

不同資料結構下異常作答指標決斷值的探討

黃財尉

摘 要

本研究的目的是建立高中低不同能力水準的受試群分別面對不同難度水準試題集時，內注意係數(Within-Ability Concern Index, W^*)與外注意係數(Beyond-Ability Surprise Index, B^*)指標達到顯著異常的標準，並考驗指標異常參照標準是否可適用於不同的資料分配。研究以虛擬資料產生九個資料矩陣，每個資料矩陣再產生 500 個排序結果，計算內外注意係數在此 500 個排序結果的第 90 百分位數，並重複 20 次，以其平均數作為最後的異常參照標準。研究結果發現：

- 一、指標於樣本矩陣上的分佈，大概呈一倒 Z 字型分布，符合實際作答現象。
- 二、高能力者答錯其能力內試題的異常情形較答對能力外試題所引發的異常情形嚴重。
- 三、答對能力外試題所引發的可能訝異情形對低能力者而言比其答錯能力內試題的惋惜情境更加嚴重。
- 四、不同能力區間的內外注意係數指標之決斷標準值大體上有明顯差異，但能力區內面對不同難度水準的情況差異就不甚明顯，甚至低能力區內與中能力區內可以共用指標的決斷標準。

關鍵字：異常指標、內注意係數、外注意係數、決斷值。

黃財尉 國立嘉義大學輔導與諮商學系 twhuang@mail.ncyu.edu.tw

前言

自從 Sato 於 1975 年提出 S-P (Student-Perform) 表及注意係數(Caution Index, SCI)以來，異常作答組型(aberrant response pattern)的相關理論與應用研究成果就方興未艾(D'Costa, 1993a, 1993b; Harnisch & Linn, 1981; Huang, 2002; Hulin, Drasgow, & Parsons, 1983; Karabatsos, 2000; Kogut, 1986; Linacre & Wright, 1994; Meijer & Sijtsma, 1999, Smith, 1982; Tatsuoka & Tatsuoka, 1983; Wright & Stone, 1979)而這異常作答組型通常會表現在一個以受試者為主體的二元計分(dichotomously scoring)測驗之中，特別當受試者整體的作答組型有不一致或不符合某一種模式期待現象之時，這種非源自受試者一致性的表現，便潛在出現異常作答現象。

隱藏在異常作答組型之後的作答行為非常複雜(Kogut, 1986)，同一異常作答組型，可能反映出多種甚至交互作用的異常作答行為，例如摸索、粗心、猜測、作答緩慢或創意思維等等；但也因為這些異常作答能將受試者的特質表現出來，因此可以進一步用來幫助了解教學者對受試者的學習認知。為單純起見，統計計量學者通常設計一些指標來偵測受試者的異常作答組型，其背後的作答行為歸屬，便不在討論之內。

在許多著名的異常作答偵測指標之中，Harnisch 與 Linn(1981)的 MCI 指標(Modified Caution Index)算是普遍被使用的，它將 Sato 的 SCI 指標修正為介在 0 與 1 間的數值，但這個修正的指標依然面臨三個困難：無法有效地辨識異常作答的程度、沖銷兩種不同的心理作答特徵、經驗法則取向但缺乏統計上的判斷標準(Huang, 2002)。

Huang (2002, 指標公式請見附錄)改進 D'Costa (1993a, 1993b)的 Bs 與 Wc 指標而提出內注意係數(Within-Ability-Concern Index, W*)與外注意係數(Beyond-Ability-Surprise Index, B*)，該指標是依據 Guttman (1944)的理想作答模式原則而設計，亦即在試題難度與受試者能

力同時順序排序之下，以差異距離(discrepancy distance)概念來計算受試者的異常作答程度，以克服 MCI 指標的三項困難；惟 Huang 的公式有二倍常數的累贅，且該研究是以一般中等能力－難度矩陣作為探討依據，並無針對其他不同能力－難度水準的情況分析，進一步的修正與研究頗值得進行。

根據以上動機，本研究欲精進先前的研究，相關的研究目的如下：

- 一、建立高、中、低能力受試群分別面對高、中、低難度試題集分配下內外注意係數指標的標準。
- 二、考驗上述指標標準是否可適用於不同的資料分配。

文獻探討

一般而言，異常作答偵測指標可歸屬為兩個領域(Meijer & Sijtsma, 1999)：群屬類(Group-based)以及 IRT 屬類，前者係因為在指標計算中涉及許多群集屬性而歸類（例如以通過率代表試題難度），而後者是因為涉及以某一 IRT 模式為異常作答的參照標準而命名。相關的群屬類異常作答偵測指標包括注意係數家族：SCI(Sato, 1975)、MCI (Harnisch & Linn, 1981)、Bs 與 Wc 指標(D'Costa, 1993a, 1993b)、B''與 W'' 指標(Cheng, 1997)，以及本研究的 B*和 W*指標(Huang, 2002)；其他如點二系列指標(point-biserial index, Donlon & Fischer, 1968)、信賴指標(Dependability Index, Kane & Brennan, 1980)、U3 與 ZU3 指標(Van der Flier, 1982)，以及常模符應指標(Norm-Conformity Index, NCI, Tatsuoka & Tatsuoka, 1982)等。另外，著名的 IRT 屬類指標亦因提供比較的參照不同而區分成三類：1.殘差法(Residual-based approaches)指標，例如 OUTFIT/INFIT 指標以及其標準化型式(Linacre & Wright, 1994, Smith, 1991)；2.可能性指標 I_0 (Levine & Rubin, 1979) 以及其標準化型式 I_z (Drasgow, Levine, & Williams, 1986)；3.共變法指標(Covariance-based

approaches)以及其標準化型式，例如延伸型注意係數 ECI2 與 ECI4 其標準化指標(Extended Caution Index, Tatsuko, 1984)。

Runder (1983)曾經調查 MCI 與幾種 IRT 屬類指標於假性高低分變動的條件之中，發現 MCI 指標的性質最為穩定。而 Drasgow、Levine 和 McLaughlin (1987)檢視 SCI 與幾種標準化的指標於不同能力水準下的變動情形，他們發現 SCI 於各種能力組別中表現最不穩定。文獻的發現確實可能說明 MCI 指標表現穩定，然而 MCI 指標確實沖銷了某些截然不同的作答組型，例如粗心與猜測，且不能表現出良好的偵測敏感性 (Huang, 2002)；同時 MCI 指標的異常判斷標準是依據經驗法則訂定 0.3 為作答異常決斷值，此決斷標準缺乏統計上的客觀依據。

