行政院國家科學委員會專題研究計畫成果報告

檢定與分析多向性測驗試題的 Item Bias/DIF
A Statistical Procedure for Assessing Item Bias/DIF in Multidimensional Item Response Data

計畫編號：NSC 87-2118-M-018-003

執行期限：86 年 8 月 1 日至 87 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(Item Bias or Differential Item Functioning)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回答此種試題，測驗的可靠性和公平性因而受到質疑,而所謂試題的單向性(Unidimensionality)，是指試題本身只測量一種主要的特質或者能力的意思。目前關於試題偏差的研究方法，大都是建立在試題是單向性的假設上，然而當試題並非只是測量一種能力時，這些檢定試題偏差的方法便無法適用。例如：根據大考中心研究人員分析，大學聯考數學科試題至少測量二種以上之能力（例如：空間及幾何）。本計劃即是希望提出一個新法，在試題是多向性時(Multidimensional, 也就是測驗二種以上能力時)，可以適用於檢驗、分析試題偏差。

我們的理論基礎是建立在一個多向性的 DIF 模式上(Multidimensional DIF Model)。我們首先找出把受試者依各種能力得分分組的方法；依照分組求得能力的估計值；計算兩個群體間的能力估計值差異；估計試題偏差；然後求導出檢定統計量以及其漸近性質。接著我們將設計不同狀況來進行模擬研究(Simulation Study)，我們計畫同時做型一錯誤(Type I error)和檢定力(Power)的研究，希望能夠充分瞭解我們所提方法的性質，進而能夠應用於實際的多向性測驗資料。

關鍵詞：試題偏差、試題差異作用、試題作答理論、單向性試題、多向性試題、SIBTEST、MULTISIB

Abstract

Most applications of DIF procedures have been based on the assumption that only one dominant latent ability is measured by the test. However, if more than one latent trait is relevant to the purpose of test, these DIF procedures may yield misleading results. In this project, we propose a statistical procedure for assessing DIF of intentionally two-dimensional test data, such as a "math" test designed to measure algebra ability and geometry ability. Our procedure, MULTISIB, is based on the multidimensional model of DIF as presented in Shealy and Stout (1993b), and is a direct extension of SIBTEST (Shealy and Stout, 1993a), its unidimensional counterpart. First, DIF is appropriately modeled to result from secondary dimensional influence from other than the two intended dimensions. A new statistic was defined and a smoothing approach was also used in estimating the variance of the statistic. A large scale simulation studies then were carried out to investigate the performance of our procedure to detect DIF in two-dimensional tests. Our specific objects include:

(1) Investigation the performance of our procedure when there is no DIF in the two-dimensional test;

(2) Investigation the performance of our procedure when there exists DIF in the test. The amounts of DIF are varied in different simulation setting;
(3) Comparisons of our procedure with other procedures;
(4) Applications of our procedure to real data like college entrance exam for example.

Keywords: **Item Bias, Differential Item Functioning (DIF), SIBTEST, Item Response Theory, Unidimensionality, Multidimensionality, MULTISIB**

二、計畫緣由與目的

The topic of *item bias/differential item functioning (DIF)* has attracted many researchers in recent years. An item is said to exhibit item bias/DIF if the item has different probabilities of correct response for examinees of the same ability but belong to different groups (usually based on race, gender, socioeconomic status, etc.), generally referred to as reference group and focal groups. Several methodologies have been developed to assess item bias/DIF, and numerous studies have been carried out to investigate the validity of these methodologies to assess item bias/DIF. Notable among these methodologies are: Mantel-Haenszel (MH) procedure (Holland & Thayer, 1988), SIBTEST procedure (Shealy & Stout, 1993a, 1993b), logistic regression procedure (Swaminathan & Rogers, 1990), and item response theory (IRT) based procedures (Shepard, Camilli, & Williams, 1985).

