國民教育研究學報，第 2 期：27～48頁
民85年，國立嘉義師院國民教育研究所

信度考驗的另一途徑：推論力理論

李　茂　能

國立嘉義師範學院

摘　要

本文旨在介紹推論力理論的基本概念與應用步驟。推論力理論是利用變異數分析方法研究行爲測量是否可靠的理論，重視的是測驗分數之推論力。應用推論力理論時涉及四個階段：①觀察階段：選擇測量層面與層次、計算均方，②估計階段：決定測量層面的抽樣模式(隨機或固定效果)、計算變異成份，③測量階段：界定測量目標與工具之層面、計算測量誤差與推論力係數，④最佳化階段：變化測量設計與改變抽樣模式等以尋找最佳之測量模式。前三個階段是屬於推論力(G)研究之範疇，最後一個階段是俗稱的決策(D)研究階段。文中並提供壹個模擬應用實例，說明G研究與D研究之流程，供研究者參考使用。

此外，文中亦探討推論力理論的幾個特點：①推論力理論之研究只重視變異數大小之估計，不進行F考驗，②推論力理論區別相對性之決策與絕對性之決策間之不同，③推論力理論一次分析就能同時考慮多重誤差變異源，④推論力理論在G研究時非常重視各種誤差變異源之相對大小，⑤傳統之信度理論是推論力理論的特例。

文末論及使用推論力理論之要領與應注意的事項，並指出推論力理論乃是眞正能允許研究者修正與掌控測量設計與品質的統計方法，是信度考驗之另一有效途徑。

壹、緒論

傳統測驗理論建立測驗分數的信度時，因誤差來源之不同而產生不同種類的信度。最常見的有重測信度、內部一致性信度、複本信度、與評分者信度。同時測驗分數的信度大小也常因資料搜集之方法不同而變動，令不少測驗的使用者感到困惑：一種測驗分數卻有不同的信度指標；而且，有時各種信度間又有極大差異，到底要根據那個較恰當？其實，因各類測驗分數的信度指標所檢驗的測量誤差的來源都不同，才會造成一種測驗分數擁有不同信度指標。例如，重測信度旨在探討時間或測驗情境(occasions of testing)對於測驗分數的影響力；內部一致性信度旨在探討測驗題目的同質性程度；複本信度旨在探討一個測驗版本(forms)或內容對於測驗分數的影響力如何；評分者信度旨在探討測驗分數隨著評分者之不同而變動之程度。以上這些測量誤差之來源在傳統信度理論中並無法同時進行研究，研究者必須設計不同之資料搜集方式分開研究之，而費去不少時間與精力。而且，這些測量誤差也可能產生交互作用而衍生額外的測量誤差，這在傳統信度理論中並無法加以檢驗。

此外，傳統測驗理論更令研究者困擾的是：

一、測驗分數的信度常隨抽取樣本之變動而變動，也就是測驗分數的信度是常隨樣本而變動(sample-dependent)。二、傳統測驗信度的平行測驗基本假設常是無法成立，也就是兩套觀察分數的眞分數與誤差分數之變異數常是不相等的。

雖然傳統測驗理論的信度考驗方法甚爲簡便易行，但由於以上諸多不便與困擾，促使很多國外教育與心理研究者開始尋求信度考驗的其它途徑。Cronbach, Rajaratnam, and Gleser(1963)首先提出了推論力理論(Generalizability Theory)，以解決上述傳統測驗信度理論的困境。之後，Cronbach, Gleser, Nanda, and Rajaratnam(1972)又出版了一本深具權威的專書，詳細介紹了推論力理論，奠定了推論力理論的理論架構。本理論雖然早在三十幾年前就已發展出來，但由於它涉及複雜的均方期望值(expected mean squares)計算，並且一般學校常不列入教材，推論力理論並未在國內測驗界與研究者間普遍被使用，殊屬可

惜。目前國外已有不少研究者應用推論力理論於內容效度評估(Crocker, et al., 1988)，社會工作(Gehlert, 1994)，心理諮商(Webb, Rowley, & Shavelson, 1988)，軍事訓練(Shavelson, Mayberry, Li, & Webb, 1990)，與臨床決斷分數等等方面上(Nugent, 1994)。從1982年到1995年九月爲止的 ERIC 資料庫中即有191篇與推論力理論有關之論文，可見推論力理論將是未來信度考驗的趨勢。

貳、推論力理論的意義

推論力理論是利用變異數分析的原理，研究行爲測量是否可靠(dependability)的理論。顧名思義，其立意重心在推論上。因此推論力的概念取代了傳統測驗信度(reliability)的概念，認爲從觀察分數推估領域分數(universe score)時，推論是否正確才是測驗分數應用者所應關切的問題(Cronbach, et al., 1972)。而領域分數是指在推論領域上之理想分數，它取代了傳統測驗理論中眞分數(true score)的概念，是測量對象(object of measurement)在推論領域上所有可能觀察值的平均數。在早期的推論力理論研究中，需先決定測量對象(通常是人)，只有測量對象上之變異是屬於領域分數的有效變異部份，而其餘的測量觀察層面(facets of measurement)，像測驗題目、評分員都屬於測量誤差中無關變異量的來源。所有測量觀察層面的聯集稱爲待測量觀察之領域(universe of admissible observations)。

