次則是醫療衛生機構，而其使用目的大多在瞭解病人的心理健康程度，這其中又以精神科的使用頻率最高；另外則是學術研究者用來作為研究工具：如在輔導學界中便常有研究者以此作為輔導成效的探究工具。此外，因為「柯氏性格量表」發行的時間甚早，因此有很多新編製的人格測驗便選擇以「柯氏性格量表」作為效標，以探究新編測驗的效度。

二、「柯氏性格量表」發展的方向

「柯氏性格量表」經過大家三十多年來的使用後，陸陸續續發現一些缺失亟待改進，而在這些演進及修訂的過程中我們發現是循著某些方向在進行的（柯永河，民 81）：

首先，「柯氏性格量表」太集中在變態或偏差內容上，且欠缺正常、積極的內容，所以不甚符合

測驗名稱「性格量表」四個字。因此理論上，應再增加正面、積極、健康的項目，才能達到測驗效度概化的基本要求。其次為相似內容的項目集中編印在一起，這將使受試者產生強烈的作答偏差趨勢，為了削弱這種偏差趨勢，宜將相似的內容分散，使作答者對每一項目都能不受作答偏差趨勢的污染，而根據每一項目所問的內容據實作答。

第三，原有的柯氏性格量表採「是」「非」二選一反應方式，而修訂後的「柯氏性格量表（1996）」的作答方式則改採「六選一」（柯永河，民86）。雖然作答方式的調整有助於信度的提昇，但相對的亦產生某些缺失，例如在「六選一」的作答方式下，心智低下或神智錯亂者是否能在六個答案中做明確的區辨而選擇與自己最相似的答案呢？這可能也是值得商榷的。

第四，「柯氏性格量表」欠缺「效度量尺」以顯示整套測驗的結果是否可靠，這應是值得進一步探究的。不過在修訂後的「柯氏性格量表（1996）」中作答可靠量尺只有2題，作答不可靠量尺只有1題，似乎是少了些，能否發揮「效度量尺」的功能有待接續探究。

第五，在「柯氏性格量表」中有許多項目是不甚符合內部一致性 α 係數的要求，應從其原屬量尺中剔除，但因整體信度係數值頗高而未加刪除，因此為了這方面的改善，宜把這類項目全部刪除，而代以較合適的項目。

第六，「柯氏性格量表」創編時，主要是為了幫助精神醫學臨床診斷，因此為了催化臨床心理師與醫師間的溝通，「柯氏性格量表」的修訂發展應考慮增加符合DSM-III-R的診斷標準。

第七，一套測驗的內容越豐富越好，而內容豐富的測驗必然含有較多的項目，但項目越多時，受試者對它所付出的時間與精力就越多，合作意願便相對的降低，但在「柯氏性格量表（1996）」中的量尺數增加到38個，這是否會產生解釋上的混淆與困擾，或者是各量尺的代表性及效度上的問題，可

能是之後研究者所需再探究的議題。

三、測驗的效度概化

為某一測驗所蒐集到的證據，除了得到這種證據的情境外，用於其它情境也有意義嗎？有關這類問題的討論即屬於「效度概化」的範疇（路君約，民80）。換句話說「效度概化」即在探究測驗結果的可推論性；基本上它是指在某一群體（樣本）中所做的效度研究能否類推到其他群體（樣本）之中、在某一群體所得的效度證據與在另一群體所得的效度證據是否相同，以及某一測驗對所有的受試者是否均具有相同的效度等相關的問題。

測驗結果對未來工作表現的預測效度有效與否，一直是心理學者和測驗學者所極力要探索的（林邦傑，民76）。因為從過去50年來人事心理學家的研究發現：測驗受情境特殊性(situational specificity)的影響甚大，同一個測驗即使施測於相類似(similar)甚至相同(identical)的情境中，其所得到的效度係數也都不同，且變異程度頗大；因此更談不上是用在因素結構差異頗大的不同工作上。所以學者們認為測驗的效度必須在每一施測的特定領域或地區中建立方才有效，特別是涉及職業預測者（范德鑫，民82; Hedges, 1988; Schmidt, 1988）。正因此，學者們普遍認為效度概化是不可能的，所以在測驗施測的每一情境中均需進行效度的驗證。所以測驗學者們便在此方面投注不少心血，以促使效度概化成為可能。如Hunter & Schmidt在1977年便發展了一個效度概化模式，來替代以往效度的「情境特定性」的限制，並用以減少小樣本研究中測驗效度偏低的現象。Hunter & Schmidt (1990)認為：各測驗的效度係數間之所以會有變異產生而無法達到概化的程度，主要是一些人為的誤差因素(artifactual sources)所造成的一如抽樣誤差、效標信度之差異、測驗信度之差異、全距限制之差異、效標的效度問題、計算及謄寫之間的錯誤甚至是電腦程式的錯誤……等等，若將這些誤差來源去除後，

則各研究間的效度係數應當相當類似。而 Hunter & Schmidt 即由此觀念進而建立起效度概化的模式。

在此模式中對於上述誤差來源的估計，是藉助統分法來完成。Hunter & Schmidt 認為如果效度真正概化，則所計算出來的殘餘變異數應該為 0，不過並非每一種人為因素均可以數量化，因此實際計算出來的殘餘變異數並非為 0。但 Hunter & Schmidt 依他們的經驗認為，若所估計的誤差可解釋的變異達總變異的 75% 即可說是具有效度概化證據，並將之視為真正的效度之間的變異為 0，顯示所蒐集到的效度係數之間是相同的，(Hunter & Schmidt, 1990; Schmidt, 1988; Wolf, 1986)，也就是達到效度概化的程度。由此可見效度概化是可能的。因此，經由測驗的效度概化分析，將擺脫過往效度「情境特殊性」的限制（路君約，民 80；林邦傑，民 76）。

四、統合分析的概念

(一)何謂統合分析？

「統合分析」是一種將過去個別研究的結果綜合起來作計量結合的技術 (Glass et al, 1981)，並在此過程中消除各種誤差來源，以發現變項間的真正關係及其強度 (Hunter & Schmidt, 1990)。因此它是一種與傳統敘述性的文獻分析 (narrative literature reviews) 相反的方法，它從個別的研究結果中運用統計的過程，以收集實證性的發現，研究的重點在效果量 (effect size) 的大小。因此我們可用最簡單的話語描述統合分析為「對研究統合的量化方法」(Wolf, 1986)。

(二)統合分析技術之發展背景與演進：

統合分析技術的產生可說是針對傳統文獻分析中敘述性資料整合 (narrative integration) 的缺失而成。因為當使用傳統的文獻分析法對相同或類似的研究主題進行評論時，通常比較依賴評論者的個人

主觀見解而定，在這種情形下當評論的研究篇數少時尚能形成有效的結論，可是一旦評論的研究篇數多時，其效果將令人質疑 (Cooper & Rosenthal, 1980)。

雖然 Meta-analysis 一詞首先由 Glass 於 1976 年提出，但 Stigler (1986) 在他「統計的歷史」(The History of Statistics) 一書中卻指出其實 Legendre 於 1805 所創造的最小平方的原則 (the principle of least squares) 實際上已能解決目前統合分析中的問題。而目前關於統合分析的後續發展仍有相當多的學者努力的探討，至少在效果量的解釋上已有數種新的方式出現 (Dunlap, 1994; McGraw & Wong, 1992)，或許在不久的將來統合分析的技術將因這些學者的廣續努力而更趨於完善，得到更為堅強的結論。

(三)統合分析法的特性

首先，統合分析是一種量化的技術，它使用大量的數字以及統計方法，並從過去大量的累積研究資料中，組織並抽取出訊息以統合這些大量研究的結果，以進行再分析的工作 (reanalysis)，而在這整個研究過程中，全為計量的程序，研究者始終保持價值中立。

其次，統合分析不以個別研究的研究品質預判研究發現。在傳統的文獻分析中，處理的面很廣，但相對的也排除很多資料，因此當研究者主觀地認為某些研究的設計較差、測量方式不當、工具使用不理想……等，便會將這研究刪除，可是從來沒有證據顯示這些所謂較差的研究會損及最後研究的發現 (Glass et al, 1981)。而在統合分析中，它會記錄下這些品質較差的研究並檢驗和這些研究發現之間的相關。因此研究的品質在此可視為較其次的問題而非首要問題，甚至是立即要排除在研究之外。

