行政院國家科學委員會專題研究計畫成果報告

傳統檢定試題偏差(DIF)方法的改良與分析

A smoothing approach for detecting differential item functioning (DIF): refinements of IRT area method

計畫編號：NSC 89-2118-M-018 -003
執行期限：88 **年** 8 **月** 1 **日至** 89 **年** 7 **月** 31 **日**

主持人：李信宏　　彰化師範大學數學系
E'mail: li@math.ncue.edu.tw

一、中文摘要

試題偏差是指試題本身包含了某些對於特定群體的受試者（例如：男性或者女性，原住民與非原住民）有利的因素，使得這些特定群體有較顯著的優勢來答對此種試題。在試題反應理論的架構之下，試題偏差可以用兩組不相等的試題特徵曲線來表示，所以傳統上，這兩組試題特徵曲線之間的差異或者是面積大小，便可以當作是檢定試題偏差的一種指標。然而，這種指標一方面受制於試題反應理論模式的選擇，另一方面缺乏適當的統計檢定理論為後盾，所以並非適用。 我們所提出的計畫，是利用 kernel smoothing **方法來估計這兩組試題特徵曲線，以避免模式選擇和適用的問題；然後藉由隨機化檢定** (randomization test) **來求導試題特徵曲線之間面積的經驗分配，並得到適當的統計量來檢定試題偏差的存在與否。計畫中，我們並將利用模擬研究來瞭解此新方法的性質，這包括了型一錯誤和檢定力的研究，我們預期能夠改善傳統面積檢查法在檢定試題偏差的缺失。**

關鍵詞：試題偏差、試題反應理論、試題特徵曲線、試題反應理論模式、隨機化檢定

Abstract

Item bias or differential item functioning (DIF) means that for two or more individuals, each with the same latent proficiency or latent ability level, the probability of answering a item correctly is not equal from individual to individual or group to group. In item response theory (IRT) terms, DIF can be represented by two non-identical item characteristic curves (ICC). Therefore, the difference or area between the two ICC's is often used as convenient measures of DIF. However, these measures are descriptive and the importance of significance tests for these measures is clear. The proposed procedure first utilizes a kernel smoothing technique to estimate the ICC's. A hypothesis-testing statistic of the area between two ICC's then is derived through randomization test method. To investigate the performance of our procedure, a large-scale simulation study will be carried out. The simulation includes Type I error study as well as power study. Also, the comparisons of the behavior of our procedure with other DIF detection procedures are analyzed and carefully discussed. Our new approach is expected to be a good refinement of the IRT-area-based methods for assessing DIF.

Keywords: Item Bias, Differential Item Functioning (DIF), Item Response Theory (IRT), Item Characteristic Curve (ICC), Kernel Smoothing, Randomization Test

二、計畫緣由與目的

在試題作答理論中，根據試題特徵曲線的不變性，不論受試團體為何，一個試題應只有一條試題特徵曲線。假設將想要研究的、作答表現可能不似一般受試者預期表現的團體，稱為焦點組(focal group)，而將用來與焦點組比較、作答表現如一般受試者表現的團體，稱為對照組(reference group)。那麼如果針對同一試題，焦點組和對照組的試題特徵曲線不同，則稱此試題具有 DIF。Rudner (1977) 提出以判斷兩估計試題特徵曲線間的面積作為檢定 DIF 的方法，稱之為面積法，當兩估計試題特徵曲線間的面積越小，則該試題具有 DIF 的可能性就越小；反之，則具有 DIF 的可能性就越大。以下為 Rudner 所使用的面積公式：

$$A = \int | P_R(\theta) - P_F(\theta) | w(\theta) \, d\theta ,$$

在這裡 $P_F(\theta)$代表焦點組對於該試題所估計的 ICC 函數值， $P_R(\theta)$則代表對照組對於該試題所估計的 ICC，而 $w(\theta)$ 是一個與能力分配有關的權數。

然而，此面積法有兩個主要缺點:(1)在估計試題特徵曲線時，面對各式各樣的受試者作答資料，研究者無法事先知道要以何種試題作答模式來估計試題特徵曲線較為合適。(2)面積法對於面積的大小並未提出任何統計理論以作為檢定的依據。這和目前常用的一些 DIF 檢定法相比，例如：Mantel-Haenszel 法(Holland & Thayer, 1988)和 SIBTEST 程序 (Shealy & Stout, 1993a, 1993b)，並無法彰顯出面積法的優越性。爾後，雖然也有些學者(Linn et al., 1981)提出以不同的面積定義作為判斷 DIF 試題的指標，但並未解決先前發生的問題。在本次研究中，將嘗試以 kernel smoothing (Ramsay, 1991) 估計試題特徵曲線，並以此兩條估計的試題特徵曲線間的面積作為判別 DIF 的指標，再配合隨機化檢定法 (Randomization Test, Edgington, 1964)，以實施 DIF 的檢定，此方法稱為改良面積法。希望藉由無母數估計方式和隨機化檢定，以改善前述的兩個缺點，期望能有好的檢定效能。簡單而言，本研究之主要目的包括以下數點：