D'Costa (1993a, 1993b)曾依據總分的落點而將受試者的作答組型切割成為兩部分：能力內關注組型(within-ability-concern pattern)與能力外驚訝組型(beyond-ability-surprise pattern)，前者之作答特徵為受試者答錯能力內該答對的試題，而後者之作答特徵則為答對能力範圍外該答錯的試題，這種分割法有助於阻止不同作答型態之間的沖銷現象，並與傳統 SCI 及 MCI 指標的偵測計算迥異。

D'Costa (1993b)利用虛擬資料檢視 Bs 與 Wc 指標以及 MCI 與總分的關係時，發現 MCI 與總分呈對稱的 U 字型關係，解釋了先前 Harnisch 與 Linn (1981)研究中所發現 MCI 與 SCI 與總分負相關的結果（左半 U 字型），以及 Jaeger (1988)所發現 MCI 與總分呈正相關（右半 U 字型）之間的矛盾。相似的結果也見於 Cheng(1997)接續 D'Costa 的實徵研究之中，但 D'Costa 的指標仍停留在指標數值大於 0.5 的經驗法則之中，其指標公式請見附錄。

本研究所指的內外注意係數(B* & W*)是建立在 D'Costa (1993a,1993b)與 Cheng (1997, 指標公式見附錄)之研究基礎上所設計的一組偵測異常作答組型指標而言，事實上這三種指標與 SCI 及 MCI 一樣，是建立在完美 Guttman 模式的虛無假設之下。依據 Guttman 模式原

則，試題難度間需具隸屬的包含關係，特別是在編製試題時所應考慮測驗內容上的包含關係而言；但對異常偵測的指標而言，所針對的是受試者已經作答的結果，因而在考量 Guttman 模式的試題難度包含關係時，通常是以受試者已經作答的資料結果先分析試題難度（如通過率），再由簡單到容易順序排列，以符合 Guttman 試題難度的包含關係。職此，完美的 Guttman 模式則是指在試題難度的順序排列下，答對困難試題的受試者就必須答對較為簡單試題的情形而言。因而在試題難度與受試者能力同時順序排列下，任何違反完美 Guttman 模式的作答型式，皆可視為受試者不同程度的異常作答型式。惟是否達統計上的顯著意義，是否可以拒絕完美 Guttman 模式的虛無假設，需要進行進一步的統計檢定，而本研究即是以排序檢定表示此異常作答組型的統計意義。

W*與 B*指標可同時用於受試者作答之分析以及試題反應之分析，細言之，內注意係數(W*)係設計來偵測受試者於其能力之內該答對某些試題卻回答錯的錯愕作答；而外注意係數(B*)係設計來偵測受試者於其能力之外該答錯某些試題卻回答對的驚奇作答，其公式如下：

內注意係數：

$$W_i^* = \frac{\sum_{j=1}^{T_i} (1 - u_{iT}) \times (q_{iT}^* - q_{\bullet j})}{[(K-1)/2]} \times 100$$

外注意係數：

$$B_i^* = \frac{\sum_{j=T_i+1}^K u_{ij} \times (q_{\bullet j} - q_{iT}^*)}{[(K-1)/2]} \times 100$$

在 B*與 W*公式中分子部分， u_{ij} 為 1 代表正確作答結果，反之為 0； $q_{\bullet j}$'s 則是試題難度值由易至難順序排列，當受試者的總分為 T 分時，所對應難度順序排列之第 T 道與第(T+1)道試題的難度平均值為 q_{iT}^* 。值得注意的是， q_{iT}^* 已經將受試者所得的原始總分映射

(mapping onto)至一個以試題難度所表示的連續能力量尺上，這個轉換是合理且實際的。以二元計分(dichotomously scored)為例，當受試者得到總分 T 分時，他（她）是被期望答對前面 T 道簡單試題而答錯剩餘高於其能力的試題，因此，此時受試者的能力可以被假定剛好高於第 T 道試題的難度而低於第 $(T+1)$ 道試題的難度，而達到這兩道試題難度平均值的實際上限。

在分母部分， $[(K-1)/2]$ 代表當試題數為 K 時，以下高斯整數符號(lower Gauss integer) 所表示的理論最大差異距離，亦即大於或等於 $(K-1)/2$ 的最小整數值。值得注意的是，此理論的最大差異距離即代表最大的作答誤差數目(number of errors)，這裡所談的作答誤差數目則係指受試者於其能力範圍內該答對卻未答對的試題數，或於其能力範圍外應答錯卻未答錯的試題數，而這兩類型的誤差試題數（能力範圍內與能力範圍外）則是已經被證明是相等且先遞增再遞減(Huang, 2002)。

以下將以 10 道測驗試題（難度從 0.1 到 1.0，以 0.1 難度差距遞增排列）與某位受試者於這 10 道試題的作答結果為實例，解說內外注意係數(B^* & W^*)的計算方式。假設某位受試者於這 10 道試題的作答結果為(1001111|101)，其中符號“|”代表該受試者能力以總分水準(7 分)為駐標(ability marker)，符號“|”左邊代表能力範圍內，右邊代表能力範圍外。由於該駐標介於第七道試題與第八道試題之間，因此，該受試者的能力可以這兩道試題的平均難度加以估計，即 $(0.7+0.8)/2=0.75$ 。下一步驟要計算差異距離，即異常作答之試題難度與此受試者之估計能力的差距總合，此例中受試者於第二題與第三題（該對未對）、第八題與第十題（該錯未錯）作答異常，因此能力內之差異距離為 $(0.75-0.2)+(0.75-0.3)=1.0$ ，而能力外之差異距離為 $(0.8-0.75)+(1.0-0.75)=0.30$ 。接著計算指標的分母部分： $[(10-1)/2]=5$ 。以上述計算之差異距離與指標分母比值的百分比，即可得內