Cardinet, Tourneur, & Allal(1976 & 1981)爲使推論力研究更具彈性，適用於各種情境，提出對稱性原則(principle of symmetry)。他們主張測量設計中的所有因子均是測量層面，而且測量對象(或目標)可與測量層面互換應用。因此目前的推論力理論研究中，並不先決定那一測量層面爲測量對象。此外，在推論力理論研究中，依決策性質亦可得到一信度指標稱爲推論力(G 或 ϕ)係數，它是領域分數與觀察分數之決定係數(coefficient of determination)。不管是早期或目前之推論力理論的應用，都需先進行推論力(generalizability)研究(簡稱G研究)，再進行決策(decision)研究(簡稱D研究)。G研究旨在研究各種變異源對於測驗分數穩

定性之影響，而D研究旨在利用G研究所得之誤差變異源訊息，研究那種測量設計最能有效率的去作決策。

參、推論力理論之特色

推論力理論具有以下幾點特色：

一、推論力理論雖係應用變異數分析原理，但並不進行各變異源之F考驗。

二、推論力理論在計算D研究之推論力係數時，對於作決策的特質是絕對性(如達到70分才能拿到駕照)或相對性(如百分等級超過70才能進入台大)有顯著差異。假如是計算絕對性的推論力係數，我們稱之爲ϕ係數；而計算相對性的推論力係數稱爲G係數。

三、不管是計算ϕ係數或G係數，主要目的在規畫出最佳之測量設計，以提高測量品質。

四、推論力理論在G研究時非常重視各種誤差變異源之相對大小，以決定如何控制這些誤差變異源。

五、推論力理論一次就能同時考慮多重誤差變異源，也就是能同時研究多重的測量觀察層面(余民寧，民82)。例如可同時研究評分者、題目、測驗情境、時間對於測量對象的影響。在傳統測驗裡我們並無法同時估計多重誤差變異源，必需一次研究一項誤差變異，因此產生了不同類別之信度指標。

六、推論力理論的架構區分推論力研究與決策研究，前者係後者研究之基礎。

此外，利用Hoyt(1941)氏所估計的信度，與使用KR-20公式或Alpha係數所估計之信度完全一樣(郭生玉，民81)。而推論力理論與它們之間亦具有密切關係。例如，就以上之單一測量觀察層面之交叉設計而言，利用推論力理論所求得之推論力係數與Hoyt變異數分析法、庫李信度(KR-20)所估計的信度應完全相同。請看下列Hoyt信度(r_H)與推論力(G_{Rel})係數公式之推演：

$r_H=(MS_{sub}-MS_{res})/MS_{sub}$ [公式一]

$G_{Rel}=\delta sub^2/(\delta sub^2+\delta res^2/ni)$

$$=\frac{(MS_{sub}-MS_{res})/ni}{(MS_{sub}-MS_{res})/ni+MS_{res}/ni}$$

$=(MS_{sub}-MS_{res})/MS_{sub}$ [公式二]

(ni表題目數目；sub表受試者；res表殘差，另根據EMS求法知：$MS_{res}=\delta_{res}2$，$\delta_{sub}2=(MS_{sub}-MS_{res})/ni$)

由此得知$r_H=G_{Rel}$，就最簡單之單一測量觀察層面之交叉設計而言，推論力理論與傳統信度理論並無不同。換句話說，傳統之信度係數是G係數的一個特例罷了！請注意在前面G係數公式分母中並未考慮題目效果(MS_i)，這是因爲在傳統測驗理論裡只著重個人分數間之差異，而且在交叉設計中每一個人均需作答所有的題目，試題難易並不會影響個人分數之相對位置。因此，在傳統測驗信度理論中，我們設定題目之變異量爲零($\delta_i^2=0$)。也由此分析得到一個結論：傳統測驗信度理論不適合作絕對性決策(absolute decision)時使用，只適合作相對性決策(relative decision)時使用。

肆、推論力理論之應用步驟

早期對於測驗之推論力研究一般區分爲三個階段：

一、界定待測量觀察之領域(universe of admissible observation)：旨在決定測量目標(或對象)與其它會影響推論力之測量層面。並決定這些測量層面的性質與設計是甚麼？是固定層面(fixed facet)？或隨機層面(random facet)？是交叉設計(crossed design)？或隔宿設計(nested design)？

二、進行推論力(G)研究：旨在估計領域分數之變異量與其它誤差源之變異量。

三、進行決策(D)研究：通常研究者先界定推論領域(universe of generalization)

