

二元計分的 BW 作答機率模式精確度研究： 蒙地卡羅法

沈采瑱¹ 黃財尉²

摘要

本研究旨在探討以異常作答偵測指標 BW 內外注意係數所建構的兩種二元計分 BW 作答機率模式：BW 內嵌模式 (BWEM) 與 BW 本位 Rasch 模式 (BWRM) 的實用性，期待經由檢視模式作答預測能力的精確度，為模式應用增添可靠的支持。本研究以命中率代表模式的作答預測能力，並作為評估模式的準則。透過比較 BWEM 與 BWRM 在不同受試者人數與測驗長度下的命中率，探討兩種作答機率模式於實務應用時需留意之處。

透過三因子混合設計變異數分析，研究結果顯示：(1)受試者人數、測驗長度及作答機率模式三者間的確存在交互作用，但是效果量很低；(2)整體、勝券在握與蜀道難登三區，命中率在不同受試者人數、測驗長度及作答機率模式間的確存在顯著差異；(3)BWEM 與 BWRM 的命中率孰高孰低，視所在的作答反應區域而定。

關鍵詞：精確度、BW 內嵌模式、BW 本位 Rasch 模式

1. 沈采瑱，千昊資訊有限公司研究員

2. 黃財尉，國立嘉義大學輔導與諮商學系教授

收件日期：2017.07.28；完成修改：2019.04.14；正式接受：2019.04.17

通訊作者：黃財尉；Email：twhuang@mail.ncyu.edu.tw

地址：嘉義縣民雄鄉文隆村 85 號 國立嘉義大學輔導與諮商學系

The Accuracy of Dichotomous BW Probabilistic Models: A Monte Carlo Study

Tsai-Chen Shen¹ Tsai-Wei Huang²

Abstract

The main purpose of this study was to examine the accuracy of two dichotomous BW probabilistic models constructed by BW aberrance indices: the BW embedded model (BWEM) and the BW based Rasch model (BWRM). This study used the hit rate as an index to evaluate the accuracy of the models. Through examining the differences between hit rates of BWEM and BWRM under different circumstances (number of examinees, test length, and probabilistic model), this study explored the characteristics of these two probabilistic models.

The results from the three-way mixed design ANOVA showed:

1. There exists a three-way interaction effect among the number of examinees, test length, and probabilistic model, but the strength of association was not strong.
2. There existed significant differences of hit rates among specific response areas in the conditions of number of examinees, test length, and probabilistic model, respectively.
3. The hit rate of BW embedded model could be higher or lower than that of BW based Rasch model, depending on which response area is concerned.

Keywords: Accuracy rate, BW based Rasch model, BW embedded model

1. Tsai-Chen Shen, Analyst, Chainhao Information Co., Ltd.

2. Tsai-Wei Huang, Professor, Department of Counseling, National Chiayi University

Received: 2017.07.28; Revised: 2019.04.14; Accepted: 2019.04.17

Corresponding Author: Tsai-Wei Huang; Email: twuang@mail.ncyu.edu.tw

Address: No.85, Wenlong Vil., Minxiong Township, Chiayi County 621, Taiwan

Department of Counseling, National Chiayi University

壹、前言

在成長過程中，幾乎所有人都有面臨各種測驗施測的經驗，其中與多數人最為相關的測驗便是學習場域的教育測驗，教師自編測驗便是一例。透過測驗，教師得以進行教學評量、診斷學習困難，幫助學生學習（余民寧，2011）。然而，測驗雖然能讓教師快速評估學生的學習狀況，但分數能代表學生真正的實力嗎？無論是古典測驗理論（classical test theory, CTT）或試題反應理論（item response theory, IRT），針對這道問題的答案，都是否定的。測驗理論的學者們強調不能只因一次測驗就斷定學生能力，因為學生可能因粗心、猜測、作弊或面臨考試焦慮而無法透過分數反應其真正的能力（余民寧，2011；Gulliksen, 1987; Harnisch & Linn, 1981; Levine & Rubin, 1979; Meijer, 1996）。因此，為了跳脫分數無法代表學生能力此一限制，如何更有效地評估學生狀態便成為許多測驗學者關注的焦點。

古典測驗理論假設每位受試者都有某種潛在能力，該能力可透過測驗表現在分數上，但因為古典測驗理論假設真實分數和實得分數之間存有誤差，再加上理論的限制，一直以來，如何更有效地解讀測驗結果已成為許多古典測驗理論支持者探索的重點，且致力於發展可檢測學生作答反應和進行試題分析的指標，以協助改善教學（黃財尉，2008，2011；D'Costa, 1993; Harnisch, 1983; Harnisch & Linn, 1981; Huang, 2002, 2006, 2011, 2012; Sato, 1975, 1980; Tatsuoka & Linn, 1983）。

至於試題反應理論，係指由個別試題的觀點解釋測驗分數（余民寧，2009），因為假設合理且嚴謹，受到許多測驗學者喜愛。已有許多學者運用此理論說明如何評估學生能力與試題分析，並發展出相關指標偵測學生作答反應，供教學評量參考（Cui & Li, 2015; Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982; Levine & Rubin, 1979; Tatsuoka, 1984）。可惜的是，試題反應理論需要堅實的數理背景為基礎，對身處第一線的教師而言，不見得能有餘力運用此理論評估學生。再加上此理論適合用於大樣本，對於班級人數頂多 25 至 30 位學生的教師而言，實用性相對較低。

雖然古典測驗理論與試題反應理論各有優缺點，但依據這兩種理論所發展的研究可知，中西方研究教育測驗與評量的學者們，莫不以如何能更有效評估學生學習狀況，找出學生在學習上需要補救之處為研究重點，期許能為

教師提供較佳的學生學習結果評估工具，提升學生學習成就。在此理念下，為了提供教學評量不同的分析工具，節省教師的時間成本與精力，亦無需擔心不懂複雜的數學模型，便可同時了解學生作答反應並分析試題，以進行教學改進，Huang（2002, 2006）整合 Sato（1975, 1980）的注意係數（Caution Index）和學生問題表（student-problem chart, SP 表），並承襲 D'Costa（1993）的留意指標（ W_c ）與驚訝指標（ B_s ）概念，另建立可同時偵測學生作答異常並分析試題的 W 指標和 B 指標。透過模擬與實徵，指出這兩項新指標有一定程度的應用價值，並且擴充成 BW 認知診斷模式（BW cognitive diagnostics model），內含本研究的重點，兩種能對受試者作答反應進行預測的作答機率模式：BW 內嵌式模式（BW embedded model, BWEM）與 BW 本位 Rasch 模式（BW Rasch-based model, BWRM）（黃財尉，2008，2011；黃財尉、盧志明，2013；Huang, 2011, 2012; Huang & Wu, 2013）。

只是回顧 BW 認知診斷模式的發展，BW 作答機率模式在預測受試者作答反應的能力一直未有機會做深入討論，這多少不利於該模式的推行與應用。因為在實務現場，面臨模式選擇時，模式的預測能力是評估標準之一（Hambleton & Swaminathan, 1985）。事實上，模式選擇的目的不只是想找到和資料適配的模式，還希望此模式能在面對不同資料時，有能力進行預測（Kang & Cohen, 2007）。一旦模式具有好的預測能力，意謂它被使用的機會將大幅提升，而且它所選用的參數也會成為關注焦點。以 Rasch 模式為例，此模式提到的受試者能力與試題難度便是了解受試者與試題的重要參考。因此，為了提升 BW 認知診斷模式的實用性，增進該模式所含參數的價值與使用者的信心，本研究認為可以針對目前已發展的兩套作答機率模式進行預測能力之討論，補足 BW 認知診斷模式在發展過程中尚未觸及之處。

