國立政治大學應用數學系

碩士學位論文

加入作答時間之試題反應模型在能力上的研究

A Study of the Ability after Incorporating Response Time in the Item Response Model

國立政治大學
National Chengchi University

指導教授：姜志銘 博士

宋傳欽 博士

研究生： 曾定柏 撰 中華

民國 109 年 7 月

DOI:10.6814/NCCU202001433

中文摘要

本研究旨在探討作答時間是否適合作為受試者能力值估計的一項因素。本文從制定一種加入作答時間的評分規則為出發點，建立一個包含作答反應與作答時間的模型，再以最大概似估計法估計能力值與難度值，並透過實際數據之計算結果，分析能力估計值在加入作答時間前與後是否有所不同。最後，探討在作答反應組型相同時，能否以此新模型進一步區分受試者能力值的高低。

透過模擬數據進行模型驗證，在估計受試者能力值時，與 IRT 模式比較，我們發現本文所建立之模型對高能力群之受試者，有較佳的估計精準度，且受試者人數越多，精準度越好。

同時，經過實證分析，本文歸納出以下三項結果：

1. 答錯時，使用本文所建立之評分規則的分數與作答時間呈現正相關。

   在篩選的題目中，無論是簡單、中等或是困難的試題，我們發現答錯之受試者的能力估計值與作答時間均呈現正相關，因此對於作答時間較短的受試者應給予較少的得分(即扣分較多)，而作答時間較長的受試者反而應給予較多的得分(即扣分較少)。

2. 作答時間宜採用伽瑪分布。

   在篩選的試題中，經由適合度檢定，我們發現部分試題的作答時間並不適合指數分布，宜改用較一般化的伽瑪分布。

3. 作答反應組型相同時，以本文所建立之模型能進一步區分受試者的能力值。

   無論是古典測驗理論或是試題反應理論，皆無法從作答反應組型相同的試題中區分受試者的能力值，但採用加入作答時間之試題反應模型後，我們不僅可經由這幾道試題進一步區分受試者的能力值，而且其與依全部試題所估計之能力值呈現高度的秩相關。

關鍵字：試題反應理論、作答時間、評分規則

DOI:10.6814/NCCU202001433

Abstract

This study aims to explore whether the response time is an important factor for estimating the abilities of examinees. After giving a scoring rule, which incorporates both item response and response time, we build a new model and can then estimate the ability of any examinee and the difficulty of any item by using the method of maximum likelihood estimation. Through the real data, we compare examinees' abilities based on the IRT and those based on our new speed-accuracy response model (NSARM). Finally, we explore whether this new model can further distinguish the abilities of two examinees when their response patterns are the same.

Through the simulations, we find that, on ability estimate, our NSARM shows more accurate than IRT model among those examinees with high ability. In addition, it is even more accurate when the number of examinees increases.

After analyzing our real data, we further summarize the following three results:

1. When an item is responded incorrectly, the score based on our new model is positively correlated with the response time.

Among the randomly selected items, no mater they are simple, medium or difficult, we found that the estimated abilities of the examinees who incorrectly answered the items, are positively correlated with their response times. Therefore, an examinee taking shorter response time should be given a lower score (i.e., deduct more additional score), and an examinee taking longer response time should be given a higher score (i.e., deduct less additional score).

2. The gamma distribution is more appropriate for modeling the response time.

Using the goodness of fit test, we found that the exponential distribution, which is used by many authors, is not appropriate to model the response time of some items in our data set. However, we further found that the gamma distribution, which is a generalization of the exponential distribution, is appropriate.

3. Our new model can further distinguish the abilities of examinees.

Neither classical test theory nor item response theory can distinguish the abilities of examinees when their response patterns are the same.

Keywords: IRT, response time, scoring rule

DOI:10.6814/NCCU202001433

目次

第一章　緒論.....4

第一節 研究背景與動機.....4

第二節 研究目的與研究問題.....5

第二章　文獻回顧.....6

第一節 古典測驗理論.....6

第二節 試題反應理論.....8

第三節 作答時間模型.....13

第三章　研究方法.....16

第一節 建立模型.....16

第二節 研究過程.....26

第四章　模型驗證.....31

第五章　實證分析.....42

第一節 符號負指數時間評分規則之探討.....42

第二節 作答反應與作答時間之獨立性.....47

第三節 作答時間符合伽瑪分布.....49

第四節 不同模型下試題難度值相關性之探討.....57

第五節 NSARM 與 IRT 單參數模型在能力值的比較.....61

第六節 作答反應組型相同時受試者能力值之探究.....67

第六章　結論與建議.....71

參考文獻.....74

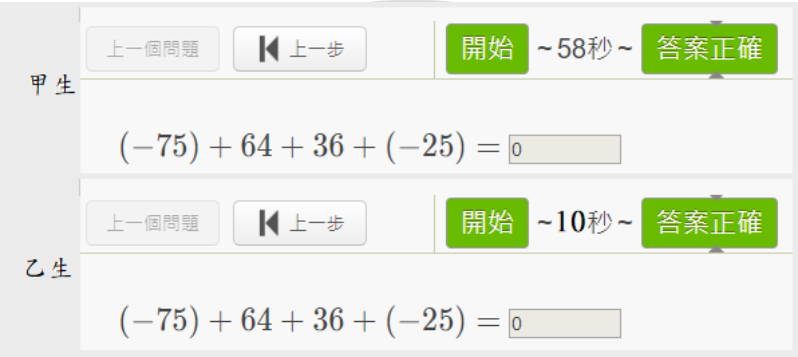
DOI:10.6814/NCCU202001433

第一章　　緒論

本章分成兩個小節，說明研究背景與動機，進而擬定研究目的來進行研究的問題與假設。

第一節 研究背景與動機

本文作者在從事教育相關工作時發現，學生的成績未必能完全反映他在該學科的精熟程度。例如我們在觀察學生的答題情形後，發現甲生與乙生皆答對某一道試題，但兩者的作答時間(response time)差異頗大，如下圖所示：



此題目的數學概念為加法結合律，乙生正確且快速使用此技巧；而甲生雖然回答正確，但從他的作答時間可以發現，其未能發覺並使用此題的數學概念進行解題。因此，將作答時間作為受試者能力估計的一項依據，或許能進一步區分甲、乙兩位受試者能力值的高低。

關於作答時間的研究，從最早 Thurstone（1937）所提出的試題反應曲面（response surface）中，就已表示了作答時間和試題難度（item difficulty）及受試者答對機率之間的關係，後來 Thissen（1983），Roskam（1987），Verhelst, Verstralen and Jansen（1997），Wang and Hanson（2005），Ferrando and Lorenzo-Seva（2007）等，亦陸續將作答時間列為估計受試者能力的重要參數之一。但是因為技術的問題，早期較無法運用電腦設備進行測驗，不易取得受試者在每一道試題的作答時間，而隨著資訊化時代的來臨，在線上測驗的過程中，電腦可以精準的記錄作答時間以利研究，有鑑於此，本文作者決定利用取得的實際資料，進行研究及探討。

DOI:10.6814/NCCU202001433

第二節 研究目的與研究問題

一、 研究目的

本研究採用 200 位國中生的實際資料求得受試者的作答反應（item response），並由試題反應理論（item response theory, IRT）估算受試者的能力值及各試題的難度參數後，與本文所建立之作答時間模型合併分析討論，旨在研究受試者能力以及試題難度的估算，作答時間是否為必要討論的參數之一。再者，受試者是否能妥善運用試題所涵蓋之學習內容進行解題，以更快速的解題技巧給予正確的作答反應，皆是評估受試者能力值的重要依據。綜合以上，研究目的為下列三項：

(一) 探討作答時間的分布

(二) 探討加入作答時間後，各參數的估計結果是否穩定

(三) 探討作答時間是否適合作為受試者能力估計的因素之一

二、 待答問題

針對作答時間，因為試題反應理論的基本假設前題之一為非速度性的測驗，即不限制受試者的作答時間，本研究亦依此原則，於施測的過程中不限制作答時間，給予受試者足夠的時間完成測驗，將受試者對於各試題的難度參數、受試者的作答反應及作答時間依研究目的做以下的問題研究：

(一) 本文模型是否合理？

(二) 對於答錯者，本文所訂定之評分規則是否合理？

(三) 受試者在每一題的作答反應與作答時間是否獨立？

(四) 是否適合將作答時間模型從指數分布修改為更一般化的伽瑪分布？

(五) 在答對（或答錯）之試題相同的情況下，探討本文模型能否進一步區分受試者能力的高低？

DOI:10.6814/NCCU202001433

第二章　　文獻回顧

本研究估算受試者能力值及各試題的參數，並將受試者的試題作答時間合併分析討論。因此，在相關文獻的部分，將針對「古典測驗理論」（classical test theory, CTT）、「試題反應理論」及「作答時間模型」進行文獻探討。

第一節 古典測驗理論

古典測驗理論主要是以整份測驗的觀點，來解釋測驗分數的涵義。它對受試者測驗分數的看法，是以各試題得分加總之後的總分作為代表。

一、　基本假設

古典測驗理論的基本假設為測驗分數等於真實分數加上誤差分數，即

$$X=T+E$$

X：測驗分數（observed score）

T：真實分數（true score）

E：誤差分數（error score）

古典測驗理論根據其基本假說，這些假設大致可以歸納如下（Allen and Yen, 2001； Crocker and Algina, 1986）：

(一) $E(X)=T$（即測驗分數的期望值等於真實分數）

(二) $\rho_{ET}=0$（即測驗分數與誤差分數之間為零相關）

(三) $\rho_{E_1E_2}=0$（即不同測驗的誤差分數之間為零相關）

(四) $\rho_{E_1T_2}=0$（即不同測驗的誤差分數與真實分數之間為零相關）

二、　試題分析的目的

試題分析的目的，主要是在診斷試題的好壞，提供試題計量特性的訊息以及達到改善測驗的品質。而分析的主要方法，有誘答分析（distractor analysis）、試題難度、試題鑑別度（item discrimination）三種，其中試題難度與試題鑑別度定義如下。

DOI:10.6814/NCCU202001433

(一) 試題難度

答對某道試題人數的百分比，也就是試題平均數（item mean）

$$P = \frac{R}{N} \times 100\%$$

其中，P 為試題的答對率，N 為試題的受試人數，R 為試題的答對人數。從這定義來看，所謂試題的難度可視為試題的容易度（item easiness），因為 P 愈大表示試題愈容易，與後續所要介紹的試題反應之難度值剛好為相反的概念，因此，在後續章節便以 $100\% - P$ 作為古典測驗理論之難度值，以方便各模式進行比較。

(二) 試題鑑別度

將受試者依總分高低排序，取極端的 25% 為高低分組，然後求出高分組與低分組在每一個試題的答對率，鑑別度即為高分組答對率減去低分組答對率

三、 優點與缺點

古典測驗理論具有上述優點，但卻因為下列諸項先天的缺失（Guion and Ironson, 1983），而逐漸喪失作為測驗理論的優勢。

(一) 古典測驗理論所採用的指標：

如難度、鑑別度等，都是一種樣本依賴（sample dependent）的指標；也就是說，這些指標會因為接受測驗之受試者樣本的不同而不同。

(二) 古典測驗的可靠性：

在古典測驗中，對可靠性的定義為測驗分數在複本測驗之間的相關性，而這些可靠性估計是具有未知偏差的。

(三) 古典測驗理論以一個相同的測量標準誤：

作為每位受試者的測量誤差指標，這種作法並沒有考慮受試者能力的個別差異情形，對於高、低能力兩極端群的受試者而言，這種指標極為不合理且不準確。

(四) 古典測驗理論忽視受試者的作答反應組型(item response pattern)：

古典測驗理論認為原始得分相同的受試者，其能力必定一樣；但其實不然，即使原始得分相同的受試者，其作答反應組型亦不見得會完全一致。

DOI:10.6814/NCCU202001433

第二節 試題反應理論

為了克服古典測驗理論的缺失，便產生理論與方法均較嚴謹的試題反應理論。試題反應理論的中心概念源自於 Tucker（1946）提出的試題特徵曲線（item characteristic curve, ICC），後來在 Lord（1980）發表了第一本以試題反應理論命名的著作後，測驗理論從此以試題反應理論為代表。因為試題反應理論包含的範圍及內容非常龐大，故以與本文研究目的相關的部份進行下列探討。

一、 基本假設

(一) 測驗的基本假設不同

古典測驗理論是真分數模式，而試題反應理論是建立在局部獨立 (local independence) 與單向度 (unidimensionality) 假設下。

(二) 不同的測驗精確度計量

在古典測驗理論中，測驗信度依賴於試題數；而試題反應理論的信度則是試題的適當性。

(三) 測驗的標準誤不同

在古典測驗理論中，測驗的標準誤是特定受試人群的所有分數計算而得；而試題反應理論中，測驗的標準誤會因為測驗分數的不同而不同，但是可以類推至所有得分相同的受試人群。

(四) 測驗試題的篩選不同

在古典測驗理論中，測驗試題的篩選是通過試題的難度、鑑別度來對試題加以選擇，通常在太困難或太簡單或是鑑別度太低的題目會被剃除；在試題反應理論中，對於試題的篩選則是透過訊息函數獲得。

$$I_i(\theta) = \frac{[P_i{}'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]} \quad i = 1, \ldots, n$$

其中，$I_i(\theta)$ 代表試題 i 在能力值為 θ 時所提供的訊息量，$P_i'(\theta)$ 為 $P_i(\theta)$在 θ 上的一階導數，而 $P_i(\theta)$ 即能力值為 θ 的受試者在試題 i 上答對的機率。

二、 試題特徵曲線

DOI:10.6814/NCCU202001433

試題反應理論 主要是以個別試題的觀點，來解釋測驗分數的涵義。它認為受試者在某一試題上的表現情形，與其背後的某種潛在特質（或能力）之間具有某種關係存在，該關係可以透過一條連續性遞增的數學函數來 加以表示，這個數學函數便稱作「試題特徵曲線」(如圖2-1所示的各條曲線)；換句話說，我們把能力不同的受試者得分點連接起來所構成的曲線，便是能力不同受試者在某一測驗試題上的試題特徵曲線。

在試題反應理論中，每一種試題反應模式都有其相對應的一條試題特徵曲線，此一曲線通常包含一個或多個參數來描述該試題的特性，以及一個或多個參數來描述受試者的潛在特質；因此，根據我們選用的試題反應模式所具有的參數個數的不同，所畫出的試題特徵曲線形狀便不相同。

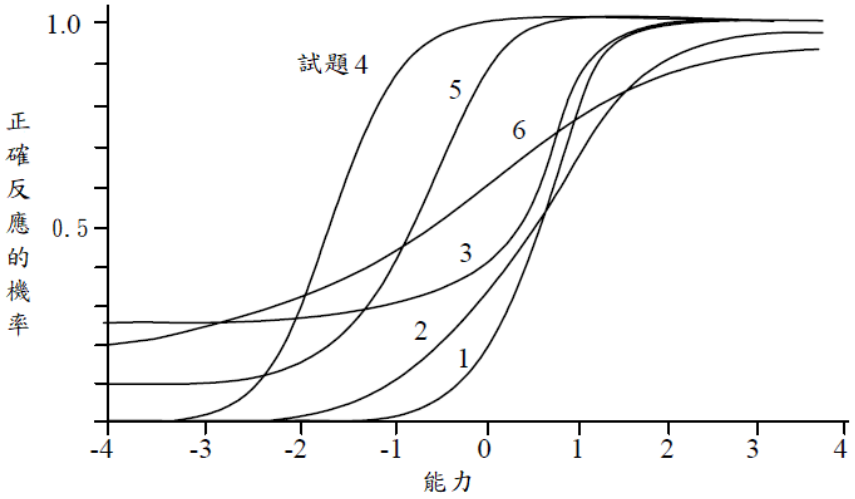


圖2-1 六條典型的三參數試題特徵曲線

資料來源：余民寧 (2009)。試題反應理論 (IRT) 及其應用。

茲舉一個典型的「三參數邏輯斯模式」(three-parameter logistic model, 3PL) 為例，說明學生的能力與試題特徵曲線之間的關係如下：

$$P_i(\theta) = c_i + \frac{(1-b_i)}{1+e^{-a_i(\theta-b_i)}} \quad i = 1, 2, \dots, n$$

表示任何一位受試者答對任何一道試題的機率，可由兩種參數來決定，一為受試者的能力值大小，另一為該試題的試題參數值大小。所以，即可簡化讀成：具有能力值為 θ 的受試者答對第 i 試題的機率，可以由該受試者的能力參數值 θ 和該試題的試題參數值 a_i、b_i、c_i 所組成的數學公式來表示。若把公式的所

DOI:10.6814/NCCU202001433

有參數值代入，算出各個答對試題的機率值，並畫出其相對應的點，便形成如圖 2-1 所表示的試題特徵曲線。只要每一道試題的試題參數值不同，所畫出來的試題特徵曲線便會不一樣，即如圖 2-1 所示（圖中的這六條試題特徵曲線，其個別具有的試題參數值，分別為：$a_1 = 1.8$、$b_1 = 1.0$、$c_1 = 0.0$，$a_2 = 0.8$、$b_2 = 1.0$、$c_2 = 0.0$，$a_3 = 1.8$、$b_3 = 1.0$、$c_3 = 0.25$，$a_4 = 1.8$、$b_4 = -1.5$、$c_4 = 0.0$，$a_5 = 1.2$、$b_5 = -0.5$、$c_5 = 0.1$，$a_6 = 0.4$、$b_6 = 0.5$、$c_6 = 0.15$，這些參數分別決定這六條試題特徵曲線的形狀各不相同）。

所以，決定每道試題特徵曲線的三個試題參數，分別為：a_i 是第 i 試題的鑑別度參數（item discrimination parameter），表示該試題能否區別出受試者能力高低的程度；b_i 是第 i 試題的難度參數（item difficulty parameter），表示該試題是否困難或容易作答的程度；而 c_i 是第 i 試題的猜測度參數（pseudo-chance parameter），表示該試題被低能力受試者隨機猜題而猜中的程度。

三、 試題反應理論參數模式

目前，在教育測驗的資料分析中，當測驗試題的作答反應只分成對或錯兩種計分方式時，常用的基本試題反應模式有下列幾種，各模式之試題特徵函數公式如下所示：

(一) 一參數邏輯斯模式（one-parameter logistic model）

其試題特性的描述僅涉及試題難度參數 b_i，機率模式為：

$$P(X_i = 1|\theta) = \frac{1}{1+e^{-(\theta-b_i)}}$$

其中 b_i 表示第 i 試題的難度參數，$X_i = 1$ 表示答對第 i 試題的反應，故 $P(X_i = 1|\theta)$ 之值表示擁有能力 θ 的受試者答對第 i 題的機率。依前述定義，給定試題難度參數分別為：$b_3 = -1$，$b_4 = 0$，$b_1 = 1$，$b_2 = 2$ 的四條試題特徵曲線，難度參數值每增加一個單位，試題特徵曲線就向右平移一個單位，如下圖所示：

DOI:10.6814/NCCU202001433

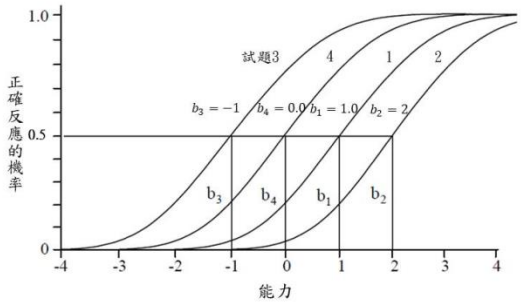


圖2-2 不同試題難度下之能力與正確反應機率函數關係圖

由定義及此圖可看出難度為 b_i 的試題特徵曲線，當正確反應機率為 0.5 時，其對應的能力值剛好為 b_i。換言之，當能力值等於試題難度參數值時 (即 $\theta - b_i = 0$)，受試者答對該題的機會只有 50%；當能力值小於試題難度參數時 (即 $\theta - b_i < 0$)，則受試者答對該題的機會便低於 50%；而當能力值大於試題難度參數時 (即 $\theta - b_i > 0$)，則受試者答對該題的機會便高於 50%。

(二) 二參數邏輯斯模式（two-parameter logistic model）

二參數邏輯斯模式比一參數多了試題鑑別度參數 a_i，最早是由 Lord (1952) 提出二參數常態肩型模式，後來由 Birnbaum (1968) 將其改為較簡單的對數模式，其機率模式為：

$$P(X_i = 1|\theta) = \frac{1}{1+e^{-a_i(\theta-b_i)}}$$

其中 a_i 表示第 i 試題的鑑別度，其餘各符號的定義與一參數模式相同 (Lord, 1952) 。二參數模式會比一參數模式更貼切實際的狀況，因為不是所有的試題都具有相同的鑑別度，在此模式下可以讓某些鑑別度不盡理想的試題依然可以發揮測量的功能。在二參數模式中，試題鑑別度參數 a_i 的數值，剛好與不同 b_i 的試題特徵曲線的斜率 (slope) 成某種比例。當 $\theta = b_i$ 時，此時 $P(\theta) = 0.5$ 且試題特徵曲線的斜率值達到最大值；換句話說，當受試者能力值等於試題難度時，此時能力的變化與答對試題機率的變化之間有最強烈的關係。斜率的大小在圖形上所代表的涵義為，在相同能力範圍情況下正確反應機率的變化率大小，換句話說，當斜率越小，越不能區別不同能力間的落差；反之，斜率越大，不同能力間的落差越明顯。如圖 2-3，這四條試題特徵曲線的試題參數分別為：$a_1 = 1.0$、$b_1 = 1.0$，$a_2 = 0.5$、$b_2 = 1.5$，$a_3 = 1.5$、$b_3 = -1$，$a_4 = 1.2$、$b_4 = 0.0$，鑑別度的不同造

DOI:10.6814/NCCU202001433

就這些曲線將不會平行。

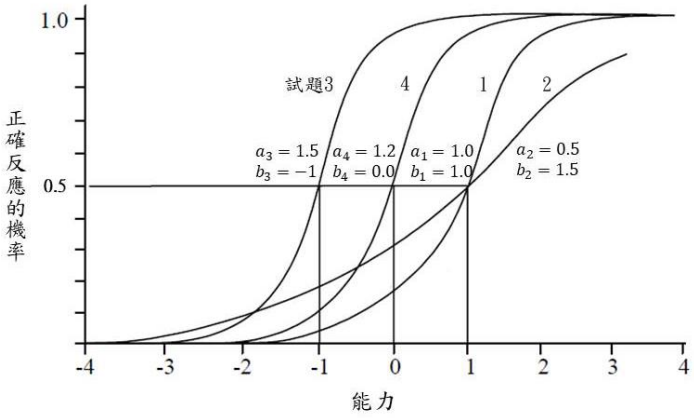


圖2-3 試題難度及鑑別度皆不同下之能力與正確反應機率函數關係圖

(三) 三參數邏輯斯模式（three-parameter logistic model）

三參數邏輯斯模式比二參數多了試題猜測度參數 c_i，是由 Lord and Novick (1968)、Lord (1980) 修改 Lord (1952) 之前所提出的三參數長態肩形模式而來，其機率模式為：

$$P(X_i = 1|\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-a_i(\theta - b_i)}}$$