，並運用G研究所得之訊息檢視各種D研究中何者最經濟而又較有高信度。在此階段並可依分數解釋之性質，計算出相對性G係數或絕對性ϕ係數。

上述推論力研究一開始即確定測量對象(即眞正變異部份)與測量觀察層面(誤差變異部份)，此種處理資料之模式限制了其應用之範圍與彈性。

Cardinet, Tourneur, & Allal(1976 & 1981)爲使推論力研究更具彈性與適用於各種情境，提出對稱性原則(principle of symmetry)。此原則說明了兩件事①測量設計中的任何一個測量層面均可被選爲測量對象(或目標)與②G理論在某一測量層面上之運作可以移轉到其它測量層面上。因此，對稱性原則可使得在資料搜集階段不必事先決定測量目標(區分面)與測量之工具層面，而且測量設計中任一個(或一個以上)測量層面均可成爲測量對象(或目標)。換言之，區分面與工具面可以互換，使得推論力理論之應用更具彈性。他們將推論力研究之基本流程細分爲以下四大階段：

一、資料蒐集(observation)階段

本階段中之所有研究因子皆稱爲層面 (facet)， 不先決定那一因子爲測量對象(object)，那些因子是測量層面。而本階段旨在

(一)確定各種測量層面之變異源，其主要工作有三：

1.選擇測量層面，

2.決定各測量層面間之關係：交叉或隔宿，

3.決定各測量層面之層次(levels)數目，

(二)計算每一測量層面之離均差平方和。

二、變異源估計(estimation)設計階段：本階段旨在

(一)確立待測量觀察之領域，

(二)決定變異數分析之適當抽樣模式：隨機模式或混合模式(隨機效果+固定效果)。

(三)依前步驟之選擇，計算變異成份(variance components)。由於推論力理論本質上是一種隨機層面測量理論，設計時至少要有一測量層面爲隨機層面。所有測量層面爲固定效果的話是無法進行研究的。假如有些測量層面爲隨機層面，而有些測量層面爲固定層面，我們稱之爲混合模式(mixed model)。

三、測量對象與測量層面區分(measurement)階段：本階段旨在

(一)指定有那些測量層面爲測量對象，有那些測量層面爲測量工具。Cardinet, Tourneur, & Allal(1981)稱前者爲區分面(face of differentiation)，後者爲工具面(face of instrumentation)。

(二)區分面與工具面上各層面抽樣模式之考慮：固定或隨機。本階段之主要工作有四：

1、測量設計中各變異成份之分派

①測量設計：界定一個以上設計(例如何者爲區分面，何者爲隨機層面)以供分析。

②連貫性之控制(control of coherence)：確定沒有區分面之層面隔宿在工具面內，以避免區分面之變異量與工具面之變異量產生混淆。

③計算動態變異量(active variance)：從總變異量中扣除與固定工具層面有關之變異量。

2、計算區分面之變異量：將動態變異量中僅含區分層面之變異量加起來。

3、計算工具面(誤差)之變異量：本項變異量之計算隨著分數之解釋方式不同，而分爲①絕對誤差變異量，與②相對誤差變異量。

4、計算推論力係數。

以上三個階段之工作，事實上就是俗稱的G研究的工作內涵。G研究的主要目的在協助研究者設計各種D研究，以找出推論力最佳而且經濟之測量設計。

四、D設計最佳化(optimization)階段：本階段旨在

(一)修正各測量層面間之關係(如交叉變爲隔宿)，

(二)工具層面測量層次數之增減，

(三)變換區分面與工具面之抽樣模式。

以上這些測量設計之修正旨在找出一測量設計既降低費用與誤差，而又能改善測量分數之信、效度。因此，研究者常需在第三與第四階段間嘗試錯誤式的來回運作。Sanders, Theunissen, & Bass(1989)與Sanders(1992)爲提高尋找最佳測量設計之精確性與效率，曾利用數學方法"branch-and-bound algorithm"設計程式尋找最佳測量設計。值得研究者參考應用。

伍、實例說明

爲使研究者具體了解推論力理論之實際應用情形，　茲舉一雙測量層面交叉設計(two-facet, crossed design)之模擬範例，按前述之四大階段一一說明如下：

假設某校今年欲舉辦一國小作文比賽，主辦者欲使作文評分符合公平，而又經濟有效之原則。但不知道要聘請幾位評分員與寫幾篇作文才能符合上述之原則? 推論力理論正是應用在此情境之最佳方法。此時研究者所關切之測量層面可以圖1表示。主辦者可根據過去辦過作文比賽之資料或隨機抽取樣本建立如表1之 G 研究原始資料。