第三，統合分析它尋求一般性的結論，因為它只是對相同研究主題的累積資料做統合，因此無法像一般調查或實驗研究可從中得到豐富的研究成果，

對未知變項做探測性的或驚人的發現。

第四，統合分析它運用平均效果量 (averaging effect sizes) 的觀念，將過往各研究的效果量予以加權平均，以探求真相，因此它不像在變異數分析中有「加和乘」(adding and multiplying) 的作用。

最後，統合分析它對研究結果不做預先的判斷、推測，只是客觀的將各研究結果統合在一起，因此人爲的操弄而使研究結果造成偏誤的情形將不復出現，所得的一般性結論將是客觀的。

(四) 統合分析法的功能：

第一，統合分析它能解決主題相同但結果卻相衝突的研究。因爲統合分析技術對於這些主題相同結果卻相異的研究，它能夠提供量化的統計分析技術，從中得到一個客觀的答案，解決這些結果相互衝突的研究，而探索事實的真相。

第二，促進理論的發展 (Hunter & Schmidt, 1990)。統合分析的技術可讓我們發現變項間的關係，且其效果量的大小更可讓我們知道變項間關係的強度及方向。因此它對促進理論的發展有相當大的貢獻。

第三，深入敘述的功能：統合分析不只將各研究結果做一般性的描述，它還可依照研究的特性加以分類及彙總，以進一步了解造成研究差異的原因，並從中找出潛在的中介變項 (鍾燕宜，民 75)。尤其在同質性的考驗呈顯著時更容易使研究者有足夠的信心去探究中介變項。

第四，促進新陳代謝的功能：由於統合分析必須蒐集許多相關資料，因而對後續研究者提供了許多寶貴的研究經驗與結果，並對問題更加了解，節省了摸索的時間，使研究者在研究計畫的重點上可以避免重複做一致性結果的研究 (Rosenthal, 1979; Smith & Glass, 1977)。

第五，再分析診斷的功能：統合分析是對各獨立的研究做再分析、再考驗的工作，因此對原本有爭論或關係模糊的問題，經過統合分析的再次診斷，

可以更清楚的了解事實的關係 (鍾燕宜，民 75)。

(五) 統合分析法的缺失與問題：

第一，有人批評統合分析並比喻其爲『蘋果、橘子不分』(mixes apples and oranges)，就像『大雜燴』一般，因爲它的發現是從不同研究中而來的，而這些研究之間並無相同的標準在，因此將這些研究結果放在一起是沒有意義的。但 Glass et al (1981) 認爲若是一系列的研究各個方面皆相同的話就沒有必要做統合了，因爲除了誤差外其餘結果皆應相同。

第二，有人批評統合分析好壞不分，認爲將「好」與「壞」的研究放在一起後，將無法區辨「好」與「壞」的研究，因此可能有利於低品質的研究產生。但 Glass et al (1981) 認爲，品質差的研究既然已經做出來了，在邏輯上就不能忽視他們，而是評估他們。關於此，Glass et al (1981) 曾提出以下的例子加以說明多個較弱的研究也可以達成一個很強的結論：例如在 100 個研究當中，有 10 個研究在實驗設計上較差；有 10 個研究在抽樣上有問題，有 10 個研究在測量上有問題，有 10 個研究在內在效度上有問題，有 10 個研究在資料分析上有問題，……，但它們卻一致的顯示實驗處理和某個變項多多少少有點正效果，則它們的結論可能是相當強的 (Glass et al, 1981)。

第三，出版偏差 (publication bias) 的問題：所謂出版的偏差是指只針對已出版的研究做統合分析，而忽略了未出版的研究。一般而言，得到出版社認同而出版的研究通常是達顯著水準的，因此，Glass et al (1981) 極力建議應將未出版的研究納入，特別是碩士及博士論文，以便比較出版的研究與未出版的研究之差別。

第四，資料不獨立 (lumpy data) 的問題：批評者認爲，統合分析將一個研究的多項結果當作獨立的研究結果來分析，將會給讀者一個錯誤的印象，以爲這些結果很一致，但其實這些研究結果之間是

不獨立的（詹志禹，民 77）。Glass et al (1981) 承認這是一個很中肯的批評，同時也提供了很多個解決方式，原則上，如果單一研究的多項結果是來自對依變項很相似的定義，則可將它們合併，否則，最好根據對依變項不同的定義，將結果歸結到不同類別的研究，分開實施統合分析。

(六) 克服統合分析缺失的方法：

1. 對於出版偏差的校正：

對於統合分析中出版偏差的現象，Rosenthal (1979) 首先提出計量的運算方法並將問題描述出來，而 Cooper (1979) 稱這種現象為「文履現象」問題。針對此問題，最佳的解決之道為計算尚須幾篇不顯著的研究方能推翻統合分析的結論，而這樣的研究稱為「Fail-Safe N」，因此若所計算的「Fail-Safe N」很大時，即表示出版偏差在此統合分析中的影響不大，但相對的，若是所得的「Fail-Safe N」很小時，便表示在此研究中出版偏差的問題很嚴重。

2. 對研究品質的校正：

針對此問題，最好的方式便是對各研究的結果進行加權 (weight)。

對於個別研究的加權方式有很多不同的型式，如 Hedges & Olkin (1985) 是使用各研究變異數的倒數為加權量、Rosenthal (1984, 1991) 則採用自由度作為加權數，而 Hunter et al (1982) 則採用樣本數為加權數。除此之外，Hunter et al 亦針對工具信度所造成的影響加以校正，以作為個別研究品質的依據。

3. 同質性的考驗 (test of homogeneity)：

在統合分析中須假設「所蒐集的各研究間是同質的」，一旦這些研究間是異質的，則不同研究間是考驗相同假設的假定將難以成立，意味著不可將這些研究放在一起做統合分析，針對此問題最佳的處理方式便是進行同質性的考驗。當進行同質性考驗後，若發現研究間是異質時，則需從各種原因中找出可能的中介變項，不可輕易地將這些研究放在

一起進行統合分析。

4. 資料不獨立的問題：

統合分析中常將同一研究的多項結果當作多個獨立的研究結果來分析，如此將給讀者錯誤的印象，以為這些結果很一致，但其實這些結果間是不同的（詹志禹，民 77）。Kulik (1983) 認為如此將膨脹統計考驗中的樣本大小與效果，並對這些結果造成混淆（引自 Wolf, 1986, p46），而對此問題 Rosenthal (1984) 建議最好是分別進行統合分析。

5. 信度與效度的問題：

所謂的信度問題乃是指統合分析過程中是否涵蓋相同的研究、而這些研究的一致性如何？此外，研究資料特徵的登錄是否一致、統合分析計算過程中的計算及記錄方式是否正確等都是信度所要探討的部份 (Wolf, 1986)。此外 Hunter et al (1982) 亦提出關於無信度的測量 (unreliability measures)、全距不一致的偏誤的校正，而這校正主要仍在提高統合分析的信度。

(七) 效果量的指標：

效果量的顯示法有相當多種，而不同的顯示法通常意味著不同的統合分析技術及理念。因此，決定以何種效果量指標為研究依據便牽涉到整個分析研究的結果。一般說來，

$$\text{考驗統計量} = \text{效果量} \times \text{樣本大小}$$

而 Rosenthal (1984, 1991) 將目前所最常使用的效果量指標分為如表一的三類。

(八) 效果量的說明

1. 以「二項式效果量顯示表」(binomial effect size display, 簡稱 BESD) 表示：

二項式效果量顯示法是 Rosenthal & Rubin (1982) 所發明的，而經過這二項式效果量顯示表的轉換，我們將可以瞭解實際效果量的重要性。而在這轉換過程中必須使用 r 值，因此若是以 Hedges & Olkin (1985) 技術進行分析者，便須先將 d 值轉

表一 三種效果量的指標

	效果量指標	定義
Product moment correlations (r) and functions of r	Pearson r r/k Zr Cohen's q	$\Sigma (ZxZy)/N$ $r/\sqrt{(1-r^2)}$ $(1/2)\log_e[(1+r)/(1-r)]$ Zr_1-Zr_2
Standardized differences Between means	Cohen's d Glass's Δ Hedges's g	$(M_1 - M_2) / \sigma$ (pooled) $(M_1 - M_2) / S(\text{control group})$ $(M_1 - M_2) / S(\text{pooled})$
Differences Between Proportions	Cohen's g d' Cohen's h	$p - .50$ $p_1 - p_2$ $p_1^b - p_2^b$