其中 c_i 表示第 i 試題的猜測度，而其他各符號的定義與二參數相同 (Lord, 1952 ; Lord and Novick, 1968)。這個參數提供試題特徵曲線一個可能大於 0 的下限值，它代表著能力很低的受試者答對某試題的機率。通常 $c_i = \frac{1}{A_i}$，A_i 代表試題 i 的選項數目。下圖 2-4，b_i 是座落在能讓試題特徵曲線之斜率呈現最大值的能力量尺上 (即 $\theta = b_i$)；此時 $P(X_i = 1|\theta) = \frac{1+c_i}{2}$。

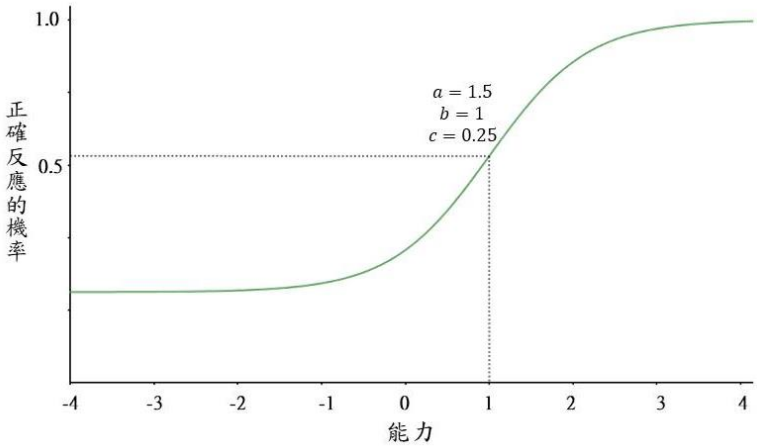


圖2-4 $a = 1.5$、$b = 1$、$c = 0.25$ 下能力與正確反應機率函數圖

DOI:10.6814/NCCU202001433

第三節 作答時間模型

依照古典測驗理論或是試題反應理論，如果受試者在相同的試題上擁有相同的作答反應，會得到相同分數或能力。但作答時間不同，仍給予相同的分數或能力似乎是不合理的，換句話說，兩位受試者在相同的能力測驗上擁有相同的測驗分數，難道他們在該能力上就完全相同嗎？雖然早在 1950 年代，測驗學者們便開始著手將作答時間納入反應模型中，並分析他們的結果，且無論是作答反應或是作答時間，兩者都與試題參數或是能力參數有關連性。然而，在電腦還不發達的年代，對於資料蒐集的困難度導致模型難以進行測試和比較。在近年電腦測驗的普及，越來越多關於作答時間的論文相應而生，其中以「作答時間和反應模型分開」以及「作答時間和反應模型合併」為主要分類，以下就兩種類型的作答時間模型分別進行討論。

一、 作答時間和反應模型分開

作答時間與反應模型分開，涉及兩種不同方式的處理：一種用於試題中的誤讀，另一種用於閱讀速度的評估。

(一) Rasch's model for misreadings

在受試者對相同性質的試題進行閱讀時發生誤讀字數的概率，可以假設其服從卜瓦松過程。在 K 個字數的試題中發生誤讀字數為 a 個的統計結果服從卜瓦松分布。

$$P(a|K) = e^{-\nu}\frac{\nu^a}{a!}$$

其中，$\nu = K\frac{b}{\theta}$ 為誤讀字數的期望值，b 為試題的難度，θ 為受試者的能力。這關係反應了一個事實，意即對一個能力較弱的受試者在一個較難的試題上會有較大的誤讀期望值 (Casella and Berger, 1990)。

(二) Rasch's model for reading speed

以恆定速度閱讀一份試題的字數假設服從卜瓦松過程。但是本模型的目的不是探討給定時間下閱讀字數，而是需要多少時間，以閱讀完成給定字數。這兩個事件有他們的相關性，意即，在給定時間 T 時，讀取字數大於等於 K 的機率相當於給定 K 個字數時讀取時間不超過 T 的機率，

$$P(k \geq K|T) = P(t \leq T|K)。$$

DOI:10.6814/NCCU202001433

因為上面等式左方服從卜瓦松分布，故等式右方的機率服從伽馬分布。亦即

$$P(t|K) = \lambda e^{-\lambda t} \frac{(\lambda t)^{K-1}}{(K-1)!}$$

其中，λ 表示閱讀速度的強度參數，在此也可稱其為速度參數。它代表在給定時間單位中，閱讀字數的期望值。此外，Rasch 還建議將受試者的閱讀速度參數 λ 分解為受試者能力參數 θ 和試題難度參數 b，如下：

$$\lambda = \frac{\theta}{b}$$

二、 包含作答時間的反應模型

以下介紹的兩種模型，係遵循 Thurstonian (1937) 的建模方式，在反應模型中結合作答時間 (或其參數)。

(一) Gaviria's model

$$ln\left(\frac{t_{ij} - T_0}{A}\right) = -a_i(\theta_j - b_i) + \varepsilon_{ij}$$

其中，t_{ij} 為第 j 位受試者在第 i 題的作答時間，A 為尺度常數，T_0 是受試者在難度最小的試題上所花費的時間，$a_i(\theta_j - b_i)$ 是來自 IRT 雙參數反應模型，誤差 ε_{ij} 服從對數常態分布。

$$\varepsilon_{ij} \sim LN(0, {\sigma_i}^2)$$

該模型為雙對數常態分布(double lognormal distribution) ，平均為 $-a_i(\theta_j - b_i)$，變異數為 ${\sigma_i}^2$ ，其中重調尺度時間 $\frac{t_{ij}-T_0}{A}$，T_0 和 A 將根據測試數據估算 (Gaviria, 2005)。

(二) Thissen's model

Thissen (1983) 提出了一個廣為人知的模型，這個模型直接給定了一些在試題反應理論上常見的參數，如下：

$$ln\, t_{ij} = \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij}$$

其中，誤差 $\varepsilon_{ij} \sim N(0, \sigma^2)$，模型中的 μ 為所有受試者和所有試題的一般程度參數， τ_j 被解釋為受試者慢度 (slowness) 參數，β_i 為試題緩慢參數，ρ 是對數作答時間對應反應參數 $a_i\theta_j - b_i$ 的迴歸斜率。

DOI:10.6814/NCCU202001433

三、 其他與作答時間有關的模式

(一) 正確試題總計剩餘時間 (correct item summed residual time, CISRT)評分規則：

Van der Maas and Wagenmakers (2005) 對國際象棋專業知識測量的反應準確性和作答時間提出了以下評分規則 (scoring rule)

$$x_{ij}(d - t_{ij})$$

其中 $x_{ij} = 0$ 或 1 分別表示不正確反應或正確反應， t_{ij} 為作答時間且 d 為時間上限。換句話說，快速正確的反應比慢速正確的反應可獲得更多的分數；而不正確的反應不會獲得任何得分。該評分規則的一個優點是受試者可以權衡速度與準確性，以取得最佳的分數。使用 CISRT 作為評分規則的一個缺點是，如果這些試題對於受試者是困難的，他們通常會迅速猜測，使得自己的得分最大化，因為猜測的反應是瞬時的（$t_{ij} = 0$）。 很明顯，對於能力水平較低的受試者來說，猜測的預期分數將高於依據實力作答的預期分數。

(二) 符號剩餘時間 (signed residual time, SRT) 評分規則

為避免程度較差的受試者會選擇迅速猜測將得分最大化，因此將 CISRT 評分規則修改如下：

$$(2x_{ij} - 1)(d - t_{ij})。$$

亦即，如果作答正確，以剩餘的作答時間值當作受試者的正分值；如果作答錯誤，以剩餘的作答時間值當作受試者的負分值。如果試題難度相對於受試者的能力有困難，且該受試者立即猜測，則他最低的可能得分將等於 $-d$，對於這樣的受試者來說，猜測是不利的。

DOI:10.6814/NCCU202001433

第三章　　研究方法

本文依據研究目的及待答問題，透過取得的實際資料，逐一分析討論。本章研究方法分別依據建立模型及研究過程，共兩節來說明。

第一節 建立模型

首先，制定包含作答反應及作答時間的評分規則；其次，依據此評分規則建立適當的統計模型，其中，統計模型分為：(一) 試題反應模型，(二) 作答時間模型，(三) 新速度準確度模型。以下先行定義本節使用的參數。

N：受試者總人數

n：測驗總題數

$$x_{ij} = \begin{cases} 0 & \text{第}j\text{位受試者答錯第}i\text{題} \\ 1 & \text{第}j\text{位受試者答對第}i\text{題} \end{cases}$$

b_i：第 i 題試題難度

θ_j：第 j 位受試者能力值

k_i：第 i 題的題目字數、正確解答字數以及誘答選項字數之總和

t_{ij}：第 j 位受試者在第 i 題的作答時間

一、　制定評分規則

在現今，如：kahoot、socrative 等許多線上學習測驗平台，相繼將受試者的作答時間納為計分標準之一，因此，如何制定一種加入作答時間的評分規則便是本章節的主要目的。

一般而言，受試者進行受測時會產生「會」、「有印象」、「不會」以上三種情形，無論是哪一種情形，皆須經過一段作答時間後才會產生答對或者答錯的反應，因此，本文作者根據實得數據並結合自身經驗，將受試者在受測當下的情況整理如下圖。

DOI:10.6814/NCCU202001433

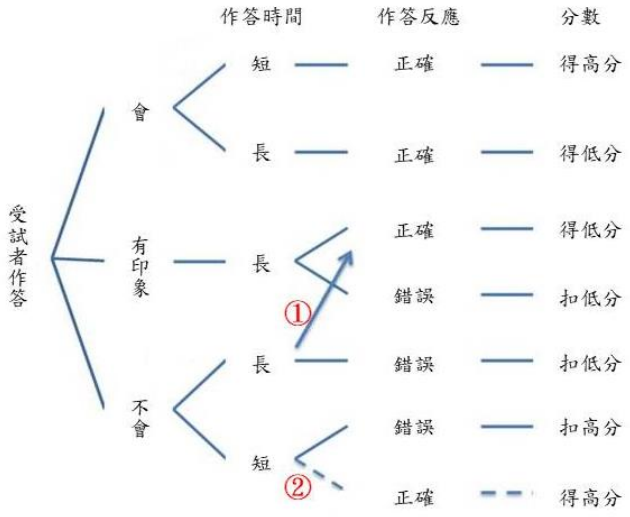


圖3-1 受試者受測時的反應樹狀圖

根據上圖 3-1，將受試者看完題目後所產生的想法分成以下三點論述：

(一) 如果「會」，必定可得到正確反應 (因為人為的疏忽而答錯，不是任何測驗理論所能顧及到的)，便可根據受試者作答時間的長短給予不同的得分。能力較高的受試者會花較短的作答時間，因此給予較多的正分；反之亦然。

(二) 如果「有印象」，此類型的受試者容易受到誘答選項的影響，無法短時間內在正確選項與誘答選項之間做選擇，因此受試者會先經過一段作答時間後，才會產生答對或者答錯的作答反應，此時便可依據其答對(或答錯)給予得正分(或負分)。無論答對或答錯，常理上此類型的受試者程度相近，差別僅在思考的解題策略不同而產生不同的作答反應，因此對於答對者(得正分)或答錯者(得負分)，兩者的分數差距必然不可相差太多。

(三) 如果「不會」，可將受試者分成兩類：

1. 受試者仍想嘗試花較多的作答時間進行解題，此時便可回到第 (二) 點的論述：若答對，如圖 3-1 的路線①給予正分；若答錯，則給予負分。

2. 受試者選擇快速猜測。在答錯的情況下，便給予較高的負分；在答對的情況下，如圖3-1的路線②，由於本文並未將猜測機率作為模型參數，因此會將此類型的受試者排除在外。

結合以上論述並參考 Maris and Van der Maas (2012) 提出的 SRT 評分規則後，將公式制定如下：

DOI:10.6814/NCCU202001433

$$s_{ij} = (2x_{ij} - 1)e^{-\frac{t_{ij}}{c}} \qquad \text{(公式 3-1)}$$

其中，c 為尺度常數，在本文第四章會針對 c 值的設定進行討論並解釋。我們稱此評分規則為符號負指數時間 (signed negative exponent time, SNET) 評分規則，圖形繪製如下，正確反應時最高為 1 分，最低為 0 分；錯誤反應時最高為 0 分最低為 -1 分。

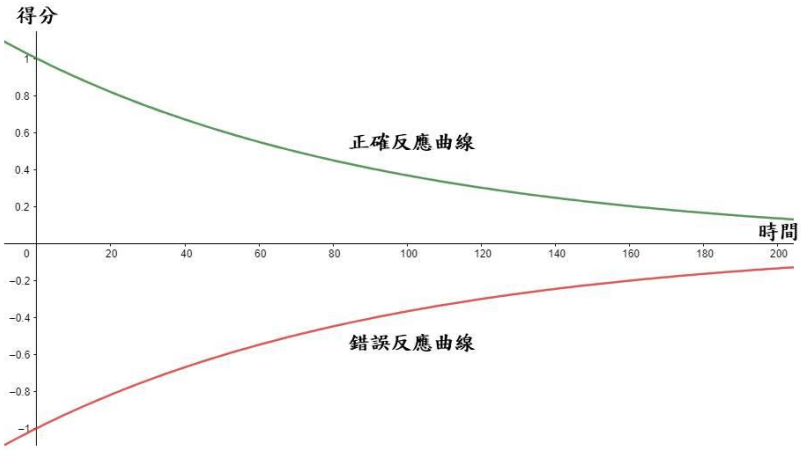


圖3-2 作答時間與受試者正(負)分之關係圖

選擇以指數形式表現出分數的加權值，用意如以下兩點

(一) 正(負)分的變化率趨緩，對於作答正確且作答時間短的受試者，可以透過較明顯的得分差距，進行能力辨別；而隨著作答時間增長，受試者的程度差距便不是那麼的容易辨別。同理，對於作答錯誤的受試者亦是如此。

(二) 當受試者完成此份測驗時測驗即終止，因此受試者就不會在作答時受到每題的時間限制所影響。

本節所建立之評分規則 (SNET)，在第四章會進行實證分析，探討 SNET 評分規則是否合理，並分析此評分規則與受試者的 CTT 測驗總分是否有差異。

二、 建立統計模型

在給定一個評分規則後，希望能藉由受試者在各試題的得分去估計各試題的難度值以及受試者的能力值。因此在本節中，將依序建立與作答反應組型和作答時間相對應的模型，以進一步了解這種新測量模型的運行特徵。

DOI:10.6814/NCCU202001433

(一) 試題反應模型

根據 IRT 一參數模式，任何一位受試者答對任何一道題目的機率，可由受試者的能力值大小和試題的難度值大小來決定。前面已經定義作答反應值 x_{ij}，它的隨機變數以 X_{ij} 表示，茲以 X_{ij} 的條件機率函數，意即試題反應函數表達式調整如下：

$$P(X_{ij}=1|\theta_j,b_i)=\frac{e^{(\theta_j-b_i)}}{1+e^{(\theta_j-b_i)}} \qquad \text{(公式3-2)}$$

公式 3-2 顯示具有能力為 θ_j 的受試者答對難度為 b_i 試題的機率， 因此答錯的機率為 $P(X_{ij}=0|\theta_j,b_i)=1-P(X_{ij}=1|\theta_j,b_i)$。

理論上來說，b_i 值的值域介於 $\pm\infty$ 之間，但在實務應用上，通常只取 ±2.5 之間的範圍值，其統計概念與標準化常態分布下的 z 分數相同。同時，在能力值亦為±2.5 之間的範圍值情況下，能力值與難度值的差最小為-5，因此先算出第 j 位受試者能力值 θ_j 與第 i 道試題難度值 b_i 的差，再加上 5 個單位，以使此值為正的 D_{ij}，亦即 $D_{ij}=\theta_j-b_i+5$，以方便後續假設作答時間模型服從伽瑪分布之建立。圖 3-3 顯示不同 D_{ij} 值的受試者之正確反應的機率： D_{ij} 分別為 9、5、3 三個正確反應機率值的情形如虛線。

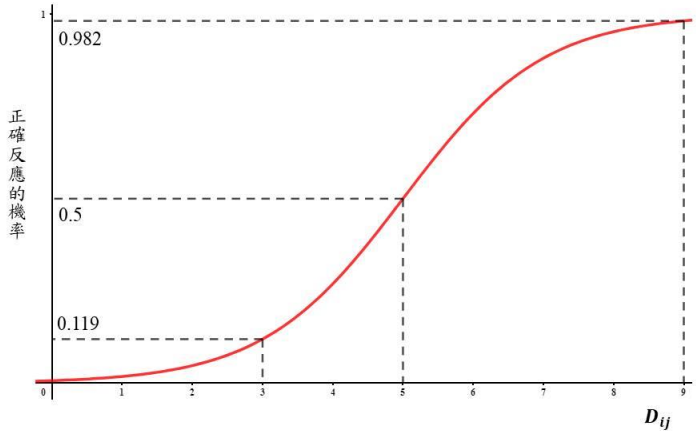


圖3-3 D_{ij}和正確反應機率的函數圖

由 $D_{ij}=9$遞減至 $D_{ij}=3$ 的過程中可得知：「當 $D_{ij}=9$ 時，答對的機率接近 1 ，而答錯的機率接近 0；當 $D_{ij}=5$ 時，答對與答錯的機率一樣都是 0.5 ；當 $D_{ij}=3$ 時，答對的機率接近0，而答錯的機率接近 1。」因此D_{ij}越大，答對的機率越高。換言之，當題目固定時，能力越好的人答對的機率越大。

DOI:10.6814/NCCU202001433

(二) 作答時間模型

在建立完試題反應模型後，我們希望找出對於作答時間的分布。Maris and Van der Maas (2012) 在其文獻中所給出的作答時間模型服從指數分布，但是考量本文以數學測驗為研究背景，因此受試者需要較長的閱讀理解時間。以難度為簡單的試題 24 為例，統計在不同時間區間的作答人數，繪製直方圖如下。

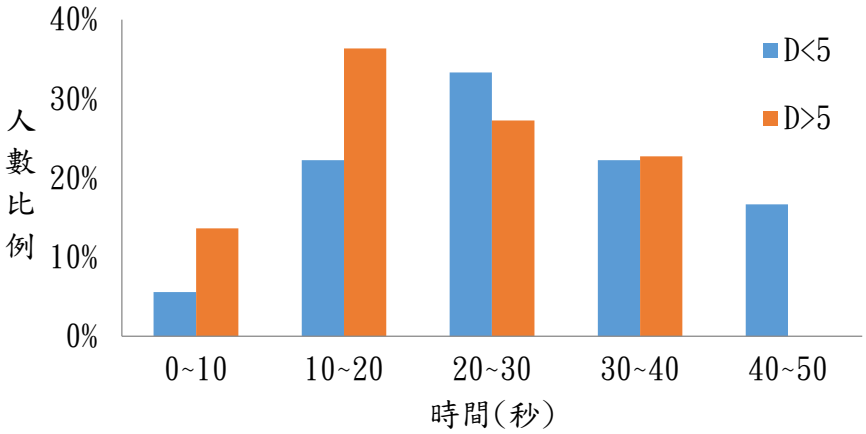


圖3-4 試題 24 各時間區間作答人數分布

其中，$D_{ij} = \theta_j - b_i + 5$，$D_{ij} < 5$ 代表能力值小於難度值的受試者；$D_{ij} > 5$ 代表能力值大於難度值的受試者。圖 3-4 顯示，作答人數從時間為 0 開始遞增至時間為 20 達到最大值。故本研究將作答時間修改為比指數分布更一般化的伽瑪分布，並會在本文第四章第三節進行伽瑪分布適合度檢定。

再者，受試者於試題作答時所花的時間，是否與運算過程的冗長程度有關，將以正確解答的字數以及誘答選項的字數來代替運算過程的冗長程度，並加上題目文字敘述的字數，一併進行研究與討論。然後，每位受試者的解答方式以及計算過程的書寫未必會相同，因此便以大考中心所公布之計算過程作為正確解答的字數 (陳茂雄, 2017)。在此先假設「閱讀字數、正確解答字數以及誘答選項字數之總和」與「伽瑪分布參數之理論值」為正相關並將其加入模型中，待後續第四章再一併進行分析討論。

若作答時間 (T) 服從伽瑪分布 gamma(α, β)，其機率密度函數如下：

$$f(\mathrm{t}) = \frac{\mathrm{t}^{(\alpha-1)} e^{(-\beta \mathrm{t})}}{\Gamma(\alpha)\beta^{(-\alpha)}}$$

其中，參數 α 稱為形狀參數 (shape parameter)，β 稱為尺度參數 (scale parameter)。

DOI:10.6814/NCCU202001433

在「閱讀字數、正確解答字數以及誘答選項字數之總和」與「伽瑪分布參數之理論值」為正相關的假設下，以字數 (k_i) 作為形狀參數；同時，以 $\theta_j - b_i + 5$ 為尺度參數後之機率函數表示如下：

$$f\left(T_{ij} = t_{ij} \middle| \theta_j, b_i, k_i\right) = \frac{{t_{ij}}^{(k_i-1)} e^{-t_{ij}(\theta_j - b_i + 5)}}{\Gamma(k_i)(\theta_j - b_i + 5)^{-k_i}} \qquad \text{(公式 3-3)}$$

當 k_i 固定為 40 時，隨著D_{ij} 值的不同，作答時間的分布圖形如下。

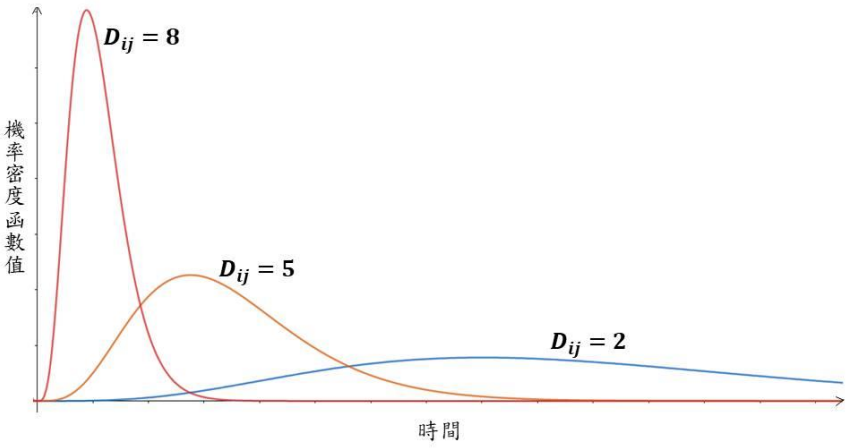


圖3-5 $k_i = 40$，D_{ij}不同時， 作答時間之分布圖

由上圖結果可知，在能力高的族群中傾向使用較少的平均作答時間完成試題且變化不大，而能力低的族群則是傾向花費較多的時間完成試題且平均作答時間變化較大。而當 D_{ij} 固定為 5 時，字數與作答時間的變化如下圖所示。