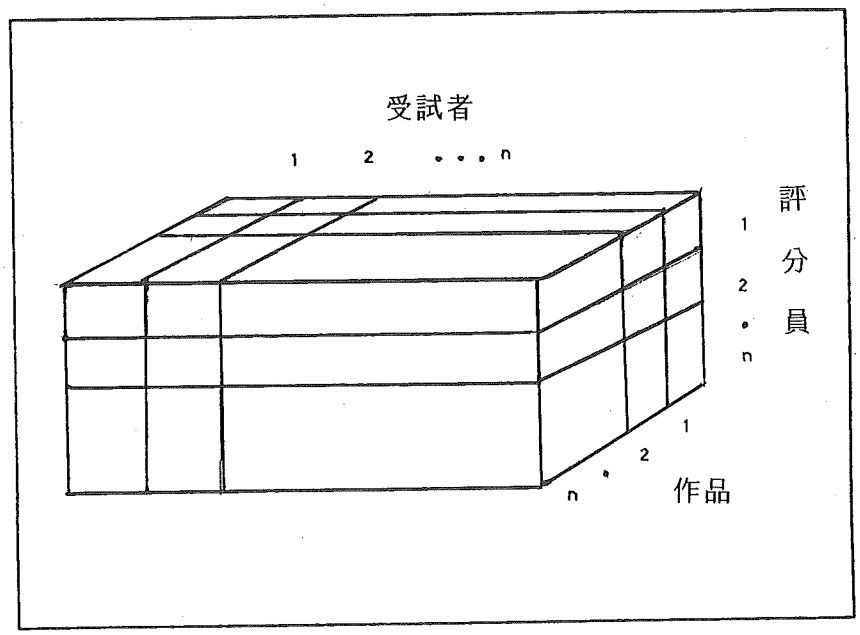

圖1 雙測量層面交叉設計

1、資料蒐集階段

首先決定測量之各種重要變異源，從表1知本資料包含三個測量層面：受試學生，評分者與測驗題目。從表1亦知本資料各測量層面爲交叉關係，且各測量層面之

觀察層次(levels)數目分別爲：學生十人(N_p=10)，作文題目二題(N_i=2)，評分員三人(N_r=3)。此種測量設計共有四種誤差源：①受試作文能力間之差異，②作文題目難度間之差異，③評分者間評分寬嚴之差異④受試作文能力與作品間之交互作用(例如學生在作品一上之作文能力與作品二上之作文能力不相同)，⑤受試作文能力與評分者間之交互作用(例如學生之作文成績之相對位置因評分者之不同而變動)。⑥作品與評分者間之交互作用(亦即顯示評分員對於學生作品之平均成績會因作品之不同而有差異)，⑦隨機性誤差，未測到之系統性誤差與以上三個測量層面之交互作用，併稱爲殘差。因此學生在某一作文題目上之觀察分數可以公式三表示。

表1 三位評分員在十位學生的兩篇作文上之評定成績(十分制)

評分員		甲		乙		丙		
作品		A	B	A	B	A	B	$\bar{X}_p$
受	1	0	2	2	4	1	3	2.0
	2	6	8	2	4	3	4	4.5
	3	4	4	2	0	2	2	2.3
試	4	2	4	0	2	0	1	1.5
	5	2	4	4	2	2	1	2.5
	6	8	8	6	8	3	4	6.2
	7	2	2	4	2	1	3	2.3
者	8	4	2	5	4	2	2	3.2
	9	8	7	5	6	5	6	6.2
	10	3	5	4	3	4	4	3.9
$\bar{X}ir$		3.9	4.6	3.4	3.5	2.3	3.0	
$\bar{X}.r$		4.25		3.45		2.65		
		$\bar{X}_{A.}$=3.20		$\bar{X}_{B.}$=3.70				

$X_{pir}=\mu$

[總平均]

$+\mu_p-\mu$

[個人效果]

$+\mu_i-\mu$
[作文題目效果]
$+\mu_r-\mu$
[評分員效果]
$+[(\mu_{pi}-\mu)-(\mu_p-\mu)-(\mu_i-\mu)]$
[個人與作文題目之交互作用]
$+[(\mu_{pr}-\mu)-(\mu_p-\mu)-(\mu_r-\mu)]$
[個人與評分員之交互作用]
$+[(\mu_{ri}-\mu)-(\mu_r-\mu)-(\mu_i-\mu)]$
[評分員與作文題目之交互作用]
$+[(X_{pir}-\mu)-(\mu_{pi}+\mu_{pr}+\mu_{ir})+(\mu_p+\mu_i+\mu_r)]$
[殘差] [公式三]

公式三中之各種測量變異源之關係爲：

$$\delta^2(X_{pir})=\delta_p^2+\delta_i^2+\delta_r^2+\delta_{pi}^2+\delta_{pr}^2+\delta_{ir}^2+\delta_{pir}^2,e$$