(引自 Rosenthal, 1991, p35)

換為 r 值。轉換這公式如下：

$$\text{Success rates} = 0.50 \pm (r/2) \quad \text{公式 2-1}$$

而在二項式效果量顯示法中，所顯示出來的是整個實驗在實驗組與控制組間成功機率差異的比率（實驗組與控制組相比所得），或謂成功率的改變 (Rosenthal, 1991)。

2. 文履現象的分析 (file drawer analysis)：

所謂文履現象乃是指對分析的主題資料蒐集不全所可能引起的誤差，也就是說對所蒐集資料完整性的懷疑 (Rosenthal, 1995)。而要回答這個問題前便須先計算文履現象，而這現象可以 Fail-safe N 指數來回答。若 Fail-safe N 值很大時，即表示研究結

果相當可信，相對的，若 Fail-safe N 值很小時，即表示統合分析後的結果很容易被推翻。

3. 效果量計算的共同語言 (a common language effect size statistic)：

效果量計算的共同語言簡稱 CL，它是 McGraw & Wong (1992) 所提出來用於解釋效果量的一種方式。它的基本理念為當效果量 ES=0 時，則 CL 為 .50，不過在結果顯示時通常將 CL × 100，因此當效果量 ES=0 時，則 CL 將顯示為 50。其實 CL 就是將效果量 d 值轉換為標準分數 z 的機率值。

4. 不重疊量數 (measures of nonoverlap)：

不重疊量數是 Cohen(1977) 所提出用以解釋效果量的方式，並經 Glass (1977) 引介至統合分析的領域來。而不重疊量數一般均用圖形來表示，Cohen 提供一個不重疊百分等級 (percentiles of nonoverlap) U3 作為效果量 d 值的考驗力分析 (power analysis)，而這 U3 也相當於 d 值在常態曲線中的位置，因此經由這常態曲線位置的對照，我們將得到實驗組超過控制組的百分比 (Wolf, 1986)。例如當平均效果量 d 值為 .54 時，參照常態曲線得知其面積為 .7054，這結果意味著實驗組接受實驗處理後其某方面的平均表現將比未接受實驗處理的控制組中 70.54% 的人來的好。因此這實驗處理被預期將使實驗組的百分等級從 50 進步到百分等級 70.54。

研究方法

一、研究架構

研究者依所蒐集到的資料性質，及本研究的特性、研究邏輯等，進而建立起本研究的架構，如圖一。

二、研究變項

在本研究中，研究者依過往「柯氏性格量表」

的使用狀況，設定了六個自變項，分別為：心理症病人、不同心理症病人、生理症病人、犯人、特殊組群、性別。並以這六個自變項為依據，探究其施測的分數以為依變項，並作為效度概化統合分析中的資料，而此六個自變項是以 Hedges & Olkin (1985) 的技術從事統合分析。此外，有關「相關係數」的研究則自成一類，以 Hunter & Schmidt (1990)、Rosenthal (1991) 技術加以分析。

三、研究對象及資料搜尋

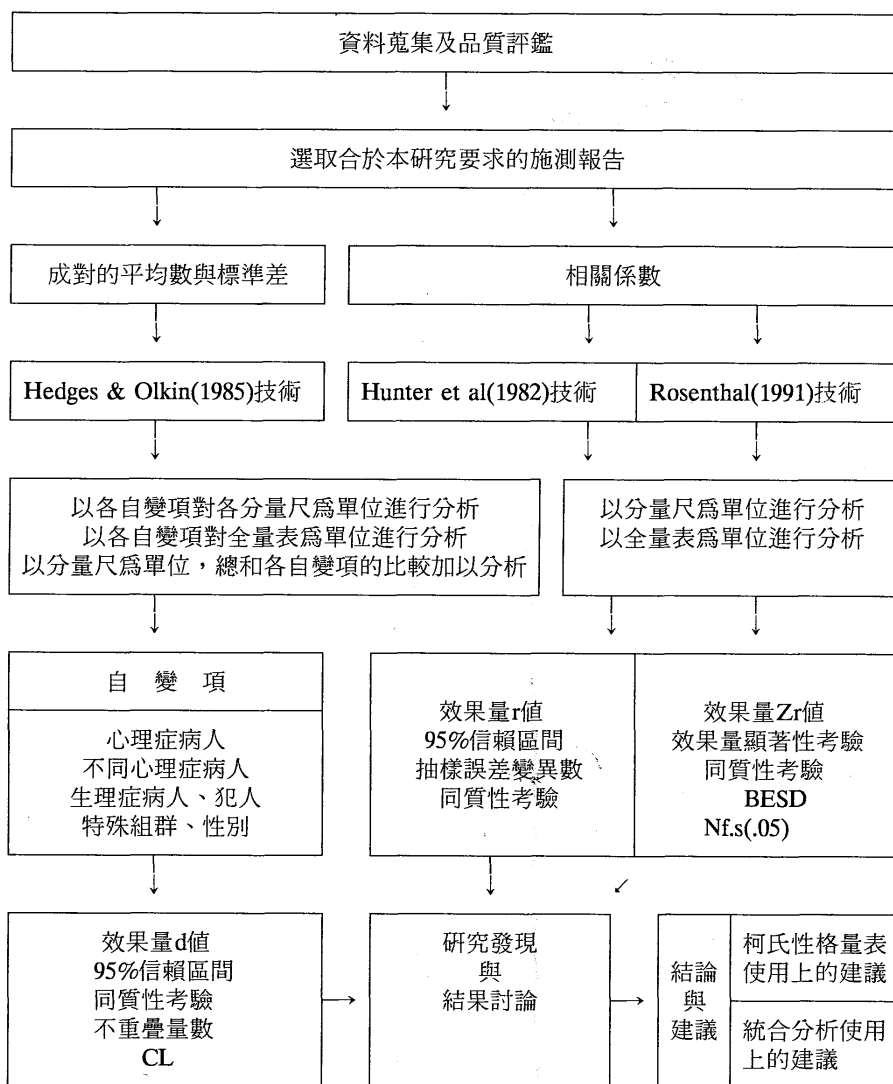
本研究的研究對象為「柯氏性格量表」自民國53年編製完成至今（民國87年4月）所有的施測結果，不論公開發表與否，全部皆為本研究的研究對象及所要搜尋的目標。

而這些「柯氏性格量表」個別施測結果的搜尋，以電腦檢索國家圖書館編製的「中華民國期刊論文索引目錄」、「中華博碩士論文資料庫」以及國立台灣師範大學圖書館編製的「期刊論文檢索系統」、

「圖書目錄」等資料庫，輸入關鍵字「柯氏性格量表」、「心理測驗」、「測驗報告」、「柯永河」、「效度概化」、「統合分析」等加以查詢。其次以溯洄法從已經蒐集到的研究中，從其所附參考文獻追溯以前的施測結果。

此外商請柯氏性格量表的編製者柯永河先生提供其所作的專案研究報告共計三冊，以作為統合分析的資料。而累積至八十七年四月一日止計檢索得87篇與「柯氏性格量表」有關的研究。

圖一 本研究之研究架構



四、研究資料的登錄

研究資料（「柯氏性格量表」的施測報告）蒐集完畢後，依下列特徵一一登錄下來：(1)研究者。(2)出版年代。(3)出版與否。(4)期刊名稱。(5)受試者人數。(6)平均數、標準差。(7)與效標間的相關係數。(8)依本研究的分類法是屬於哪一自變項。(9)測驗目的。(10)其它的計量指標，如 t ， F 等等。(11)主試者的學經歷。(12)研究性質，是學位論文或一般研究報告。而之所以須登錄如此多的特徵，除作為統合分析計量結合之用外，主要是為便於研究後半階段從事中介變項探索之用。

最後則將這些經登錄後的資料，依自變項的平均數、一般人（或常模）的施測平均數、自變項的標準差、一般人（或常模）的施測標準差、自變項的受試人數、一般人（或常模）的受試人數等順序（在相關係數資料中則以相關係數、自變項的受試人數、一般人的受試人數等順序）輸入電腦中成為資料檔，以待電腦 Meta-HC 程式的執行。

五、效果量的計算

目前最常使用的效果量指標如表一所示共有三種，在本研究中將就所蒐集到的施測結果，因其性質有別而採用不同的指標。在成對的平均數與標準差中，我們將採用 Hedges & Olkin (1985) 技術，所以是以 d 值作為效果量指標，其求法如公式 3-1。