注意係數(W^*)為 $(1.0/5)*100=20$ ，外注意係數(B^*)為 $(0.30/5)*100=6$ 。

根據 Huang (2002)的研究發現，內外注意係數(B^* & W^*)具有以下幾點性質：1. B^* 和 W^* 指標與試題長度無關；2. B^* 和 W^* 指標與受試者能力範圍內（或外）所錯誤作答的試題個數呈直線正相關；3. B^* 和 W^* 指標與受試者能力呈先增後減的曲線相關；4.在排除錯誤作答試題個數以及受試者能力的影響之下， B^* 和 W^* 呈現互為無關的現象；5.從 B^* 和 W^* 的指標分析中，高能力受試者傾向比低能力者表現出較高的 W^* 指標值與較低的 B^* 指標值，代表高能力受試者傾向比低能力者表現出較高的能力內關注(Within-Ability/Difficulty Concern)異常作答組型以及較輕微的能力外驚訝(Beyond-Ability/Difficulty Surprise)異常作答組型；6.經主成分分析， B^* 指標和 W^* 指標在受試者能力條件下確實偵測到不同類型的異常作答組型。

然 Huang (2002)的研究是以中等能力受試者參與中等難度測驗的資料矩陣來求得指標決斷值，此指標決斷值是否可適用於不同能力與難度資料結構，值得進一步探索。基於此，本研究實想了解高中低不同能力水準的受試群分別面對不同難度水準試題集時，其內外注意係數指標達到異常的參照標準是否有所不同？若有，則建立不同參照群體的指標標準便有其必要性。

研究方法與實施程序

根據研究背景的分析，本研究之主要目的在於檢視內注意係數(W^*)與外注意係數(B^*)的指標能力，透過異常參照標準的建立，比較不同群體參照標準的異同。

一、研究設計

本研究設計以能力比(ability ratio)與錯誤比(error ratio)所組成細格克服試題數變動對指標所造成的可能干擾。所謂能力比是指受試者答對試題個數(T)相對於測驗試題總數(K)的比

值($t_i = T/K$)，而錯誤比則是指受試者於其所答對試題數之內或之外，答錯試題個數(E)於試題總數的比值($e_i = E/K$)，在第 90 百分位數的設定下，能力比 t_i 可分成十個等級： $0 < t_1 \leq 0.1$ ， $0.1 < t_1 \leq 0.2$ ， $\dots$ ， $0.9 < t_1 < 1.0$ ，並依次定為能力比 1、2、 $\dots$ 、10；錯誤比 e_i 可分成五個等級： $0 < e_1 \leq 0.1$ ， $0.1 < e_1 \leq 0.2$ ， $\dots$ ， $0.4 < e_1 \leq 0.5$ ，並依次定為錯誤比 1、2、 $\dots$ 、5。這是因為受試者所犯錯誤作答的試題數目（能力範圍內該對不對或能力範圍外該錯不錯）最大僅為試題長度的一半，而且受制於受試者能力水準的影響，錯誤比的值不能超過能力比的值。例如在一個 10 道試題的測驗中，得 5 分的受試者最多只能有五道錯誤作答的試題，得 4 分或 6 分的受試者最多只能有四道錯誤作答的試題，得 3 分或 7 分的受試者最多只能有三道錯誤作答的試題， $\dots$ ，依此類推。依據此原理並稍作調整，在 10 個能力比與五個錯誤比的設計中，能力比為 5 者，其可能的最大錯誤比為 5（試想不論原先測驗是偶數或奇數的情況），所以能力比在 5 含以下者，其可能的最大錯誤比可依次為 5、4、3、2 以及 1；其次，能力比為 6 者，其可能的最大錯誤比為 5（原先測驗是奇數的情況）或 4（原先測驗是偶數的情況）。為涵蓋所有可能的組合，設計以 5 為最大的錯誤比，同理，能力比為 7 至 10 者，其可能的最大錯誤比可依次為 4、3、2 以及 1，於是總共設計出 30 個 $[2 \times (5+4+3+2+1)]$ 細格。

二、排序原理

排序檢定(permutation test)的邏輯非常直接(Edgington, 1995; Hays, 1996; Pohlmann, Perkins, & Brutten, 1999)，在進行虛無假設檢定時並不需要知道所欲計算統計量數的理論抽樣分配(theoretical sampling distribution)，這個優點在內外注意係數的標準化型式未發展完成前，將有助於實徵指標標準的建立；換言之，

本研究傾向以數值方法建立內外注意係數的指標標準。又為避免當樣本數增加所衍生龐大的排序個數，本研究以 Edgington (1969) 所提出近似排序檢定(approximate permutation test)作為虛無假設的檢定依據。根據 ver den Brink 與 ver den Brink (1990)的研究，近似排序檢定很適合用在小樣本受試者或小題數測驗的情境之中。而在實務進行時，通常定義等於或大於該觀察統計量數之所有可能統計量數的數目於所抽取代表性母群體數目的比例為其顯著檢定的 p 值(Fisher, 1935; Hayes, 1996)，即當 p 值小於 .10 時，我們即宣稱至少具 90%的信心水準說該檢定達到顯著。因此本研究希望找到的第 90 百分等級的指標決斷值，可以理解為在所有可能指標值中排序至少為第 90 百分位數所顯示的指標數值。

三、研究步驟

本研究的設計因是要產生虛擬的資料結構，先就試題數與樣本數的設計說明如下。首先是試題數，因為指標公式的設計將試題數(K)控制在分母，且 Huang (2002)的研究也發現 B^* 和 W^* 指標與試題長度無關，加上能力比與錯誤比克服試題數變動對指標所造成的可能影響，因此，本研究便不對試題數加以操弄；至於樣本人數的設計，因探討的是受試者的異常作答情形、非試題的異常反應情形，且在相關的研究中顯示，操弄影響受試者的異常作答情形的變項多為試題長度，而非受試者人數（例如 Emons, Glas, & Meijer, 2003; Emons, Meijer, & Sijtsma, 2002; Meijer, 1994; Meijer & van Krimpen-Stoop, 1998; Nering & Meijer, 1998; Orlamdo & Thissen, 2000;），再加上 B^* 和 W^* 指標的設計適用於小樣本的真實班級情境，因而本研究僅以一組樣本數進行虛擬分析。

本研究以盧志明、黃財尉 (2004) 以 Visual Basic 程式語言所撰寫的「內外注意係數指標虛擬排序程式」來產生排序矩陣並計算 B^* 與 W^* 值，詳細實施步驟如下：