今將作文題目、評分員、與受試學生間之各種測量變異源圖示如下：

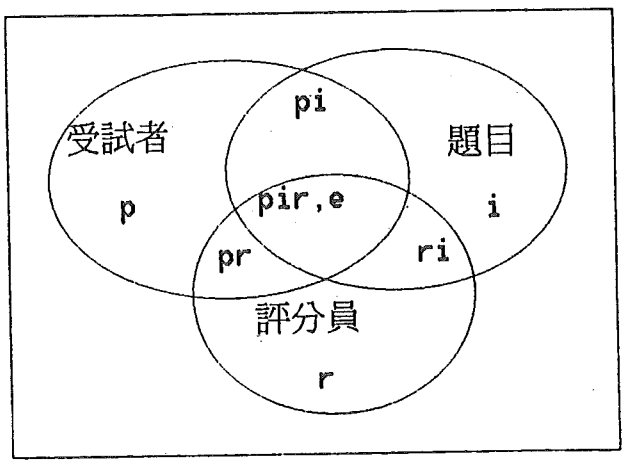

圖2 $p_x i_x r$測量變異源之關係

其次，表1之資料可使用SAS PROC VARCOMP副程式(參見附錄一)或直接運用推論力理論專用程式GENOVA(Crick & Brennan,1982)計算各種測量變異之離均差平方和，輸出之結果摘要如表2所示。

表2 變異數分析之結果

變異源	離均差平方和 (SS)	自由度 (df)	均方 (MS)	均方期望值* (EMS)
學生(p)	152.35	9	16.93	$\delta_e^2+3\delta_{pi}^2+2\delta_{pr}^2+6\delta_p^2$
作品(i)	3.75	1	3.75	$\delta_e^2+10\delta_{ri}^2+3\delta_{pi}^2+30\delta_i^2$
評分員(r)	25.60	2	12.80	$\delta_e^2+10\delta_{ri}^2+2\delta_{pr}^2+20\delta_r^2$
$p_x i$	14.75	9	1.64	$\delta_e^2+3\delta_{pi}^2$
$p_x r$	52.40	18	2.91	$\delta_e^2+2\delta_{pr}^2$
$r_x i$	1.20	2	0.60	$\delta_e^2+10\delta_{ri}^2$
殘差	14.80	18	0.82	δ_e^2

* 意指在相同之研究設計下，針對同樣之母群與領域進行重覆抽樣時，所有均方之平均(期望)值。

2、變異源估計之設計階段

由於我們最終之推論對象並不限於此十位學生，三位評分者與二題作文題目，因此這三個測量層面上之樣本僅是由母群或領域中隨機抽樣而來。因此這三個測量層面：受試學生、評分者、與作文題目均爲隨機模式設計。接著，我們需依此隨機模式計算各變異源之變異成份。讀者可令表3中均方值(MS)等於均方期望值(EMS)，即可求得如表3之各項變異成份。

表3　變異成份估計之結果

變異源	均方 (MS)	變異成份 (δ)	百分比 %
學生(p)	16.93	2.200	44.7
作品(i)	3.75	0.078	10.3
評分員(r)	12.80	0.506	1.6
p x i	1.64	0.272	5.5
p x r	2.91	1.044	21.2
r x i	0.60	-0.022 *	0.0
殘差	0.82	0.822	16.7

*負值設定爲零。

在表3中交互作用ｒｘｉ之變異成份爲負數(-.022)，不符變異數分析之基本概念(變異數不可能爲負值)。產生負值的可能原因爲：

1、樣本過小，

2、抽樣誤差(sampling error)，尤其當負值很大時，

3、測量模式界定錯誤(model misspecification)，

4、極端值，

5、測量層面之層次太少(Calkins, et al., 1978)。

解決負值之道有：①當負值很小而接近零時，通常將負值設定爲零，但仍使用其原有之負值計算其它之變異成份大小(Brennan, 1992)。②當負值很大時，通常需重新界定G研究測量模式，再估計各變異成份之大小。③增加樣本或測量層次，④去除極端值，⑤使用貝氏(Bayesian)ANOVA估計法。本範例之δ_{ri}^2之負值很小(-.022)，故將其設定爲零(見表4)。由此值亦可推知學生作文成績之相對位置不因評分員之不同而改變，其未具交互作用之效果可由圖3顯現出來。