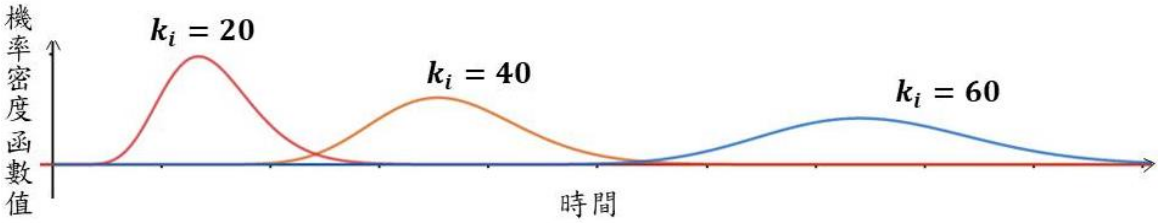


圖3-6 $D_{ij} = 5$、k_i 不同時，作答時間之函數圖

由圖 3-6 可知，當字數為 20 時，作答時間較短且集中；而隨著字數的增加，受試者的作答時間除了也會隨之增加外，作答時間也較為分散，整體會更趨近於常態分布。

DOI:10.6814/NCCU202001433

(三) 新速度準確度模型

Maris and Van der Maas (2012) 在其文獻中的模型為速度-準確度模型，但是本文將其作答時間由指數分布修改為伽瑪分布，因此重新命名為新速度-準確度模型(new speed-accuracy response model, NSARM)，模型建立如下：

在作答反應和作答時間兩個隨機變數獨立假設下，NSARM 可透過試題反應模型 (公式 3-2) 和作答時間模型 (公式 3-3) 所組成，同時，依據公式 3-1 評分規則，分數的隨機變數 (S_{ij}) 可表示如下：

$$S_{ij} = (2X_{ij} - 1)e^{-\frac{T_{ij}}{c}}，c \text{ 為尺度常數'}$$

以下將推導出S_{ij}的機率密度函數。首先，

$$\begin{cases} S_{ij} = -e^{-\frac{T_{ij}}{c}} & if\ X_{ij} = 0 \\ S_{ij} = e^{-\frac{T_{ij}}{c}} & if\ X_{ij} = 1 \end{cases}$$

將評分規則透過變數變換得

$$\begin{cases} T_{ij} = -c\ ln(-S_{ij}) & if\ S_{ij} < 0 \\ T_{ij} = -c\ ln(S_{ij}) & if\ S_{ij} \geq 0 \end{cases}$$

此時可將模型的分布表達如下：

$$f(S_{ij} = s_{ij}|\theta_j, b_i, k_i)$$

$$= \begin{cases} f(T_{ij} = -c\ ln(-s_{ij})\ |X = 0, k_i, \theta_j, b_i)\ P(X = 0|\theta_j, b_i)\ |J(T_{ij})| & if\ s_{ij} < 0 \\ f(T_{ij} = -cln(s_{ij})|X = 1, k_i, \theta_j, b_i)\ P(X = 1|\theta_j, b_i)\ |J(T_{ij})| & if\ s_{ij} \geq 0 \end{cases}$$

其中，在 $s_{ij} < 0$ 和 $s_{ij} \geq 0$ 皆需計算其 Jacobian，得 $|J(T)| = \frac{dT_{ij}}{ds_{ij}} = \frac{c}{s_{ij}}$ ，整理後如下：

$$f(S_{ij} = s_{ij}|\theta_j, b_i, k_i)$$

$$= \begin{cases} \frac{(-c \cdot ln(-s_{ij}))^{(k_i-1)}(-s_{ij})^{c(\theta_j - b_i + 5)}}{\Gamma(k_i)(\theta_j - b_i + 5)^{-k_i}} \frac{1}{1 + e^{(\theta_j - b_i)}} \frac{c}{s_{ij}} & if\ s_{ij} < 0 \\ \frac{(-c \cdot ln(s_{ij}))^{(k_i-1)} s_{ij}{}^{c(\theta_j - b_i + 5)}}{\Gamma(k_i)(\theta_j - b_i + 5)^{-k_i}} \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} \frac{c}{s_{ij}} & if\ s_{ij} \geq 0 \end{cases}$$

$$= \frac{c(-c\ ln|s_{ij}|)^{(k_i-1)}|s_{ij}|^{(c(\theta_j - b_i + 5)-1)}(\theta_j - b_i + 5)^{k_i}}{\Gamma(k_i)} \frac{e^{[(\theta_j - b_i)\delta(s_{ij})]}}{1 + e^{(\theta_j - b_i)}} \qquad \text{(公式3-4)}$$

其中，$\delta(s_{ij}) = \begin{cases} 0 & if\ s_{ij} < 0 \\ 1 & if\ s_{ij} \geq 0 \end{cases}$ 。

假設各個受試者的作答反應是相互獨立的，同一受試者對各試題間的反應

DOI:10.6814/NCCU202001433

也是相互獨立，基於試題分數分布已知，可以進行推斷其概似函數為：

$$\mathrm{L}(s_{ij}|\theta_j, b_i) = \prod_{j=1}^{N}\prod_{i=1}^{n} \frac{c\left(-c\cdot\ln\left|s_{ij}\right|\right)^{k_i-1}}{\Gamma(k_i)\left|s_{ij}\right|} \frac{(\theta_j - b_i + 5)^{k_i}\left|s_{ij}\right|^{c\theta_j - cb_i + 5c - 1} e^{[(\theta_j - b_i)\delta(s_{ij})]}}{1 + e^{\theta_j - b_i}}$$

由於 $\mathrm{L}(s_{ij}|\theta_j, b_i)$ 與 $\ln[\mathrm{L}(s_{ij}|\theta_j, b_i)]$ 同時達到最大，我們通常都是討論對數概似函數，簡化為：

$$\ln\left[\mathrm{L}\left(s_{ij}\middle|\theta_j, b_i\right)\right] = \sum_{j=1}^{N}\sum_{i=1}^{n}\left[\ln\left[\frac{c\left(-c\cdot\ln\left|s_{ij}\right|\right)^{k_i-1}}{\Gamma(k_i)\left|s_{ij}\right|}\right] + k_i\cdot\ln\left(\theta_j - b_i + 5\right) + (c\theta_j - cb_i + 5c - 1)\cdot ln\left|s_{ij}\right| + \left(\theta_j - b_i\right)\delta\left(s_{ij}\right) - \ln\left(1 + e^{\theta_j - b_i}\right)\right] \quad \text{(公式3-4 b)}$$

以下同樣以平移後受試者能力與試題難度的差 $D_{ij} = \theta_j - b_i + 5$ 作為參數，以方便討論模型中參數值的大小變化。在選擇字數為 40 的情況下，將此試題分數分布的圖形依照 $s_{ij} \geq 0$ 和 $s_{ij} < 0$ 繪製如下：

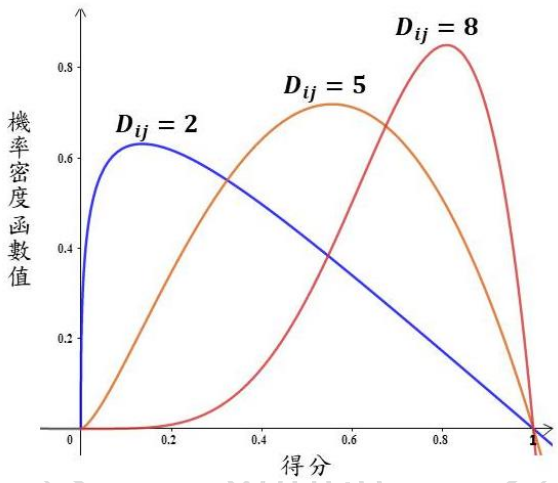


圖3-7 不同 D 值所獲得分數($s_{ij} \geq 0$)之函數圖

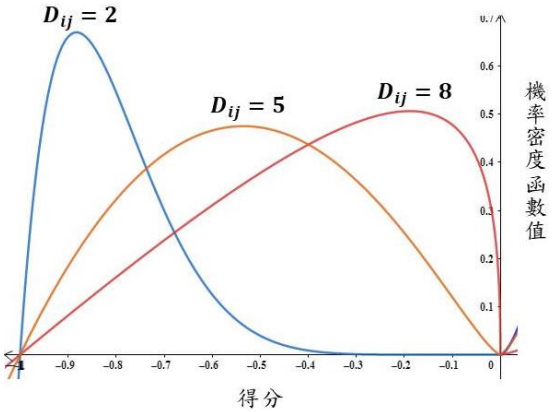


圖3-8 不同 D 值所失去分數($s_{ij} < 0$)之函數圖

DOI:10.6814/NCCU202001433

由圖 3-7 可知，$D_{ij}=8$、$D_{ij}=5$、$D_{ij}=2$ 的受試者獲得高分的機率隨著 D 值的下降而遞減；而根據圖 3-8，$D_{ij}=8$、$D_{ij}=5$、$D_{ij}=2$ 的受試者失去高分的機率隨著 D 值的下降而遞增。因此，在同一道試題的情況下，由圖 3-7 可得知，D 的值越大，即能力值較高的受試者獲得高分數的機率較大，而當 D 的值越小，即能力值下降時，獲得高分數的機率也會隨之遞減。同理，失去分數的情況如圖3-8。

為了解不同能力值所能得到的預期分數，先找出 s_{ij} 的期望值 $E(S_{ij})$。在找出 s_{ij} 的期望值之前，先行探討以下公式。

引理 3-1.

$$\int_0^1(-c\cdot ln\,s)^{k-1}s^{cD}ds=\frac{c(k-1)}{cD+1}\int_0^1(-c\cdot ln\,s)^{k-2}s^{cD}ds$$

證：重複使用分部積分可得結果如下：

$$\int_0^1(-c\cdot ln\,s)^{k-1}s^{cD}ds$$

$$=\frac{c^{k-2}\Gamma(k)}{(cD+1)^{k-2}}\int_0^1(-c\cdot ln\,s)^{k-(k-1)}s^{cD}ds$$

$$=\frac{c^{k-1}\Gamma(k)}{(cD+1)^{k}} \qquad \text{（公式 3-5）}$$

以下根據公式 3-5 的化簡結果。

$$E(S_{ij})=\int_0^1 s_{ij}\cdot f(s_{ij})ds_{ij}+\int_{-1}^0 s_{ij}\cdot f(s_{ij})ds_{ij}$$

1. 當 $s_{ij}\geq 0$ 時

$$\int_0^1 s_{ij}\cdot f(s_{ij})ds_{ij}$$

$$=\frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)}\frac{e^{(D_{ij}-5)}}{1+e^{D_{ij}-5}}\int_0^1\left(-c\cdot ln\,s_{ij}\right)^{k_i-1}s_{ij}{}^{cD_{ij}}ds_{ij}$$

$$=\frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)}\frac{e^{(D_{ij}-5)}}{1+e^{D_{ij}-5}}\frac{c^{k_i-1}\Gamma(k_i)}{(cD_{ij}+1)^{k_i}}$$

2. 當 $s_{ij}<0$ 時，

DOI:10.6814/NCCU202001433

$$\int_{-1}^{0} s_{ij} \cdot f(s_{ij}) ds_{ij}$$

$$= \frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)} \frac{1}{1+e^{D_{ij}-5}} \int_{-1}^{0} \left(-c \cdot ln(-s_{ij})\right)^{k_i-1} (-s_{ij})^{cD_{ij}} ds_{ij}$$

$$= \frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)} \frac{1}{1+e^{D_{ij}-5}} \frac{c^{k_i-1}\Gamma(k_i)}{(cD_{ij}+1)^{k_i}}$$

因此，

$$\mathrm{E}(s_{ij}) = \frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)} \frac{e^{(D_{ij}-5)}}{1+e^{D_{ij}-5}} \frac{c^{k_i-1}\Gamma(k_i)}{(cD_{ij}+1)^{k_i}} + \frac{cD_{ij}{}^{k_i}}{\Gamma(k_i)} \frac{1}{1+e^{D_{ij}-5}} \frac{c^{k_i-1}\Gamma(k_i)}{(cD_{ij}+1)^{k_i}}$$

$$= \frac{(cD_{ij})^{k_i}}{(cD_{ij}+1)^{k_i}}$$

當 c 和 k_i 固定時，D_{ij} 越大，預期分數的平均越高；D_{ij} 越小，預期分數的平均越低。亦即能力越好的受試者，預期分數會越高。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

第二節 研究過程

一、 研究對象

本研究使用線上教學平台的題庫，挑選出七年級一元一次方程式的會考試題共 15 題作為前測，並挑選八年級一元二次方程式的會考試題共 15 題作為後測，前測與後測的難度及鑑別度具有一致性 (題型以及大考中心答對率相似，並且按題序一一對應，如：試題 1 對應試題 16；試題 2 對應試題 17，以此類推)，並依大考中心所定義之答對率將試題難度分為簡單、中等、困難。同時，從前後測皆完成的所有使用者當中，隨機抽樣共計 200 位八年級生 (其中包含 109 位男性、91 位女性)， 並蒐集這些受試者的作答反應與作答時間作為本文模型分析之實際數據。在本文第四章將以前測作為受試者能力的初步分類，並利用後測進行驗證。

二、 研究工具

本研究施測所採用的測驗平台對於每個登入帳號之受試者，採用隨機順序出題，每次僅會出現一道試題讓受試者進行作答，受試者一但提交答案即立刻進入下一道試題，無法再回去查閱及修改，且每道試題皆必須作答。測驗過程中系統將記錄受試者在每一試題自題目出現開始至完成選項作答為止的時間。

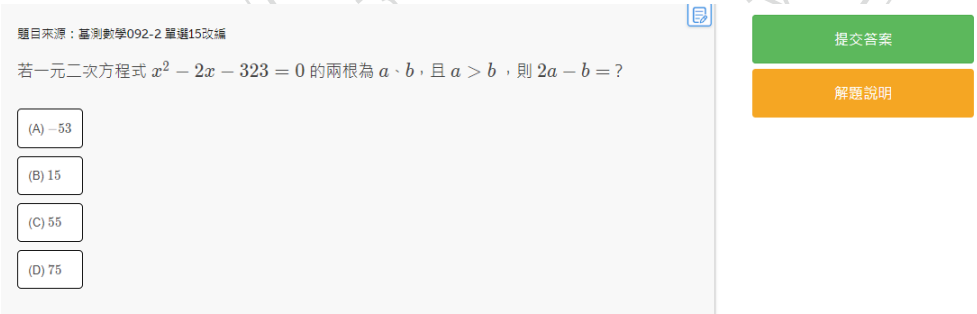


圖 3-9 線上測驗施測畫面

本研究需要估算之數據為 200 位受試者分別於 30 道試題中的所有作答反應及時間。在當今 IRT 領域適用的電腦軟體程式，由於這些程式有的已經商業化，像是 Bilog-mg、Parscale 等等，有的仍停留在實驗研究階段。再者，這些軟體主要是 IRT 用來分析二元計分 (非對即錯的反應) 及多元計分 (答題結果有很多種

DOI:10.6814/NCCU202001433

可能性) 資料的程式，在經過多方考量後便以下列電腦軟體進行運算：

(一) Excel 軟體

使用 Microsoft Excel 2010 軟體，整理受試者於各試題之作答反應資料，匯入下列軟體，進行各項數據之演算及圖表之繪製。

(二) R 軟體

依據受試者於各試題之作答時間及作答反應資料，使用 catR、ltm、TAM 等 package 對受試者進行參數估計

(三) Matlab 軟體

本研究亦使用 Matlab R2010a 軟體，依據施測者於各試題之作答時間，及作答反應資料，來進行參數估計。

三、 程式驗證

本研究使用 Matlab 軟體作為本文模型 (NSARM) 的參數估計軟體，在進行數據分析之前，希望透過模擬的方式探討程式的穩定性。將受試者人數以及試題數作為控制變因，藉由模擬次數的增加，評估程式的可信度。

模擬過程分為兩部分，介紹如下：

第一部分 (Excel 軟體)：先隨機給定 k_i (字數)、b_i (試題難度) 及 θ_j (受試者能力)，其中b_i 範圍界於 -2.5~2.5 ，θ_j 範圍界於 -2.5~2.5 。

利用 Excel 軟體，先將這些給定的 b_i 及 θ_j 值代入公式 3-2 得出 $P(X_{ij}=1|\theta_j,b_i)$ 的機率值 p，再隨機選出介於 0~1 的亂數值，亦即均勻隨機變數 Y 的值，若 Y ≤ p 則生成 $x_{ij}^{(1)}=1$；否則生成 $x_{ij}^{(1)}=0$ 。重複上述步驟進行 M 次，可得作答反應 $x_{ij}^{(m)}$，其中 $m=1,\dots,M$。

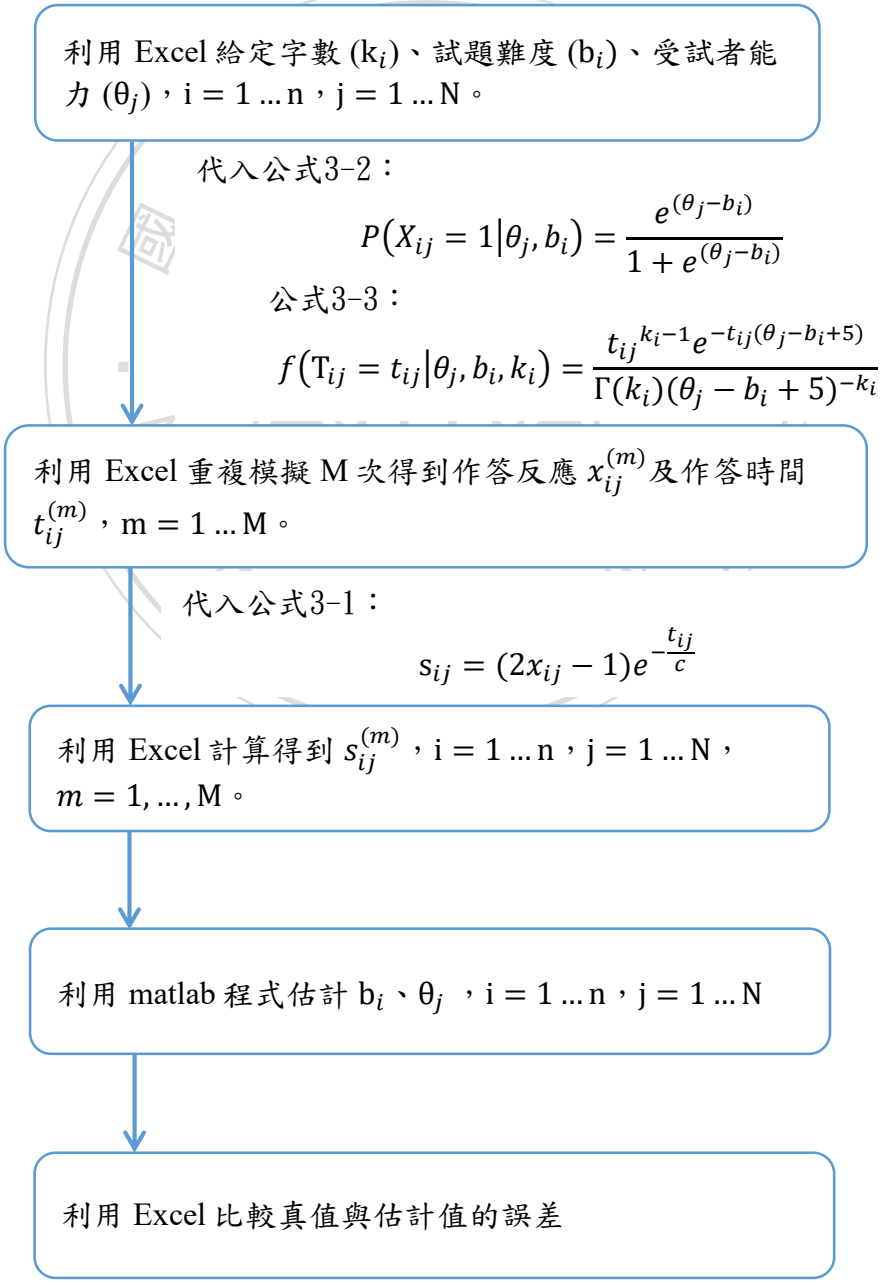
同上述方式模擬產生均勻隨機變數 Y 的值，再將此值及隨機給定的 k_i、θ_j 、b_i 值透過 Excel 的 GAMMA.INV 指令產生作答時間$t_{ij}^{(1)}$ (該指令可以傳回伽瑪累積分配 $F(T_{ij}=t_{ij}|\theta_j,b_i,k_i)$的反函數值$t_{ij}$)。重複上述步驟進行 M 次，可得作答時間 $t_{ij}^{(m)}$，其中 $m=1,\dots,M$。

把第 m 次的第 i 道試題第 j 為考生相關的模擬值 $x_{ij}^{(m)}$ 及 $t_{ij}^{(m)}$ ，代入 $s_{ij}=(2x_{ij}-1)e^{-\frac{t_{ij}}{c}}$ (公式 3-1) ，利用 Excel 算出分數 $s_{ij}^{(m)}$ 。

DOI:10.6814/NCCU202001433

第二部分 (Matlab 軟體)：將第一部分的 $s_{ij}^{(m)}$ 彙整成檔案後作為輸入的數據，並利用 Matlab 處理多變量函數的指令 fminsearch 透過公式 3-4 b 估計θ_j 、b_i 的值。該指令能夠從一個初始值開始，找到目標函數的最小值(或將目標函數取負號可得最大值)，例如：x ＝ fminsearch (function, x_0) 從 x_0 開始，找到目標函數 function 中的局部最小值 x。fminsearch 的第一個參數為目標函數，如果目標函數裡有多個變數，則將其中需要估計的參數以匿名函數 (anonymous function) 的方式呈現在目標函數之前，如果僅有一個變數則不需要，例如x ＝ fminsearch (@(待估參數) function, x_0)；第二個參數是初始值的設定。