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s} \quad \text{公式 3-1}$$

其中分母 s 是採用聯合標準差。

此外在相關係數的部份則採用 r 值作為效果量的指標，並將此值轉換成 Z_r 值，因為在 Rosenthal 的觀念裡，未經 Fisher 轉換過的 r 值會產生非線性的偏誤。其中 Rosenthal 認為 Z_r 值仍有些許的偏差，因此需要再加以校正，而本研究即採用 Rosenthal 的建議，其校正公式如公式 3-2 所示（此外若要再更精確則可將校正過的 Z_r 再轉換為 r 再多校正一

次，不過第二次校正時所使用的仍是原始的 Z_r ，但 r 值則是第一次校正並經 Z_r 轉換）：

$$Z_{r_{ci}} = Z_{r_i} - \frac{r_i}{2(N-1)} \quad \text{公式 3-2}$$

其中 $N = n_1 + n_2$ 。

此外若所蒐集的施測報告是以其它的統計量數如 t 、 F 或其它檢定值表示時，則便加以轉換（詳細內容請參閱 Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Lyons, 1997; Rosenthal, 1991）。

六、研究工具、材料

(一) 柯氏性格量表：

「柯氏性格量表」是柯永河（民 53）為了多方瞭解個人與團體心理健康與不健康而發展出來的量表。修訂前後有十一分量尺是相同的，而這十一個分量尺也是本研究所要探究的目標，這十一個分量尺的名稱分別為疑心量尺、慮病量尺、離群量尺、信心量尺、自卑量尺、焦慮量尺、強迫性格量尺、性壓抑量尺、攻擊量尺、自我強度量尺以及獨立量尺。

(二) 統合分析技術：

本研究參酌目前統合分析方法的發展後，決定採行目前最主要的三種技術作為本研究的依據，並將三種技術所得的結果相互比較，以探究不同技術之間的運算結果是否有所差異，而這三技術簡述如下：

1. Hedges & Olkin (1985)

此技術最大的特性為以「成對的平均數、標準差」為分析資料，以變異數的倒數為加權數。它的效果量指標是 d 值，而 d 值它是每一對比較的標準化，經此標準化的過程後，不同單位的原始資料即可進行計量性的結合，以探究累積資料的效果量為何。

2. Rosenthal (1991)

Rosenthal 技術最主要的特徵在於以相關係數經 Fisher's z 轉換後的 Z_r 為分析資料；以自由度為加權數，並特別強調同質性的分析。Rosenthal 指出在作結合之前必須先作比較的工作，若比較的結果顯示其同質性達顯著，則不可進行結合，須進行中介變項的探究；若是同質性考驗不顯著後才可作結合的工作。

3. Hunter & Schmidt (1990)

此技術最大的特徵是以相關係數為分析資料，以樣本數為加權數，並有多種校正誤差的技術，包括抽樣誤差、測量誤差、全距限制等。此外此技術對於效果量它不採用顯著性的考驗，而是採用信賴區間的方式進行，主要是因為 Hunter & Schmidt 認為信賴區間它所犯的第一類錯誤可以控制在 5%，而顯著性考驗則不行。

七、本研究所使用的電腦運算程式：

本研究的研究工具為電腦運算程式 Meta-HC，此程式為研究者與張沼澤共同發展而成。Meta-HC 在執行時，它會先出現三個選項：

1. 有方向性考驗。
2. 無方向性考驗。
3. 相關係數資料。

在選項 1、2 有方向性及無方向性的考驗中，它們所處理的是成對的平均數與標準差的資料。其差別主要在於使用無方向性的檢定求取每一組比較的效果量 d 值時，會取其絕對值，而有方向性的檢定則無。在本研究中的自變項「特殊群體」、「不同心理症病人」的考驗是採用無方向性的方法。而在「心理症病人」、「生理症病人」、「犯人」及「性別」中是採用有方向性考驗方法。選項 3 則專門處理相關係數的資料，它是以 Hunter & Schmidt (1990) 與 Rosenthal (1991) 的技術為主。

八、實施程序

本研究是以「柯氏性格量表」過往的累積研究，使用「統合分析方法」以探究其效度概化的程度。研究之初，先將本研究的領域定義清楚—界定為「柯氏性格量表」自民國 53 年編製完成後至今所有施測的結果，以進行「柯氏性格量表」各分量尺及「柯氏性格量表」全量表效度概化程度的檢驗。其次建立「柯氏性格量表」的評鑑標準。

之後便開始尋找「柯氏性格量表」過往的施測結果。而「柯氏性格量表」累積研究的收集是以電腦檢索、溯洄法及商請柯永河先生提供相關研究等三種方式同時進行。而經上述程序後共計蒐集到 87 篇有關「柯氏性格量表」的相關研究，但合於本研究者只有 52 篇。

之後則以研究假設中所列的六個自變項為標準將此 52 篇施測報告作歸類，以知悉目前「柯氏性格量表」的使用情況。然後再決定本研究所使用效果量的種類， d 值與 r 值，以及所使用的統合分析方法—Hedges & Olkin (1985) 技術、Hunter & Schmidt (1990) 技術、Rosenthal (1991) 技術。

在資料搜集完成並登錄編碼後，再商請台中縣海墘國小兩位教師予以覆核登錄，而此二位覆核者的登錄結果與研究者的登錄結果其一致性達 98%，其差異處主要在於筆誤與登錄的位置有所失誤所致，而這情形在登錄信度檢核後已予以改正。

資料（個別「柯氏性格量表」施測結果）登錄及分類（自變項）後，則依個別研究所使用的統計量數（ \bar{X} 、 t 、 r 、 F ）及其它特徵進行轉換，然後依研究架構所述將每一筆資料輸入電腦後，再使用 Meta-HC 進行統合分析。

最後則依研究的結果提出本研究的結論與建議以為後續研究者的參考。

結果與討論

經過統合分析後，各分量尺的效果量如如表二所示：

表二 「柯氏性格量表」各量尺效果量摘要表

	量尺一	量尺二	量尺三	量尺四	量尺五	量尺六
心理症病人	0.7127*	0.9083*	0.6785*	-0.3537*	0.8820*	0.7425*
不同心理症病人	0.4917*	0.2988*	0.3851*	0.3827*	0.4999*	0.4411*
生理症病人	0.0393	0.5806*	0.2247*	-0.3902*	0.3985*	0.3008*
犯人	0.4842*	0.5187*	0.2947*	-0.1429*	0.4503*	0.3596*
特殊組群	0.5609*	0.5990*	0.5947*	0.3173*	0.7092*	0.6238*
男女	0.1068*	-0.0489*	0.0369*	0.4210*	-0.1591*	-0.2327*
相關 Hunter 技術	.2778*	.2542*	.2904*	.1416*	.3157*	.3312*
相關 Rosenthal技術	0.2916*	0.2675*	0.3070*	0.1482*	0.3389*	0.3558*
	量尺七	量尺八	量尺九	量尺十	量尺十一	
心理症病人	0.4461*	0.4495*	0.2815*	-0.6288*	-0.3944*	
不同心理症病人	0.4748*	0.2626*	0.3089*	0.4849*	0.5307*	
生理症病人	0.3182*	0.4832*	-0.1811*	-0.2501*	-0.0766	
犯人	0.4139*	0.1242*	1.2070*	-0.1659*	-0.0567	
特殊組群	0.3048*	0.2787*	0.4103*	0.4533*	0.3453*	
男女	0.0674*	-0.6982*	0.3946*	0.1723*	0.3621*	
相關 Hunter 技術	.1619*	.1482*	.1932*	.2546*	.3001*	
相關 Rosenthal技術	0.1648*	0.1510*	0.1984*	0.2662*	0.3209*	

* $p < .05$

註：1. 除不同心理症病人、特殊組群部份使用無方向性考驗外，其餘均使用有方向性考驗。

2. 各分量尺的效果量指標為d值。
3. 在Hunter技術中各分量尺的效果量指標為r值。
4. 在Rosenthal技術中各分量尺的效果量指標為Zr值。
5. 表中的顯著是統計上的意義，而實際的意義可參考Cohen所提供的數值。
6. 根據Cohen所提供關於效果量d值的參考標準而言：效果量在0.8以上者表示效果量高，效果量在0.5左右表示效果量中等，效果量在0.2左右表示效果量低。
7. 根據Cohen所提供關於效果量r值的參考標準而言：效果量在0.5以上者表示效果量高，效果量在0.3左右表示效果量中等，效果量在0.1左右表示效果量低。