為了探討兩種 BW 作答機率模式的預測能力，在整理有關測驗模式比較的文獻後，發現相關研究多著重在模式適配度與作答異常指標偵測力的討論，分析內容則視其目的，實徵資料或是模擬資料皆有，常見的作法是檢視實際值與模式預測值的殘差（Ludlow, 1985, 1986; Smith, 2000）。然而，作答正確與否是二元概念，模式所求得的預測值卻是連續概念，由二元與連續的差異評估模式的預測能力，合理性似乎存在討論空間。若能將預測值轉為二元概念再與實際結果做比較，會不會更能貼近模式有「好的預測能力」所欲表達的意思。故本研究參考邏輯斯迴歸（logistic regression）的作法，以模式正確預測受試者作答反應的比例，即命中率（hit rate），作為預測能力的評

估標準，衡量作答機率模式在預測受試者作答反應上的精確度。

另外，為了證明模式值得使用，無論是模擬或實徵研究，除了決定研究標的外，自變數的選擇也是比較不同模式優劣時需要考量的重點。考量教學現場，以及模式適配度與作答異常指標偵測力的相關研究（盧志明、黃財尉、方將任，2007；Drasgow, 1982; Huang, 2011, 2012; Meijer & Sijtsma, 1995, 2001; Stone & Zhang, 2003; St-Onge, Valois, Abdous, & Germain, 2011），本研究選擇由受試者人數與測驗長度的變化著手，探討 BWEM 與 BWRM 的預測能力，以協助教師運用 BW 認知診斷模式在分析學生和試題時，面臨作答機率模式的選擇，能有參考基準。期待透過比較 BWEM 與 BWRM 的命中率，為學生作答反應分析提供另一個實用工具，以呼應 Karabatsos（2003）所言，精準地衡量學生學習結果以協助教育決策的健全。

貳、BWEM 與 BWRM

BWEM 與 BWRM 分別是 Huang 與 Wu（2013）以及黃財尉、盧志明（2013）運用 BW 認知診斷模式的八項指標，設計用來預測受試者作答反應的兩種作答機率模式。其中，BW 認知診斷模式的核心指標 BW 內外注意係數是改進 D'Costa（1993）的指標而來，此指標除了運用 Sato（1975, 1980）的 SP 表，更修正 Sato 的指標無法有效辨識異常作答程度、會沖銷兩種不同心理作答特徵（如粗心大意或猜測），以及經驗法則取向但缺乏統計上判斷標準之困難（黃財尉，2011；Huang, 2002）。以下就 BW 認知診斷模式的八項指標與兩種作答機率模式做簡單說明。

一、BW 認知診斷模式的八項指標

BW 認知診斷模式的用意透過分析受試者的作答反應，了解受試者在學習過程中的哪些方面需要留意，以及經由試題分析，思考如何改進試題品質。它包含受試者作答與試題反應兩面向，分別對應 Person-fit BW 與 Item-fit BW 兩組指標，指標各自所代表的意義如表 1 所示（Huang & Wu, 2013）。

表 1 BW 認知診斷模式的指標意義對照表

Person-fit BW 指標		Item-fit BW 指標	
指標	意義	指標	意義
<i>W</i>	粗心度 (carelessness)	<i>w</i>	提示度 (hint)
<i>B</i>	猜測度 (guess)	<i>b</i>	干擾度 (disturbance)
<i>C</i>	精熟度 (capability)	<i>c</i>	真難度 (difficulty)
<i>M</i>	迷思度 (misconception)	<i>m</i>	簡單度 (easiness)

表 1 的 *W* 為粗心度指標，評估受試者在作答時是否出現粗心；*B* 為猜測度指標，檢視受試者是否以猜測方式作答；*C* 為精熟度指標，透過衡量能力內的受試者正確作答情形，檢視受試者對試題的精熟程度；*M* 為迷思度指標，衡量能力外的受試者作答錯誤情形，了解受試者對試題的理解是否因盲點而陷入迷思。

w 為提示度指標，檢視試題是否因具提示效果，導致在此難度內無法難倒理應被考倒的受試者；*b* 為干擾度指標，檢視試題是否對受試者的作答反應有所干擾，導致原本不應被難倒的受試者卻被考倒；*c* 為真難度指標，衡量試題難度；*m* 為簡單度指標，評估試題是否過度簡單，導致試題不具鑑別力。至於 Person-fit BW 指標以及 Item-fit BW 指標的詳細計算公式，請見 Huang 與 Wu (2013)。

二、BW 內嵌模式 (BW embedded model)

Huang 與 Wu (2013) 運用 BW 認知診斷模式的八項指標，建立可代表受試者能力與試題難度的參數，分別為：

$$P_i = [C_i \times (1 - W_i)] \times [1 - (1 - B_i) \times M_i] \quad (1)$$

$$P_j = [c_j \times (1 - w_j)] \times [1 - (1 - b_j) \times m_j] \quad (2)$$

第(1)式的 $C_i \times (1 - W_i)$ 表示受試者面對試題時，真的會且沒有粗心； $(1 - B_i) \times M_i$ 是指受試者沒有猜題且存在迷思，即真不會，以 1 減之去除真不會後，求得兩項之乘積 P_i 。此值代表第 i 位受試者就整份試卷正確作答的能力，是第 i 位受試者的整體能力，數值愈大表示能力愈高。

第(2)式的 $c_j \times (1 - w_j)$ 表示試題真的難且沒有提示； $(1 - b_j) \times m_j$ 是指試題沒有干擾且簡單，以 1 減之去除試題真簡單之情況，求得兩項之乘積 P_j 。

此值代表第 j 題考倒受試者的強度 (power)，第 j 題的整體強度數值愈大，表示考倒的受試者愈多。這裡所指的試題強度與傳統的試題難度之概念相似，但因計算方式不同，故改以強度稱之。

結合第(1)式與第(2)式，並以 Sato (1975) 的 S - P 表的 S 曲線為分界，設定受試者於能力內與能力外對每道試題的作答機率公式，如下所示：

$$P_{ij}^{EM} = \begin{cases} \{[P_i \times (1 - P_j)]^{u_{ij}} \times [(1 - P_i) \times P_j]^{(1-u_{ij})}\}^{1/4} & \text{當 } T_i \geq j \\ 1 - \{[P_i \times (1 - P_j)]^{u_{ij}} \times [(1 - P_i) \times P_j]^{(1-u_{ij})}\}^{1/4} & \text{當 } T_i < j \end{cases} \quad (3)$$