步驟 1：資料矩陣的產生

1. 首先以.5 的機率隨機產生一個 300×75 的二元(0, 1)資料矩陣，因如此各變項出現"1"的機率大約在.4 至.6 之間，為避免難度集中於如此小的範圍，進行下一個步驟。
2. 將上述資料矩陣的受試者能力排序並分隔成 10 個階層，每個階層 30 人，且階層之間以約 7.5 分 (75/10) 的差距間隔，再分別以.5 的機率隨機產生整數"1"到每位受試者的作答中，如此在 75 題的作答當中，第一個階層的受試者大約可產生一至七個 "1"的分佈，而第二個階層的受試者大約有八至十五個"1"的分佈，，第 10 個階層的受試者大約有 68 至 75 個"1"的分佈等，如此可以擴大試題難度的分佈。
3. 合併上述 10 個 30×75 資料矩陣成為一個 300×75 的新二元(0,1)資料矩陣。
4. 針對所產生的資料矩陣，依試題難度與受試者能力依序再次排列，以便分割成九個子矩陣(Hh、Hm、Hl、Mh、Mm、Ml、Lh、Lm、Ll)，其中第一個大寫字母代表受試者能力，而第二個小寫字母代表試題難度。

步驟 2：資料矩陣的排序

步驟 1 所產生的九個原始矩陣將作為資料的排序種子(seeds for permutations)，從每個原始資料矩陣中都將產生 500 個排序矩陣結果，因為本研究設計了 30 個 $t_i \times e_i$ 細格，且每個排序矩陣結果的受試者人數皆為 $\times 100$ ，因此在每個 $t_i \times e_i$ 細格將預計平均各有 $\times 1667$ 個($500 \times 100 / 30$)B*與 W*值產生，以這個數字作為指標的代表性母群體應該是足夠的。

步驟 3：確定 B*與 W*指標的決斷值(cutoffs)

將步驟 2 中所產生每一細格的 1667 個 B*與 W*值分別由低至高排列，並找出第 90 百分位數相對應的 B*與 W*值，這些值即為 B*與 W*指標的決斷值。

步驟 4：重複步驟 2 與步驟 3 的過程

為了消除隨機誤差，本研究擬重複操作步驟 2 與步驟 3 的過程 20 遍，以在第 90 百分位數與每一細格的條件下，產生 20 個不同 B*與 W*指標的決斷值。

步驟 5：計算指標決斷值的平均數

本步驟將計算在步驟 4 中所產生的 20 個不同 B*與 W*指標決斷值的平均數與標準差，這些在每一細格與每一個百分位數中之指標決斷值的平均數將作為 B*與 W*指標的最後決斷值。

研究結果

一、指標矩陣分佈

在未確認指標決斷值之前，B*與 W*指標於各樣本矩陣中的出現情形值得檢視。在表 1 中，能力比－錯誤比細格被平均分為三區，每區 10 個細格，在受試者能力依高、中、低順序的區隔下，指標於樣本矩陣上的分佈情形，大概呈一倒 Z 字型分佈（如表 1 反黑加框強調之處）。由表可見高能力者面對不同難度試題時，比較會呈現高分數分配情形（在細格 7_1-10_1 出現比例高達 82.83%）；而低能力者面對不同難度試題時，則比較容易呈現低分數的分配情形（在細格 1_1-4_4 出現比例高達 98.67%）；但中等能力者面對不同難度試題時，其分數的分配就顯得平均（在各細格區間出現的比例均接近八成）。

這表示此虛擬結果也與學習心理的現象相符，一般高能力者在面對試題時常會表現出自我挑戰的本能，尤其是高難度試題，故其得分明顯為高分；而低能力者於面對試題則常會表現出自我放棄的情形，因而其得分則明顯低落，即使至所面對是容易的試題；至於中等能力者的答題情形則介於上述兩者之間，因而其得分的分配高、中、低都有。

二、矩陣內指標分析

以下以 90%信賴區間為例，分別說明內注

意係數與外注意係數在上述倒 Z 型矩陣中的變動趨勢。首先是高能力受試者面對高、中、低難度試題集的情境，如表 2 所示。

由表 2 可看出，內注意係數的值最大發生在細格 7_4 的位置 ($W^* = 7.35$ 與 4.36)，這與 Huang (2002) 的研究一致；又在外 Hh 資料矩陣中，內注意係數的值 (W^*) 高於外注意係數的值 (B^*)，這顯示高能力者面對高難度試題且得分高時，其所犯於能力內應答對卻未答對的異常作答情形，會比其於能力外應答錯卻答對的異常作答情形嚴重；這現象也見於中低度難度試題且高得分的情況 (Hm 與 Hl 中細格 8_1 以上)，此顯示高能力者答錯其能力內試題的異常情形較答對能力外的試題所引發的異常情形更嚴重。

其次，檢視低能力受試者面對高、中、低難度試題集情境 (資料結構 Lh、Lm 與 Ll)，結果如表 3 所示。

由表 3 可看出，外注意係數的值最大發生在細格 4_4 的位置 ($B^* = 2.85$ 、 2.60 與 2.57)，這也與 Huang (2002) 的研究一致。值得注意的是，在低得分的情況下，低能力者面對所有難度試題時，其所犯於能力外應答錯卻答對的異常作答情形會比其於能力內應答對卻未答對的異常作答情形嚴重 ($B^* > W^*$)，這顯示對低能力者其答對能力外的試題所引發的可能訝異情形較其答錯能力內試題的惋惜情境更嚴重。

最後，當中等能力受試者在面對高、中、低難度試題集情境時，情況會如何呢？

由表 4 中可見最大的內注意係數 (W^*) 除

表 1 指標於樣本矩陣的分布情形

細格分佈	H (h, m, l)	M (h, m, l)	L (h, m, l)
1_1-4_4	34 (5.67%) ^a	487(81.17%)	592(98.67%)
5_1-6_5	191(31.83%)	480(80.00%)	163(27.17%)
7_1-10_1	497(82.83%)	474(79.00%)	0(0.00%)

a. 分母以 10 細格 $\times$ 3 難度 $\times$ 20 次重複計算。

表 2 高能力受試者內注意係數 (W^*) 與外注意係數 (B^*) 的 90th 百分位數均值

細格	Hh				Hm				Hl			
	W^*		B^*		W^*		B^*		W^*		B^*	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
7_1	2.18	0.55	1.62	0.47	1.26	0.39	1.34	0.54	0.96	0.18	1.96	0.18
7_2	4.26	1.07	3.11	0.91	2.39	0.68	2.44	0.93	1.83	0.24	2.70	1.10
7_3	5.75	1.53	4.14	1.31	3.27	0.83	3.22	1.32	2.75	--	4.67	--
7_4	7.35	2.40	5.05	1.82	4.36	0.42	2.67	0.79	--	--	--	--
8_1	2.57	0.55	1.31	0.27	1.60	0.57	1.07	0.41	1.04	0.39-	0.92	0.40
8_2	4.68	1.01	2.52	0.59	2.86	1.07	1.91	0.70	1.77	0.67	1.55	0.67
8_3	6.24	1.57	3.45	1.03	3.80	1.55	2.75	1.02	3.38	0.85	2.78	1.70
9_1	2.90	0.48	0.97	0.36	1.94	0.68	0.73	0.33	1.44	0.40	0.57	0.32
9_2	4.46	0.79	1.42	0.43	2.92	1.02	1.15	0.54	2.13	0.61	1.00	0.47
10_1	3.10	0.42	0.40	0.23	2.02	0.61	0.37	0.26	1.31	0.46	0.26	0.19