一、第一分量尺：疑心量尺

在第一分量尺中，心理症病人、不同心理症病人、犯人以及特殊組群中的效果量均達到 Cohen 標準中的中到高的效果量，可見柯氏性格量表的第一分量尺對於前述這些族群具有顯著區辨能力，而這其中又以在心理症病人最具成效。但在生理症病人與一般人之間以及男女之間的區辨力則不甚理想。

疑心量尺在心理症病人的測量中，其加權平均效果量 d 值為 0.7127，此結果告訴我們，疑心量尺在心理症病人與正常人間是具顯著區辨力的。若將 d 值轉換成不重疊量數為 .7621，顯示心理症病人接受疑心量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 76.21。再將效果量 d 值轉換成 CL 值為 .6915，這意味著當將疑心量尺施測於心理症病人時，每一百次施測中將有 69.15 次得到較正常人為高的測驗分數。

另外在此部份的 $Nf.s(.05)$ 值高達 2338，比 Rosenthal (1991) 所提出的容忍數 115（此值等於 $5K+10$ ， K 為成對的比較數）大的多，因此此結果的有效性應是相當理想的。

另外再從相係數值部份來看：由 Hunter 技術所得到的效果量 r 值為 .2778，Rosenthal 技術的 Zr 值為 0.2916。若以效果量 d 值約為效果量 r 值與 Zr 值的兩倍來看，不論是成對的資料或是相關係數的資料，所得的結果均是相當接近，由此更加深我們對此分量尺效果量所代表的意義的信心。

二、第二分量尺：慮病量尺

在第二分量尺中，心理症病人、生理症病人、犯人以及特殊組群的效果量均達到 Cohen 標準中的中到高的效果量，可見柯氏性格量表的第二分量尺對於前述這些族群具有顯著區辨能力，而這其中又以在心理症病人最具成效，其加權平均效果量高達 0.9083，但在生理症病人與一般人之間以及男女之間的區辨力則不甚理想。

此結果告訴我們，慮病量尺在心理症病人與正常人之間是具有頗高的區辨力，且具有積極的效度證據。而在此部份不重疊量數為 .8186，顯示心理症病人接受慮病量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 81.86。而 $CL=.7389$ 意味著將慮病量尺施測於心理症病人時，每一百次將有 73.89 次得到較正常人為高的測驗分數。因此，慮病量尺在心理症病人的區辨功能上是相當顯著的。

另外在此部份的 $Nf.s(.05)$ 值高達 3585，比 Rosenthal 所提出的容忍數 115（此值等於 $5K+10$ ）大的多，因此前述結果的有效性應是相當理想的。

然而在此同時我們發現，雖然此分量尺在心理症病人的加權平均效果量高達 0.9083，但在不同心理症病人中的加權平均效果量卻只有 0.2988，由此可見慮病量尺雖可將心理症病人區辨出來，但若進一步要求區辨不同心理性疾病則力有未逮。

另外再從相係數值部份來看：由 Hunter 技術所得到的效果量 r 值為 .2542，Rosenthal 技術的 Zr 值為 0.2675。若以效果量 d 值約為效果量 r 值與 Zr 值的兩倍來看，不論是成對的資料或是相關係數的資料，所得的結果均是相當接近，由此更加深我們對此分量尺效果量所代表的意義的信心。

三、第三分量尺：離羣量尺

在第三分量尺中，只有心理症病人及特殊組群的效果量達到 Cohen 標準中的中到高的效果量，而其它四個自變項中的效果量都只達 Cohen 的低度標準，而這與其它分量尺比較起來是屬於區辨力較弱的分量尺，也就是說其效度概化的程度較低。

此部份效果量最高的是在心理性病人上，其加權平均效果量 d 值為 0.6785，轉換成不重疊量數為 .7517，顯示心理症病人接受離羣量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 75.17。而 $CL=.6844$ 意味著將離羣量尺施測於心理症病人時，每一百次將有 68.44 次得到較正常人為高的測驗分數。

另外此部份的 $Nf.s(.05)$ 值高達 2071，比 Rosenthal 所提出的容忍數 115 (此值等於 $5K+10$) 大的多，因此前述結果的有效性應是相當理想的。

四、第四分量尺：信心量尺

在第四分量尺中，並未有自變項的加權平均效果量超過 0.5，但卻都集中在 0.35 ~ 0.45 之間，或許我們可說第四分量尺是具有效度概化的程度，因為他在不同群體之間的有效性都差不多，只不過其概化的程度不高。另外再從相係數值部份來看：由 Hunter 技術所得到的效果量 r 值為 .1416，Rosenthal 技術的 Zr 值為 0.1482。若以效果量 d 值約為效果量 r 值與 Zr 值的兩倍來看，不論是成對的資料或是相關係數的資料，所得的結果均是相當接近，可見前述的論述是可靠的。

其中加權平均效果量最大的是在男女的比較中，其 $d = 0.4210$ ，。而此結果告訴我們，信心量尺在男生與女生之間是具有區辨力的。此外，此部份的不重疊量數為 .6628，顯示男生接受信心量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 66.28。而 $CL = .6179$ 意味著將信心量尺施測於男生時，每一百次將有 61.79 次得到較女生為高的測驗分數。

而此部份的 $Nf.s(.05)$ 值高達 42700，比 Rosenthal 所提出的容忍數 440 (此值等於 $5K+10$) 大的多，因此前述結果的有效性應是相當理想的。正因此，更加支持我們前述的看法，顯示信心量尺在男女之間有其一定的區辨效度。

五、第五分量尺：自卑量尺

在第五分量尺中，除了在男女比較中效果量僅為 -0.1519 外，其餘自變項的效果量均達中高程度，顯見柯氏性格量表的第五分量尺是具有理想的區辨能力，而這其中又以在心理症病人最具成效，效果量高達 0.8820。

這結果告訴我們，自卑量尺在心理症病人與正

常人之間是具有區辨力的，因此我們可說自卑量尺在心理症病人與正常人之間具有積極的效度證據。此外，此部份的不重疊量數為 .8106，顯示心理症病人接受自卑量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 81.06。而 $CL = .7324$ 意味著將自卑量尺施測於心理症病人時，每一百次將有 73.24 次得到較正常人為高的測驗分數。因此，自卑量尺在心理症病人的施測上是效度是相當理想的。

另外，此部份的 $Nf.s(.05)$ 值高達 3689，比 Rosenthal 所提出的容忍數 115 (此值等於 $5K+10$) 大的多，因此前述結果的有效性應是可接受的，且受出版偏誤的影響相當小。

六、第六分量尺：焦慮量尺

在第六分量尺中，加權平均效果量最小的自變項是男女，但是其 d 值仍有 -0.2327，這與其它分量尺比較起來是屬於較少見的，顯示焦慮量尺的區辨能力頗強，而這結果與柯永河的臨床經驗是相同的，均顯示焦慮量尺在這十一個分量尺中是效度最高的一個。

而這可再從相係數值部份來看：由 Hunter 技術所得到的效果量 r 值為 .3312，Rosenthal 技術的 Zr 值為 0.3558。若以效果量 d 值約為效果量 r 值與 Zr 值的兩倍來看，不論是成對的資料或是相關係數的資料，所得的結果均是相當接近，且其數值均相當大，顯示其具有積極的效度證據，由此更加深我們對此分量尺的信心。

而焦慮量尺加權平均效果量最大者為心理症病人，其 d 值為 0.7425，此結果告訴我們，焦慮量尺在心理症病人與正常人之間是具有頗高的區辨力，且其效度證據相當理想。

而此部份的不重疊量數為 .7703，顯示心理症病人接受焦慮量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 77.03。而 $CL = .7019$ 意味著將焦慮量尺施測於心理症病人時，每一百次將

有 70.19 次得到較正常人為高的測驗分數。因此，焦慮量尺在心理症病人的施測上是相當有效的，且有相當積極的效度證據。

另外在此部份的 $Nf.s(.05)$ 值高達 4553，比 Rosenthal 所提出的容忍數 115 (此值等於 $5K+10$) 大的多，因此前述結果的有效性應是可接受的，且受出版偏誤的影響相當小。

因此我們可說焦慮量尺它是一個頗值得信賴的量尺，且其效度概化的程度頗為理想。

七、第七分量尺：強迫性格量尺

在第七分量尺中，並未有自變項的加權平均效果量超過 0.5，其中除了自變項男女的 d 值為 0.0674 外，其它的自變項其加權平均效果量 d 值都集中在 0.30 ~ 0.45 之間，或許我們可說第七分量尺是具有效度概化的程度，因為他在不同群體之間的有效性都差不多，只不過其概化的程度不高。