P_{ij}^{EM} 為第 i 位受試者對第 j 題作出正確反應的機率，此機率值的計算考量實際作答結果，是對受試者作答能力的一種相信 (belief)。其中， u_{ij} 為受試者的原始作答結果， u_{ij} 為 1 表示答對，為 0 表示答錯。此機率模式屬間斷模式，為嵌入式機率模式 (embedded pattern probabilistic model)，以受試者的能力所及 ($T_i \geq j$) 和力有未逮 ($T_i < j$) 作劃分，內含兩種作答策略。如此劃分之因在於單一方程式不足以詮釋受試者的應試策略，分別是面對難度在能力內的試題，勢在必得的攻勢；超乎能力時，減少失分的守勢。再者，考量作答過程中存在能力內答錯之情況，進一步細分為受試者主導 (examinees dominate) 與試題主導 (items dominate) 兩種作答策略。

第(3)式第一條方程式為能力所及時的作答機率公式，其中 $[P_i \times (1 - P_j)]^{u_{ij}}$ 為受試者主導作答， $[(1 - P_i) \times P_j]^{(1-u_{ij})}$ 表示試題主導作答。若 u_{ij} 為 1，會由受試者主導作答機率的產生；若為 0，作答機率受試題影響，改由後者求得作答機率值。第(3)式第二條方程式則是詮釋受試者力有未逮時之情境，和第一條方程式為互補關係，同樣地，作答機率會受受試者主導與試題主導影響。

接著，以表 2 的受試者 4 號回答試題 Q3 與 Q5 為例，說明作答機率值的計算。將受試者 4 號的各項 Person-fit BW 指標值 ($W_4 = .08$, $B_4 = .02$, $C_4 = .21$, $M_4 = .03$) 代入第(1)式，可得受試者 4 號的整體能力 (P_4) 為 .19。將試題 Q3 和試題 Q5 的各項 Item-fit BW 指標值 ($w_3 = .04$, $b_3 = .06$, $c_3 = .02$, $m_3 = .34$; $w_5 = .10$, $b_5 = .04$, $c_5 = .40$, $m_5 = .00$) 代入第(2)式，可得 P_{Q3} 為 .01， P_{Q5} 為 .36。因為試題 Q3 在受試者能力內，由第(3)式第一條方程式可得受試者正確答對試題 Q3 的機率為 .66。試題 Q5 理論上已非受試者能力所及，機率公式改為第(3)式第二條方程式，求得作答機率值為 .26。

表 2 BW 指標說明 S-P 表

座號 \ 試題	Q3	Q9	Q10	Q8	Q6	Q7	Q1	Q4	Q2	Q5	T	t _i
5	1	勝券在握	1	1	1	1	1	1	1	0	9	.9
4	1	勝券在握	1	0	1	0	1	1	1	軒 1 不分	7	.7
10	1	勝券在握	1	1	0	1	0	0	軒 1 不分 (I)	1	7	.7
9	1 (II)	0	1	1	1	0	0	1	0	0	5	.5
1	0	1	1	0	1	1	1	0	0	0	5	.5
2	1	1	1	1	0	0	0	0	0	0	4	.4
7	0	1	1	0	0	1	1	0	蜀道難登	0	4	.4
8	0	天 1 勝負定負	0	0	1	0	0	0	0	0	2	.2
6	1 (III)	0	0	0	0	0	0	0	0 (IV)	1	2	.2
3	1	0	0	0	1	0	0	0	0	0	2	.2
Q	3	3	4	5	5	6	6	7	7	8		
q _j	.3	.3	.4	.5	.5	.5	.6	.7	.7	.8		

註：—S 曲線；---P 曲線。

另外，為能清楚區別表 2 中 S-P 曲線所劃分出的四區特色，本研究根據 Sato (1980) 的設計為四區做定義，勝券在握 (II) 代表受試者在能力內且答對試題，勝負天定 (III) 代表受試者仍處在能力內，但被試題考倒。這兩區若是採 BWEM 計算作答機率，使用的公式為第(3)式「能力所及」的部分，而且勝券在握 (II) 偏向由受試者主導作答機率產生。不分軒輊 (I) 表示受試者處於能力外但答對試題，蜀道難登 (IV) 的受試者在能力外，且被試題考倒。這兩區若是採 BWEM 計算作答機率，使用的公式為第(3)式「力有未逮」的部分，而且蜀道難登 (IV) 偏向由試題主導作答機率產生。進一步根據各區的作答屬性，將不分軒輊 (I) 與勝負天定 (III) 歸為模糊作答區，勝券在握 (II) 與蜀道難登 (IV) 則同屬明確作答區。由受試者 4 號回答試題 Q3 與 Q5 的結果並配合表 2 的分區定義，可知受試者處於勝券在握 (II) 時，作答正確的機率較高；處於不分軒輊 (I) 時，作答正確的機率會下降。

三、BW 本位 Rasch 模式 (BW based Rasch Model, BWRM)

在二元計分下，黃財尉、盧志明 (2013) 依據 Rasch 模式的特性設計具 Rasch 特質的 BW 本位 Rasch 模式 (BWRM)，用以預測作答機率。BWRM

的數學公式如下：

$$P_{ij}^{RM} = \frac{e^{10(\theta - b_j)}}{1 + e^{10(\theta - b_j)}} \quad (4)$$

P_{ij}^{RM} 為作答機率模式為 BWRM 時，第 i 位受試者對第 j 題作出正確反應的機率。 θ 值與 b_j 值則是代入第(1)式與第(2)式所求得之能力值和強度值。同樣分別以受試者 4 號回答最簡單的試題 Q3，與回答最難的試題 Q5 為例，改用 BWRM 時，受試者 4 號答對 Q3 的機率為 .85，答對 Q5 的機率為 .15。

由 BWEM 和 BWRM 的公式可知，BWEM 在計算機率值的過程中，雖然和 BWRM 一樣皆考量能力和難度，但是 BWEM 運用實際作答結果，且將預測作答機率的過程細分為受試者主導和試題主導，此舉背後的意涵迥異於 BWRM 所表達的概念。BWEM 除了關心作答前與作答中的不完全與不完美訊息，例如：粗心、迷思、試題具干擾性或提示性等特質，更將實際反應會影響作答機率的產生列入考慮，使得 BWEM 求得的機率值具有「實然」概念，即考量實際作答結果而得到的實際機率值，而 BWRM 的預期機率值則屬「應然」概念，是單純只考量受試者能力與試題強度而得到的機率值。直覺上，實然機率值包含更多訊息，若此值判斷受試者應答對但實際卻答錯，理應可推論受試者的真正實力並未反應在最終作答結果上，因而存在作答異常。由此看來，似乎 BWEM 比 BWRM 更具應用價值，但果真如此嗎？BWRM 計算機率時雖未再次考量實際作答結果，公式卻較為簡潔。就模式選擇的簡潔原則來看，理應選擇簡潔又能解釋資料的模式（Kang & Cohen, 2007）。BWRM 的複雜度較低，會不會較具應用優勢呢？此點顯示 BWEM 與 BWRM 的比較有其必要。

參、研究方法

目前已知蒙地卡羅法能在短時間內依研究需要產生大量資料供分析使用，所以在數學、物理乃至於社會科學領域皆有許多應用（黃財尉，2008；鄭文吉，2013；盧志明等人，2007；Huang, 2011）。為了在討論 BWEM 與 BWRM 的適用環境時，不受其他因素，如受試者所處地區、試題學科等變項，可能帶來的影響，研究者選擇蒙地卡羅模擬法（Monte Carlo Simulation）產生分析用資料。以下說明研究設計、研究資料及資料分析方式。