依據上述兩個部分，重新整理出模擬流程圖如下：



DOI:10.6814/NCCU202001433

首先，以受試者人數與試題數作為控制變因(即設定 n=8，N=6)，重複模擬 40 次後(即 M=40)，將參數 k、b 以及 θ 的估計值彙整如下：

表3-1 k 和 b 之真值與估計值的比較 (M=40)

n=8，N=6	真值		估計值		相對誤差	
	k	b	k	b	K	b
模擬題一	20	-1	20.008	-1.000	-0.038%	0.043%
模擬題二	40	1	39.981	1.000	0.047%	0.050%
模擬題三	35	0.6	35.022	0.600	-0.063%	0.032%
模擬題四	45	0.2	44.979	0.200	0.046%	0.045%
模擬題五	25	-0.2	25.013	-0.200	-0.052%	0.055%
模擬題六	50	1.4	49.978	1.399	0.045%	0.064%
模擬題七	55	-0.6	54.982	-0.600	0.033%	0.055%
模擬題八	30	-1.4	30.016	-1.399	-0.053%	0.051%

表3-2 θ 之真值與估計值的比較 (M=40)

n=8，N=6	真值	估計值	相對誤差
	θ	θ	θ
模擬受試者一	0.3	0.300	0.063%
模擬受試者二	0.9	0.899	0.060%
模擬受試者三	-1.5	-1.499	0.059%
模擬受試者四	-0.3	-0.300	0.053%
模擬受試者五	1.5	1.500	0.033%
模擬受試者六	-0.9	-0.900	0.050%

表3-1、3-2 結果呈現 k、b、θ 三者的真值與估計值的差異情形。接著我們同樣固定 n=8，N=6 時，將模擬次數從 40 調整至 60 後，結果如下表所示：

DOI:10.6814/NCCU202001433

表3-3 k 和 b 之真值與估計值的比較 (M=60)

n=8，N=6	真值		估計值		相對誤差	
	k	b	k	b	k	b
模擬試題一	20	-1	20.003	-1.000	-0.015%	0.016%
模擬試題二	40	1	39.994	1.000	0.016%	0.013%
模擬試題三	35	0.6	35.007	0.600	-0.019%	0.015%
模擬試題四	45	0.2	44.992	0.200	0.018%	0.015%
模擬試題五	25	-0.2	25.004	-0.200	-0.015%	0.011%
模擬試題六	50	1.4	49.992	1.400	0.017%	0.014%
模擬試題七	55	-0.6	54.992	-0.600	0.015%	0.018%
模擬試題八	30	-1.4	30.005	-1.400	-0.017%	0.009%

表3-4 θ之真值與估計值的比較 (M=60)

n=8，N=6	真值	估計值	相對誤差
	θ	θ	θ
模擬受試者一	0.3	0.301	0.020%
模擬受試者二	0.9	0.903	0.023%
模擬受試者三	-1.5	-1.499	0.021%
模擬受試者四	-0.3	-0.302	0.023%
模擬受試者五	1.5	1.501	0.019%
模擬受試者六	-0.9	-0.897	0.017%

觀察表 3-1 至表 3-4 的結果，發現當 n=8，N=6 固定時，隨著模擬次數從 40 次增加到 60 次，真值與估計值的相對誤差隨之下降。換言之，隨著模擬次數的增加，估計值與真值越趨一致。

DOI:10.6814/NCCU202001433

第四章　　模型驗證

本章透過符合 NSARM 的模擬資料(即包含作答反應和作答時間)，分別在 IRT 及 NSARM 的模式下估計能力值和難度值，並計算其均方根誤差(RMSE)，以說明使用 NSARM 之合理性。

本研究模擬次數 (M) 設定為 40 次，模擬資料共設計兩個研究變項。

1. 受試者樣本大小，包括 $\mathrm{N}=500$ 和 $\mathrm{N}=1000$ 位受試者。
2. 試題數，包括 $\mathrm{n}=10$ 題與 $\mathrm{n}=20$ 題 。此設定是參考 de la Torre and Song (2009) 及郭伯臣、謝典佑、吳慧珉、林佳樺 (2012)。

模擬過程分為四部分，介紹如下：

第一部分 (Excel 軟體)：先利用 Excel 軟體隨機模擬 k_i (字數)、θ_j (受試者能力)，其中 k_i~U[15,45]，θ_j~N(0,0.75) 範圍界於 -2.5~2.5 。再模擬 b_i (試題難度) ，範圍界於 -2.5~2.5，將其區分成 3 等分後，分成以下四種情形分別進行模擬：

1. 若 b_i 範圍界於 -2.5~2.5 ，為所有難度可能值，其中 b_i~N(0,0.75)。
2. 若 b_i 範圍界於 -2.5~-0.8 ，定義為簡單試題，其中 b_i~N(-1.67,0.25)。
3. 若 b_i 範圍界於 -0.8~0.8 ，定義為中等試題，其中 b_i~N(0,0.25)。
4. 若 b_i 範圍界於 0.8~2.5 ，定義為困難試題，其中 b_i~N(1.67,0.25)。

首先，再利用 Excel 軟體，將這些隨機給定的 b_i 及 θ_j 值代入公式 3-2 得出 $P\left(X_{ij}=1\middle|\theta_j,b_i\right)$ 的機率值 p，並隨機選出介於 0~1 的亂數值，亦即均勻隨機變數 Y 的值，若 $\mathrm{Y}\leq\mathrm{p}$ 則生成 $x_{ij}^{(1)}=1$；否則生成 $x_{ij}^{(1)}=0$ 。重複上述步驟進行 M 次，可得作答反應 $x_{ij}^{(m)}$，其中 $m=1,\ldots,M$。

同上述方式模擬產生均勻隨機變數 Y 的值，再將此值及隨機給定的 k_i、θ_j、b_i 值透過 Excel 的 GAMMA.INV 指令產生作答時間$t_{ij}^{(1)}$ (該指令可以傳回伽瑪累積分布 $\mathrm{F}\left(T_{ij}=t_{ij}\middle|\theta_j,b_i,k_i\right)$的反函數值$t_{ij}$)。重複上述步驟進行 M 次，可得作答時間 $t_{ij}^{(m)}$，其中 $m=1,\ldots,M$。

第二部分 (Matlab 軟體)：先將第一部分由 Excel 軟體所模擬出的字數、作答反應及作答時間值，分別由 Matlab 軟體讀取後，再把第 m 次的第 i 道試題第

DOI:10.6814/NCCU202001433

j 位考生相關的模擬值 $x_{ij}^{(m)}$ 及 $t_{ij}^{(m)}$ ，代入 $s_{ij} = (2x_{ij} - 1)e^{-\frac{t_{ij}}{c}}$ (公式 3-1) ，並透過 Matlab 算出分數 $s_{ij}^{(m)}$作為數據，利用這些數據在 NSARM 模式下進行參數 (θ_j, b_i) 的估計，可得到 M 對 $(\theta_j{}^{(m)}, b_i{}^{(m)})$ 。

第三部分(R 軟體)：先將第一部分由 Excel 軟體所模擬出的作答反應值，由 R 軟體讀取後，再把第 m 次的第 i 道試題第 j 位考生的模擬值 $x_{ij}^{(m)}$作為數據，利用這些數據在 IRT 模式下進行參數 (θ_j, b_i) 的估計，可得到 M 對 $(\theta_j{}^{(m)}, b_i{}^{(m)})$ 。

第四部分 (Excel 軟體)：本研究所探究的估計效能是指真值與估計值間的差異，數值小代表參數估計越正確，估計效能越佳。參考 de la Torre and Song (2009) 的評估指標，使用均方根誤差 (root-mean-square error, RMSE)，公式詳列於下：

$$\text{能力值之 } RMSE(j) = \sqrt{\frac{\sum_{m=1}^{M}\left(v_j - e_j^{(m)}\right)^2}{M}} \quad \mathrm{j} = 1 \ldots \mathrm{N}$$

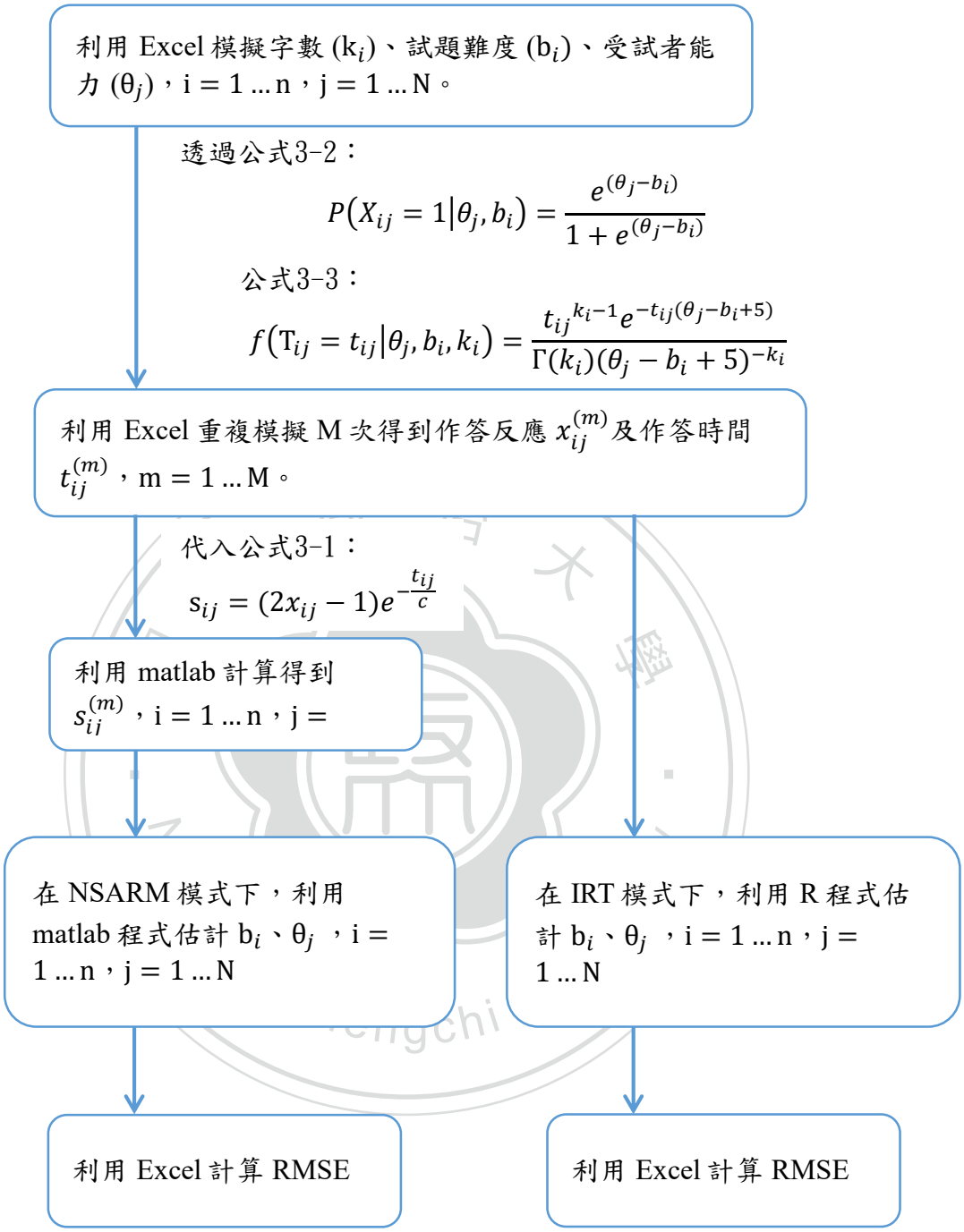
$$\text{難度值之 } RMSE(i) = \sqrt{\frac{\sum_{m=1}^{M}\left(v_i - e_i^{(m)}\right)^2}{M}} \quad \mathrm{i} = 1 \ldots \mathrm{n}$$

其中，v_j 、v_i為真值，$e_j^{(m)}$、$e_i^{(m)}$為估計值，M 為模擬次數。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

依據上述四個部分，重新整理出模擬流程圖如下：



DOI:10.6814/NCCU202001433

在相同受試者下，使用該受試者每一次的 (共 M 次)能力估計值與真值計算 RMSE，依此類推，總共會有 N 個 RMSE 值。根據本文所有難度可能值的範圍，將兩種估計法於能力值估計之 RMSE 繪製成折線圖如下。

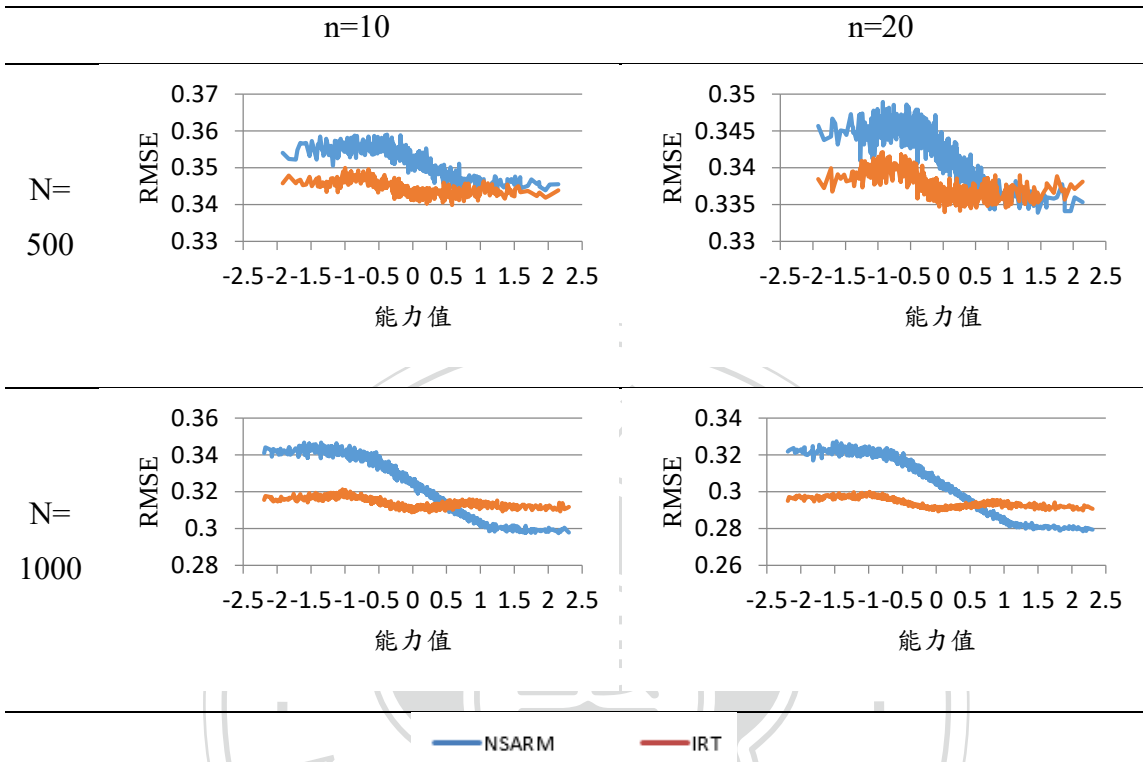


圖4-1 兩種估計法於能力值估計之 RMSE 圖(依所有試題)

觀察圖4-1可發現以下兩點：

1.當試題數固定時，隨著受試人數的增加，在估計能力時 NSARM 之 RMSE 呈現下降的趨勢，且能力越高的受試者，RMSE 下降的趨勢越明顯，甚至當能力值大於一定數值時，NSARM 之 RMSE 比 IRT 之 RMSE 還小。

2.當受試人數固定時，隨著試題數的增加，可以從縱坐標的數值發現，NSARM 之 RMSE 及 IRT 之 RMSE 均有明顯的減少。

以下分別對簡單、中等以及困難的試題，重複上述方式，計算兩種模式在能力值估計的 RMSE，並將結果依序彙整如下圖4-2至圖4-4。

DOI:10.6814/NCCU202001433

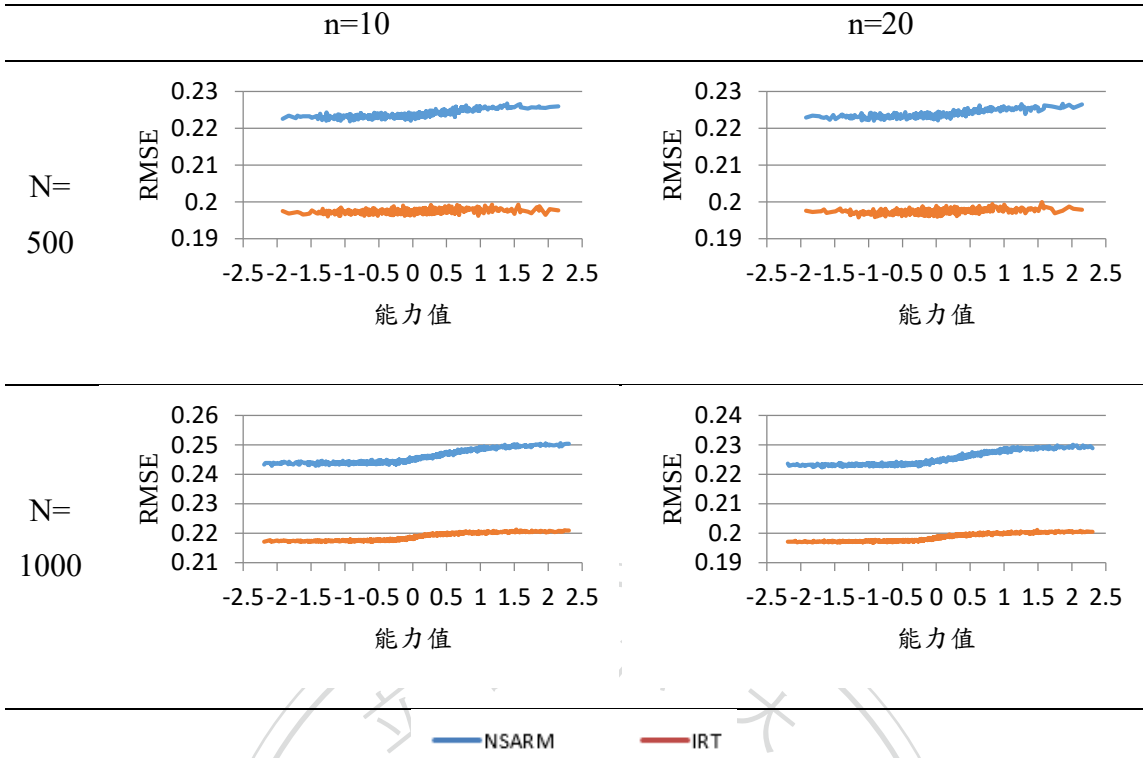


圖4-2 兩種估計法於能力值估計之 RMSE 圖(依簡單試題)

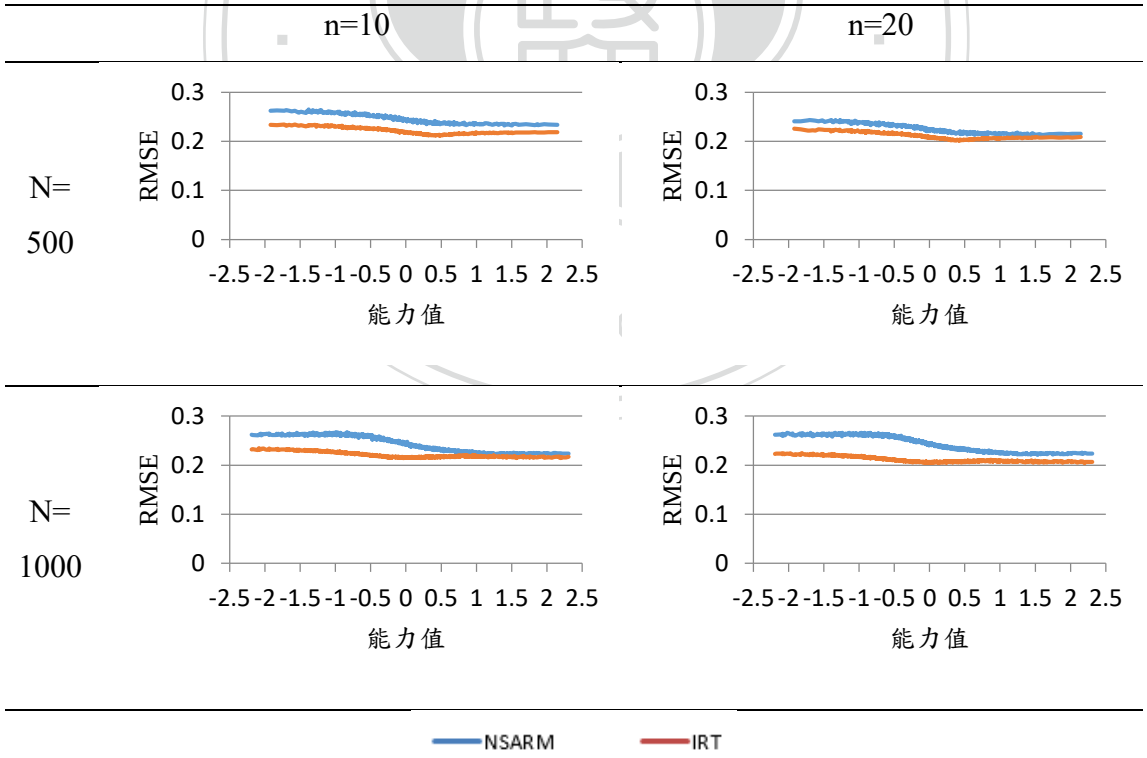


圖4-3 兩種估計法於能力值估計之 RMSE 圖(依中等試題)