1. 證明改良面積法具有良好的檢定力
2. 證明改良面積法具有適當的型一錯誤率
3. 比較改良面積法和其它 DIF 偵測方法的優缺點
4. 應用改良面積法於實際的測驗資料

三、結果與討論

當顯著水準為 0.01 時，改良面積法發生型一錯誤的機率平均約在 0.05 到 0.3 之間，研究結果顯示檢定效能並不好。以下將從受試人數、測驗長度、以及猜測參數的有無等三方面來分析改良面積法發生型一錯誤機率過高的原因。

由研究結果得知，當焦點組/對照組受試人數為 1500/1500 時, 改良面積法發生型一錯誤的機率平均為 0.25；但是當受試人數為 5000/5000 時, 型一錯誤的機率平均則為 0.075；由此可以得知，當受試人數越多時，改良面積法發生型一錯誤的機率也就越低，這是因為當樣本越大時，估計的試題特徵曲線就越準確，因此，改良面積法的檢定效能也就相對的提升。但是，發生型一錯誤的機率仍然高達 0.075，比顯著水準仍然高出許多。這個結果仍然凸顯出改良面積法檢定效能的不佳。

其次，在測驗長度方面，當測驗題數分別為 40 題和 75 題時，發生型一錯誤的機率平均為 0.17 和 0.15，結果顯示測驗長度對於改良面積法的檢定效能並沒有明顯的影響。在猜測參數方面，根據模擬研究的結果得知，當沒有猜測參數時，發生型一錯誤的機率平均為 0.17；當猜測參數不為 0 時，發生型一錯誤的機率平均為 0.165。由此可以推斷，猜測參數的有無並不會影響改良面積法的檢定效能，這和先前預期的結果是蠻一致的。

接著是檢定力模擬研究，在顯著水準

為 0.01 時，改良面積法在檢定力的表現大致上不錯。但是，根據前面模擬研究的結果，改良面積法發生型一錯誤的機率極高，因此，檢定力提高很有可能是隨著型一錯誤機率提高的緣故。同樣地，以下將從受試人數、測驗長度和 DIF 的大小等三方面來探討影響改良面積法檢定力的因素。

在受試人數方面，當人數為 500/500 時，改良面積法的平均檢定力大約 0.74；而當人數為 3000/3000 時, 檢定力的平均值為 1.0；由此可以推測，當受試人數越多時，檢定的效能也就越好，這蠻符合大樣本估計較好的原則。而在測驗長度方面，當測驗題數分別為 30 題和 80 題時，檢定力平均為 0.7 和 0.8。由此可以得知，測驗長度的多寡並不會影響改良面積法的檢定效能。至於在 DIF 的大小程度方面，當 $|\Delta_{mh}| = 0.5$ 時，檢定力為 0.7。但當 $|\Delta_{mh}| = 1.5$ 時，檢定力卻高達 0.95。由此可以發現，假若試題的 DIF 程度越大時，改良面積法的檢定力就越高，這也和先前預期的結果相同的。

四、計劃成果自評

本次研究嘗試從試題作答理論出發，以估計的試題特徵曲線間的面積作為檢定 DIF 的指標，再配合隨機化檢定，以實施 DIF 試題的檢測。然而經由兩個模擬研究的結果可以得知，改良面積法的檢定效能並不好。在型一錯誤率方面，改良面積法的發生機率偏高了許多；而在檢定力方面則有不錯的表現，但是這有可能是型一錯誤率提高的緣故。以下是對於後續研究的幾項建議：

1. 本次研究中使用 TESTGRAF (Ramsay, 1993) 程式來估計試題特徵曲線。然而因為程式的限制，在一些變數的使用上較無法自由更動。爾後可以自撰電腦程式，以改善改良面積法的檢定效能。
2. 本研究中, 僅以每組 12 個函數估計值來進行隨機化檢定，但是檢定的效果不佳。因此，建議以更多的函數估計值來進行隨機化檢定，如此對於提高改良面積法的檢定效能可能會有幫助。
3. 在隨機化檢定的統計量方面，建議可以嘗試以其它的面積公式做為檢定 DIF 試題的統計量，例如加權面積等。

五、參考文獻

[1] Edgington, E. S. (1980). *Randomization test.* New York and Basel: Marcel Dekker, Inc.

[2] Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

[3] Linn, R. L., & Harnisch D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109-118.

[4] Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation, *Psychometrika, 56,* 611-630.

[5] Ramsay, J. O. (1993). *TESTGRAF*: a program for the graphical analysis of multiple choice test and questionnaire.

[6] Rudner, L. M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York.

[7] Shealy, R., & Stout, W. F. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

[8] Shealy, R., & Stout, W. F. (1993b). An item response theory model for test bias. In P.W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum Associates.