註：-- 表指標未出現於該細格，故無平均數與標準差的計算。

表 3 低能力受試者內注意係數(W*)與外注意係數(B*)的 90th 百分位數均值

細格	Lh				Lm				Ll			
	W*		B*		W*		B*		W*		B*	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
1_1	0.19	0.11	1.17	0.20	0.18	0.13	1.12	0.23	0.16	0.11	1.06	0.18
2_1	0.35	0.14	1.14	0.17	0.36	0.18	1.09	0.19	0.31	0.13	1.03	0.18
2_2	0.63	0.23	1.83	0.32	0.67	0.29	1.74	0.31	0.60	0.23	1.66	0.30
3_1	0.51	0.16	1.00	0.18	0.50	0.18	0.91	0.20	0.43	0.16	0.90	0.19
3_2	0.96	0.30	1.92	0.34	0.93	0.34	1.76	0.40	0.78	0.26	1.74	0.38
3_3	1.28	0.44	2.51	0.45	1.16	0.45	2.31	0.53	0.98	0.33	2.24	0.49
4_1	0.65	0.18	0.91	0.23	0.63	0.18	0.92	0.22	0.51	0.12	0.84	0.19
4_2	1.23	0.36	1.69	0.30	1.11	0.32	1.65	0.34	0.98	0.22	1.60	0.38
4_3	1.60	0.44	2.24	0.37	1.47	0.42	2.17	0.45	1.28	0.27	2.09	0.53
4_4	2.03	0.53	2.85	0.49	1.77	0.52	2.60	0.51	1.56	0.29	2.57	0.69

Mh 發生在細格 8_3 之外(4.61)，其餘均發生在 7_4 的位置 (4.51 與 4.78)，這與 Huang (2002) 的研究大致相似；此外，在 Mh 矩陣中細格 6_4 之前、Mm 矩陣中細格 5_4 之前、Ml 矩陣中細格 4_4 之前的外注意係數均大於內注意係數 ($B^* > W^*$)，其餘細格則呈現相反結果，此乃因為中等能力受試者綜合表 3 與表 4 的結果，在低得分情況，答對能力外的試題的異常作答情形較答錯能力內試題更值得重視；而高得分的情況下，答錯其能力內試題應較答對能力外的試題更具異常的作答現象。

三、指標標準差異分析

表 5 與表 6 是以依變項為每一細格內之指標決斷值平均數（分別為 W^* 與 B^* 指標），自變項為九種不同資料結構，所作單因子變異數分析在 .05 顯著水準下依變項平均數同質子集 (homogeneous subsets) 的歸類情形。兩個表中所呈現的數字為子集代號，數字較大者代表其內的指標平均數較大，同一子集數字代表指標決斷值於不同資料結構中並無顯著差異（例如表 5 細格 1_1 中 W^* 指標決斷值，在資料結構 Lh、Lm、與 Ll 中並無不同，子集數字皆為 1）。另外，表中也呈現子集歸屬的顯著性 ($p_i, i = 1 \dots 4$)，並以 p 值的大小決定子集的主要與次要

歸屬，例如在細格 3_2 中，出現在 Lh 與出現在 Lm 及 Ll 的 W^* 數值並無顯著差異（主要歸屬的子集皆為 1），但其中 Lh 的 W^* 數值亦可歸入子集 2，只是較不明顯（括號內的 2）。最後，表格中的 HN 代表指標重覆次數的樣本調和平均值 (Harmonic Mean sample size)。

整體而言，由表 5 可見， W^* 指標決斷值出現的分佈（有歸屬子集數字之處）呈現類似倒 Z 字型的圖形，表 6 的 B^* 指標決斷值分佈也有這種現象，符應表 1 中所陳述的現象。而細部言之，於表 5 細格 1_1-4_4 之中，所出現的 W^* 指標決斷值大致可歸屬於中能力區 (M) 和低能力區 (L)，從主要子集數值的不同可知，存在於中能力區 (M) 的 W^* 指標決斷值明顯不同於存在低能力區 (L) 的 W^* 指標決斷值；但區間之內子集歸屬較為一致，代表 W^* 指標決斷值在中能力區與低能力區之內並無顯著差異。這種情形也出現在表 6 的 B^* 指標決斷值分佈中，但後者的分佈區隔較清楚，比較沒有次要子集的類別歸屬的模糊情況，亦即中能力區與低能力區之間的 B^* 指標決斷值明顯不同。

至於細格 5_1-6_5 之中的 W^* 指標決斷值，則是分佈於高能力區、中能力區與低能力區之間，代表三個區間的 W^* 指標決斷值具明顯差異，表 6 的 B^* 指標決斷值分佈也有這種情