另外再從相係數值部份來看：由 Hunter 技術所得到的效果量 r 值為 .1619，Rosenthal 技術的 Zr 值為 0.1648。若以效果量 d 值約為效果量 r 值與 Zr 值的兩倍來看，不論是成對的資料或是相關係數的資料，所得的結果均是相當接近，可見前述的論述是可靠的，都顯示強迫性格量尺其區辨能力只具中等程度，但是卻具有效度概化的雛形。

此部份較為特殊的情形發生在男女的比較中，因為其加權平均效果量與其它自變項相形之下實在是太小了，只有 0.0674。此結果告訴我們，強迫性格量尺在統計上，男生與女生之間是具有區辨力的，但若將 $d = 0.0674$ 參照 Cohen 所提的標準來說它是屬於相當低的效果量。因此，強迫性格量尺在男女之間雖然在統計上達顯著，但是其效度證據相當薄弱，而這情形可從不重疊量數及 CL 值中再加以說明：

此部份的不重疊量數為 .5279，顯示男生接受強迫性格量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 52.79。而 $CL = .5199$ 意味著

將強迫性格量尺施測於男生時，每一百次將有 51.99 次得到較女生為高的測驗分數。由這些論述中可知，雖然此部份的平均效果量達顯著水準，但是其所代表的意義相當有限，且若將 $d = 0.0674$ 轉換成效果量 r 值，則只有 .0104，可說與 0 相當接近。此外再從不重疊量數值為 .5279 來看，其區辨力 z 也將接近 0，所以我們可說當以強迫性格量尺做為男女之間的區辨工具時其效度證據相當薄弱。而此部份的 $Nf.s(.05)$ 值高達 7581，比 Rosenthal 所提出的容忍數 430 (此值等於 $5K+10$) 大的多，因此前述結果的有效性應是可接受的。正因此，更加支持我們前述的看法。

八、第八分量尺：性壓抑量尺

在性壓抑量尺中，加權平均效果量最大者為男女比較。這現象較為少見，從其它分量尺的例子可知，自變項男女的加權平均效果量通常是最小的，因此在性壓抑量尺中便顯得較為特殊，而這結果與柯永河先生編製時的構念是一致的，顯示統合分析的結果相當正確，亦顯示柯永河先生編製性壓抑量尺時，從構念的形成至試題的編製過程是頗為理想的。

由表二中可知性壓抑量尺在男女差異的施測中，其加權平均效果量 $d = -0.6982$ ，顯著性考驗及信賴區間分別小於臨界值及不包含 0，可見此值是顯著地異於 0。此結果告訴我們，性壓抑量尺在女生與男生之間是具有區辨力的，而 $d = -0.6982$ 在 Cohen 所提的標準中是屬於較高的效果量，因此性壓抑量尺在男女之間的區辨力頗強，這同時也顯示性壓抑量尺在男女間具有積極的效度證據。而此部份的不重疊量數為 .2451，顯示男生接受性壓抑量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 75.49。而 $CL = .3121$ 意味著將性壓抑量尺施測於男生時，每一百次將只有 31.21 次得到較女生為高的測驗分數，由此顯示當以性壓抑量尺做為男女之間的鑑別工具時其效度頗為理想。而此

部份的 $Nf.s(.05)$ 值高達 110574，遠比 Rosenthal 所提出的容忍數 430（此值等於 $5K+10$ ）大的多，因此前述結果的有效性應是可接受的。正因此，更加支持研究者的論點，顯示性壓抑量尺在男女之間具有頗高的區辨力，效度證據相當理想。

九、第九分量尺：攻擊量尺

在攻擊量尺中，加權平均效果量最大者為犯人中的 1.2070，而這 1.2070 也是整個研究中所產生的最大加權平均效果量值，顯示攻擊量尺在犯人的施測中最具成效，效度意義最顯著。但在其它自變項中，其加權平均效果量則不甚理想，顯示其區辨能力並不強。

由表二中可知攻擊性量尺在犯人的測量中，其加權平均效果量高達 $d = 1.2070$ ，其顯著性考驗及信賴區間分別大於臨界值及不包含 0，可見此值是顯著地異於 0。此外若將 $d = 1.2070$ 參照 Cohen 所提的標準，它是屬於相當高的效果量，比 Cohen 所提高效率量的 $d = 0.8$ 還大許多。因此從這結果中可知，攻擊量尺在犯人與一般人之間是具有積極的效度證據，顯示其區辨能力相當強。因此我們可說當以攻擊量尺作為犯人與一般人之間的鑑別工具是相當合適的。因此，我們可說攻擊量尺在犯人的施測中其區辨力是相當強烈的，甚至可說在整個「柯氏性格量表」中，以攻擊量尺施測於犯人時的效度最高。

而此部份的不重疊量數為 .8869，顯示犯人接受攻擊量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 88.69。而 $CL = .8023$ 意味著將攻擊量尺施測於犯人時，每一百次將有 80.23 次得到較一般人為高的測驗分數。此外，此部份的 $Nf.s(.05)$ 值為 370 也比 Rosenthal 所提出的容忍數 35（此值等於 $5K+10$ ）大不少，顯示此部份的結果相當穩固。因此出版偏誤對此部份的影響力甚小。

十、第十分量尺：自我強度量尺

在自我強度量尺中，加權平均效果量達到中度或高度水準的有心理症病人、不同心理症病人以及特殊組群。相對的在生理症病人，犯人以及男女比較中，其加權平均效果量都只具低度的水準。由此顯見自我強度量尺他在不同群體中的效度是差別頗大的。最高從心理症病人的加權平均 d 值 -0.6288，到最低男女比較的加權平均 d 值 0.1723，差異相當顯著。

在心理症病人的測量中，其加權平均效果量 d 值為 -0.6288，此結果告訴我們，自我強度量尺在心理症病人與正常人之間是具有區辨力的，由此顯示當以自我強度量尺作為心理症病人與正常人之間的區辨工具時，其效度證據頗為理想。相對的，在男女差異的施測中，其加權平均效果量 $d = 0.1723$ ，雖然這在統計上是達顯著，表示在男女之間能具有區辨力，可是在實用上的價值不高，因為要達到統計上的顯著只要增加樣本數即可。這可再從不重疊量數及 CL 值的比較中印證。

在心理症病人中的不重疊量數為 .2643，顯示心理症病人接受自我強度量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 73.57。而 $CL = .3300$ 意味著將自我強度量尺施測於心理症病人時，每一百次將只有 33.00 次得到較正常人為高的測驗分數，這樣的結果顯示出測驗對研究者的助益是存在的。

而在男女比較中，不重疊量數為 .5714，顯示男生接受自我強度量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 57.14。而 $CL = .5517$ 意味著將自我強度量尺施測於男生時，每一百次將有 55.17 次得到較女生為高的測驗分數。由此可看出當以自我強度量尺作為男女之間的區辨工具時，其效益是相當微弱的。

而前述這些論述經 $Nf.s(.05)$ 的考驗後，均比 Rosenthal 所提出的容忍數大的多，因此其立論應是可靠的。

十一、第十一分量尺：獨立量尺

在獨立量尺中，其加權平均效果量 d 值大於 0.5 的只有自變項「不同心理症病人」一個，在心理症病人、特殊組群以及男女比較中，其加權平均 d 值都只有 0.35 左右，而在自變項「犯人」及「生理症病人」中，其加權平均 d 值更只有 -0.0766 及 -0.0567。

由這樣的結果中我們可知，獨立量尺他最適用於不同心理性病人的區辨上，若使用於男女之間及特殊組群則效度將不高，一旦做為犯人及生理性病人的診斷工具，則其效度將趨近於 0。

而在加權平均效果量最大的不同心理症病人上，其加權平均效果量 $d = 0.5307$ ，此結果告訴我們，獨立量尺在不同心理症病人之間是具有區辨力的，顯見獨立量尺在不同性質的精神病患之間所提供的效度證據頗為理想。此外，此部份的不重疊量數為 7019，顯示不同心理症病人接受獨立量尺後，其區辨力將由未接受施測前的百分位 50 進步到百分位 70.19。而 $CL = .6480$ 意味著將獨立量尺施測於不同心理症病人時（此時須假設是固定的兩種不同心理性疾病），每一百次將有 64.80 次得到較另一種心理疾病患者為高的測驗分數。