一、研究設計

本研究設定受試者能力和試題難度皆為常態分配，在四種受試者人數（30 人、40 人、50 人、200 人）、三種測驗長度（20 題、30 題、40 題），共 12 種情境下，利用 WinGen 3（Han, 2007; Han & Hambleton, 2007），每種情境重複產生 20 組用於計算命中率的受試者作答反應矩陣，並運用 WBstar 1.0 求得不同情境下的作答反應矩陣運用 BWEM 與 BWRM 時的命中率，供分析模式預測能力之使用。

二、研究資料

命中率的計算運用邏輯斯迴歸的分析概念，設定受試者答對該題的機率大於或等於 .5，視為答對該題；低於 .5，視為答錯該題。然後，與實際作答資料比對，機率值大於 .5 且實際值為 1，表示此模式命中受試者的作答反應；若預測受試者答錯該題，即機率值小於 .5 且實際值為 0，亦表示命中受試者的作答反應。最後，統計命中的題數占所有作答資料筆數的比例，即為命中率。本研究在四種受試者人數、三種測驗長度及兩種作答機率模式的設定下，每種情境又各有 20 筆命中率，故共有 480 筆命中率。

三、資料分析

本研究的作答機率模式為相依樣本，受試者人數與測驗長度為獨立樣本，屬混合設計（mixed design），透過三因子混合設計變異數分析探討不同受試者人數、測驗長度及作答機率模式對命中率的影響。由於吳明隆（2010）建議社會科學研究不應只著重統計上是否顯著，而忽略研究結果實際上的重要性。因此，本研究參考王國川（2002）的建議以及盧志明等人（2007）的作法，以淨 η^2 值評估自變項對依變項的解釋強度，並根據 Cohen（1988）的標準選擇淨 η^2 值大於 .138，與命中率具有強度關聯的交互作用為分析重點。接著，依序說明整體與各區命中率在不同設定下的表現。另外，如需事後比較，則選擇 Bonferroni 法校正。

肆、研究結果

以下先針對交互作用的強度檢視，之後再針對整體及各作答區的命中率說明如下。

一、交互作用的強度檢視

本小節的重點在檢視因子間的交互作用，故僅呈現整體、勝券在握（I）、勝負天定（II）、不分軒輊（III）、蜀道難登（IV）五種三因子以及二因子交互作用的結果，如表 3 所示。

表 3 三因子混合設計變異數分析交互作用摘要表

變異來源	命中率	df	MS	F	淨 η^2
N×TL	整體	6	2.64E-04	1.067	.027
	不分軒輊（I）	6	.008	4.047***	.096
	勝券在握（II）	6	.002	2.269*	.056
	勝負天定（III）	6	.007	1.252	.032
	蜀道難登（IV）	6	.003	3.123**	.076
N×PM	整體	3	.005	39.193***	.340
	不分軒輊（I）	3	.017	10.825***	.125
	勝券在握（II）	3	.010	22.466***	.228
	勝負天定（III）	3	.181	36.027***	.322
	蜀道難登（IV）	3	.009	26.275***	.257
TL×PM	整體	2	.005	36.759***	.244
	不分軒輊（I）	2	.074	47.052***	.292
	勝券在握（II）	2	.021	46.069***	.288
	勝負天定（III）	2	.079	15.778***	.122
	蜀道難登（IV）	2	.019	54.949***	.325
N×TL×PM	整體	6	3.73E-04	3.027**	.074
	不分軒輊（I）	6	.002	1.504	.038
	勝券在握（II）	6	.002	5.431***	.125
	勝負天定（III）	6	.012	2.286*	.057
	蜀道難登（IV）	6	.001	1.516	.038

註：N：受試者人數；TL：測驗長度；PM：作答機率模式。

* $p < .05$. ** $p < .01$. *** $p < .001$.

（一）三因子交互作用

表 3 顯示，整體、勝券在握（II）與勝負天定（III）的受試者人數、測驗長度與作答機率模式三者間存在交互作用，即受試者人數、測驗長度及作答機率模式各自對命中率的影響會受到其他兩項因子影響，其中以勝券在握（II）的淨 η^2 值（ $\eta^2=.125$ ）為最高，但仍未達 .138 的強度標準。

（二）二因子交互作用

表 3 指出，受試者人數（N）與測驗長度（TL）二因子交互作用只在不分軒輊（I）、勝券在握（II）以及蜀道難登（IV）呈現顯著，但淨 η^2 值最高只達 .096。至於受試者人數與作答機率模式（N×PM），以及測驗長度與作答機率模式（TL×PM）的二因子交互作用在整體與分區皆為顯著，淨 η^2 值最低為 .122，次低為 .125，其餘之值皆在 .138 之上（.228～.340）。因此，本研究由二因子交互作用切入，探討命中率在不同變項的差異。

已知無論是整體或分區，受試者人數與測驗長度（N×TL）的交互作用對命中率的解釋程度低，看似可以直接檢視這兩項變數的主要效果。然而，由其他二因子交互作用的結果可知，命中率在不同受試者人數和測驗長度的變化會受作答機率模式之影響，所以得進行單純主要效果考驗，並探討命中率在不同受試者人數、測驗長度及作答機率模式的差異。

二、整體命中率

已知受試者人數與作答機率模式（N×PM），以及測驗長度與作答機率模式（TL×PM）的二因子交互作用不僅顯著，且與整體命中率間屬強度關聯，故由單純主要效果考驗著手討論整體命中率在不同受試者人數、測驗長度及作答機率模式的差異。

（一）受試者人數×作答機率模式（N×PM）

表 4 為進行受試者人數與作答機率模式（ $F_{(3, 228)}=39.193$ ， $p<.001$ ，淨 $\eta^2=.340$ ）的單純主要效果考驗之平均數及標準差。

表 4 整體命中率在不同受試者人數與作答機率模式的平均數與標準差

相依因子	獨立因子	受試者人數			
		30 人	40 人	50 人	200 人
作答機率模式	BWEM	.796 (.016)	.801 (.015)	.787 (.015)	.791 (.012)
	BWRM	.747 (.021)	.731 (.023)	.747 (.023)	.731 (.018)

註：括號內的值為標準差。

表 4 顯示，整體命中率平均數介於 .731 至 .801 之間，標準差介於 .012 至 .023 之間。單純主要效果考驗的統計結果，如表 5 所示。

表 5 整體受試者人數與作答機率模式的單純主要效果考驗

變異來源	df	MS	F	事後比較
作答機率模式				
在受試者人數 30 人	1	.073	594.429***	BWEM>BWRM
在受試者人數 40 人	1	.147	1197.000***	BWEM>BWRM
在受試者人數 50 人	1	.050	407.143***	BWEM>BWRM
在受試者人數 200 人	1	.106	863.143***	BWEM>BWRM
作答機率模式×群內受試	228	1.23E-04		
受試者人數				
在 BWEM	3	.002	12.667***	30 人>50 人 40 人>50 人 40 人>200 人
在 BWRM	3	.005	25.333***	30 人>40 人 30 人>200 人 50 人>40 人 50 人>200 人
細格內誤差	456	1.84E-04		

*** $p < .001$.

表 5 顯示，在受試者人數 30 人、40 人、50 人與 200 人時，型一錯誤率採族系錯誤率 ($\alpha_{FW} = .05/6 = .0083$)，命中率在不同作答機率模式皆達到顯著差異，且 BWEM 的整體命中率皆顯著高於 BWRM。