DOI:10.6814/NCCU202001433

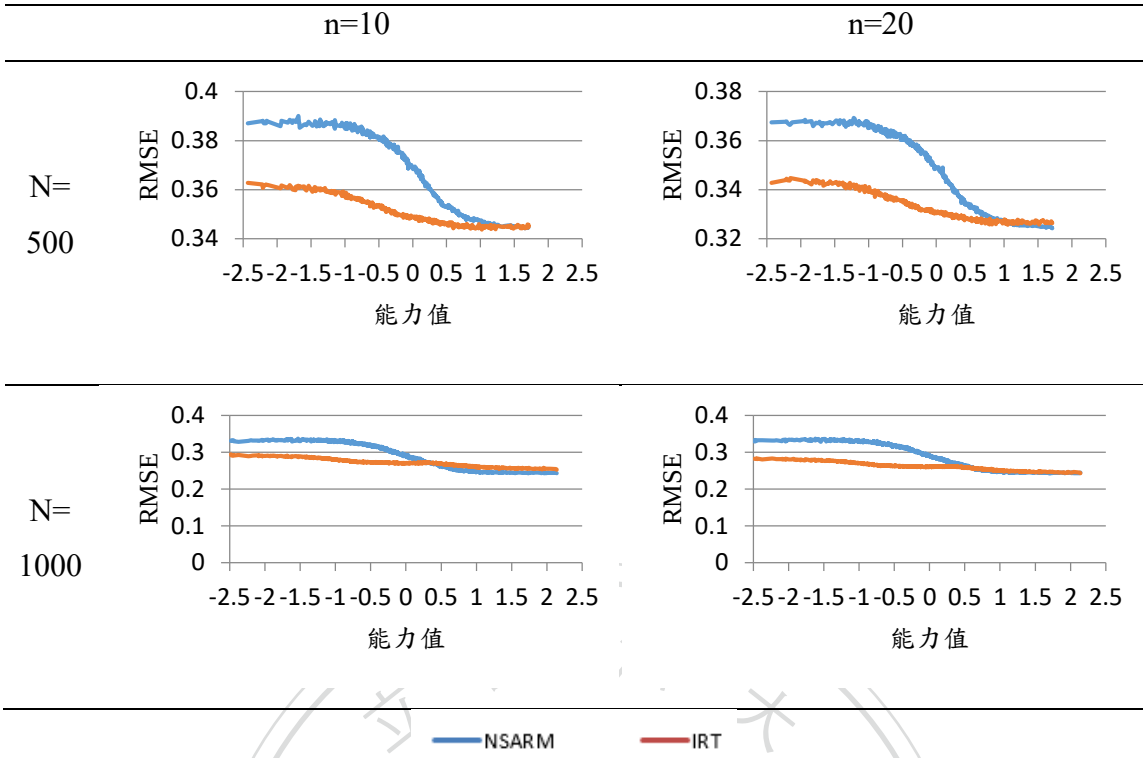


圖4-4 兩種估計法於能力值估計之 RMSE 圖(依困難試題)

觀察圖4-2至圖4-4可發現，依簡單試題進行模型驗證時，IRT 之 RMSE 比 NSARM 之 RMSE 還小，但隨著試題難度的增加，NSARM 之 RMSE 呈現下降的趨勢，且能力越高的受試者，下降的趨勢越明顯，與圖4-1的結果一致。綜合以上結論，能力較好的受試者在進行難度較高的測驗時，可以採用 NSARM 進行參數估計。

上述為兩種模式在能力值估計之 RMSE 的比較，後續我們將探討兩種模式在難度值估計之 RMSE 是否也存在類似的特徵。

在相同試題下，使用該試題每一次的 (共 M 次) 難度估計值與真值計算 RMSE，依此類推，總共會有 n 個 RMSE 值。根據本文所有難度可能值的範圍及簡單、中等、困難三種不同難度值的範圍，兩種模式在難度值估計之 RMSE 的折線圖如下：

DOI:10.6814/NCCU202001433

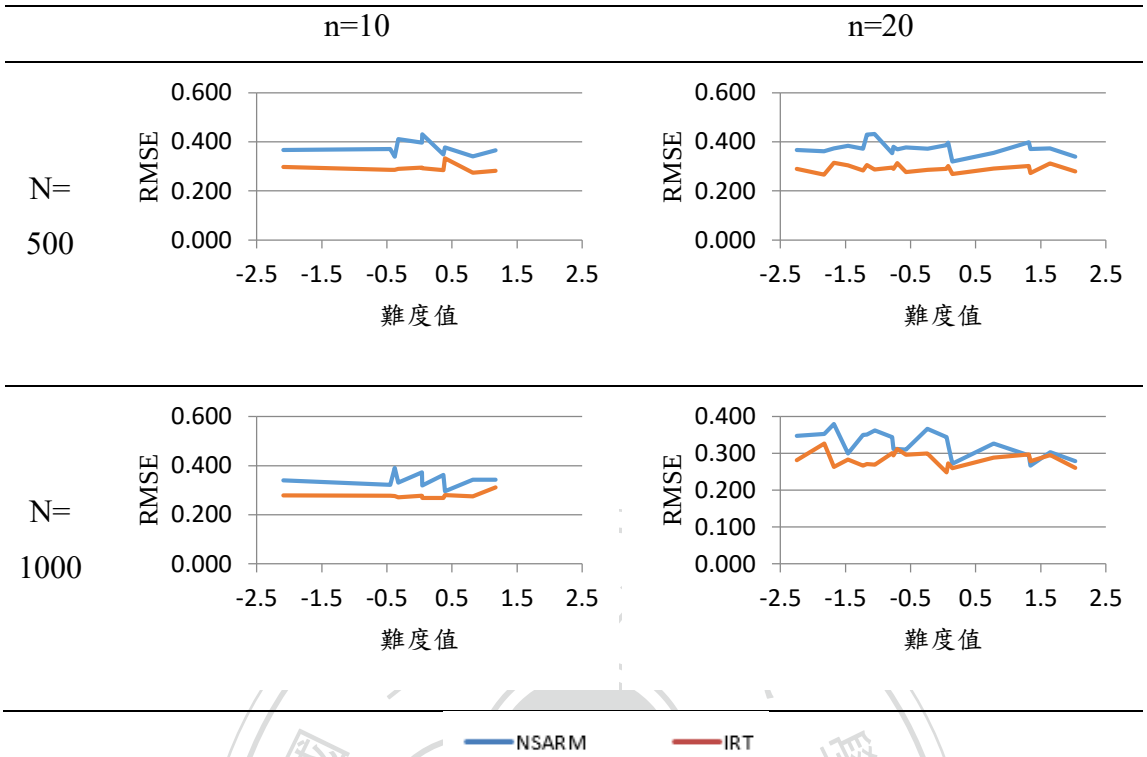


圖4-5 兩種估計法於難度值估計之 RMSE 圖(依所有試題)

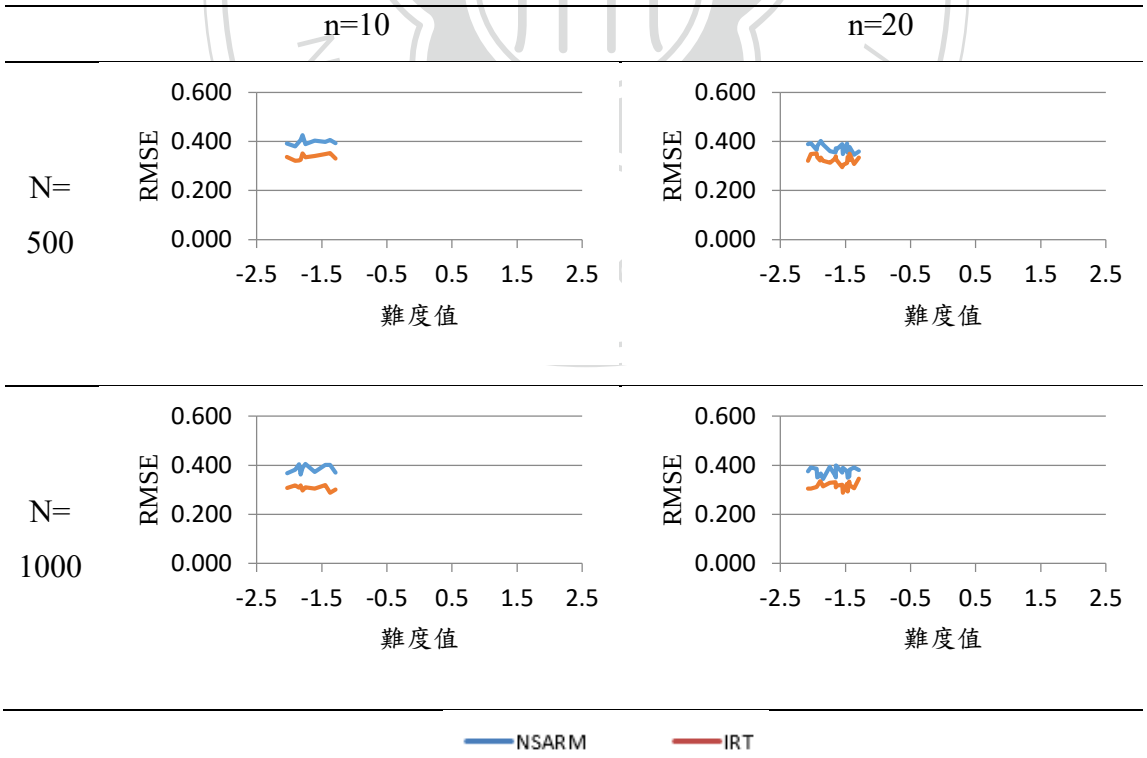


圖4-6 兩種估計法於難度值估計之 RMSE 圖(依簡單試題)

DOI:10.6814/NCCU202001433

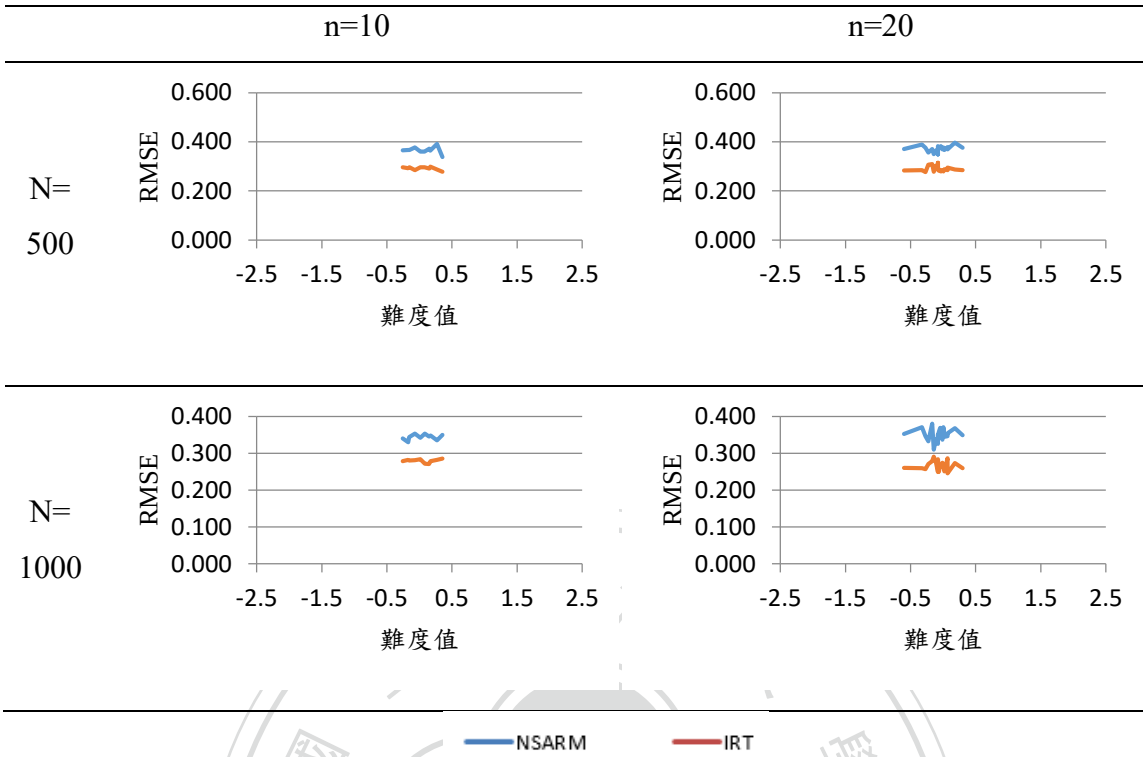


圖4-7 兩種估計法於難度值估計之 RMSE 圖(依中等試題)

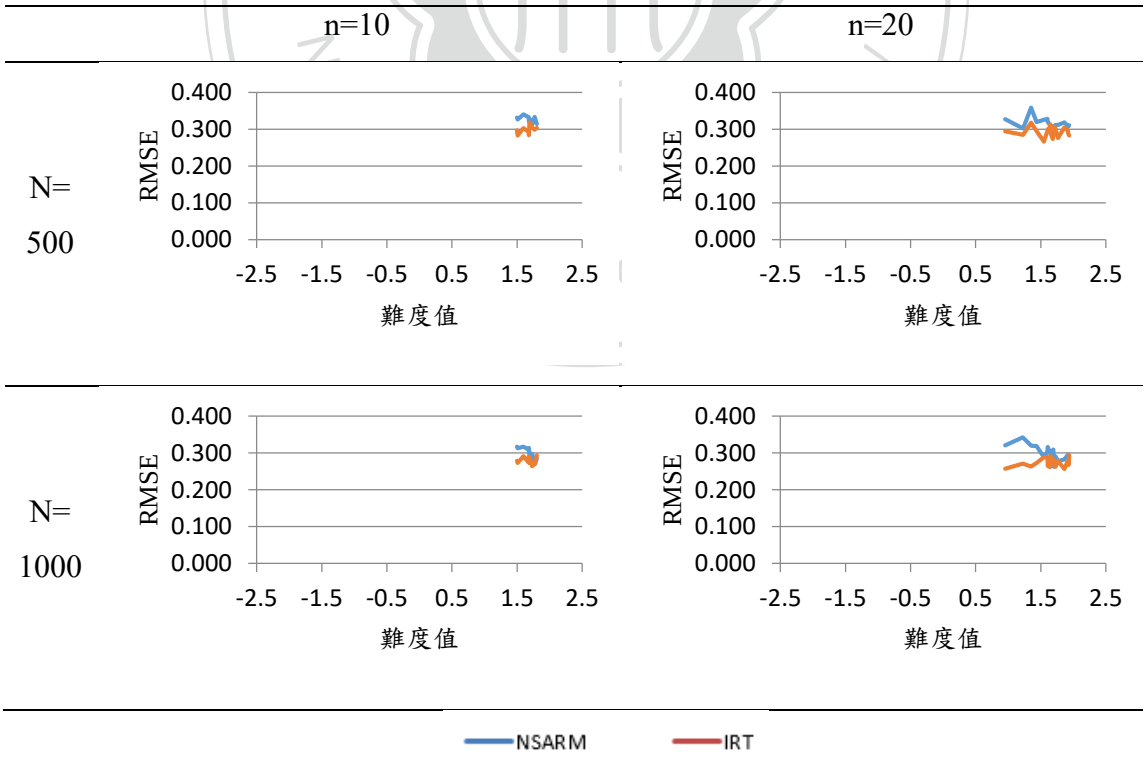


圖4-8 兩種估計法於難度值估計之 RMSE 圖(依困難試題)

DOI:10.6814/NCCU202001433

觀察圖4-5可以發現，受試人數及試題數較多，且部分難度較高的試題，NSARM 之 RMSE 比 IRT 之 RMSE 還小，其餘情形則是 IRT 之 RMSE 呈現出比較好的結果。從圖4-6至圖4-8可發現，在簡單、中等難度的試題時，亦是 IRT 之 RMSE 呈現出比較好的結果，但在困難的試題時，反而是部分 NSARM 之 RMSE 呈現較好的結果。

以上透過符合 NSARM 及 IRT 的模擬資料(即包含作答反應和作答時間)，比較兩種模型的估計誤差。接著將以更貼近一般測驗中，簡單、中等以及困難試題之分配比例，進行題數分布的模擬，以進一步說明使用 NSARM 之合理性。

仿照上述模型驗證的方式，但此次研究變項變更如下。

1. 受試者樣本大小，包括 N = 100、N = 500、N = 1000 位受試者。
2. 試題數，包括 n = 25、n = 50、n = 100 題。

隨機模擬 k_i (字數)、θ_j (受試者能力) 及 b_i (試題難度)。通常設定 θ_j 的平均數為 0、標準差為 1，亦即θ_j~N(0,1) 範圍界於 −2.5~2.5 (余民寧, 2009)；b_i 參數的分配亦為常態分布，平均數為 0、標準差為 0.67，其中標準差的設定則是因為一般測驗的試題難度按照簡單、中等、困難的分配比例約為 1：2：1。

在兩種模式下，將能力值估計之 RMSE 繪製成折線圖如下：

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

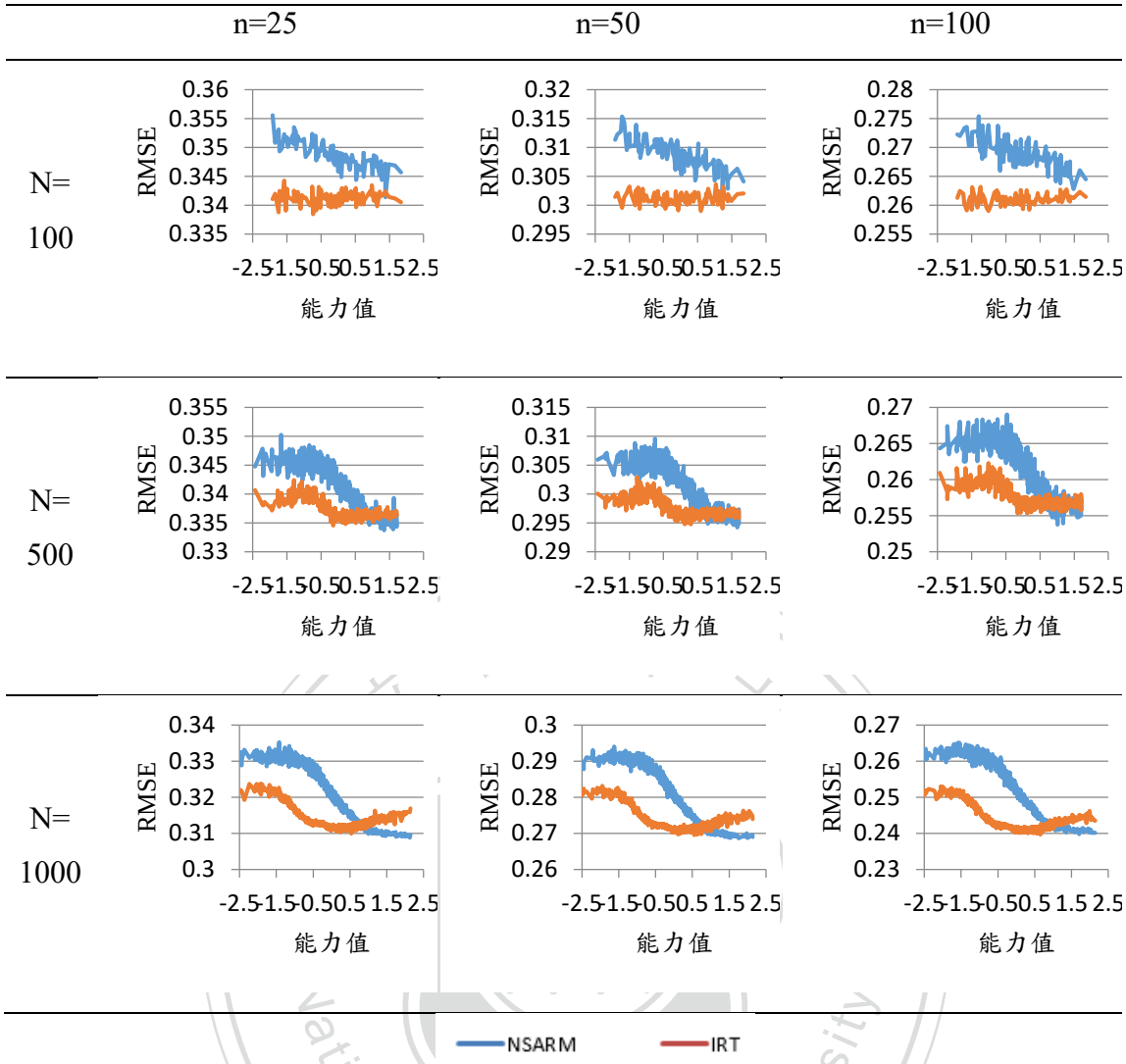


圖4-9 兩種估計法於能力值估計之 RMSE 圖

觀察圖4-9可發現以下兩點：

1. 當試題數固定時，隨著受試人數的增加，NSARM 之 RMSE 呈現下降的趨勢，且能力越高的受試者，下降的趨勢越明顯，甚至當能力值大於一定數值時，NSARM 之 RMSE 比 IRT 之 RMSE 還小。

2.當受試人數固定時，隨著試題數的增加，可以從縱坐標的數值發現，NSARM 之 RMSE 及 IRT 之 RMSE，均有明顯的減少。

綜合以上模擬結果，我們發現無論試題難易度數量的比例分配是否符合一般測驗的要求，只要受試人數夠多的話，採用 NSARM 估計較高的能力值時，其結果會比 IRT 來得更好。後續我們將探討在兩種模式下，難度值估計之 RMSE 是否也存在類似的特徵。