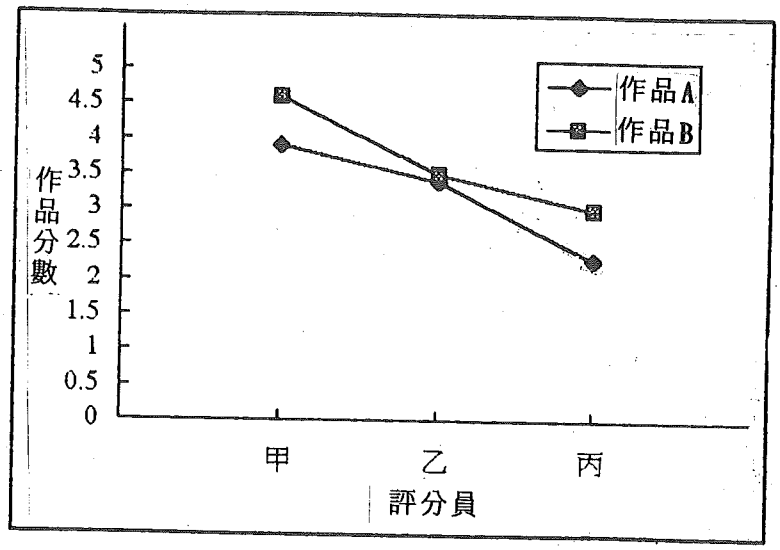

圖3 作品與評分員之交互作用

3、測量對象與測量層面區分階段

本範例旨在了解學生之作文分數是否具有信度，因而學生爲測量對象，作文題目與評分員爲測量工具，亦即學生爲區分面，作文題目與評分員爲工具面。而且推論對象並不限於這些樣本學生、評分員、與作文題目，因此區分面與工具面均爲隨機抽樣模式。接著我們要確定沒有區分面之層面隔宿在工具面內，以避免區分面之變異量與工具面之變異量產生混淆。因爲本範例爲一交叉設計，因此並無連貫性之問題，且三個測量層面均爲一隨機效果模式，並不需要計算動態變異量。我們可直接計算區分面與工具面之變異量。這三個層面之變異成份請參見表3。本例旨在希望利用這次作文比賽之分數決定學生在作文成績上的優先順序，此種分數之應用方式爲相對性之解釋。因此我們必須計算如表四之相對性誤差變異量 (δ_{Rel}^2) 與其推論力(G)係數。

表4 G研究與各種D研究之變異成份分析與推論力係數

變異源	G研究變異成份之估計值	D研究變異成份之估計值							
nr'=	1	1	2	2	2	3	3	3	4
ni'=	1	2	1	2	3	1	2	3	1
學生(p)	2.200	2.200	2.200	2.200	2.200	2.200	2.200	2.200	2.200
(區分面)									
評分員 (r)	.5056	.5056	.2528	.2528	.2528	.1685	.1685	.1685	.1264
作文題目(i)	.0778	.0389	.0778	.0389	.0258	.0778	.0389	.0258	.0778
p x r	1.044	1.044	.5222	.5222	.5222	.3481	.3481	.3481	.2611
p x i	.2722	.1361	.2722	.1361	.0907	.2722	.1361	.0907	.2722
r x i	0	0	0	0	0	0	0	0	0
殘差	.8222	.4111	.4111	.2055	.1203	.2740	.1203	.0913	.2055
δ_{Rel}^2	2.138	1.592	1.210	.8638	.7332	.8943	.6045	.5301	.7388
δ_{Abs}^2	2.720	2.136	1.536	1.155	1.012	1.141	.8119	.7244	.9430
G係數(相對性)	.507	.580	.650	.718	.750	.711	.784	.806	.749
ϕ係數(絕對性)	.447	.507	.590	.656	.685	.659	.730	.752	.700

$\delta_{Rel}^2=\delta_{pr}^2/nr'+\delta_{pi}^2/ni'+\delta_{pri}^2,_e/(nr'.ni')$

$\delta_{Abs}^2=\delta_{Rel}^2+\delta_r^2/nr'+\delta_i^2/ni'+\delta_{ri}^2/(nr'ni')$

4. D設計最佳化階段

從表4知假如D研究設計之層面與層次與G研究設計相同時(三個評分員與二道作文題目)，其G與ϕ係數分別爲.784與.730。不過當作文題目只有一題與評分者只有一位時，其相對性之G係數只有0.507。顯然這樣的信度是不夠的。到底需要多少位評分員與多少題作文題目端視下列三要素而定：①測量誤差大小，②所需推論力大小，與③實用上之考慮。由表八中知評分員的變異量(.5056)遠超過作文題目

之變異成份(.0778)。而且從評分員與學生之交互作用之大小(1.044)知學生之作文成績之相對位置亦隨評分員而有差異。因此增加評分員比增加作文題目更能有效改善推論力層次。例如從表4之D研究知使用兩位評分員與一題作文題目之推論力(.65)大於使用一位評分員與兩題作文題目之推論力(.58)。假如我們希望測驗推論力高於.75時，就實用性與效率性來說，我們會應用三位評分員及使用二道作文題目而不是使用二位評分員及使用三道作文題目(.784 >.750)。此時傳統相對性之測量標準誤(以原始總分爲單位)之不偏估計值爲1.555(即$\sqrt{2x2x0.6045}$)，而其絕對性之測量標準誤爲1.80(即$\sqrt{2x2x0.8119}$)。根據此測量標準誤可以建立測驗分數之適當的信賴區間。

陸、應用推論力理論之要領與應注意之事項

一、在D研究中不能包含有G研究階段中未加以研究之測量層面。

二、由於推論力理論本質上是一種隨機層面測量理論，設計時至少要有一測量層面爲隨機層面。

三、取樣之樣本勿過少，以免產生變異成份估計值(estimates of variance components)不穩定現象；甚至產生負值之變異成份。根據Webb,Rowley, & Shavelson(1988)之建議，取樣人數應至少大於20，而每一測量層面之層次應大於2。