在作答機率模式為 BWEM 時，整體命中率在不同受試者人數之間達顯著差異，事後比較發現 40 人的命中率 ($M = .801$) 高於 50 人的命中率 ($M = .787$)。在作答機率模式為 BWRM 時，則出現相反結果。但整體趨勢顯示大樣本人數 ($N = 200$) 時命中率最低。

(二) 測驗長度×作答機率模式 ($TL \times PM$)

表 6 為進行測驗長度與作答機率模式 ($F_{(2, 228)} = 36.759$, $p < .001$, 淨 $\eta^2 = .244$) 的單純主要效果考驗之所需資料。

表 6 整體命中率在不同測驗長度與作答機率模式的平均值與標準差

相依因子	獨立因子	受試者人數		
		20 題	30 題	40 題
作答機率模式	BWEM	.803	.786	.792
		(.015)	(.015)	(.013)
	BWRM	.760	.726	.730
		(.017)	(.016)	(.016)

註：括號內的值為標準差。

表 6 顯示，整體命中率平均值介於 .726 至 .803 之間，標準差介於 .013 至 .017 之間。單純主要效果考驗的統計結果，如表 7 所示。

表 7 整體測驗長度與作答機率模式的單純主要效果考驗

變異來源	<i>df</i>	<i>MS</i>	<i>F</i>	事後比較
作答機率模式				
在測驗長度 20 題	1	.073	594.492***	BWEM>BWRM
在測驗長度 30 題	1	.141	1148.143***	BWEM>BWRM
在測驗長度 40 題	1	.156	1270.286***	BWEM>BWRM
作答機率模式×群內受試	228	1.23E-04		
測驗長度				
在 BWEM	2	.006	32.571***	20 題>30 題 20 題>40 題 40 題>30 題
在 BWRM	2	.028	152.000***	20 題>30 題 20 題>40 題
細格內誤差	456	1.84E-04		

*** $p < .001$.

表 7 顯示，在測驗長度 20 題、30 題與 40 題時，型一錯誤率採族系錯誤率 ($\alpha_{FW} = .05/5 = .01$)，命中率在不同作答機率模式間皆達顯著差異，且 BWEM 的表現皆優於 BWRM。

在作答機率模式為 BWEM 時，整體命中率在不同測驗長度間存在顯著差異，事後比較發現 20 題的命中率 ($M=.803$) 最高。在作答機率模式為 BWRM 時，命中率在不同測驗長度間也存在顯著差異，事後比較也是 20 題的命中率 ($M=.760$) 為最高。

三、不分軒輊區命中率

不分軒輊 (I) 中受試者人數與作答機率模式 ($N \times PM$) 的二因子交互作用 ($F_{(3, 228)} = 10.825, p < .001$, 淨 $\eta^2 = .125$)，以及測驗長度與作答機率模式 ($TL \times PM$) 的二因子交互作用 ($F_{(2, 228)} = 47.052, p < .001$, 淨 $\eta^2 = .292$) 皆達顯著，不過只有後者高於 .138。因此，將單獨檢視受試者人數的主要效果，並由測驗長度與作答機率模式的單純主要效果考驗切入，檢視此區的命中率在不同測驗長度及作答機率模式間是否存在差異。

(一) 受試者人數 (N)

受試者人數的主要效果雖顯著 ($F_{(3, 228)} = 10.899, p < .001$)，但淨 η^2 值為 .125，未達強度關聯的標準 ($> .138$)，效果量不夠高，命中率的差異探討參考價值較低。

(二) 測驗長度 \times 作答機率模式 ($TL \times PM$)

表 8 為進行測驗長度與作答機率模式 ($F_{(2, 228)} = 47.052, p < .001$, 淨 $\eta^2 = .292$) 的單純主要效果考驗之所需資料。

表 8 不分軒輊命中率在不同測驗長度與作答機率模式的平均值與標準差

相依因子	獨立因子	測驗長度		
		20 題	30 題	40 題
作答機率模式	BWEM	.547 (.057)	.508 (.040)	.483 (.039)
	BWRM	.500 (.058)	.520 (.036)	.521 (.034)

註：括號內的值為標準差。

表 8 顯示，不分軒輊 (I) 區的命中率平均值介於 .483 至 .547 之間，標準差介於 .034 至 .058 之間。單純主要效果考驗統計結果，如表 9 所示。

表 9 不分軒輊區測驗長度與作答機率模式的單純主要效果考驗

變異來源	<i>df</i>	<i>MS</i>	<i>F</i>	事後比較
作答機率模式				
在測驗長度 20 題	1	.087	55.100***	BWEM>BWRM
在測驗長度 30 題	1	.005	3.167	
在測驗長度 40 題	1	.056	35.467***	BWRM>BWEM
作答機率模式×群內受試	228	.002		
測驗長度				
在 BWEM	2	.083	47.848***	20 題>30 題 20 題>40 題 30 題>40 題
在 BWRM	2	.010	5.765**	30 題>20 題 40 題>20 題
細格內誤差	456	.002		

** $p < .01$. *** $p < .001$.

表 9 顯示，不分軒輊(I)區的类型一錯誤率採族系錯誤率($\alpha_{FW} = .05/5 = .01$)，測驗長度為 20 題與 40 題時，命中率在不同作答機率模式間存在顯著差異，測驗長度 20 題時，BWEM ($M = .547$) 的命中率顯著高於 BWRM ($M = .500$)；40 題時則呈現相反結果，BWRM ($M = .521$) 的命中率高於 BWEM ($M = .483$)。

在作答機率模式為 BWEM 時，命中率在不同測驗長度的差異達到顯著，事後比較顯示測驗長度 20 題的命中率最高 ($M = .547$)。然而，當作答機率模式改為 BWRM，命中率在不同測驗長度的差異也達到顯著，事後比較結果卻是測驗長度 40 題的命中率最高 ($M = .521$)。

四、勝券在握區命中率

由於受試者人數與作答機率模式 ($N \times PM$)，以及測驗長度與作答機率模式 ($TL \times PM$) 的二因子交互作用在勝券在握(II)區不僅顯著且與命中率有高度關聯，所以進行單純主要效果考驗，檢視此區的命中率在不同受試者人數、測驗長度，以及作答機率模式間是否存在差異。

(一) 受試者人數×作答機率模式 ($N \times PM$)

表 10 為進行受試者人數與作答機率模式 ($F_{(3, 228)} = 22.466$, $p < .001$, 淨 $\eta^2 = .228$) 的單純主要效果考驗之所需資料。

表 10 勝券在握命中率在不同受試者人數與作答機率模式的平均值與標準差

相依因子	獨立因子	受試者人數			
		30 人	40 人	50 人	200 人
作答機率模式	BWEM	.857 (.038)	.852 (.054)	.799 (.037)	.813 (.040)
	BWRM	.788 (.025)	.763 (.023)	.754 (.040)	.751 (.020)

註：括號內的值為標準差。

表 10 顯示，勝券在握（II）區的命中率平均值介於 .751 至 .857 之間，標準差介於 .020 至 .054 之間。受試者人數和作答機率模式進行單純主要效果考驗統計結果，如表 11 所示。