而此部份的 $Nf.s(.05)$ 值達 5431，遠比 Rosenthal 所提出的容數 290（此值等於 $5K+10$ ）大的多，所以出版偏誤對此結果的影響相當有限，因此我們有相當的信心支持前述的說明。

而在加權平均效果量最小的犯人中，其加權平均效果量 d 值為 -0.0567，其顯著性考驗 z 值為 -0.8220，並未達顯著水準，且信賴區間 [-0.1918, 0.0785] 又包含 0，因此可見效果量 $d = -0.0567$ 是隨機而得的，與 0 並無顯著差異，且以效果量 $d = -0.0567$ 來看，也不具實用上的價值，因為它跟 $d = 0$ 幾乎是相同的。而這結果告訴我們，若以獨立量尺作為犯人與一般人之間的區辨工具是不甚理想的。此外，此部份的 $Nf.s(.05)$ 值為 7，比 Rosen-

thal 所提出的容數 35（此值等於 $5K+10$ ）還小，顯示此部份的結果較不穩定。

總括而言，獨立量尺除了在不同心理症病人的區辨上具有理想的效度證據外，用於其它自變項時其對研究者的助益並不大。

在本節裡我們從各分量尺在各自變項的效果量大小中分析發現，各分量尺在各自變項中其效度證據差異頗大，加權平均效果量最大可達 1.2070，但最小卻只有 0.0393，可見若將「柯氏性格量表」各分量尺視為獨立的單位看待，其效度的狀況是頗受情境特殊性的限制的，顯示其具有頗高的區辨能力。如在疑心量尺中效果量最低的是生理症病人， d 值僅有 0.0393，意味著以疑心量尺作為生理性病人與正常人之間的診斷工具是不恰當的；此外在性壓抑量尺中，效果量最大的是在男女的比較中，其 d 值達 0.6982，顯示性壓抑在男女之間是最具區辨力的。雖有這些差異，但若從整體看來，柯氏性格量表在心理症病人中的加權平均效果較大，顯示這些分量尺在心理症病人中的效度證據較高，或許這樣的結果最符合柯永河於編製「柯氏性格量表」之初的目的；此外若以分量尺為單位來看的話，則類似前述這種趨勢便較不明顯。

此外若從不同的資料及不同的統合分析技術的所產生的結果中加以研判，將會發現結果是相當雷同的，而這樣的結果更加提高我們對本研究結果的信心。因為當資料是成對的平均數與標準差時，使用 Hedges & Olkin 技術所得的結果，與使用相關係數的資料，運用 Hunter & Schmidt 及 Rosenthal 技術所得的統合分析結果相差不大，皆顯示「柯氏性格量表」是一個具有頗高區辨力的人格測量工具。因此經由不同的資料（成對的平均數與標準差、相關係數），不同的技術 (Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1991) 卻能得到相近的答案，可見這結果的可信度相當高。因此我們可說：「柯氏性格量表」是一個區辨效度頗高的人格測驗。

結論與建議

一、結論

根據研究結果，本研究獲致下列結論：

(一)柯氏性格量表各分量尺在不同群體間的效果量及效度：

1. 在整個研究的的 66 個（11 個分量尺×6 個自變項）效果量中，以在自變項「犯人」中的第九分量尺：攻擊量尺的效果量最大，其 d 值達 1.2070。因此可說在整個「柯氏性格量表」中，以攻擊量尺在犯人中的效度證據最為積極、顯著。

2. 效果量在 0.70 以上的有：「心理症病人」中的第一、二、五、六分量尺；「特殊組群」中的第五分量尺以及「男女」中的第八分量尺等六個。因此我們可說，當以這六個分量尺在各自的自變項上，作為受試者之間的區辨工具時，其效度證據是相當顯著的，所得到的分數也值得大家的信賴，因此這六部份是相當適合作為人格診斷的測量工具。

3. 效果量 d 值在 0.40 至 0.70 之間的有：「心理症病人」中的第三、七、八、十分量尺；「不同心理症病人」中的第一、五、六、七、十、十一分量尺；「生理症病人」中的第二、八分量尺；「犯人」中的第一、二、五、七分量尺；「特殊組群中」中的第一、二、三、六、九、十分量尺以及「男女」中的第四分量尺等 23 個。而效果量 d 值在 0.40 至 0.70 之間所代表的意義為：其所具有的區辨力顯著，效度證據也充分。因此我們可說這 23 個分量尺在各自的自變項中是相當合適的人格診斷工具。

4. 而效果量 d 值在 0.10 至 0.40 之間的有：「心理症病人」中的第四、九、十一分量尺；「不同心理症病人」中的第二、三、四、八、九分量尺；「生理症病人」中的第三、四、五、六、七、九、十分量尺；「犯人」中的第三、四、六、八、十分

量尺；「特殊組群中」中的第四、七、八、十一分量尺及「男女」中的第一、五、六、九、十、十一分量尺等 30 個。而效果量 d 值在 0.10 至 0.40 之間是屬於較低程度的水準。因此我們可說在這 30 個分量尺中其區辨能力不大，在各自的自變項中以之作為人格測驗工具的效度意義不高。

5. 此外尚有六個分量尺的效果量 d 值在 0.00 至 0.10 之間，分別是「生理症病人」中的第一、十一分量尺；「犯人」中的第十一分量尺以及「男女」中的第二、三、七分量尺。而效果量在 0.00 至 0.10 之間所代表的意義常是與 0 無顯著差異，因此這六個分量尺在各自的自變項中其區辨力幾乎是 0。

(二)從不同群體看柯氏性格量表各分量尺的效度證據：

1. 疑心量尺在心理症病人的效度證據最高，而在生理症病人的效度證據最低，接近於 0。

2. 慮病量尺在心理症病人的效度證據最高，效果量 d 值達 0.9083，而在男女之間的區辨效度證據最低。

3. 離群量尺在心理症病人的效度證據最高，而在男女之間的區辨效度證據最低。

4. 信心量尺在生理症病人的效度證據最高，而在犯人的效度證據最低。

5. 自卑量尺在心理症病人的效度證據最高，效果量 d 值達 0.8820，而在男女之間的區辨效度證據最低。

6. 焦慮量尺在心理症病人的效度證據最高，而在男女之間的區辨效度證據最低。

7. 強迫性格量尺在不同心理症病人的效度證據最高，而在男女之間的區辨效度證據最低。

8. 性壓抑量尺在男女之間的效度證據最高，而在犯人的效度證據最低。

9. 攻擊量尺在犯人的效度證據最高，效果量值直達 1.2070，而在生理症病人的效度證據最低。

10. 自我強度量尺在心理症病人的效度證據最高，而在犯人的效度證據最低。

11. 獨立量尺在不同心理症病人的效度證據最高，而在犯人的效度證據最低。

二、建議

(一)對「柯氏性格量表」使用上的建議：

1. 在本研究中充份顯示「柯氏性格量表」效度概化的程度頗高，是一個理想的人格測量工具。因此，當須要以測驗作為人格上的診斷工具時，「柯氏性格量表」將是一個區辨力強的工具。

2. 整體而言「柯氏性格量表」是一個有效的人格測驗，但在研究中我們也發現在某些「自變項」的某些分量尺上其效果量甚低，區辨力也不足。因此當使用「柯氏性格量表」表時應特別注意效果量較低的分量尺，在參考解釋上，值得再加以斟酌採用。

3. 「柯氏性格量表」在自變項「男女」中的效果量偏低，因此欲以「柯氏性格量表」對性別傾向作鑑別時，最好再佐以其它資料，以免做出錯誤的判斷。

4. 在本研究的「不同心理症病人」中，顯示其效度概化程度頗高，區辨力也不差，但因為現有資料的限制，同一種心理病症的研究數量相當少，未能獨自成為一自變項以為分析的單位，因此本研究此部份是以不同的心理症病人作為比較的依據，來作為「柯氏性格量表」效度概化的規準。但在此情況下便無法對「柯氏性格量表」在不同心理疾病間的區辨力作一明確的判斷，只能籠統地說「柯氏性格量表」在「不同心理症病人」間是具有區辨力的，因此在個別心理病間的診斷的功能大失。也就是說目前所擁有的是測驗的訊息，但診斷的功能則不足。而這可說是本研究最大的缺憾，也是日後進一步研究的主要範疇。