四、測量設計與變異成份之估計具有密切關係，研究者需對研究設計中各資料搜集模式之EMS估計有正確之認識，否則易產生錯誤。

五、應用多重特質與多重方法模式，推論力理論亦可進行效度分析(P. 153, Suen, 1992)。例如每一位學生均需寫六篇作文，其中論說文、抒情文、記敘文各寫兩篇。此種測量設計允許我們探究不同寫作方法間之推論效度，而同一種方法內之兩篇作文分數的一致性代表著測驗分數之信度。因此這種仿效多重特質與多重方法模式之G研究設計能讓我們研究測驗分數之信、效度。這也說明了信、效度不分家之概念：改善了信度即能改善效度，反之亦然。Kane(1982)提供了詳

細之計算公式，有興趣之讀者請自行參閱。

六、D 研究之最佳化可透過下列途徑達成：①增大樣本與測量層次，②改變抽樣模式（例如將隨機模式改爲固定模式），③重新界定推論領域或區分面之母群與④改變測量設計（例如將交叉設計改爲隔宿設計）。

七、絕對性的分數應用所產生之誤差比相對性的分數應用所產生之誤差來得大，因此計算決斷分數（cut-off score）之信賴區間需使用絕對性之 δ_{Abs}^2 去計算測量標準誤。

八、在 G 研究階段最好使用交叉測量設計。因爲如果使用隔宿設計，在 D 研究階段時，如欲更改測量設計爲交叉設計，將無法取得正確之變異成份。

九、報告測驗分數之信度時，請同時詳細說明測量結構，推論領域範圍，樣本與測量層次之大小，及抽樣模式。因爲這些因素常會嚴重影響推論力之強弱，只報告係數大小易誤導測驗使用者（Brennan, 1992b）。

十、當測量對象擁有多重領域分數（例如施測中華智力量表後，可得到兩種智商：語文與非語文智商）且又關切整個量表之信度時，需使用多變項推論力理論（Multivariate generalizability theory）。有需要之讀者可參閱 Cronbach, et al.(1972); Shavelson, & Webb(1981); & Brennan(1992a) 等人之論著。

柒、結語

簡言之，推論力理論乃是利用變異數分析方法研究行爲測量之可靠度的理論，重視的是測驗分數之推論力；它將是未來信度考驗之另一有效途徑。應用時涉及四個階段：①觀察階段：選擇測量層面與層次、計算均方，②估計階段：決定測量層面的抽樣模式（隨機或固定效果）、計算變異成份，③測量階段：界定測量目標與工具之層面、計算測量誤差與推論力係數，④最佳化階段：變化測量設計與改變抽樣模式等以尋找最佳之研究設計。

此外，推論力理論具有以下幾個特色或優點：①推論力理論之研究與變異數分析

具有密切關係。前者重變異數大小之估計，而後者重F值之考驗，②傳統信度係數只適合於常模參照測驗上之相對性決策使用，推論力理論不管常模參照測驗或效標參照測驗上之絕對性決策都適用，③推論力理論一次就能同時考慮多重誤差變異源，④傳統之信度理論只不過是推論力理論之一特例而已，⑤推論力理論在G研究時非常重視各種誤差變異源之相對大小，以決定如何控制這些誤差變異源(如題目要出幾題，評分者要用幾位)而設計出一個誤差小、效率高之測量工具。因此，推論力理論乃是一眞正能允許研究者修正與掌控測量設計與品質的統計方法，值得測驗界與研究者推廣應用。

參考文獻

余民寧(民82)。測驗理論的發展趨勢。載於中國測驗學會主編：心理測驗的發展與應用(pp.23-62)。台北：心理。

郭生玉(民81)。心理與教育測驗。台北：精華。

Brennan, R. L.(1992a). Elements of generalizability theory. Iowa City: ACT.

Brennan,R.L.(1992b). Generalizability theory.Educational Measurement: Issues and practice, 11(4), 27-34.

Calkins, D. S., Erlich, O., Marston, P. T., & Malitz, D. (1978). An empirical investigation of the distributions of generalizability coefficients and various estimates for an application of generalizability theory. Paper presented at the Meeting Of the American Educational Research Association, Tornoto, March.

Cardinet, J., Tourneur, Y., & Allal, L.(1976) The symmetry of generalizability theory:Applications to educational measurement. Journal of Educational measurement, 13, 119-135.

Cardinet, J., Tourneur, Y., & Allal, L.(1981). Extension of generalizability theory and its applications in educational measurement. Journal of Educational measurement, 18(4), 183-204.

Crick, J. E., & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system(FORTRAN IV computer program and manual). Dorchester, Mass: Computer facilities, University of Massachusetts at Boston.

Crocker, L., et al. (1988). The generalizability of content validity ratings. Journal of Educational Measurement, 25(4), 287-299.

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

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A librealization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Gehlert, S. (1994). The applicability of generalizability theory to social work research and practice. Journal of Social Service Research, 18,73-87.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Kane, M. T. (1982). A sampling model of validity. Applied Psychological measurement, 6, 125-160.