表 11 勝券在握區受試者人數與作答機率模式的單純主要效果考驗

變異來源	df	MS	F	事後比較
作答機率模式				
在受試者人數 30 人	1	.142	317.412***	BWEM>BWRM
在受試者人數 40 人	1	.239	534.235***	BWEM>BWRM
在受試者人數 50 人	1	.060	134.118***	BWEM>BWRM
在受試者人數 200 人	1	.117	261.529***	BWEM>BWRM
作答機率模式×群內受試	228	4.47E-04		
受試者人數				
在 BWEM	3	.048	69.747***	30 人>50 人 30 人>200 人 40 人>50 人 40 人>200 人
在 BWRM	3	.016	23.570***	30 人>40 人 30 人>50 人 30 人>200 人
細格內誤差	456	6.93E-04		

*** $p < .001$.

表 11 顯示，在受試者人數 30 人、40 人、50 人與 200 人時，型一錯誤率採族系錯誤率（ $\alpha_{FW} = .05/6 = .0083$ ），命中率在不同作答機率模式間存在顯著差異，BWEM 的命中率皆高於 BWRM 的命中率。

在作答機率模式為 BWEM 時，此區的命中率在不同受試者人數間達到顯著差異，且事後比較指出 30 人（ $M = .857$ ）和 40 人（ $M = .852$ ）的命中率

明顯高於 50 人 ($M=.799$) 和 200 人 ($M=.813$) 的命中率。

在作答機率模式為 BWRM 時，命中率在不同受試者人數間亦達顯著差異。事後比較可知 30 人 ($M=.788$) 的命中率高於其他三種受試者人數的命中率。

(二) 測驗長度 \times 作答機率模式 ($TL \times PM$)

表 12 為用於測驗長度與作答機率模式 ($F_{(2, 228)} = 46.069$, $p < .001$, 淨 $\eta^2 = .288$) 的單純主要效果考驗之所需資料。

表 12 勝券在握命中率在不同測驗長度與作答機率模式的平均值與標準差

相依因子	獨立因子	測驗長度		
		20 題	30 題	40 題
作答機率模式	BWEM	.805 (.040)	.815 (.042)	.871 (.037)
	BWRM	.764 (.025)	.742 (.032)	.786 (.019)

註：括號內的值為標準差。

表 12 顯示，勝券在握 (II) 區的命中率平均值介於 .742 至 .871 之間，標準差介於 .019 至 .042 之間。測驗長度和作答機率模式的單純主要效果考驗結果，如表 13 所示。

表 13 勝券在握區測驗長度與作答機率模式的單純主要效果考驗

變異來源	df	MS	F	事後比較
作答機率模式				
在測驗長度 20 題	1	.068	152.000***	BWEM>BWRM
在測驗長度 30 題	1	.210	469.412***	BWEM>BWRM
在測驗長度 40 題	1	.291	650.471***	BWEM>BWRM
作答機率模式 \times 群內受試	228	4.47E-04		
測驗長度				
在 BWEM	2	.101	145.747***	40 題>20 題 40 題>30 題
在 BWRM	2	.038	54.114***	20 題>30 題 40 題>20 題 40 題>30 題
細格內誤差	456	6.93E-04		

*** $p < .001$.

表 13 顯示，在測驗長度 20 題、30 題與 40 題時，型一錯誤率採族系錯誤率 ($\alpha_{FW} = .05/5 = .01$)，此區的命中率在不同作答機率模式間存在顯著差異，BWEM 的命中率高於 BWRM。

在作答機率模式為 BWEM 時，命中率在不同測驗長度間達到顯著差異，事後比較指出 40 題 ($M = .871$) 的命中率最高。在作答機率模式為 BWRM 時，命中率在不同測驗長度間達到顯著差異，事後比較也顯示 40 題 ($M = .786$) 的命中率最高。

五、勝負天定區命中率

勝負天定 (III) 只有受試者人數與作答機率模式 ($N \times PM$) 的二因子交互作用達顯著且與命中率間具有強度關聯，所以由受試者人數與作答機率模式的單純主要效果考驗出發，檢視命中率在不同受試者人數與作答機率模式間是否存在差異。至於測驗長度與作答機率模式 ($TL \times PM$) 的二因子交互作用，因為與命中率之間未達強度關聯，故謹單獨檢視測驗長度的主要效果。

(一) 受試者人數 \times 作答機率模式 ($N \times PM$)

進行受試者人數與作答機率模式 ($F_{(3, 228)} = 36.027, p < .001$ ，淨 $\eta^2 = .322$) 的單純主要效果考驗所需資料，如表 14 所示。

表 14 勝負天定命中率在不同受試者人數與作答機率模式的平均值與標準差

相依因子	獨立因子	受試者人數			
		30 人	40 人	50 人	200 人
作答機率模式	BWEM	.488 (.114)	.515 (.109)	.342 (.094)	.430 (.066)
	BWRM	.553 (.065)	.576 (.048)	.567 (.044)	.576 (.027)

註：括號內的值為標準差。

表 14 顯示，勝負天定 (III) 區的命中率平均值介於 .342 至 .576 之間，標準差介於 .027 至 .114 之間。受試者人數和作答機率模式的單純主要效果考驗結果，如表 15 所示。

表 15 勝負天定區受試者人數與作答機率模式的單純主要效果考驗

變異來源	<i>df</i>	<i>MS</i>	<i>F</i>	事後比較
作答機率模式				
在受試者人數 30 人	1	.125	24.826***	BWRM>BWEM
在受試者人數 40 人	1	.113	22.443***	BWRM>BWEM
在受試者人數 50 人	1	1.521	302.080***	BWRM>BWEM
在受試者人數 200 人	1	.638	126.711***	BWRM>BWEM
作答機率模式×群內受試	228	.005		
受試者人數				
在 BWEM	3	.351	66.295***	30 人>50 人 30 人>200 人 40 人>50 人 40 人>200 人 200 人>50 人
在 BWRM	3	.007	1.323	30 人>40 人 30 人>50 人 30 人>200 人
細格內誤差	456	.005		

*** $p < .001$.

表 15 顯示，在受試者人數 30 人、40 人、50 人與 200 人時，型一錯誤率採族系錯誤率（ $\alpha_{FW} = .05/6 = .0083$ ），命中率在不同作答機率模式間的差異皆達到顯著，BWRM 的命中率皆高於 BWEM 的命中率。

在作答機率模式為 BWEM 時，命中率在不同受試者人數間存在顯著差異，事後比較發現，在不考慮測驗長度下，BWEM 在受試者人數 50 人（ $M = .342$ ）的命中率为最低。在作答機率模式為 BWRM 時，命中率在不同受試者人數間的差異未達顯著，意謂在勝負天定（III），BWRM 的命中率高和低和受試者人數多少無關。

（二）測驗長度（*TL*）

勝負天定（III）區的三因子混合設計變異數分析結果顯示，測驗長度的主要效果雖顯著（ $F_{(2,228)} = 9.095, p < .001$ ），但是淨 η^2 值為 .074，未達強度關聯（ $> .138$ ），效果量較不高，參考價值較低。