DOI:10.6814/NCCU202001433

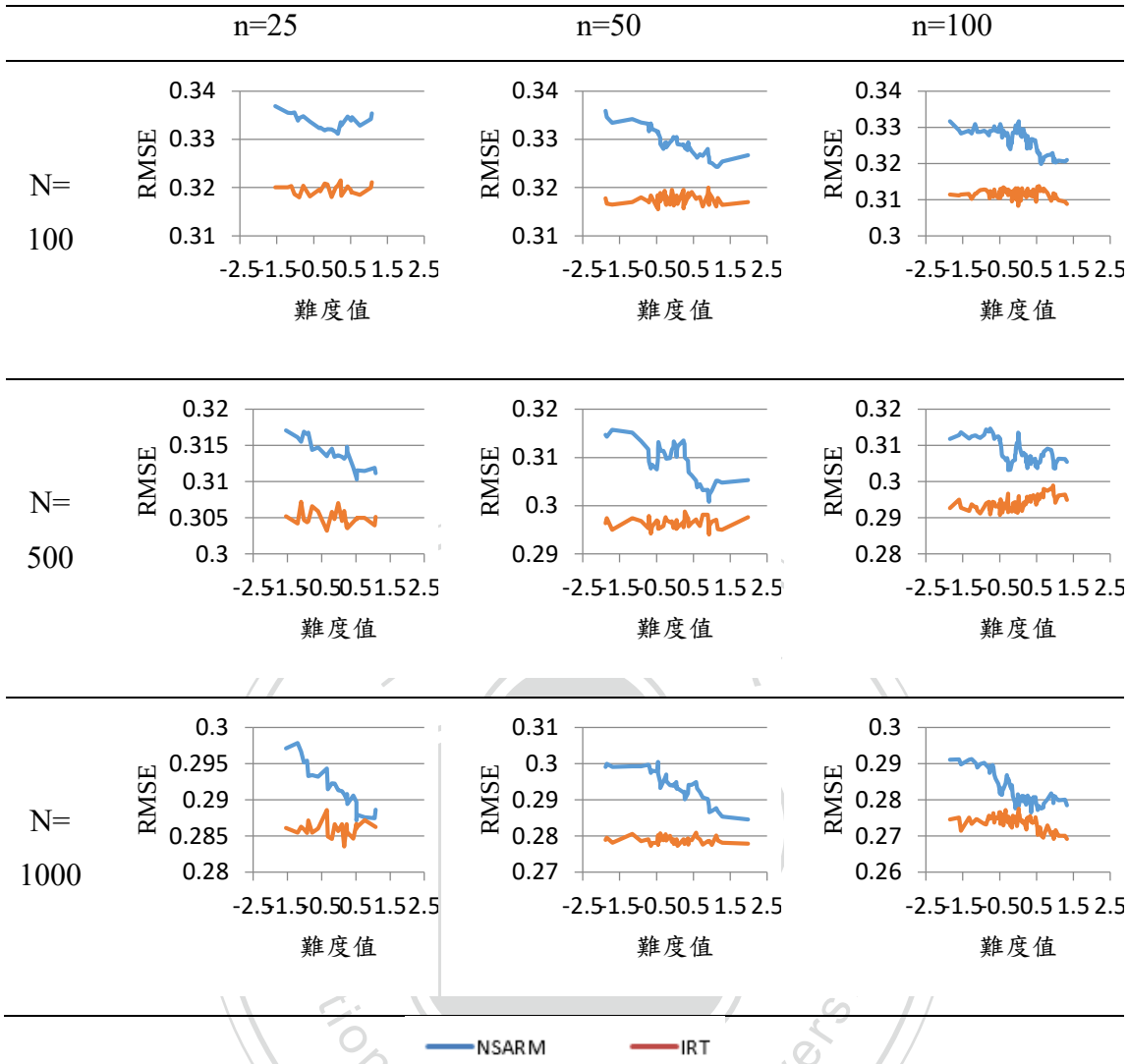


圖4-10 兩種估計法於難度值估計之 RMSE 圖

觀察圖4-10可以發現以下兩點：

1.當試題數固定時，隨著受試人數的增加，可以從縱坐標的數值發現，兩種模式於難度值估計之 RMSE 皆呈現下降的趨勢。

2.當受試人數增加、試題數也增加的情況下，可以發現 NSARM 之 RMSE 有逐漸下降的趨勢。

綜合以上模擬結果，我們發現無論試題難易度數量的比例分配是否符合一般測驗的要求，皆是採用 IRT 模式進行難度值的估計，會呈現較好的結果。

DOI:10.6814/NCCU202001433

第五章　　實證分析

本章利用所獲取的資料，分別依前面所敘述的模型，進行分析與討論，並細分為六節，敘述如下。

第一節　符號負指數時間評分規則之探討

根據第三章所建立符號負指數時間 (SNET) 評分規則 (見公式 3-1)，本節使用：一、比較兩種評分規則與 IRT 能力值的秩。 二、作答時間與能力值之相關性。探討 SNET 評分規則是否適合作為試題的評分方式。

一、　加入作答時間後兩種評分規則之比較

以答對者而言，隨著作答時間的增加，獲得的分數應該要越少，意即解題速度越慢的受試者給予的分數自然就要越少，這部分較無爭議；但是以答錯者而言，隨著作答時間的增加，失去的分數應該要越多還是越少，便會隨著第三者的角度不同而有不同的看法，因此，我們僅針對答錯時以甲 (亦即公式 3-1 中，$x = 0$)、乙兩種評分方式做比較，如下圖。

評分規則甲：$-e^{-\frac{t_{ij}}{c}}$

評分規則乙：$e^{-\frac{t_{ij}}{c}} - 1$

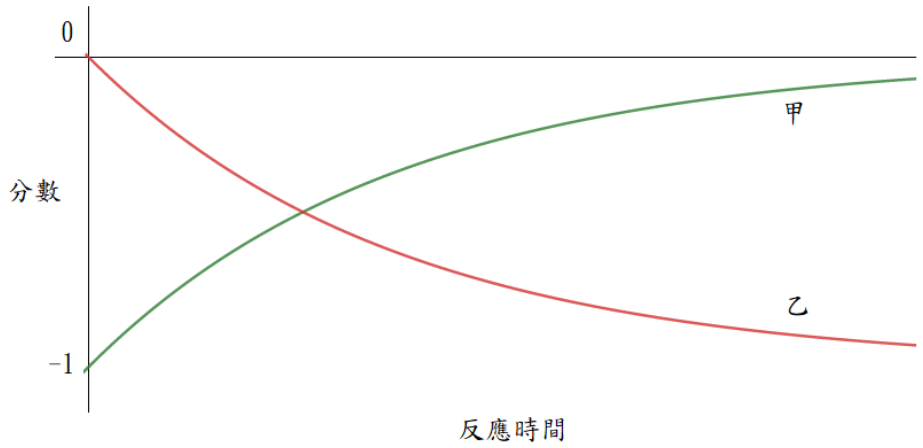


圖5-1 答錯時的兩種評分規則

DOI:10.6814/NCCU202001433

評分規則甲之得分(即得負分)隨著作答時間增加而增加；評分規則乙之得分(即得負分)隨著作答時間增加而減少。我們以「依後測的 IRT 能力值」與「分別使用甲或乙兩種評分規則所得之分數」做比較，判斷何者與 IRT 能力值較為一致。

首先，從難度為簡單的試題中，隨機挑選一題(試題 24)，並從 200 位受試者中篩選出答錯試題 24 的受試者共 37 位，利用甲、乙兩種評分規則分別計算受試者在試題 24 的分數，此時每位受試者在試題 24 皆有甲、乙兩種分數，其次，分別計算這 37 位受試者「依後測的 IRT 能力值」與「使用甲評分規則時試題 24 的得分」之秩相關係數；以及「依後測的 IRT 能力值」與「使用乙評分規則時試題 24 的得分」之秩相關係數。在難度分別為中等或困難的試題中，同樣各隨機挑選一題，依序為試題 16 及試題 26，再依上述方式分別計算秩相關係數後，將表格彙整如下，其中 IRT-θ為試題反應理論所估計之能力值。

表5-1 不同難度下 IRT 能力值與兩種評分規則得分之秩相關係數

	簡單試題-甲規則	簡單試題-乙規則
IRT-θ	0.673	-0.417
	中等試題-甲規則	中等試題-乙規則
IRT-θ	0.747	-0.334
	困難試題-甲規則	困難試題-乙規則
IRT-θ	0.917	-0.215

觀察上表發現，無論何種難度，受試者「依後測之 IRT 能力值」的秩與「甲評分規則之分數」的秩均呈現高度正相關，並且相關性隨著試題難度上升而呈現增加的趨勢。本文作者推測，隨著試題難度的增加，會有更多程度較差的受試者選擇快速猜測以進入下一道試題，但仍然會有部分程度較好的受試者選擇花更多的時間思考如何解題。根據以上結果，相較於乙評分規則，採用甲評分規則作為計分方式更為合理，並且難度較高的試題較能夠根據作答時間辨別受試者能力值的高低。

DOI:10.6814/NCCU202001433

二、 作答時間與能力值之相關性

我們想知道在不同的試題難度下，答對者或答錯者的作答時間與試題反應理論(IRT) 依後測估計之能力值的相關性，因此，先從難度為簡單、中等、困難分別挑選鑑別度較高的三題，其中簡單為試題 17、21、23；中等為試題 16、25、27；困難為試題 20、22、26，以下將依答對者或依答錯者分別探討不同難度下作答時間與 IRT 能力值之相關性。

(一) 答對難度為簡單的試題 (試題 17、21、23)

兩百位受試者中，此 3 題全對的受試者共 77 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 17 (或 21、或 23) 的作答時間」之相關係數，共三組數據彙整如下表，其中 IRT-θ為本文模型依後測所估計之能力值，RT 為作答時間。

表5-2 答對簡單試題下 IRT 能力值與作答時間之相關係數

	試題17-RT	試題21-RT	試題23-RT
IRT-θ	-0.671	-0.713	-0.634

(二) 答對難度為中等之試題(試題 16、25、27)

兩百位受試者中，此 3 題全對之受試者共 31 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 16 (或25、或27) 的作答時間」之相關係數，共三組數據彙整如下表。

表5-3 答對中等試題下 IRT 能力值與作答時間之相關係數

	試題16-RT	試題25-RT	試題27-RT
IRT-θ	-0.721	-0.817	-0.793

(三) 答對難度為困難之試題 (試題 20、22、26)

兩百位受試者中，此 3 題全對之受試者共 19 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 20 (或22、或26)的作答時間」之相關係數，共三組數據彙整如下表。

DOI:10.6814/NCCU202001433

表5-4 答對困難試題下 IRT 能力值與作答時間之相關係數

	試題20-RT	試題22-RT	試題26-RT
IRT-θ	-0.872	-0.911	-0.887

觀察表 5-2 至表 5-4可 發現，無論何種試題難度，在答對的情況下 IRT 能力值與作答時間皆呈現負相關，意即答對者的能力值越高，作答時所花的時間就越少；答對者的能力值越低，作答時所花的時間就越多。同時，隨著試題難度的增加，可發現相關係數越接近 -1，也就是越困難的試題較能說明程度越好的學生，其作答時間會越少。根據推論，答對困難題目的受試者其平均能力值一定比答對簡單題目的受試者來得更高，因此在困難的題目，作答時間較能突顯出這些受試者能力值的高低差異。

(四) 答錯難度為簡單之試題(試題 17、21、23)

兩百位受試者中，此 3 題全錯之受試者共 27 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 17 (或21、或23)的作答時間」之相關係數，共三組數據彙整如下表。。

表5-5 答錯簡單試題下 IRT 能力值與作答時間之相關係數

	試題17-RT	試題21-RT	試題23-RT
IRT-θ	0.587	0.514	0.611

(五) 答錯難度為中等之試題 (試題 16、25、27)

兩百位受試者中，此 3 題全錯之受試者共 46 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 16 (或25、或27) 的作答時間」之相關係數，共三組數據彙整如下表。

表5-6 答錯中等試題下 IRT 能力值與作答時間之相關係數

	試題16-RT	試題25-RT	試題27-RT
IRT-θ	0.699	0.679	0.713

DOI:10.6814/NCCU202001433

(六) 答錯難度為困難之試題 (試題 20、22、26)

兩百位受試者中，此 3 題全錯之受試者共 78 位，我們計算這些受試者「依後測的 IRT 能力值」與其在「試題 20 (或22、或26) 的作答時間」之相關係數，共三組數據彙整如下表。

表5-7 答錯困難試題下 IRT 能力值與作答時間之相關係數

	試題20-RT	試題22-RT	試題26-RT
IRT-θ	0.781	0.744	0.817

觀察表 5-5至表 5-7可發現，無論何種試題難度，在答錯的情況下 IRT 能力值與作答時間皆呈現正相關，意即答錯者的能力值越高，作答時所花的時間就越多；答錯者的能力值越低，作答時所花的時間就越少。同時，隨著試題難度的增加，可發現相關係數越接近 1，也就是越困難的試題較能說明程度越差的學生，其作答時間會越少。根據推論，困難的試題會有更多程度較差的受試者選擇用快速猜測的，但仍然會有程度較好的受試者選擇花更多的時間思考如何解題，既使最後同樣都是答錯，但還是能利用作答時間區分出受試者程度的高低。

綜合以上兩種方式，如果我們考慮的是受試者的能力與試題的難度相當時，受試者勢必會經過一段時間之後才作答，因此，即使想了很久之後還答錯，也不應給予太多懲罰(得負分)，同理，想了很久之後才答對，也不應給予太高的評價(得正分)。

DOI:10.6814/NCCU202001433

第二節 作答反應與作答時間之獨立性

本節將以卡方獨立性檢定 (The Chi-Squared Test of Independence) 檢驗作答反應與作答時間是否獨立。由於假設作答時間服從伽瑪分布，因此在統計了每道試題的平均作答時間和標準差後，利用這些樣本數據推算母體參數，並且比對伽瑪分布假設模型中的支撐集 (support)，發現將所有受試者在各試題的作答時間改以 5 秒為一個單位與本文實際數據最為適配，因此在公式3-1中將 c 值則設定為 5。

首先，在試題難度為：簡單、中等、困難各隨機抽取一題，依序為試題 24、16、26，並統計出各時間區間答對者和答錯者的人數，如下表所示：

表5-8 試題 24 作答反應人數分布

時間區間*	0~2	2~4	4~6	6~8	8~10	總和
答對人數	5	13	12	6	3	39
答錯人數	21	48	39	30	19	157
總和	26	61	51	36	22	196

*時間以 5 秒為一單位

以時間區間為 0~2 之答對人數為例，在作答時間和作答反應獨立假設下。

$$P(\text{時間區間}0\sim2 \cap \text{答對人數}) = P(\text{時間}0\sim2\text{單位})P(\text{答對人數}) = \frac{26}{196}\frac{39}{196}$$

$$= \frac{5.173}{196}$$

期望人數$= P(\text{時間}0\sim2\text{單位} \cap \text{答對人數}) \times \text{總人數} = \frac{5.173}{196} \times 196 = 5.173$

同理，在作答時間和作答反應獨立假設下，其餘各項期望人數如下表所示：

表5-9 試題 24 作答反應期望人數

時間區間*	0~2	2~4	4~6	6~8	8~10
答對人數	5.173	12.138	10.148	7.163	4.378
答錯人數	20.827	48.862	40.852	28.837	17.622

*時間以 5 秒為一單位

DOI:10.6814/NCCU202001433

根據表 5-8、5-9，計算卡方值如下：

$$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}=\frac{(5-5.173)^2}{5.173}+\frac{(13-12.138)^2}{12.138}+\frac{(12-10.148)^2}{10.148}$$
$$+\frac{(6-7.163)^2}{7.163}+\frac{(3-4.378)^2}{4.378}+\frac{(21-20.827)^2}{20.82}+\frac{(48-48.862)^2}{48.862}$$
$$+\frac{(39-40.852)^2}{40.852}+\frac{(30-28.837)^2}{28.837}+\frac{(19-17.622)^2}{17.622}=1.283$$

其中O_i為觀察次數，E_i為期望次數。對獨立性之檢定，在顯著水準為0.05的情況下，因 $\chi^2=1.283<\chi^2_{4,0.05}=9.488$，故試題24的作答時間與作答反應無關。同樣，將試題16、26依各作答時間區間內的人數分布統計如下表。

表5-10 試題16作答反應人數分布

時間區間*	2~4	4~6	6~8	8~10	10~12	12~14	14~16	總和
答對人數	2	10	15	28	21	11	5	92
答錯人數	3	11	16	31	24	15	8	108
總和	5	21	31	59	45	26	13	200

*時間以 5 秒為一單位

表5-11 試題26作答反應人數分布

時間區間*	6~8	8~10	10~12	12~14	14~16	16~18	18~20	20~22	22~24	24~26	總和
答對人數	1	6	13	21	45	34	25	12	6	2	165
答錯人數	0	1	2	4	9	7	5	2	0	0	30
總和	1	7	15	25	54	41	30	14	6	2	195

*時間以5秒為一單位

根據表5-10、5-11，計算χ^2 =0.6643和χ^2 =5.2551，依序小於$\chi^2_6=12.5916$和$\chi^2_9=16.919$。因此，本文所得資料的作答時間與作答反應獨立。

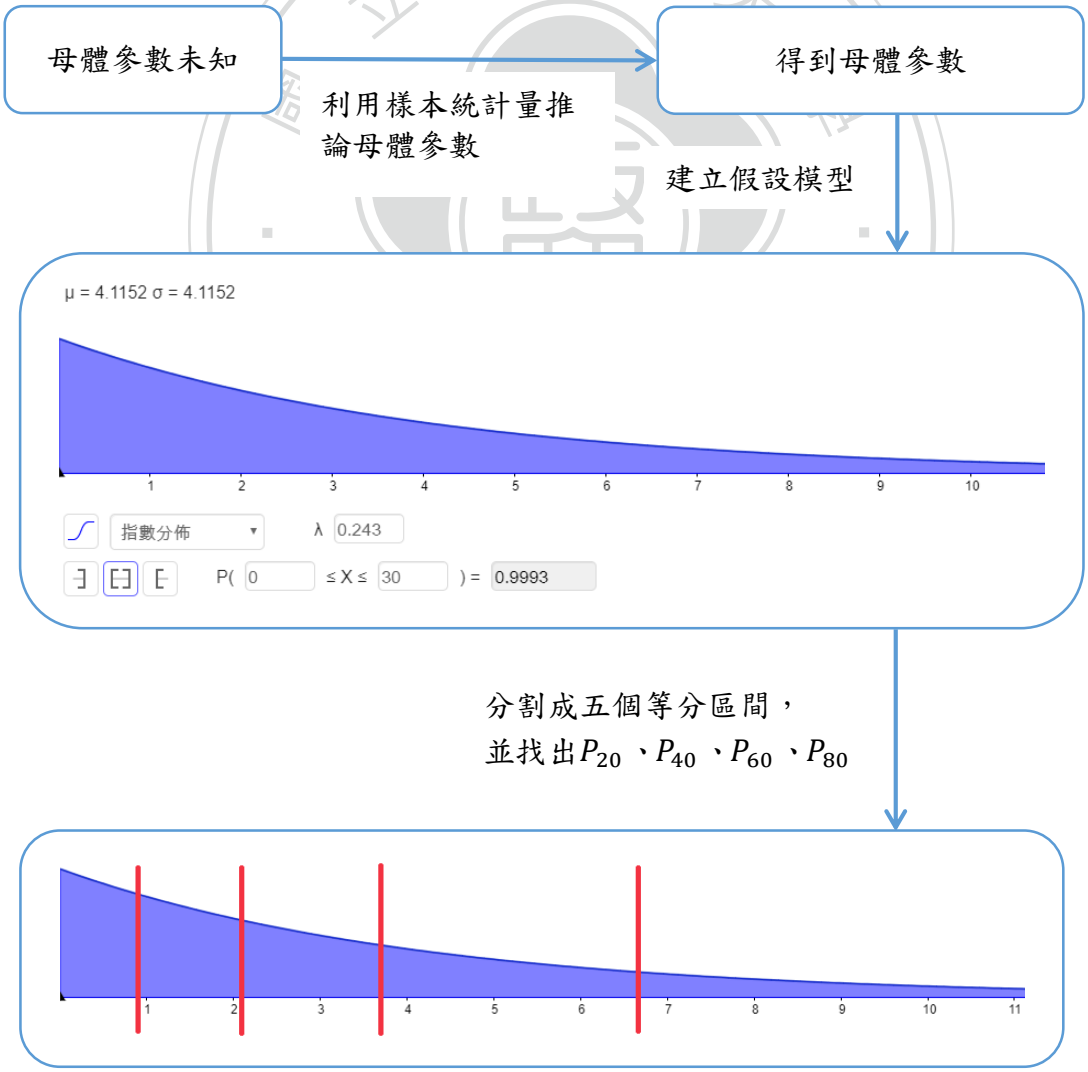
DOI:10.6814/NCCU202001433

第三節 作答時間符合伽瑪分布

以下先從本研究實得數據中，分析受試者作答時間分布是否服從指數分布；再進一步探討作答時間是否適合修改為更一般化的伽瑪分布。

一、 以適合度檢定檢驗作答時間是否服從指數分布

當作答時間服從指數分布，即 $f(t) = De^{-Dt}$，由於我們無法知道指數分布的母體參數 D，因此先利用樣本資料(即作答時間)的平均去估計母體參數，並建立指數分布的假設模型，再根據假設模型找出第二十、四十、六十、八十百分位數 (P_{20}、P_{40}、P_{60}、P_{80})，接著依照這四個百分位數將時間分成五個等分區間，如此每個區間的期望人數便會相同，最後便可以每個區間的觀察人數與期望人數進行適合度檢定，檢定流程圖整理如下。



DOI:10.6814/NCCU202001433

以試題24為例，先利用樣本資料推論母體參數，由於指數分布之平均μ和標準差σ分別為$\begin{cases}\mu=\frac{1}{D}\\ \sigma=\frac{1}{D}\end{cases}$，故$\begin{cases}D=\frac{1}{\mu}\\ D=\frac{1}{\sigma}\end{cases}$，此時可分別由平均作答時間或標準差推論母體參數 D 值，如下表所示：

表5-12 D 值的兩種估計方式 (依試題24)

總人數	平均作答時間*	標準差	依平均數推論	依標準差推論
200	4.117	0.891	0.243	1.122

*時間以 5 秒為一單位

由於母體參數有兩種可能值，以下便分成兩種情況進行適合度檢定。

(一) 依平均作答時間推論 D 值為 0.243

當 D 值已知後，便可建立假設模型找出$P_{20}=0.929$、$P_{40}=2.099$、$P_{60}=3.776$、$P_{80}=6.634$，如下表以試題 24 為例。

表5-13 觀察人數與期望人數分布表(依平均數推論 D 值)

時間區間*	0~0.929	0.929~2.099	2.099~3.776	3.776~6.634	6.634~15	總人數
觀察人數	8	12	98	77	5	200
期望人數	40	40	40	40	40	200

*時間以 5 秒為一單位

利用表5-13，計算卡方值如下：

$$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}=194.15$$

其中O_i為觀察人數，E_i為期望人數。對適合度之檢定，在顯著水準為0.05的情況下，檢定結果如下：$\chi^2=194.15>\chi^2_{4,0.05}=9.488$，因此有證據拒絕受試者在試題 24 的作答時間服從指數分布。

DOI:10.6814/NCCU202001433

(二) 依標準差推論 D 值為1.122

當 D 值已知後，便可建立假設模型找出$P_{20} = 0.201$、$P_{40} = 0.457$、$P_{60} = 0.820$、$P_{80} = 1.440$，如下表以試題 24 為例。

表5-14 觀察人數與期望人數分布表(依標準差推論 D 值)

時間區間*	0~0.201	0.201~0.457	0.457~0.820	0.820~1.440	1.440~7	總人數
觀察人數	0	1	3	11	185	200
期望人數	40	40	40	40	40	200

*時間以 5 秒為一單位

利用表5-14，計算卡方值如下：

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = 658.9$$

其中O_i為觀察人數，E_i為期望人數。對適合度之檢定，在顯著水準為0.05的情況下，適合度檢定結果如下：$\chi^2 = 658.9 > \chi^2_{4,0.05} = 9.488$，因此有證據拒絕受試者在試題 24 的作答時間服從指數分布。

綜合以上甲、乙兩點，無論是以平均作答時間推論母體參數進行適合度檢定，亦或是以標準差推論母體參數進行適合度檢定，試題 24 的作答時間並不適合指數分布。

以上我們是對所有的受試者進行檢定，接著我們改以 D 值(以本文模型估計能力值與難度值之差)相同者進行檢定，觀察結果是否會有不同，為避免在相同 D 值下的受試者人數過少，因此將 D 值以3<D<4、4<D<5、5<D<6、6<D<7四個區間作分類，並統計受試者之 D 值落在該區間的人數，如下表以試題24為例。