Nugent, W. R. (1994). An investigation of the dependability of clinical cutting scores using generalizability theory. Journal of Social Service Research, 18, 89-107.

Sanders, P. F. (1992). The optimization of decision studies in generalizability theory. Master's thesis, University of Amsterdam.

Sanders,P.F., Theunissen, T.J.J.M.,& Bass, S.M.(1989). Minimizing the number of observations: A generalization of the Spearman

-Brown formula. Psychometrika, 54(4), 587-598.

Shavelson, R. J., & Webb, N. M.(1981). Generalizability theory:1973-1980. The British Journal of Mathematical and Statistical Psychology, 34,133-166.

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine corps rifleman. Military Psychology, 2(3), 129-144.

Shavelson, R. J.,& Webb, N. M.(1991). Generalizability theory: A primer.Newbury Park: SAGE.

Suen, H. K. (1992). Principles of test theory. New Jersey: Lawrence Erlbaum Associates.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. Measurement and evaluation in counseling and development, 21, 81-90.

附錄一: 範例—SAS PROC VARCOMP之程式設計

```
DATA GSTUDY;
INPUT SUBJECT RATER ESSAY SCORE @@;
CARDS;
01 1 1 0 01 1 2 2 01 2 1 2 01 2 2 4 01 3 1 1 01 3 2 3
02 1 1 6 02 1 2 8 02 2 1 2 02 2 2 4 02 3 1 3 02 3 2 4
03 1 1 4 03 1 2 4 03 2 1 2 03 2 2 0 03 3 1 2 03 3 2 2
04 1 1 2 04 1 2 4 04 2 1 0 04 2 2 2 04 3 1 0 04 3 2 1
05 1 1 2 05 1 2 4 05 2 1 4 05 2 2 2 05 3 1 2 05 3 2 1
06 1 1 8 06 1 2 8 06 2 1 6 06 2 2 8 06 3 1 3 06 3 2 4
07 1 1 2 07 1 2 2 07 2 1 4 07 2 2 2 07 3 1 1 07 3 2 3
08 1 1 4 08 1 2 2 08 2 1 5 08 2 2 4 08 3 1 2 08 3 2 2
09 1 1 8 09 1 2 7 09 2 1 5 09 2 2 6 09 3 1 5 09 3 2 6
10 1 1 3 10 1 2 5 10 2 1 4 10 2 2 3 10 3 1 4 10 3 2 4
```

```
;*表5之資料矩陣必須重整如以上之格式;
PROC VARCOMP METHOD=TYPE1;
*如係不等組設計要使用METHOD=MIVQUE0;
CLASS SUBJECT RATER ESSAY;
    MODEL SCORE=SUBJECT RATER ESSAY SUBJECT*
RATER SUBJECT*ESSAY
    RATER*ESSAY/;
* 上述模式爲隨機效果模式 ( 內定值不必設定 )，如模式中第一個效果(ESSAY)
* 爲固定效果模式，需使用下行控制敘述
* MODEL SCORE=ESSAY SUBJECT RATER SUBJECT*RATER
* SUBJECT*ESSAY RATER*ESSAY/FIXED=1;
PROC SORT DATA=GSTUDY;
    BY RATER ESSAY;
PROC MEANS;
    VAR SCORE;
    BY RATER ESSAY;
PROC MEANS;
    VAR SCORE;
    BY RATER;
PROC SORT DATA=GSTUDY;
    BY ESSAY;
PROC MEANS;
    VAR SCORE;
    BY ESSAY;
```

An Alternative Approach to Reliability Analysis: Generalizability Theory

Mao-Neng Li

National Chia-yi Teachers Colledge

Abstract

Basic concepts in generalizability theory and its procedures for practical applications are introduced and exemplified using a hypothetical data set. Generalizability theory is a powerful technique for investigating the dependability of behavioral measurements in the sense that how accurately an observed score can be generalized to the universe score. Generalizability theory consists of two stages of analysis: generalizability(G) study and decision (D) study. Conducting a G study, aiming at the estimation of magnitude of error variance, involves three steps: (1.) observation,(2.) estimation, and (3.) measurement. The optimization step concludes the D study that uses the information from G study to determine the best measurement design to get the most reliable scores in the most efficient way.

Several unique features of generalizability theory are also pinpointed.

1. Although generalizability theory is based on ANOVA, it ignores the usual F test for each of the variance components.
2. Generalizability theory distinguishes between relative decisions and absolute decisions.
3. Generalizability theory allows the simultaneous estimation of multiple sources of error variance.
4. Generalizability theory puts much more emphasis on the relative contribution of each source of error variance than the magnitude of summary generalizability

coefficients.

5. Generalizability theory subsumes classical reliability theory as a special case.

Finally, several useful guidelines are suggested for appropriate uses of generalizability theory. The paper concludes that generalizability theory is an effective and efficient alternative tool for reliability analysis of behavioral measurements.