六、蜀道難登區命中率

蜀道難登（IV）區的受試者人數與作答機率模式（ $N \times PM$ ），以及測驗長度與作答機率模式（ $TL \times PM$ ）的二因子交互作用不僅顯著，且與命中率有強度關聯。故透過單純主要效果考驗，檢視此區的命中率在不同受試者人數、測驗長度及作答機率模式的差異。

（一）受試者人數 \times 作答機率模式（ $N \times PM$ ）

表 16 為進行受試者人數與作答機率模式（ $F_{(3,228)} = 26.275$ ， $p < .001$ ，淨 $\eta^2 = .257$ ）的單純主要效果考驗之所需資料。

表 16 蜀道難登命中率在不同受試者人數與作答機率模式的平均值與標準差

相依因子	獨立因子	受試者人數			
		30 人	40 人	50 人	200 人
作答機率模式	BWEM	.852 (.044)	.862 (.048)	.908 (.029)	.895 (.031)
	BWRM	.782 (.041)	.777 (.039)	.810 (.027)	.786 (.033)

註：括號內的值為標準差。

表 16 顯示，蜀道難登（IV）區的命中率平均值介於 .777 至 .908 之間，標準差介於 .027 至 .048 之間。受試者人數與作答機率模式的單純主要效果考驗統計結果，如表 17 所示。

表 17 蜀道難登區受試者人數與作答機率模式的單純主要效果考驗

變異來源	<i>df</i>	<i>MS</i>	<i>F</i>	事後比較
作答機率模式				
在受試者人數 30 人	1	.146	426.769***	BWEM>BWRM
在受試者人數 40 人	1	.215	628.462***	BWEM>BWRM
在受試者人數 50 人	1	.289	844.769***	BWEM>BWRM
在受試者人數 200 人	1	.361	1055.231***	BWEM>BWRM
作答機率模式×群內受試	228	3.42E-04		
受試者人數				
在 BWEM	3	.043	66.630***	50 人>30 人 50 人>40 人 200 人>30 人 200 人>40 人
在 BWRM	3	.013	19.781***	50 人>30 人 50 人>40 人 50 人>200 人
細格內誤差	456	.001		

*** $p < .001$.

表 17 顯示，在受試者人數 30 人、40 人、50 人與 200 人時，型一錯誤率採族系錯誤率（ $\alpha_{FW} = .05/6 = .0083$ ），命中率在不同作答機率模式皆達顯著，BWEM 的命中率皆高於 BWRM 的命中率。

在作答機率模式為 BWRM 時，命中率在不同受試者人數的差異達顯著，事後比較發現受試者人數 50 人（ $M = .908$ ）的命中率高於 30 人（ $M = .852$ ）和 40 人（ $M = .862$ ）的命中率，200 人（ $M = .898$ ）的命中率也高於 30 人和 40 人的命中率。在作答機率模式為 BWEM 時，命中率在不同受試者人數間呈顯著差異，事後比較指出 50 人（ $M = .810$ ）的命中率最高。

（二）測驗長度×作答機率模式（ $TL \times PM$ ）

表 18 為進行測驗長度與作答機率模式（ $F_{(2, 228)} = 54.949$ ， $p < .001$ ，淨 $\eta^2 = .325$ ）的單純主要效果考驗之所需資料。

表 18 蜀道難登命中率在不同測驗長度與作答機率模式的平均值與標準差

相依因子	獨立因子	測驗長度		
		20 題	30 題	40 題
作答機率模式	BWEM	.892 (.031)	.899 (.041)	.847 (.044)
	BWRM	.825 (.024)	.789 (.021)	.752 (.023)

註：括號內的值為標準差。

表 18 顯示，蜀道難登（IV）區的命中率平均值介於 .752 至 .899 之間，標準差介於 .021 至 .044 之間。測驗長度和作答機率模式的單純主要效果考驗結果，如表 19 所示。

表 19 蜀道難登區測驗長度與作答機率模式的單純主要效果考驗

變異來源	df	MS	F	事後比較
作答機率模式				
在測驗長度 20 題	1	.180	526.154***	BWEM>BWRM
在測驗長度 30 題	1	.482	1408.923***	BWEM>BWRM
在測驗長度 40 題	1	.359	1049.385***	BWEM>BWRM
作答機率模式×群內受試	228	3.42E-04		
測驗長度				
在 BWEM	2	.063	98.786***	20 題>40 題 30 題>40 題
在 BWRM	2	.105	163.973***	20 題>30 題 20 題>40 題 30 題>40 題
細格內誤差	456	.001		

*** $p < .001$.

表 19 顯示，在測驗長度 20 題、30 題與 40 題時，型一錯誤率採族系錯誤率（ $\alpha_{FW} = .05/5 = .01$ ），命中率在不同作答機率模式間達到顯著，BWEM 的命中率皆高於 BWRM 的命中率。

在作答機率模式為 BWEM 時，命中率在不同測驗長度的差異達到顯著，事後比較顯示，不考量受試者人數下，40 題（ $M = .847$ ）的命中率最低。在作答機率模式為 BWRM 時，命中率在不同測驗長度的差異亦達到顯著，事後比較亦顯示 40 題（ $M = .752$ ）的命中率为最低。

伍、結論與建議

整理研究結果產生的結論如下。

一、無論受試者人數多寡、測驗長度長短，BWEM 在整體與明顯作答區的命中率皆高於 BWRM

無論是 30 人、40 人、50 人或 200 人，BWEM 在整體、勝券在握區與蜀道難登區的命中率皆顯著高於 BWRM；而題數方面，無論是 20 題、30 題或 40 題，BWEM 在整體、勝券在握區與蜀道難登區的命中率皆顯著優於 BWRM，且二模式命中率皆在七成以上。

二、整體作答上，二模式於小樣本人數時呈現較佳的命中率，但分區不盡一致

BWEM 模式在整體和勝券在握區，皆出現 30 人的命中率顯著高於 50 人，40 人的命中率顯著高於 50 人與 200 人；在勝負天定區，30 人的命中率顯著高於 50 人與 200 人，40 人的命中率顯著高於 50 人與 200 人，且 200 人的命中率顯著高於 50 人；在蜀道難登區，50 人的命中率顯著高於 30 人與 40 人，200 人的命中率顯著高於 30 人和 40 人。

而 BWRM 模式在整體方面，30 人的命中率顯著高於 40 人與 200 人，50 人的命中率顯著高於 40 人與 200 人；在勝券在握區，30 人的命中率最高；在蜀道難登區，50 人的命中率最高。

三、整體作答上，二模式於短題數情境時呈現較佳的命中率，但分區不同

BWEM 模式在整體和不分軒輊區，皆以 20 題的命中率最高；在勝券在握區，40 題的命中率最高；在蜀道難登區，則是 40 題的命中率為最低。

BWRM 模式在整體方面，20 題的命中率最高；在不分軒輊區，則是 20 題的命中率最低；在勝券在握區，40 題的命中率最高；在蜀道難登區，20 題的命中率最高。

雖然目前仍缺少討論 BWEM 與 BWRM 精確度的相關文獻，但本研究結果和其他討論模式預測能力或異常指標偵測力的文獻一樣，皆指出不同受試者人數或測驗長度可能會影響命中的精確度。因此，根據上述結論，本研究在實務應用上的建議如下。