Most studies conducted thus far have studied item bias/DIF of intentionally unidimensional tests. That is, it was presumed that the intent of the test was to assess a unidimensional ability. In order to assess item bias/DIF in such tests, the general procedure is to divide examinees from the two groups (reference and focal) into subgroups based on a single score derived from a subset of the test items referred to as the matching subtest. In each of the matching subgroups, the performance of examinees from reference and focal groups is compared on the item(s) of interest and aggregated across subgroups to arrive at a statistic that is used to assess item bias/DIF. Throughout this proposal a "*multidimensional test*" *will mean* a test designed to measure two or more dominant dimensions. In particular, a two-dimensional test is a test designed to measure two dominant dimensions. Item bias / DIF is caused when nuisance abilities other than those intended are part of the multidimensional test.

Few studies have investigated item bias/DIF for such multidimensional test data. In applications, often there are situations when the tests are deliberately constructed to be multidimensional. That is, items intentionally tap more than one dominant dimension. In such situations it is unrealistic to presume that a unidimensional test score will adequately represent a valid matching subtest. For example, in a general science test, generally there is more than one dominant ability influencing examinee responses. In these cases, using total test score as a matching criterion can lead to misleading results. For example, let us suppose that in a two-dimensional test with no contaminating secondary dimensions that half the items tap ability 1 (θ1) and the other half tap ability 2 (θ2). Then, for a subgroup consisting of examinees with 50% correct on the total test, it is evident that such examinees could obtain 50% correct on the test in several ways: by answering correctly all of the items tapping (θ1), or all of the items tapping (θ2), or any combination of the two sets of items. If we are comparing the performance of reference and focal groups on an item based on single score matching, it could lead to inconsistent and misleading results about item fairness. Therefore, in order to obtain accurate results, it is important to match examinees so that they are comparable on all abilities to be measure by the test.

Most existing procedures can only be used for assessing item bias/DIF of unidimensional tests. In order to utilize them for assessing item bias/DIF in the context of multidimensionality, these procedures need to be modified and/or new procedures need to be developed. The purpose of the present study has been to develop a statistical test MULTISIB to detect item bias/DIF of two dimensional test data, and to study its performance with respect to the Type I error and power of simulated two-dimensional tests as well as real two-dimensional tests.

三、結果與討論

A few simulation studies were conducted to assess the performance of MULTISIB to detect DIF in two-dimensional test data. In these simulation studies, MULTISIB demonstrated good Type I error behavior and reasonable power across a wide range of sample size. Comparisons of true DIF parameter value with the average estimated DIF in both Type I error and power studies showed that MULTISIB also displayed minimal statistical bias. This behavior extended to conditions analogous to conditions in which unidimensional DIF procedures might be expected to perform less well.

Studies also showed that the MULTISIB performance observed was not dependent on the matching subtests exhibiting orthogonal simple structure (i.e., not all items in the matching subtest must be pure measures of $\theta 1$ or $\theta 2$). The behavior of MULTISIB was maintained quite well even when using matching subtests with items that had more realistic angular spreads.

Finally, our studies also demonstrated the importance of matching examinees on both matching subtest scores separately instead of matching on the basis of one total matching subtest score. That is, when assessing DIF with two-dimensional data, MULTISIB should be used rather than the unidimensional SIBTEST or MH procedures.

四、計劃成果自評

This study demonstrated the utility of MULTISIB for assessing DIF in two-dimensional test data. Further studies should attempt to extend MULTISIB to assess DIF to test data in which the number of intentional underlying dimensions is greater than two. For this situation, matching will be more difficult because of the exponentially increasing number of score cells. It also may be worthwhile to investigate further the performance of MULTISIB under other conditions of test multidimensionality, such as varying the correlation between dimensions, the underlying trait distribution, or the multidimensional model for item response functions by which data are generated. Finally, applications of MULTISIB using real data should be conducted.

五、參考文獻

[1] Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Education Measurement, 29*, 67-91.

[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] Li, H., & Stout, W. F. (1996). A new procedure for detection of crossing DIF/bias. *Psychometrika, 61*.

[4] Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.

[5] Narayanan, P., & Swaminathan, H. (1997). Performance of the Mantel-Haenszel and simultaneous item bias procedure for detecting differential item functioning. To appear in *Applied Psychological Measurement.*

[6] 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.

[7] 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.

[8] Shepard, L. A., Camilli, G., & Williams, D. M.

(1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22,* 77-105.

[9] Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.