DOI:10.6814/NCCU202001433

表5-15 不同區間下 D 值的兩種估計方式

D 值區間	總人數	平均作答時間*	標準差	依平均數推論	依標準差推論
3<D<4	0				
4<D<5	13	5.556	1.108	0.180	0.903
5<D<6	98	4.383	0.905	0.228	1.105
6<D<7	89	3.714	0.756	0.269	1.323

*時間以 5 秒為一單位

根據上表，在5<D<6及6<D<7的總人數皆大於 30 人，因此便以此二區間為例，同時，由於母體參數有兩種可能值，以下便分成四種情況進行適合度檢定。

(一) 5<D<6時，依平均作答時間推論 D 值為0.228

當 D 值已知後，便可建立假設模型找出$P_{20} = 0.987$、$P_{40} = 2.266$、$P_{60} = 4.063$、$P_{80} = 7.076$，如下表以試題 24 為例。

表5-16 觀察人數與期望人數分布表 (5<D<6且依平均數推論 D 值)

時間區間*	0~0.987	0.987~2.266	2.266~4.063	4.063~7.076	7.076~15	總人數
觀察人數	4	7	48	37	2	98
期望人數	19.6	19.6	19.6	19.6	19.6	98

*時間以 5 秒為一單位

適合度檢定結果如下：$\chi^2 = 92.918 > \chi^2_{4,0.05} = 9.488$，因此有足夠證據拒絕受試者在試題 24 的作答時間服從指數分布。

(二) 5<D<6時，依標準差推論 D 值為1.105

當 D 值已知後，便可建立假設模型找出$P_{20} = 0.205$、$P_{40} = 0.460$、$P_{60} = 0.832$、$P_{80} = 1.456$，如下表以試題 24 為例。

DOI:10.6814/NCCU202001433

表5-17 觀察人數與期望人數分布表 (5<D<6且依標準差推論 D 值)

時間區間*	0~0.205	0.205~0.460	0.460~0.832	0.832~1.456	1.456~15	總人數
觀察人數	0	0	1	6	91	98
期望人數	19.6	19.6	19.6	19.6	19.6	98

*時間以5秒為一單位

適合度檢定結果如下：$\chi^2 = 326.388 > \chi^2_{4,0.05} = 9.488$，因此有足夠證據拒絕受試者在試題 24 的作答時間服從指數分布。

(三) 6<D<7時，依平均作答時間推論 D 值為0.269

當 D 值已知後，便可建立假設模型找出$P_{20} = 0.872$、$P_{40} = 1.924$、$P_{60} = 3.407$、$P_{80} = 5.999$，如下表以試題24為例。

表5-18 觀察人數與期望人數分布表 (6<D<7且依平均數推論 D 值)

時間區間*	0~0.872	0.872~1.924	1.924~3.407	3.407~5.999	5.999~15	總人數
觀察人數	3	6	43	34	3	89
期望人數	17.8	17.8	17.8	17.8	17.8	89

*時間以5秒為一單位

適合度檢定結果如下：$\chi^2 = 82.854 > \chi^2_{4,0.05} = 9.488$，因此有足夠證據拒絕受試者在試題 24 的作答時間服從指數分布。

(四) 6<D<7時，依標準差推論 D 值為1.323

當 D 值已知後，便可建立假設模型找出$P_{20} = 0.171$、$P_{40} = 0.393$、$P_{60} = 0.695$、$P_{80} = 1.219$，如下表以試題 24 為例。

DOI:10.6814/NCCU202001433

表5-19 觀察人數與期望人數分布表 (6<D<7且依標準差推論 D 值)

時間區間*	0~0.171	0.171~0.393	0.393~0.695	0.695~1.219	1.219~15	總人數
觀察人數	0	0	0	5	84	89
期望人數	17.8	17.8	17.8	17.8	17.8	89

*時間以5秒為一單位

適合度檢定結果如下：$\chi^2 = 308.809 > \chi^2_{4,0.05} = 9.488$，因此有足夠證據拒絕受試者在試題24的作答時間服從指數分布。

綜合以上甲、乙、丙、丁四點，既使以 D 值落在相同區間的受試者進行檢定，同樣可以拒絕受試者在試題 24 的作答時間服從指數分布。

二、 以適合度檢定檢驗作答時間是否服從伽瑪分布

利用樣本資料檢定母體是否符合伽瑪分布，由於我們無法知道母體參數，因此先利用實際資料估計 k 及 D 之理論值並建立伽瑪分布的假設模型，同樣將 D 值以3<D<4、4<D<5、5<D<6、6<D<7四個區間作分類，並統計受試者之 D 值落在該區間的總人數，如下表以試題 24 為例。

表5-20 不同區間下 D 值的兩種估計方式

D 值區間	總人數	平均作答時間*	標準差	k 理論值	D 理論值
3<D<4	0				
4<D<5	13	5.556	1.108	25.141	4.525
5<D<6	98	4.383	0.905	23.438	5.348
6<D<7	89	3.714	0.756	24.117	6.494

*時間以5秒為一單位

觀察上表，在 5<D<6 及 6<D<7 的總人數皆大於 30 人，因此便以此二區間為例。由於伽瑪分布之平均μ和標準差 σ 分別為$\begin{cases} \mu = \frac{k}{D} \\ \sigma = \frac{\sqrt{k}}{D} \end{cases}$，故$\begin{cases} k = \frac{\mu^2}{\sigma^2} \\ D = \frac{\mu}{\sigma^2} \end{cases}$。

DOI:10.6814/NCCU202001433

便可透過當 5<D<6 時之平均作答時間及標準差推算母體參數 k=23.438 及 D=5.348 並建立伽瑪分布的假設模型找出 $P_{20}=3.568$、P_{40}=4.040、P_{60} =4.484、P_{80}=5.056；以及當 6<D<7 時之平均作答時間及標準差推算母體參數 k=24.117及 D=6.494 找出 $P_{20}=2.830$、P_{40}=3.227、P_{60} =3.589、4.045。統計各時間區間的實際人數以及期望人數，如下表所示：

表5-21 簡單試題之觀察人數與期望人數分布表

時間區間*	0~3.568	3.568~4.040	4.040~4.484	4.484~5.056	5.056~8	總人數
5<D<6之觀察人數	16	18	24	22	18	98
5<D<6之期望人數	19.6	19.6	19.6	19.6	19.6	98
時間區間*	0~2.83	2.830~3.227	3.227~3.589	3.589~4.045	4.045~7	總人數
6<D<7之觀察人數	14	16	22	19	18	89
6<D<7之期望人數	17.8	17.8	17.8	17.8	17.8	89

*時間以5秒為一單位

根據上表，計算卡方值如下：

當5<D<6，$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}=2.204$

當6<D<7，$\chi^2=\sum_{i=1}^{n}\frac{(O_i-E_i)^2}{E_i}=2.067$

其中O_i為觀察次數，E_i為期望次數。對適合度之檢定，在顯著水準為0.05的情況下，適合度檢定結果如下： 5<D<6時，$\chi^2=2.204<\chi^2_{4,0.05}=9.488$；當6<D<7時，$\chi^2$ =2.067< $\chi^2_{4,0.05}=9.488$。據檢定結果，不拒絕試題24的作答時間服從伽瑪分布。以相同方式檢定難度中等(試題16)以及難度困難(試題26)，統計如下表。

DOI:10.6814/NCCU202001433

表5-22 中等試題之觀察人數與期望人數分布表

時間區間*	0~9.209	9.209~10.042	10.042~10.801	10.801~11.738	11.738~17	總人數
4<D<5之觀察人數	10	13	18	16	13	70
4<D<5之期望人數	14	14	14	14	14	70
時間區間*	0~9.209	9.209~10.042	10.042~10.801	10.801~11.738	11.738~17	總人數
5<D<6之觀察人數	20	22	28	25	24	119
5<D<6之期望人數	23.8	23.8	23.8	23.8	23.8	119

*時間以5 秒為一單位

表5-23 困難試題之觀察人數與期望人數分布表

時間區間*	0~10.04	10.04~11.25	10.04~12.39	12.39~13.8	13.8~21	總人數
3<D<4之觀察人數	15	17	23	20	19	94
3<D<4之期望人數	18.8	18.8	18.8	18.8	18.8	94
時間區間*	0~13.69	13.69~14.76	14.76~15.76	15.76~17.02	17.02~23	總人數
4<D<5之觀察人數	15	17	25	20	20	97
4<D<5之期望人數	19.4	19.4	19.4	19.4	19.4	97

*時間以 5 秒為一單位

根據上表，計算試題 16 當4<D<5、5<D<6時以及試題26當3<D<4、4<D<5時之卡方值依序為2.714、1.546、1.957、2.948，皆小於 $\chi_4^2 = 9.488$，同樣不拒絕受試者的作答時間服從伽瑪分布。因此在篩選的試題中，經由適合度檢定，我們發現部分試題的作答時間並不適合指數分布，宜改用較一般化的伽瑪分布。同時，根據我們也發現試題24、16、26這三道試題字數的理論值與估計結果顯示相關係數為0.861，符合第三章作答時間模型的前提假設。

DOI:10.6814/NCCU202001433

第四節 不同模型下試題難度值相關性之探討

本節將根據所有受試者在全部 30 題的作答反應，分成下列二種方式來分析與討論，首先，依全部 30 題探討各模式彼此間的試題難度值之相關性；其次，分別探討各模式「依前測(共 15 題)的試題難度值」與「依後測(共 15 題)的試題難度值」之相關性。

一、 不同模型下試題(依全部 30 題)難度值相關性之比較

我們先比較本文模型(NSARM)與試題反應理論(IRT)、古典測驗理論(CTT)之難度值的相關性。為避免試題已被受試者做過或是受試者在使用線上測驗時因其他因素未能正常完成，因此將受試者在各試題的作答時間偏離中位數超過四分位距之 1.5 倍的離群值淘汰後，以 IRT、NSARM 兩種模式估計 30 道試題的難度值，並且和 CTT、大考中心做比較，整理如下表，其中 NSARM-b 表示本文模型的難度值，IRT-b 表示試題反應理論的難度值，CTT-b 表示古典測驗理論的難度值(以“1－答對率“表示)，大考中心-b 表示由大考中心所公布之答對率換算成難度值(以“1－答對率“表示)。

表5-24 各模式之試題難度值

試題	NSARM-b	IRT-b	CTT-b	大考中心-b
1	-0.23	-0.32	0.45	0.46
2	-1.27	-1.76	0.21	0.28
3	1.50	2.08	0.73	0.76
4	1.04	1.44	0.67	0.68
5	1.32	1.84	0.75	0.73
6	-2.02	-2.81	0.16	0.15
7	1.73	2.40	0.81	0.80
8	-1.09	-1.52	0.33	0.31
9	-1.73	-2.40	0.19	0.20
10	-0.46	-0.64	0.41	0.42
11	2.02	2.84	0.86	0.85
12	0.46	0.64	0.55	0.58

DOI:10.6814/NCCU202001433

13	0.17	0.24	0.56	0.53
14	-0.81	-1.12	0.39	0.36
15	-1.56	-2.16	0.24	0.23
16	-0.19	-0.33	0.44	0.46
17	-1.27	-1.75	0.23	0.28
18	1.50	2.01	0.72	0.76
19	1.02	1.39	0.68	0.68
20	1.32	1.83	0.73	0.73
21	-2.02	-2.77	0.18	0.15
22	1.73	2.29	0.80	0.80
23	-1.09	-1.59	0.34	0.31
24	-1.66	-2.35	0.17	0.20
25	-0.42	-0.64	0.39	0.42
26	2.30	2.75	0.85	0.85
27	0.52	0.62	0.57	0.58
28	0.27	0.31	0.61	0.53
29	-0.73	-1.12	0.41	0.36
30	-1.51	-2.18	0.25	0.23

根據上表，NSARM 的難度值在 -2.02 至 2.30 之間，IRT 的難度值在 -2.81 至 2.84 之間，整份測驗的試題如同大考中心的難度值，從簡單至困難的試題皆有。雖然各模式在計算難度值的方式皆不相同，但四種模式的難度值排序應當一致。另外，常見的三種模式 (NSARM、IRT、CTT) 難度值之間的相關係數如下表所示。

表5-25 不同模式間試題難度值的相關係數(依全部 30 題)

	CTT-b	IRT-b	NSARM-b
CTT-b			
IRT-b	0.8519		
NSARM-b	0.6771	0.6988	

觀察上表，NSARM 的試題難度值分別與 IRT 及 CTT 的試題難度值探討相

DOI:10.6814/NCCU202001433

關性，它們的相關係數分別為 0.6988 及 0.6771，顯示皆具有中度正相關。更為明確表示當試題愈困難時，NSARM-b 同樣愈高。而對於 CTT-b 與 IRT-b 值而言，兩者的相關係數為 0.8519，顯示具有高度正相關。

二、 不同模型下試題(依前測與依後測)難度值相關性之比較

由於所蒐集的前、後測，受試者受測時間點相差一年左右，因此受試者在這段時間各方面的能力值均應該有所提升，且前、後測依序挑選符合他們當下正在學習的數學概念之類似試題。我們以 NSARM 依前測與依後側進行參數估計，並以題號順序將難度值繪製如下圖 5-2。以相同方式可得 IRT 難度值如下圖 5-3。

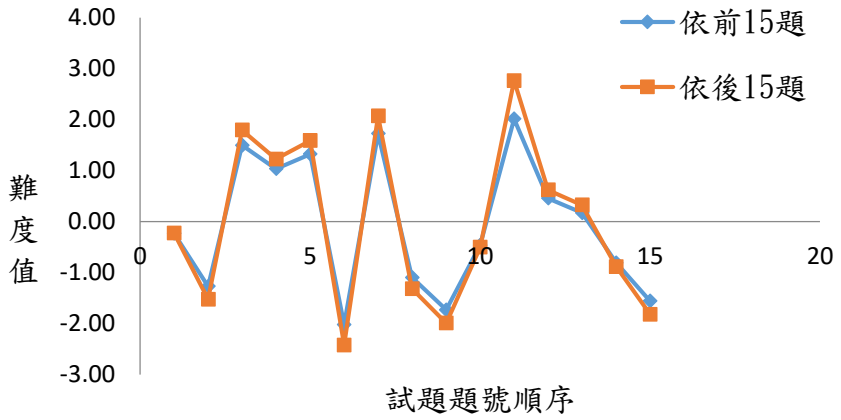


圖5-2 前、後測之 NSARM 難度值

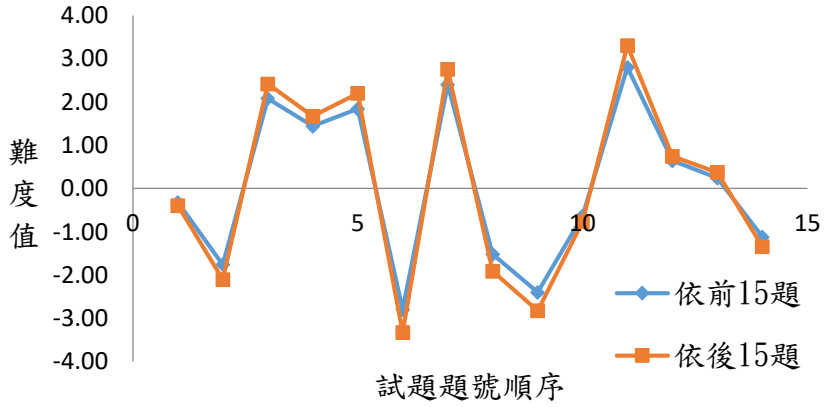


圖5-3 前、後測之 IRT 難度值

DOI:10.6814/NCCU202001433

由上圖可知，不論是 NSARM 或是 IRT，在前、後測相對應試題的難度估計值均非常相似。以下分別計算兩者之相關係數。

表5-26 兩種模式前、後測難度值間之相關係數

	前測之 NSARM-b	前測之 IRT-b
後測之 NSARM-b	0.7431	
後測之 IRT-b		0.8081

由上表可知，不管是 NSARM 還是 IRT，依前測的試題難度值與依後測的試題難度值，均為高度正相關，但是相關性並不如預期的高。本文作者推測，由於前、後測並不是在相似的時間點之下所完成的，因此會受到重測信度(test-retest reliability)的大小所影響，重測信度公式如下。

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

其中，x和y分別是同一受試者在兩次測驗的分數，$\bar{x}$和$\bar{y}$分別為兩次測驗的平均分數，所有測驗的重測信度都介於 0 和 1 之間。計算兩次測驗的重測信度為 0.7975，受試者可能因為兩次施測的環境不同，而產生測量誤差。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

第五節 NSARM 與 IRT 單參數模型在能力值的比較

在上一節，分析了 NSARM 在試題難度估計上與其餘模式皆為正相關。本節以下兩種方式比較 NSARM 及 IRT 在能力估計上的差異：首先依後測探討 NSARM 能力值的秩與 IRT 能力值的秩之相關性；其次，分別探討 NSARM 及 IRT「依前測之能力值」的秩與「依後測之能力值」的秩之相關性。

一、 NSARM 與 IRT 在能力值(依後測)的比較

為了探討本文模型 (NSARM) 及試題反應理論 (IRT) 在受試者能力估計上的相關性，透過計算兩模式所估計之能力值的秩相關係數 (Spearman's rank correlation coefficient)以便進一步探討。先把樣本數為 n 個的原始數據 X_i,Y_i 轉換成秩 (rank) 數據 x_i, y_i，再計算兩模式能力值的秩相關係數，公式如下：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

其中，$d_i = x_i - y_i$。

在下表，IRT-θ表示試題反應理論之受試者能力值，NSARM-θ表示本文模型之受試者能力值，並且依大小順序排名。

表5-27 IRT 及 NSARM 之受試者能力值及排名

受試者	IRT-θ	排名	NSARM-θ	排名	受試者	IRT-θ	排名	NSARM-θ	排名
1	1.047	36	1.126	37	101	1.054	34	0.818	34
2	-0.806	185	-0.839	185	102	0.210	104	0.249	104
3	1.197	27	0.786	25	103	0.438	89	0.342	89
4	0.832	56	0.814	57	104	-0.019	126	-0.249	126
5	0.856	55	0.886	54	105	1.005	41	0.942	41
6	0.190	106	0.069	106	106	0.514	86	0.776	86
7	1.003	43	0.753	43	107	1.259	18	1.094	18
8	0.345	96	0.408	96	108	-0.817	186	-0.970	186
9	0.854	54	0.928	55	109	0.995	49	1.221	47
10	-0.649	175	-0.731	175	110	0.967	52	0.916	52

DOI:10.6814/NCCU202001433

11	1.000	44	0.834	44	111	-0.339	156	-0.355	155
12	-0.197	144	-0.307	144	112	-0.352	151	-0.535	156
13	0.141	112	-0.132	111	113	-0.583	171	-0.203	170
14	0.011	121	0.163	121	114	1.198	24	1.405	24
15	1.031	39	1.070	39	115	-0.245	155	-0.252	151
16	1.234	21	0.860	21	116	-0.423	161	-0.289	161
17	-1.237	198	-1.219	198	117	1.237	19	1.206	19
18	1.049	37	0.844	36	118	0.984	49	1.076	49
19	0.094	115	0.214	115	119	1.402	9	1.621	9
20	-0.839	188	-0.962	188	120	-0.011	125	0.313	125
21	-0.746	180	-0.792	179	121	-0.073	130	-0.215	130
22	0.543	81	0.918	81	122	-0.217	147	-0.423	147
23	1.423	4	1.199	4	123	0.689	69	0.786	69
24	0.128	110	-0.028	112	124	1.411	7	1.410	7
25	1.411	6	1.181	6	125	1.385	13	1.436	13
26	1.403	8	1.524	8	126	-0.590	174	-0.629	171
27	-0.008	124	0.669	124	127	0.535	82	0.637	82
28	-0.710	176	-0.634	176	128	-0.636	171	-0.626	174
29	0.051	118	-0.058	118	129	-0.723	178	-0.465	177
30	-0.320	153	-0.253	153	130	-0.182	142	-0.347	142
31	1.400	10	1.371	10	131	-0.533	164	-0.170	164
32	-0.223	141	-0.134	149	132	0.157	111	-0.009	110
33	-0.181	149	0.286	141	133	-0.732	177	-0.650	178
34	-0.561	168	-0.548	168	134	0.172	109	0.089	109
35	-1.021	194	-1.177	194	135	1.016	40	0.865	40
36	1.051	35	1.319	35	136	0.697	68	0.525	68
37	1.169	27	0.856	27	137	-0.787	183	-0.571	183
38	0.806	59	0.495	59	138	-0.025	127	-0.218	127
39	-0.978	190	-1.000	190	139	0.978	50	0.720	50
40	-0.312	152	-0.672	152	140	-0.984	192	-0.871	192
41	0.178	108	0.159	108	141	0.736	63	0.814	63
42	0.719	65	0.951	65	142	0.408	91	0.872	91

DOI:10.6814/NCCU202001433

43	1.234	20	1.363	20	143	-0.379	160	-0.586	158
44	0.238	102	-0.193	102	144	-0.594	172	-0.588	172
45	0.852	56	0.642	56	145	-1.272	199	-1.279	199
46	1.092	31	1.334	31	146	-0.945	189	-1.134	189
47	0.991	48	1.246	48	147	1.421	5	1.550	5
48	0.303	99	0.840	98	148	0.501	87	0.502	87
49	0.546	79	0.911	79	149	0.498	88	0.225	88
50	0.269	98	-0.090	99	150	-0.210	146	0.382	146
51	0.112	114	-0.010	114	151	0.399	93	0.331	93
52	0.409	90	0.216	90	152	0.219	103	0.280	103
53	0.706	67	0.678	67	153	-0.430	161	-0.363	162
54	1.168	28	1.196	29	154	1.276	15	1.338	16
55	-1.408	200	-1.310	200	155	0.533	83	0.683	83
56	1.210	24	1.547	23	156	1.265	17	1.045	17
57	0.545	79	0.471	80	157	1.122	30	1.112	30
58	1.005	42	1.257	42	158	-1.184	197	-1.359	197
59	1.391	10	1.253	11	159	0.808	58	0.667	58
60	1.038	38	1.105	38	160	0.195	105	0.101	105
61	0.707	67	1.262	66	161	0.682	71	0.691	71
62	-0.375	157	-0.487	157	162	-0.140	135	-0.222	135
63	0.587	77	0.330	77	163	0.012	120	-0.005	120
64	-1.038	196	-1.130	196	164	0.077	117	0.009	117
65	-0.066	129	-0.299	129	165	-0.120	132	-0.159	132
66	-0.786	182	-1.012	182	166	1.059	34	0.838	33
67	0.244	100	0.171	102	167	-0.535	165	-0.294	165
68	0.402	92	0.689	92	168	-0.323	154	-0.279	154
69	0.684	70	0.821	70	169	-0.829	187	-0.874	187
70	-0.005	123	0.318	123	170	0.862	53	0.971	53
71	1.191	26	1.144	26	171	1.075	32	1.143	32
72	0.976	51	1.069	51	172	1.213	22	1.107	22
73	-0.512	163	-0.291	163	173	-0.405	162	-0.287	160
74	0.521	84	0.616	85	174	0.600	73	0.567	75