表 4 中能力受試者內注意係數(W*)與外注意係數(B*)的 90th 百分位數均值

細格	Mh				Mm				Ml			
	W*		B*		W*		B*		W*		B*	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1_1	0.40	0.32	2.23	0.72	--	--	--	--	--	--	--	--
2_1	0.93	0.66	1.98	0.31	0.81	0.42	1.99	0.68	--	--	--	--
2_2	1.47	1.00	3.27	0.51	1.23	0.56	3.12	0.70	0.98	0.91	2.77	0.21
3_1	1.00	0.51	1.76	0.33	0.91	0.36	1.69	0.45	--	--	--	--
3_2	1.87	1.02	3.30	0.59	1.56	0.54	2.98	0.71	2.10	0.59	3.11	0.71
3_3	2.29	1.23	4.33	0.79	1.90	0.65	3.90	0.94	2.26	0.72	3.92	0.84
4_1	1.15	0.48	1.62	0.29	1.01	0.30	1.46	0.35	1.32	0.48	1.40	0.36
4_2	2.18	0.91	3.11	0.61	1.83	0.56	2.82	0.63	2.39	0.64	2.62	0.60
4_3	2.89	1.24	4.04	0.82	2.35	0.73	3.68	0.84	3.10	0.91	3.43	0.79
4_4	3.49	1.45	4.90	1.01	2.80	0.90	4.52	1.05	3.77	1.12	4.20	0.97
5_1	1.26	0.51	1.44	0.26	1.06	0.37	1.27	0.32	1.47	0.44	1.21	0.26
5_2	2.47	0.92	2.81	0.58	2.10	0.65	2.61	0.62	2.66	0.59	2.31	0.51
5_3	3.20	1.18	3.68	0.77	2.74	0.86	3.42	0.80	3.48	0.81	3.02	0.66
5_4	3.92	1.48	4.55	0.99	3.37	1.05	4.18	0.98	4.27	1.05	3.69	0.82
5_5	--	--	--	--	--	--	--	--	--	--	--	--
6_1	1.32	0.41	1.44	0.39	1.29	0.33	1.13	0.33	1.41	0.41	1.05	0.26
6_2	2.61	0.74	2.72	0.63	2.57	0.63	2.24	0.69	2.85	0.68	2.03	0.46
6_3	3.40	1.01	3.57	0.87	3.35	0.84	2.93	0.91	3.71	0.92	2.65	0.62
6_4	4.18	1.28	4.36	1.11	4.10	1.04	3.61	1.17	4.55	1.16	3.22	0.78
6_5	--	--	--	--	--	--	--	--	--	--	--	--
7_1	1.53	0.48	1.29	0.46	1.47	0.35	0.98	0.31	1.50	0.34	0.98	0.28
7_2	2.91	0.68	2.52	0.70	2.84	0.68	1.84	0.63	2.98	0.65	1.81	0.47
7_3	3.89	0.94	3.32	1.03	3.71	0.93	2.39	0.84	3.89	0.87	2.29	0.59
7_4	4.40	1.01	3.51	0.93	4.51	1.13	2.80	1.03	4.78	1.14	2.70	0.71
8_1	--	--	--	--	1.72	0.35	0.79	0.25	1.70	0.30	0.85	0.25
8_2	3.26	0.87	2.26	0.75	3.10	0.72	1.49	0.49	3.09	0.52	1.59	0.49
8_3	4.61	1.24	2.62	0.93	4.01	1.02	1.86	0.63	4.01	0.75	2.04	0.69
9_1	--	--	--	--	1.88	0.37	0.64	0.27	1.87	0.34	0.57	0.25
9_2	--	--	--	--	3.09	0.99	0.83	0.43	3.01	0.46	0.78	0.32
10_1	--	--	--	--	--	--	--	--	2.19	0.38	0.30	0.19

註：--表指標未出現於該細格，故無平均數與標準差的計算。

形出現。值得注意的是在細格 7_1-10_1 之中，W*指標決斷值在高能力區與中能力區差異的明確性（次要子集較少）高於 B*指標決斷值於這兩能力區的分佈，對照於前述細格 1_1-4_4

之中前者的差異明確性低於後者的現象，或許可說明針對不同的資料結構時，應用 W*指標與 B*指標決斷值時，宜有不同的決斷值水準。

表 5 W*指標的子集歸類分佈

細格	Hh	Hm	HI	Mh	Mm	MI	Lh	Lm	Ll	p_1	p_2	p_3	p_4	HN
1_1				2			1	1	1	.97	1.00			20
2_1				2	2		1	1	1	.99	.91			18.42
2_2				2	2	1	1	1	1	.99	.76			20
3_1				2	2		1	1	1	.95	.94			20
3_2				3	3(2)	3	1(2)	1	1	.96	.05	.13		20
3_3				3	3(2)	3	1(2)	1(2)	1	.89	.08	.74		18
4_1				2	2	2	1	1	1	.86	.12			19.83
4_2	4			3	3(2)	4(3)	1(2)	1(2)	1	.96	.07	.30	.32	19.51
4_3	4			3(4)	3(2)	3(4)	1(2)	1(2)	1	.97	.16	.34	.15	19.09
4_4				3	3(2)	3	1(2)	1(2)	1	.80	.05	.08		18
5_1	2			1	1	2(1)				.07	.18			17.63
5_2	4			3(4)	2(3)	3(4)	1(2)	1	1	.67	.31	.28	.09	18.57
5_3	4			3	3(2)	3(4)	1(2)	1	1	.98	.08	.30	.28	18.26
5_4	4			3(4)	3(2)	3(4)		1(2)	1	.99	.15	.54	.10	13.61
5_5														
6_1	2	1		1	1	1				.96	1.00			14.58
6_2	3	2		2	2	2			1	1.00	.12	1.00		18.62
6_3	3	2		2	2	2			1	1.00	.29	1.00		18.43
6_4	2			1	1	1				.77	1.00			20
6_5														
7_1	2	1		1	1	1				.46	1.00			18.57
7_2	3	1(2)	1	2	2	2				.56	.48	1.00		14.40
7_3	2	1		1	1	1				.49	1.00			20
7_4	2			1	1	1				.91	1.00			17.63
8_1	3	2	1		2	2				1.00	.95	1.00		20
8_2	3	2	1	2	2	2				1.00	.81	1.00		20
8_3	2	2	2	2	2	2				.27	1.00			13.53
9_1	2	1	1	1		1				.06	1.00			18.06
9_2	3	2(1)	1		1	1				.06	.98	1.00		20
10_1	3	2	1		2					1.00	.73	1.00		20

另外，將上述指標決斷值在統計上顯著差異的次數累計情形呈現於表 7 之中。不論是 W* 或 B* 指標的決斷值於同一能力區間內的差異情形($H \neq H$, $M \neq M$, $L \neq L$)其實都不大，但在相鄰能力區間的差異就頗為明顯，例如在高能力

區與中能力區的比較上($H \neq M$)，W* 與 B* 指標決斷值分別累計 52 次與 50 次的顯著差異，而在中能力區與低能力區的比較上($M \neq L$)，兩個指標決斷值更是累計分別高達 93 次與 104 次的顯著差異。