5. 目前有關「柯氏性格量表」的研究大多只在探討受試者施測後的平均數與標準差，甚少作其它處理，所以本研究中的設計便無法將效度概化創始者 Hunter et al 的特色發揮極至。因此，日後「柯氏性格量表」應致力於發展與其它已有成效的人格量表（如 MMPI）間的相關研究，如此方能增進「柯氏性格量表」的效度證據。

6. 從本研究中發現，效度概化程度較低的量尺大多是測量較正面的心理特質，而這與「柯氏性格量表」初編製時的理念相同，主要乃針對精神異常者所設計。因此，從這樣的結果中我們應建議「柯氏性格量表」再增強測量正向心理特質题目的效度，如此對於人格量表而言將是較為完整。

(二)對「統合分析」使用上的建議：

1. 統合分析法發展至今其技術仍持續在演進之中，尤其是在效果量的解釋上。因此，當在作研究報告時，應特別注意效果量的說明，以使得大家容易明白，如所謂的「效果量的共同語言」(common language effect size, 簡稱 CL)，便是 90 年代以後所發明的。因為只有將統合分析法通俗化、簡易化，其影響力才會擴大，如此社會科學才能有長足的進步，成為真正的「科學之學」。

2. 統合分析法不像一般統計理論只有一種作法，而是有多種不同但並行的技術存在，因此當以統合分析法從事研究時，應先考慮研究的目標，以及所擁有的資料型態，再決定採用何種技術。一般說來，當資料是成對的平均數與標準差時，使用 Hedges & Olkin 的技術最方便；若資料是與某測驗或效標之間的相關係數時則使用 Hunter & Schmidt、Rosenthal 較理想。

3. 當以統合分析法從事測驗效度概化的研究時，最佳的方式應是採用 Hunter & Schmidt 的技術，因為效度概化的觀念乃 Hunter et al 等人所提出來的。不過在 Hunter & Schmidt 技術中所要求的資料常無法獲得，常常只能做到抽樣誤差的校正，而

這樣的結果比 Hedges & Olkin 技術的功效還差。因此，此時應將效度概化作為對測驗效度檢驗的觀念，轉而以其它的技术為之，如本研究即採用 Hedges & Olkin 技術，相信這樣的作法是可以被接受的。

4. 在 Hunter & Schmidt 的技術中，其效果量 r 值並未作 Fisher 的 z 轉換，因此會產生非線性偏誤，所以當使用 Hunter & Schmidt 技術時，最好將其效果量 r 值轉換為 Zr ，如此的結果會是較為正確的。

參考文獻

- 方小萍 (民 84) : 台灣大型會計師事務所女性從業人員人格特質之研究。國立台灣大學會計研究所碩士論文。
- 王順正 (民 84) : 效度概化在運動科學上的應用。中華體育, 9(2), 8-14。
- 呂傳欽、闕清模、卓良珍、藍忠亮 (81) : 原發性纖維肌肉病病患之人格特質。中華醫學雜誌, 49(4), 264-270。
- 林邦傑 (民 76) : 整合分析的理論及其在國內的應用。教育與心理研究, 10, 1-38。
- 吳裕益 (民 71) : 認知發展與道德判斷、角色取替能力、創造力及人格的關係。教育學刊, 4, 257-300。
- 柯永河、楊思根、蔡玉玲 (民 53) : 柯氏心理健康質量表效度之分析。中國測驗學會測驗年刊, 11, 34-38。
- 柯永河 (民 66) : 柯氏性格量表手冊。台北: 中國行為科學社。
- 柯永河 (民 80) : 柯氏性格量表需要脫胎換骨嗎? 中國測驗學會測驗年刊, 38, 17-30。
- 柯永河、吳瑞屯、吳振能、張玉潔、危芷芬、程千芳 (民 81) : 柯氏性格量表脫胎換骨之後—KMHQ (1991)。中國測驗學會測驗年刊, 39, 139-153。
- 柯永河 (民 83) : 蛻變中的柯氏性格量表。學生輔導, 31, 22-31。
- 柯永河 (民 86) : 修訂後之柯氏性格量表(KMHQ1996)內容, 信效度常模及其使用說明。中國測驗學會測驗年刊, 44(1), 3-28。
- 范德鑫 (民 82) : 心理測驗的發展趨勢。載於中國測驗學會主編, 心理測驗的發展與應用—中國測驗學會成立六十週年慶論文集。台北: 心理出版社。
- 詹志禹 (民 77) : 後設分析量化的文獻探討法。思語言, 26(4), 1-15。
- 路君約 (民 80) : GATB 的效度類化問題的評介。載於行政院勞委會職業訓練局主編, 職業心理測驗使用手冊 (修訂版)。台北市: 行政院勞委會職業訓練局。
- 鍾燕宜 (民 75) : 我國工作滿足研究的整合分析。私立東海大學企業管理研究所碩士論文。
- Auastas, A. (1997). *Psychological testing*. New York: Prentice Hall.
- Cohen, J. (1977). *Statistical power analysis for the behavioral science*. New York: Academic Press.
- Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 37, 131-146.
- Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.
- Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin*, 116(3), 509-511.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Education researcher*, 5, 3-8.
- Glass, G. V. (1977). Integrating finding: The meta-analysis of research. *Review of Educational Research*, 5, 351-379.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.
- Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490-499.
- Hedges, L. V. (1988). *The meta-analysis of test validity studies*. In H. Wainer, & H. Braun. (Eds). *Test Validity* (pp191-212). Hillsdale, NJ: Lawrence Erlbaum.
- Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-analysis*. Orlando, FL: Academic Press.
- Hunter, J. E., & Schmidt, F. L. (1977). A critical analysis of the statistical and ethical implications of various definitions of test fairness. *Psychological Bulletin*, 83, 1053-1071.
- Hunter, J. E., Schmidt, F. L. & Jackson (1982). *Methods of Meta-Analysis: Cumulating research findings across studies*. Newbury Hill, CA: Sage.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

- Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of Three Major Meta-Analytic Approaches. *Journal of applied psychology, 80*(1), 94-106.
- Kraemer, H. C., & Andrew, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological bulletin, 91*, 404-412.
- Lyons, L. C. (1997). Meta-Analysis : Methods of Accumulating Results Across Research Domain. [WWW document]. URL <http://www.mnsinc.com/solomon/MetaAnalysis.html>.
- McGraw, K. O., & Wong, S. P. (1992). A common language effect size. *Psychological Bulletin, 111*(2), 361-365.
- Rosenthal, R. (1979). The file drawer problem tolerance for null results. *Psychological Bulletin, 86*, 638-641.
- Rosenthal, R., & Rubin, D. (1982). A simple, general purpose display of magnitude of experiment effect. *Journal of Educational Psychology, 74*, 166-169.
- Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research (Rev.ed.)*. Beverly Hills, CA: Sage.
- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*(2), 183-192. *Psychology, 38*, 697-79.
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*, 473-485.
- Schmidt, F. L. (1988). *Validity generalization and the future of criterion-related validity*. In H. Wainer., & H. Braun. (Eds). *Test Validity* (pp173-189). Hillsdale, NJ: Lawrence Erlbaum.
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.
- Stigler, S. M. (1986). *The History of Statistical: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard University Press.
- Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

收稿日期：民國 87 年 9 月 18 日

修正日期：民國 87 年 10 月 19 日

接受日期：民國 87 年 10 月 19 日

A Meta-Analysis of KMHQ's Validity Generalization

Sieh-Hwa Lin Pao-Yuan Huang

Educational Psychology and Counseling
National Taiwan Normal University

Abstract

This research aims at objective and statistical analysis on the investigative data of Ko's Mental Health Questionnaire (KMHQ) in the past thirty-four years to get some idea about the extent of KMHQ's validity generalization. This research employs the computer program Meta-HC as the analyzing tool. The research reveals : Among the six independent variables, the weighted average effect size of the psychological-invalids is the highest, meaning that KMHQ's distinguishing capability being the strongest. The next three independent variables are special clans, nonpsychological-invalids and criminals. However, the effect size on physical-invalids and sex only reach mid or low degree, showing that KMHQ's distinguishing capability on physical-invalids and sex is not high. To sum up, the utmost conclusion of this research is that, the extent of KMHQ's validity generalization is pretty high, its distinguishing capability is pretty idealized. Therefore, KMHQ is a personality test that deserves to be further developed, studied as well as popularly adopted.

Keywords: validity generalization, meta-analysis, KMHQ