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

75	-0.171	137	-0.221	138	175	0.647	78	0.358	73
76	-0.388	160	-0.666	159	176	0.996	46	1.254	46
77	0.785	61	0.540	61	177	0.564	75	0.574	78
78	1.169	28	1.533	27	178	1.281	15	1.140	15
79	-0.772	181	-0.768	182	179	-0.092	137	0.071	131
80	0.241	101	0.241	102	180	-0.142	131	-0.057	137
81	0.399	94	0.055	94	181	0.322	97	0.538	97
82	-0.125	133	-0.100	134	182	0.044	116	-0.072	119
83	-1.018	191	-1.148	193	183	-0.560	167	-0.442	167
84	-0.980	193	-0.739	191	184	0.087	119	0.071	116
85	1.432	2	1.659	2	185	0.181	107	0.385	107
86	0.522	84	0.746	84	186	0.126	113	0.188	113
87	1.316	14	1.189	14	187	-0.191	148	-0.157	143
88	0.679	72	0.969	72	188	0.392	95	0.514	95
89	0.740	62	0.965	62	189	-0.574	169	-0.658	169
90	-0.124	134	0.061	133	190	-0.788	184	-0.881	184
91	-0.242	150	-0.453	150	191	-0.218	140	0.047	148
92	-0.199	145	0.154	145	192	-0.174	139	-0.367	139
93	0.727	64	0.794	64	193	0.786	59	0.679	60
94	1.442	1	1.051	1	194	-0.141	136	0.115	136
95	0.010	122	0.195	122	195	-0.595	172	-0.001	173
96	1.390	12	1.498	12	196	-0.766	181	-0.463	180
97	0.588	77	0.587	76	197	-0.547	167	-0.877	166
98	0.998	46	0.905	45	198	-0.178	143	-0.126	140
99	-1.025	196	-0.812	195	199	1.427	4	1.342	3
100	0.642	75	0.484	74	200	-0.035	128	0.079	128

根據上表之排序，透過 Excel 計算兩模式能力值之秩相關係數為0.9671，表示兩模型估計的受試者能力值在排名上有很大程度的關聯。

DOI:10.6814/NCCU202001433

二、 NSARM 與 IRT 在能力值(依前測與依後測)的一致性

在上文探討了 NSARM 及 IRT 兩種模型之能力估計值的秩依後測上的相關性，接著我們想了解 NSARM 及 IRT 在能力估計上是否具有一致性(依前測與依後測)。首先，以 IRT 模式分別依前測以及依後測估計受試者能力值並將結果繪製如下所示：

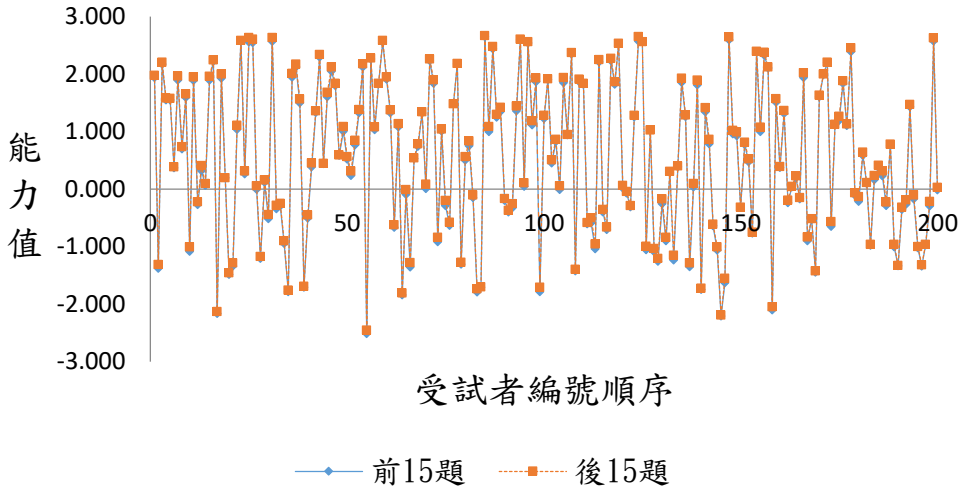


圖5-4 前、後測之 IRT 能力值

觀察圖表可發現，IRT 模式在依前測與依後測所估計之受試者能力值並無明顯差異。接著以 NSARM 模式分別依前測以及依後測估計受試者能力值並將結果繪製如下所示：

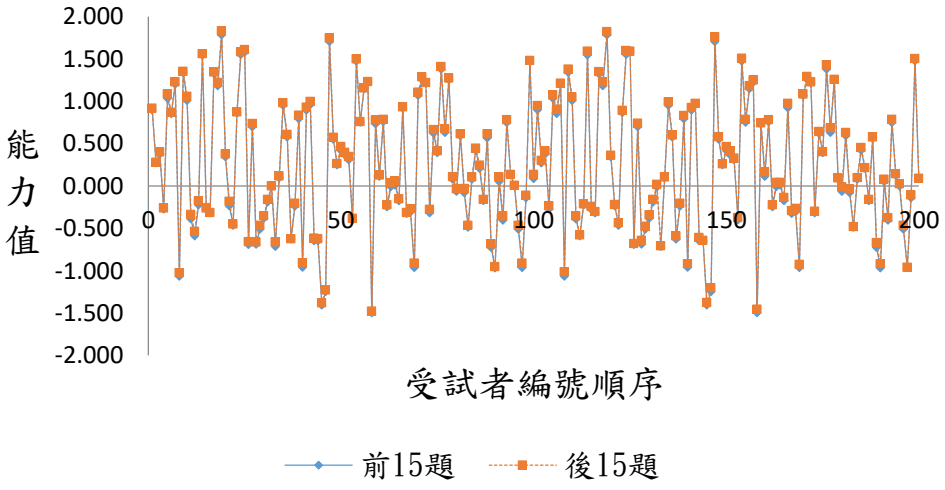


圖5-5 前、後測之 NSARM 能力值

DOI:10.6814/NCCU202001433

由上圖可發現，NSARM 模式依前測與依後測所估計之受試者能力值同樣並無明顯差異。因此進一步計算兩種模式依前測與依後測所估計受試者能力值之秩相關係數如下表所示，

表5-28 兩種模式前、後測能力值間之秩相關係數

	前測之 NSARM-θ	前測之 IRT-θ
後測之 NSARM-θ	0.791	
後測之 IRT-θ		0.884

由上表可知，不管是 NSARM 還是 IRT，依前測之能力值的秩與依後測之能力值的秩均為高度正相關，簡言之，無論是 IRT 亦或是 NSARM 所估計之能力值，在兩份試題類型相似的測驗具有高度一致性的關係，但是兩次測驗同樣會受到施測時間、環境、人為因素等影響，數據準確性還有待探究。

國立政治大學 National Chengchi University

DOI:10.6814/NCCU202001433

第六節 作答反應組型相同時受試者能力值之探究

本節先比較 CTT 測驗總分與 SNET 總分(依後測)是否有顯著差異，以便進一步判斷加入作答時間後對能力估計的影響；其次，受試者在相同試題反應組型下，是否可利用 NSARM 進一步區分受試者能力值。

一、 CTT 測驗總分與 SNET 總分(依後測)之差異性

由於同一群受試者使用兩種不同計分方式屬於成對樣本，故以成對 t 檢定進行檢定，方式如下：

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

第 j 對樣本值令為(x_{1j}, x_{2j})，$j=1 \ldots N$，計算各試題檢定統計量。

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

其中，$d_j = x_{1j} - x_{2j}$、$\bar{d} = \frac{\sum d_j}{n}$、$s_d = \sqrt{\frac{n\sum d_j^2 - (\sum d_j)^2}{n(n-1)}}$

利用 Excel 的 T.TEST 函數在顯著水準為 0.05 的情況下，檢定受試者在後測的傳統測驗總分與 SNET 評分規則總分，兩者的平均總分是否有差異。

表5-29 後測試題總分之成對 t 檢定表

	CTT 測驗總分	SNET 總分
平均數	7.79	1.135
變異數	5.785	10.280
觀察值個數	200	200
皮耳森相關係數	0.990	
自由度	199	
t 統計	105.693	
P(T<=t) 單尾	4.3893E-177	
臨界值：單尾	1.653	
P(T<=t) 雙尾	8.7786E-177	
臨界值：雙尾	1.972	

DOI:10.6814/NCCU202001433

觀察上表，雙尾 P 值小於0.05，故我們拒絕H_0，表示該群學生的 CTT 測驗總分與 SNET 總分，兩者平均分數有顯著差異。同時，依簡單、中等、困難三種難度之試題分別檢驗平均總分是否有差異，得結果如表5-30至表5-32。

表5-30 簡單試題總分之成對 t 檢定表

	CTT 測驗總分	SNET 總分
平均數	3.845	2.20387038
變異數	1.257261307	3.388044529
觀察值個數	200	200
皮耳森相關係數	0.999244087	
自由度	199	
t 統計	32.16545376	
P(T<=t) 單尾	4.48932E-81	
臨界值：單尾	1.652546746	
P(T<=t) 雙尾	8.97865E-81	
臨界值：雙尾	1.971956544	

表5-31中等試題總分之成對 t 檢定表

	傳統測驗總分	SNET 總分
平均數	2.58	0.131444041
變異數	1.651859296	2.566202336
觀察值個數	200	200
皮耳森相關係數	0.99712809	
自由度	199	
t 統計	103.4156384	
P(T<=t) 單尾	3.1007E-175	
臨界值：單尾	1.652546746	
P(T<=t) 雙尾	6.2013E-175	
臨界值：雙尾	1.971956544	

DOI:10.6814/NCCU202001433

表5-32 困難試題總分之成對 t 檢定表

	傳統測驗總分	SNET 總分
平均數	1.365	-1.200646391
變異數	1.12741206	1.280163348
觀察值個數	200	200
皮耳森相關係數	0.995665191	
自由度	199	
t 統計	293.6631901	
P(T<=t) 單尾	9.6752E-265	
臨界值：單尾	1.652546746	
P(T<=t) 雙尾	1.935E-264	
臨界值：雙尾	1.971956544	

根據表5-30至5-32，在雙尾 P 值同樣皆小於 0.05 情況下，表示不管是總體而言，亦或是在不同難度的部分試題下，兩種評分方式，其平均分數皆有明顯差異。

二、 利用 NSARM 進一步區分受試者能力值

本章第一節驗證了受試者在答對相同的試題時，作答時間與 IRT 能力值呈現負相關；在答錯相同的試題時，作答時間與 IRT 能力值呈現正相關，因此我們分別以答對相同試題之受試者；以及答錯相同試題之受試者做分析。首先，從 5 題簡單的試題中抽選 3 題並篩選出此 3 題全對(及此3 題全錯)的受試者，接著比較「依此 3 題的 NSARM 能力值」與「依後測的 NSARM 能力值」。中等及困難的試題同樣以此方式比較，彙整如以下兩點。

(一) 答對相同試題

難度為簡單之試題(試題17、21、23)皆答對之受試者共 77 位，計算「依此 3 題的 NSARM 能力值」與「依後測的 NSARM 能力值」兩者的秩相關係數。答對中等試題共 31 位及答對困難的試題共 19 位，同樣以此方式計算秩相關係數，三組結果彙整如下。

DOI:10.6814/NCCU202001433

表5-33 不同難度下試題全對時的 NSARM 能力值與其後測能力值之秩相關係數

	依簡單試題 NSARM-θ	依中等試題 NSARM-θ	依困難試題 NSARM-θ
依後測 NSARM-θ	0.813	0.911	0.974

觀察上表之結果，依各難度之試題全對時 NSARM-θ的秩與依後測 NSARM-θ的秩均呈現高度正相關，且隨著試題難度的增加相關性越接近 1。

(二) 答錯相同試題

答錯簡單試題共 27 位；答錯中等試題共 31 位；答錯困難的試題共 19 位。使用相同方式計算兩者之秩相關係數並整理如下表。

表5-34 不同難度下試題全錯時的 NSARM 能力值與其後測能力值之秩相關係數

	依簡單試題 NSARM-θ	依中等試題 NSARM-θ	依困難試題 NSARM-θ
依後測 NSARM-θ	0.791	0.834	0.897

觀察上表之結果，依個別難度之試題全錯時 NSARM-θ的秩與依後測 NSARM-θ的秩呈現高度正相關，且隨著試題難度的增加相關性越接近 1。

根據以上兩點之結果，無論是全部答對還是全部答錯，亦或是不同難度試題，受試者在 CTT 之測驗總分相同，甚至 IRT 能力值也相同的情況下，採用 NSARM，我們不僅可經由這幾道試題進一步區分受試者的能力值，而且其秩與依全部試題所估計之能力值的秩呈現高度的正相關。

綜合本章分析之結果，在一般的情況下，無論是使用 CTT、IRT 或是 NSARM 做為主要的測驗依據，估計結果差異並不大，但是在作答反應組型相同時的特殊情況下，我們便能將受試者的作答時間作為其能力估計的一項因素，進一步區分出受試者的能力值之高低，並且越困難的題目，越適合採用 NSARM 進行參數估計。

DOI:10.6814/NCCU202001433

第六章　　結論與建議

本研究目的在探討評分規則的合理性、作答時間長度的分布情形、作答時間的長度與透過古典測驗理論或試題反應理論所估算之受試者能力的相關程度以及作答反應組型相同者(依答對、答錯)的能力值與作答時間長度之相關性。依據第四章的模型驗證及第五章的分析，整理出如下結論。

一、　本文模型是否合理？

在第四章中，我們透過符合 NSARM 的模擬資料(即包含作答反應和作答時間)，分別在 IRT 及 NSARM 的模式下估計能力值和難度值，並計算其 RMSE，可得如下之結論：

(一) 隨著受試人數增加，兩種模式的能力及難度估計值之 RMSE，均呈現下降的趨勢。

(二) 在受試人數較多的情況下，能力值較高的受試者，其 NSARM 的能力估計值之 RMSE 小於 IRT 的能力估計值之 RMSE。

因此透過模擬數據進行模型驗證，在估計受試者能力值時，與 IRT 模式比較，我們發現本文所建立之模型對高能力群之受試者，有較佳的估計精準度，且受試者人數越多，精準度越好。

二、　對於答錯者，本文所訂定之評分規則是否合理？

本文評分規則(SNET)對於受試者答錯某一題之得分 (即負分)，係再依該題作答時間之長或短，分別給予高 (即負分少) 或低 (即負分多) 之分數，透過實際資料，經由5.1節的分析，可得如下之結論：

(一) 對於答錯者，「依後測的 IRT 能力值」與「使用本文評分規則的得分」呈現正相關，並且相關性隨著試題難度上升而呈現增加的趨勢。

(二) 對於答錯者，「依後測的 IRT 能力值」與「作答時間」呈現正相關，並且相關性隨著試題難度上升而呈現增加的趨勢。

依據以上兩點之相關性，可以顯示 SNET 評分規則之合理性。

DOI:10.6814/NCCU202001433

三、　受試者在每一題的作答反應與作答時間是否獨立？

在5.2節中，我們從簡單、中等或困難之試題中，各隨機抽取一題進行獨立性檢定，皆顯示作答反應與作答時間互相獨立。

四、　是否適合將作答時間模型從指數分布修改為更一般化的伽瑪分布？

在篩選的試題中，透過直方圖發現一般受試者作答時間並非文獻中所指的指數分布，因此在5.3節我們從簡單、中等或困難之試題中，各隨機抽取一題，對該題各受試者的作答時間是否為指數分布，作適合度檢定，發現各題皆拒絕虛無假設。然而若將分布修改為比指數分布更一般化的伽瑪分布，再作適合度檢定，可發現各題皆不再拒絕虛無假設，顯示作答時間宜改用較一般化的伽瑪分布。

五、　在答對(或答錯)之試題相同的情況下，探討本文模型能否進一步區分受試者能力的高低？

在5.6節中，無論是古典測驗理論或是試題反應理論，皆無法從作答反應組型相同的試題區分受試者的能力值，但採用本文加入作答時間之試題反應模型後，我們不僅可經由這幾道試題進一步區分受試者的能力值，而且其秩與依全部試題所估計能力值的秩，呈現高度的正相關。

綜合本文之研究結果，歸納出以下兩點：

(一) 依全部試題所得之 IRT 能力值、CTT 之測驗分數或是 NSARM 所估計能力值的秩，兩兩之間皆呈現高度的正相關，顯示 NSARM 為合理之模型。

(二) 當受試者在部分試題的作答反應組型相同時，僅依這些試題所得之 IRT 能力值，或 CTT 所得之測驗總分，皆相同；而 NSARM 僅依這些試題仍可進一步區別受試者能力值的高低，且合理。因此，透過 NSARM，可以讓受試者作較少的試題，便能有效的達到區別受試者能力之目的。

DOI:10.6814/NCCU202001433

根據本文之研究結果，我們提出未來可繼續研究的建議如下：

1. NSARM 的參數僅涉及試題難度與受試者能力值，因此本文採用 IRT 單參數模式與 NSARM 進行比較，但是現實中的測驗，每道試題的鑑別度並不會相同。因此建議未來研究者可以採用 IRT 雙參數模式，甚至 IRT 三參數模式與 NSARM 進行比較。

2. 本文第四章，以難度參數為常態分布，平均數為 0、標準差為 0.67 進行數據的模擬，其中標準差的設定則是因為一般測驗的試題難度按照簡單、中等、困難的分配比例約為 1：2：1。根據圖4-9可發現，IRT 能力值之 RMSE 在中間較低且兩端較高，是否與試題數為 1：2：1的分配有相關，可再進一步探究。另外，從圖4-9亦可發現 NSARM 並不受到試題數為 1：2：1的影響，同樣有待讀者探究。

3. 根據本文第四章的分析結果，大部分的情況下，IRT 模式之 RMSE 比較小，但是隨著樣本數的增加，NSARM 模式之 RMSE 呈現下降的趨勢。因此，可進一步探究兩種模式之 RMSE，差距的合理範圍。

4. 本文於第五章所分析的實證資料，利用適合度的檢定結果，拒絕一般常用的指數分布作為作答時間的模型假設，而應採用較一般的伽瑪分布，即 $T \sim \Gamma(k, D)$。未來可研究，若真實的作答時間模型為伽瑪分布，但卻用特殊的指數分布作估計，則估計值誤差將增加多少。亦即作答反應時間為非指數分布 (即$k \neq 1$) 時，但 NSARM 若以指數分佈 (即$k = 1$) 時的公式，去估計能力值 (θ_j) 及難度值 (b_i) 時，他們的誤差會增加多少。

DOI:10.6814/NCCU202001433

參考文獻

中文文獻

(1) 余民寧 (2009)。試題反應理論(IRT)及其應用。臺北：心理出版社股份有限公司。

(2) 陳俊宏、陳淑英 (2010)。試題作答時間在電腦適性測驗上的應用---改善具洩題資訊受試者的能力估計精確性。測驗學刊第五十七輯第四期 P.459~483。

(3) 陳茂雄 (2017)。作答時間與答題機率關係的探討~以小數的乘法單元為例。國立台中教育大學，教育資訊與測驗統計研究所。

(4) 黃聖筠、陳淑英 (2009)。「試題作答時間」在洩題偵測上的應用。測驗學刊第五十六輯第四期 P.543~571。

(5) 張雅媛 (2007)。融合 kernel smoothing 之 MMLE 法於 IRT 參數估計之應用。國立台中教育大學，教育測驗統計研究所。

(6) 郭伯臣、謝典佑、吳慧珉、林佳樺 (2012)。一因子高層次試題反應理論模式之評估。測驗學刊第五十九輯第三期 P.329~348。

英文文獻

(1) Allen, W. J. and Yen, W. M. (2001). *Introduction to measure theory* (2nd ed.). Monterey, CA: Brooks/ Cole.

(2) Bock, R. D. (1972). *Estimating item parameters and latent ability when responses are scored in two or more nominal categories*. Psychometrika, 37, 29–51.

(3) Birnbaum, A. (1968). *Some latent trait models and their use in inferring an examinee's ability*. In: Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, 397-479.

(4) Casella, G., and Berger, L. (1990). *Statistical inference*. Pacific Grove, CA: Brooks/Cole.

(5) Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

(6) de la Torre, J. and Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. Applied Psychological Measurement 33 (8), 620-639.

(7) Dennis, I. and Evans, J. (1996). *The speed-error trade-off problem in*

DOI:10.6814/NCCU202001433

psychometric testing. British Journal of Psychology, 87, 105–129.

(8) Ferrando, P. J., and Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement*, 31, 525–543.

(9) Gaviria, J. L. (2005). *Increase in precision when estimating parameters in computer assisted testing using response times*. Quality & Quantity, 39, 45–69.

(10) Guion, R. M. and Ironson, G. H. (1983). Latent trait theory for organizational research. *Organizational Behavior and Human Performance*, 31, 54-87.

(11) Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph, No. 7.*

(12) Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

(13) Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

(14) Maris, G. and Van der Maas, H. L. (2012). *Speed-Accuracy Response Models: Scoring Rules based on Response Time and Accuracy*. Psychometrika, 77(4), 615–633.

(15) Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago:University of Chicago Press.

(16) Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam and R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North Holland.

(17) Thissen, D. (1983). *Timed testing: An approach using item response theory*. In D. J. Weiss(Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing.(pp. 179–203) New York: Academic Press.

(18) Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, 2, 249–254.

(19) Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

(20) Van der Linden, W. J. (2007). *A hierarchical framework for modeling speed and accuracy on test items*. Psychometrika,72, 287–308.

(21) Van der Linden, W. J. (2009). *Conceptual Issues in Response-Time Modeling*. Journal of Educational Measurement, Vol. 46, No. 3, pp. 247–272.

(22) Van der Maas, H. L. and Wagenmakers, E. J. (2005). *A psychometric analysis of chess expertise*. The American Journal of Psychology, 118(1), 29–60.

DOI:10.6814/NCCU202001433

(23) Verhelst, N. D., Verstralen, H. H. F. M., and Jansen, M. G. (1997). A logistic model for timelimit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). New York: Springer.

(24) Wang, T., and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29, 323–339.

(25) Wickelgren, W. (1977). *Speed-accuracy tradeoff and information processing dynamics*. Acta Psychologica, 41, 67–85.

國立政治大學
National Chengchi University

DOI:10.6814/NCCU202001433