表 6 B*指標的子集歸類分佈

細格	Hh	Hm	HI	Mh	Mm	MI	Lh	Lm	LI	p_1	p_2	p_3	p_4	HN
1_1				2			1	1	1	.86	1.00			20
2_1				2	2		1	1	1	.91	1.00			18.42
2_2				2	2		1	1	1	.86	.89			20
3_1				2	2		1	1	1	.88	.96			20
3_2				2	2	2	1	1	1	.96	.64			20
3_3				2	2	2	1	1	1	.93	.63			18
4_1				2	2	2	1	1	1	.98	.32			19.83
4_2	3			2	2	2	1	1	1	1.00	.47	1.00		19.51
4_3	3			2	2	2	1	1	1	.99	.45	1.00		19.09
4_4				2	2	2	1	1	1	.96	.29			18
5_1	2			1	1	1				.31	1.00			17.63
5_2	3			2	2	2	1	1	1	1.00	.42	1.00		18.57
5_3	4			3	3	3(2)	1(2)	1	1	.99	.08	.45	1.00	18.26
5_4	3			2	2	2		1	1	.98	.45	1.00		13.61
5_5														
6_1	1	1		2(1)	1	1				.12	.08			14.58
6_2	3	2		2	2	2			1	1.00	.15	1.00		18.62
6_3	3	2		2	2	2			1	1.00	.16	1.00		18.43
6_4	2			1	1	1				.07	1.00			20
6_5														
7_1	2	2(1)		2(1)	1	1				.16	.24			18.57
7_2	2	2(1)	2(1)	2(1)	1	1				.10	.36			14.40
7_3	2	2(1)		2(1)	1	1				.06	.12			20
7_4	2			1	1	1				.28	1.00			17.63
8_1	2	2(1)	1		1	1				.14	.25			20
8_2	2	1(2)		2	1	1				.49	.10			20
8_3	2	2(1)	2(1)	2(1)	1	1				.27	.38			13.53
9_1	2	1(2)	1		1	1				.65	.26			18.06
9_2	2	1(2)	1(2)		1	1				.16	.08			20
10_1	3	2(1)	1		2	2				.06	.98	1.00		20

表 7 累計指標決斷值之顯著差異次數

	細格	H ≠ H	M ≠ M	L ≠ L	H ≠ M	M ≠ L
W*	1_1--4_4	0	2	0	3	66
	5_1--6_5	3	0	0	18	27
	7_1--10_1	17	0	0	31	0
	Sum	20	0	0	52	93
B*	1_1--4_4	0	0	0	6	75
	5_1--6_5	2	4	0	21	29
	7_1--10_1	4	0	0	23	0
	Sum	6	0	0	50	104

結論與建議

雖然本研究僅侷限在內外注意係數 90% 等第的決斷值分析，但卻為日後建立 95% 與 99% 等第的決斷值奠定基礎。同時本研究結果也提供我們一些關於異常作答現象的發現：1. 指標於樣本矩陣上的分佈，大概呈一倒 Z 字型分佈，符合實際作答現象；2. 高能力者答錯其能力內試題的異常情形較答對能力外的試題所引發的異常情形嚴重；3. 答對能力外試題所引發的可能訝異情形對低能力者而言比其答錯能力內試題的惋惜情境更加嚴重；4. 大體上說，不同能力區間的內外注意係數指標之決斷標準值有明顯差異，但能力區內面對不同難度水準的情況差異就不甚明顯，甚至低能力區內與中能力區內可以共用指標的決斷標準。

從這些發現也意味著內外注意係數指標決斷值的建立可能需因不同的能力難度組合而調整，有適合高能力的資優測驗，也有適合低能力的精熟測驗，以及適合一般能力的普通測驗，惟可將指標決斷表依倒 Z 字型建構，高分區的高能力群、低分區的低能力群，以及適用一般能力之全部分數區的建構。另外，若加深考量異常作答判定的重要性，未來研究亦可針對第 95 與 99 百分等級的指標決斷值進行整理，以降低錯誤歸類。然因本研究站在輔導與診斷角度，因而在指標決斷值的判定上容許較多的誤判。

最後在輔導與諮商或其他領域的實務應用上，W* 與 B* 指標決斷值比較適用於認知結構的測驗之中（試題具備正確或最佳答案）。例如在課室內學生的數學學習情境中，使用者可以根據受試者在數學測驗中作答情形，先排序試題難度（由易至難）與作答者能力（由高至低），再以受試者答對題數（得分）為駐標，依據公式分別計算 W* 與 B* 指標值，並依能力比與錯誤比細格，對照比較指標決斷值，最後檢視受試者的作答是否達於異常，以進一步進行必要的學習輔導處遇。另外，對於輔導與諮商領域中較多的情意測驗情境，使用者也可以針對受試

者之知覺、想法或態度上的迷失情形，將情意測驗先轉成類似認知型測驗，再應用 W* 與 B* 指標來協助學生釐清迷失概念。

致謝

研究者感謝國科會科教處之經費補助 (NSC 92-2511-S-415-006)，特別感謝兩位期刊匿名審查委員的寶貴意見，在此一併感謝致意。

參考文獻

- 盧志明、黃財尉（2004）。「內外注意係數指標虛擬排序程式（Visual Basic 6.0）。未出版程式。
- Cheng, W. D. (1997). *An exploratory study examining the relationships among total score and different caution and fit indices*. Unpublished thesis. The Ohio State University.
- D'Costa, A. (1993a, April). *Extending the Sato caution index to define the within and beyond ability caution indexes*. Paper presented at convention of National Council for Measurement in Education, Atlanta, GA.
- D'Costa, A. (1993b, April). *The validity of the W, B and Sato Caution indexes*. Paper presented at the Seventh International Objective Measurement Conference, Atlanta, GA.
- Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement*, 28, 105-113.
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11, 59-79.
- Drasgow, F., Levine, M. V., & William, E. A. (1986). Appropriateness measurement with

- polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Edgington, E. S. (1969). Approximate randomization tests. *Journal of Psychology*, 72, 143-149.
- Edgington, E. S. (1995). *Randomization tests* (3rd ed.). NY: Marcel Dekker, Inc.
- Emons, W. H., Glas, C. A., & Meijer, R. R. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement*, 27, 457-478.
- Emons, W. H., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of U_3 person-fit statistic. *Applied Psychological Measurement*, 26, 88-108.
- Fisher, R. A. (1935). *The design of experiments*. London: Oliver & Boyd.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18(3), 133-146.
- Hayes, A. (1996). Permutation test is not distribution free: Testing $H_0 : \rho = 0$. *Psychological Methods*, 1, 184-198.
- Huang, T. W. (2002). *The power of the Beyond-Ability-Surprise Index and the Within-Ability-Concern Index for detecting person/item aberrances: A Monte Carlo study*. Unpublished doctorate dissertation, The Ohio State University.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood, IL: Dow Jones-Irwin.
- Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. *Applied Measurement in Education*, 1(1), 17-31.
- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-setting judgments. *Applied Measurement in Education*, 1, 17-31.
- Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1(2), 152-176.
- Kogut, J. (1986). *Review of IRT-based indices for detecting and diagnosing aberrant response patterns* (Research Report No. 86-4). Enschede, Netherlands: University of Twente, Department of Education.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Linacre, J. M., & Wright, B. D. (1994). Chi-square fit statistics. *Rasch Measurement Transactions*, 8(2), 360-361.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful Person-fit statistic. *Applied Psychological Measurement*, 18, 311-314.
- Meijer, R. R., & Sijtsma, K. (1999). *A review of methods for evaluating the fit of item score patterns on a test* (Research Report No. 99-01). Enschede, Netherlands: University of Twente, Faculty of Educational Science and Technology.
- Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (1998). *Simulating the null distribution of person-fit statistics for conventional and adaptive tests*. (Research Report No. 98-02). Faculty of Educational Science and Technology.
- Nering, M. L., & Meijer, R. R. (1998). A com-

- parison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, 22, 53-69.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50-64.
- Pohlmann, J. M., Perkins, K., & Brutten, S. (1999, October). *Permutation tests in principal component analysis*. Paper presented at the meeting of the Mid-Western Education Researcher Association, Chicago, IL.
- Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, 18, 171-182.
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo: Meiji Tosho.
- Smith, R. M. (1982). *Detecting measurement disturbances with the Rasch model*. Unpublished dissertation. University of Chicago.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49, 95-110.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, 7, 215-231.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20, 221-230.
- van den Brink, W. P., & van den Brink, S. G. J. (1990). A modified approximate permutation test procedure. *Computational Statistics Quarterly*, 3, 241-247.
- van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13 (3), 267-298.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

附 錄

1.Bs & Wc Indices (D'Costa, 1993a, 1993b)

$$W_i = \frac{\text{Sum of p's for items got wrong within ability level}}{\text{Sum of p's for all items within ability level}} = \frac{\sum_{j=1}^{T_i} (1 - u_{ij}) p_{\bullet j}}{\sum_{j=1}^{T_i} p_{\bullet j}},$$

$$B_i = \frac{\text{Sum of q's for items got right beyond ability level}}{\text{Sum of q's for all items beyond ability level}} = \frac{\sum_{j=T_i+1}^K u_{ij} q_{\bullet j}}{\sum_{j=T_i+1}^K q_{\bullet j}}$$

where

i = 1, 2, ..., N, refer to subjects,

$j = 1, 2, \dots, K$, refer to items,

$u_{ij} = 1$ if subject i answered item j correctly,

$= 0$ if subject i answered item j wrongly,

T_i = total score earned by subject i ,

K = test length,

$p_{\bullet j}$ = the proportion of all persons who got the item j correct,

$q_{\bullet j}$ = the proportion of all persons who got the item incorrect, and $q_{\bullet j} = 1 - p_{\bullet j}$.

2. W'' and B'' indices (Cheng, 1997)

$$W'' = \frac{\text{Sum of (p's of items wrong within ability - p of Item}_{(T+1)})}{\text{The average p times (K/2)}},$$

$$B'' = \frac{\text{Sum of (p of Item}_{(T)} - \text{p's of items right beyond ability)}}{\text{The average p times (K/2)}}$$

3. W^* and B^* indices (Huang, 2002)

$$\begin{aligned} W_i^* &= \frac{\text{Sum of the discrepancy distances of difficulty within a person's ability level}}{\text{Half of maximum value of the numerator}} \times 100 \\ &= \frac{\sum_{j=1}^{T_i} (1 - u_{ij}) \times (q_{iT}^* - q_{\bullet j})}{[(K-1)/2] \times 1/2} \times 100 \end{aligned}$$

$$\begin{aligned} B_i^* &= \frac{\text{Sum of the discrepancy distances of difficulty beyond a person's ability level}}{\text{Half of maximum value of the numerator}} \times 100 \\ &= \frac{\sum_{j=T_i+1}^K u_{ij} \times (q_{\bullet j} - q_{iT}^*)}{[(K-1)/2] \times 1/2} \times 100. \end{aligned}$$

A Study of Cutoffs for Aberrant Indices under Different Data Structures

Tsai-Wei Huang

Abstract

Unexpected response patterns can provide substantial information for individual responses. Many aberrant indices were designed to detect examinees' aberrant response patterns. Two newly developed aberrance indices, the Within-Ability-Concern Index (W^*) and the Beyond-Ability-Surprise Index (B^*), were classified as the group-based index family and were applicable to practical classrooms. The purpose of this study was to establish and examine the cutoffs of these two indices through a nonparametric test technique (permutation test) on different levels of individual ability and item difficulty. Nine simulated data matrices, including all combinations of high-median-low levels of ability and difficulty, were generated and permuted 500 times. The 90th percentiles of permutations in each ability-error-ratio cell were calculated and served as cutoffs for each index. After twenty repeats, an average of each index cutoff was used as a point of reference for the W^* and B^* indices, respectively. Findings showed that cutoffs of indices mapped by ability-difficulty matrices and ability-error-ratio cells from low to high showed a reverse Z-shape distribution. This was in line with real responses. As high achievers encounter different difficulty levels of items, they tend to have a high-score distribution. Low achievers tend to have a low-score distribution, and intermediate achievers tend to have an average distribution. Second, the high achievers were found to perform more aberrantly on their within-ability items than on their beyond-ability items. Third, the low achievers performed more aberrantly on their beyond-ability items than on their within-ability items. Fourth, the cutoffs of the two indices differed significantly between ability districts, but minimally different within each ability district. Finally, several tables and suggestions were also provided as the resolution of the issues.

Keywords: Aberrant indices, Within- Ability-Concern Index (W^*), Beyond-Ability-Surprise Index (B^*), Cutoffs

Tsai-Wei Huang Department of Counseling, National Chiayi University twhuang@mail.ncyu.edu.tw