一、作答機率模式的選擇可視關注的作答反應區而定

已知高命中率表示該模式猜中的作答反應愈多，隱含該作答機率模式所提供的作答機率值貼近實際作答結果，故選擇預測能力高的作答機率模式較有助於教師了解學生實際的答對機率，供分組補救教學或下次測驗前能力評估的參考。然而，模擬結果顯示 BWEM 和 BWRM 各自有其表現較好的作答反應區域，所以在評估學生作答機率時，建議先利用 S-P 表將作答反應如本研究的作法分成四區，不同區域選擇預測能力較高的作答機率模式評估學生，讓評估結果能更貼近學生的實際能力。

二、無論選用何種作答機率模式，皆需留意受試者人數在不同作答反應區的表現

面對 BWEM，受試者人數一旦超過 40 人，整體預測能力會下降，但分區的命中率可能變低，也可能提高。也就是說，某些區域的受試者作答反應將無法有效評估，不利於了解受試者真實的作答能力。若改用 BWRM 評估受試者，受試人數超過 40 人可以有較佳的預測能力。然而，與 BWEM 一樣，並不是所有區域都一面倒向支持特定的受試者人數。

由整體命中率來看，BWEM 和 BWRM 皆可用於小班教學，但如果關注的是特定區域的受試者，那得留意受試者人數所帶來的分析限制，作答機率模式所得的機率值可能猜錯受試者的作答反應，而不足以詮釋受試者的真正能力。

三、無論選用何種作答機率模式，皆需留意測驗長度在不同作答反應區的表現

由整體命中率來看，BWEM 和 BWRM 皆出現 20 題的預測能力最佳。就形成性評估而言，此測驗長度既不易加重老師的閱卷負荷，也不會使學生面臨考試範圍過大的準備壓力。實務現場的教師無論是採何種作答機率模式評估學生的學習成果，皆可考慮將測驗長度設為 20 題。

然而，若是分區來看，作答機率模式在 20 題的預測能力有時並非最佳。一旦教師更關心的是特定區域的受試者，可能得增加測驗題數。假若測驗題數不宜增加，就得注意此作答機率模式所提供的機率值，可能因為題數不足，而無法貼近受試者的真實作答反應。

本研究透過分析模擬的作答反應檢視 BWEM 與 BWRM 的應用時機，初步證實兩種作答模式的預測力的確存在差異，但是模擬資料的優點同時也是缺點。為了提高模式的應用性，未來研究方向有以下三點建議。

一、改變受試者能力與試題難度的分配

實務現場的受試者能力和試題難度並不一定是常態分配，為了豐富 BWEM 和 BWRM 的討論，建議未來可以討論作答機率模式在不同受試者能力和試題難度分配的表現。除了提高作答機率模式的價值，亦協助使用者在面對作答機率模式的選擇時，能有更多的判斷依據。

二、模式的修正

目前，BWEM 在計算作答機率時，利用實際作答結果作為判斷是受試者主導或試題主導的依據，產生作答機率。考量作答機率可能同時受到受試者和試題影響，而不是像目前的間斷概念，建議可採連續概念，將 BWEM 的計算公式中決定受試者主導或試題主導的參數 u_{ij} 改為介於 0 到 1 之間的常數，也可以讓它是受試者能力與試題難度所形成的函數，或是 S-P 表中分區實際答對的細格數占該區細格數的比例。此法不只可以將作答機率公式變為連續概念，還可考慮不同作答反應區有不同的參數值。透過權重的給予，找出預測能力更佳的模式。

三、以實際資料評估模式

由於本研究為排除其他變數的影響，先以模擬資料進行作答機率模式的比較。若資料改為真實作答反應，兩種作答機率模式的預測能力會有何差異，仍需進一步評估。畢竟，模式的可用性若能有實徵資料支持將更具說服力，所以建議未來改以實際資料討論模式的選擇，以提高模式的應用價值。

參考文獻

中文部分

- 王國川（2002）。圖解 SAS 在變異數分析上的應用。臺北市：五南。
- 余民寧（2009）。試題反應理論及其應用。臺北市：心理。
- 余民寧（2011）。教育測驗與評量：成就測驗與教學評量（第三版）。臺北市：心理。
- 吳明隆（2010）。SPSS 操作與應用：變異數分析實務。臺北市：五南。
- 黃財尉（2008）。不同資料結構下異常作答指標決斷值的探討。輔導與諮商學報，30（1），1-16。
- 黃財尉（2011）。內外注意係數異常作答診斷空間的建立與探證。測驗學刊，58（1），1-27。
- 黃財尉、盧志明（2013）。學習診斷分析軟體 1.0 介紹。發表於 2013 測驗年會暨心理教育學術研討會，臺中教育大學。
- 鄭文吉（2013）。漫談蒙地卡羅法的原理及其應用。行政院農業委員會高雄區農業改良場研究彙報，23（1），16-41。
- 盧志明、黃財尉、方將任（2007）。Guttman 型異常作答指標偵測力之比較。測驗學刊，54（1），147-174。

英文部分

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement*, 39(3), 223-238.
- D'Costa, A. (1993, 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.
- Dragow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6(3), 297-308.
- Dragow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with

- polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67-86.
- Gulliksen, H. (1987). *Theory of mental test*. Hillsdale, NJ: Lawrence Erlbaum Associates. (Originally published in 1950 by New York, NY: John Wiley & Sons)
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer · Nijhoff Publishing.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459.
- Han, K. T., & Hambleton, R. K. (2007). *User's manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.
- Harnisch, D. L. (1983). Item response patterns: Application for educational practice. *Journal of Educational Measurement*, 20(2), 191-206.
- 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.
- 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* (Order No. 3049040). Available from ProQuest Dissertations & Theses A&I. (305579942). Retrieved from <http://search.proquest.com/docview/305579942?accountid=10076>
- Huang, T.-W. (2006). Aberrant response diagnoses by the Beyond-Ability-Surprise index B and the Within-Ability-Concern index W. *Proceedings of 2006 Hawaii International Conference on Education* (pp. 2853-2856). Honolulu, Hawaii.
- Huang, T.-W. (2011). Robustness of BW aberrance indices against test length. *Knowledge Management & E-Learning: An International Journal (KM&EL)*, 3(3), 310-318.
- Huang, T.-W. (2012). Aberrance detection powers of the BW and person-fit indices. *Educational Technology & Society*, 15(1), 28-37.
- Huang, T.-W., & Wu, P.-C. (2013). Classroom-based cognitive diagnostic model for a teacher-made fraction-decimal test. *Educational Technology & Society*, 16(3), 347-361.
- Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, 31(4), 331-358.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, 16(4), 277-298.
- Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35(1),

42-56.

- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4(4), 269-290.
- Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement*, 45(4), 851-859.
- Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10(3), 217-229.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3-8.
- Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent development. *Applied Measurement in Education*, 8(3), 261-272.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135.
- Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meiji Toshō. (In Japanese)
- Sato, T. (1980). The S-P chart and the caution index. *NEC Educational Information Bulletin*, 80(1), C&C Systems Research Laboratories, Nippon Electric Co., Ltd., Tokyo, Japan. (In Japanese)
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1(2), 199-218.
- Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4), 331-352.
- St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2011). Accuracy of person-fit statistics: A monte carlo study of the influence of aberrance rates. *Applied Psychological Measurement*, 35(6), 419-432.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, 49(1), 95-110.
- Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detection unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*, 7(1), 81-96.

