

縱貫性研究中度量化的一些議題： 以症狀檢核表 SCL-90-R 為例

Scaling Issues in Longitudinal Studies: The Symptom Checklist-90-Revised as an Empirical Example

王文中 吳齊殷

本研究利用Rasch模式(Rasch, 1960)及其延伸模式,分析吳齊殷(1997, 1998, 1999)所蒐集的SCL-90-R之縱貫性資料,以說明如何確保縱貫性研究心理變項的等距特性,以及變項意義在不同時間點的穩定性。該研究持續追蹤了1101位學生達五年之久,每年利用SCL-90-R的四個分量表以調查其心理症狀的變化。受試者每年受測的題目不盡相同。結果發現這四個分量表大致吻合Rasch模式,因此可以獲致等距量尺。不過題目的閾值過高,因此信度不佳。五點的量尺可以簡化成爲兩點,而不失其信度。測驗建構基本上保持著時間的穩定性,因此可以進行改變的測量。研究並發現受試者在體化症、焦慮上,國二比國一時來得低一些,但到國三時又揚升起來。在憂鬱程度上,國一和國二沒有差別,但國三時則稍微增加。敵意的變化在國二時略微揚升。一旦國中畢業(第四個時間點),受試者在這四種心理症狀上,均大幅的減輕,尤其敵意降得最爲明顯。除此之外,這四個時間點的相關爲中度相關,介於.42至.76之間。

關鍵詞：Rasch 模式、試題反應理論、心理症狀、改變測量、差異試題功能。

This study demonstrates how to obtain an interval scale and to check whether test construct remains stable over time in longitudinal studies through analyzing a longitudinal dataset of the SCL-90-R collected by Wu (1997, 1998, 1999), where 1101 seven-graders took four subtests of the SCL-90-R in four years, once a year. Items administered were not identical over the four time points. To obtain an interval scale for each subtest and to put scores of the four time points onto the same scale so that changes in psychological disorders could be measured, the Rasch model (Rasch, 1960) as well as its extensions was employed. Generally, the four subtests fitted the model's expectation fairly well so that interval scales could be obtained. However, the items were far too difficult for these subjects, which reduced test reliability. Five-point scales could be collapsed to two-point scales without scarifying their reliabilities substantially. The test construct remained unchanged over time so that change measurement was applicable. On Somatization and Anxiety, the mean magnitudes were reduced from seven- to eight-grade, but were increased from eight- to nine-grade; On Depression, they remained unchanged from seven- to eight-grade but were increased from eight- to nine-grade; On Hostility, they were increased from seven- to eight-grade and then remained unchanged from eight- to nine-grade. After graduation from junior high schools (the fourth time point), the mean magnitudes on all the four subtests were decreased significantly, especially on Hostility. The correlations between the four time points for the four subtests were all positively moderate, ranging from .42 to .76.

Key words: Rasch model, item response theory, psychological symptom, change measurement, differential item functioning.

作者：王文中現爲中正大學心理系教授，National Chung Cheng University，E-mail: psywew@ccu.edu.tw。

吳齊殷爲中央研究院社會學研究所副研究員，Academia Sinica。

收稿：2002年8月13日；接受：2003年10月27日

一、緒論

在縱貫性研究裡，受試者接受數個時間點的調查或測量，據以衡量其改變(或成長)的趨勢。以探討國小學童身高而言，必須滿足以下兩個條件才能探討身高的改變。第一，身高必須是量的變項(quantitative variable)，且對身高的測量必須是等距量尺(interval scale)，如此才能進行加減的運算，進一步衡量改變的大小。第二，身高這個概念必須保持不變。如果身高這個概念隨著測量的時間不同，而產生「質變」，那就無從探討身高的量變。對於物理變項，如身高、體重、血壓等，上述兩個條件都成立。但是若換做心理變項，如智力、動機、自我概念、心理症狀等，上述兩個條件恐怕就未必成立。因此在心理變項的縱貫性研究中，必須先確保這兩個條件成立後，才能衡量改變的趨勢。本研究以Derogatis(1983)發展的症狀檢核表(the Symptom Checklist-90-Revised, 簡稱SCL-90-R)的縱貫性調查為例，說明如何去判斷這兩個條件是否成立。讀者可以將此方法運用至其他心理變項或其他量表的縱貫性研究。

SCL-90-R共有90道自陳題。包括九個向度：體化症(somatization)、強迫症(obsessive-compulsive)、人際敏感(interpersonal sensitivity)、憂鬱(depression)、焦慮(anxiety)、敵意(hostility)、恐懼焦慮(phobic anxiety)、偏執狂(paranoid ideation)、精神病(psychoticism)。其功用在於檢測受試者具有這些症狀的強度，當作進一步療育的參考。吳齊殷(1997, 1998, 1999)在一項關於「青少年藥物濫用之起因：一個社會學習模型」研究計畫，自1996年實地調查國一學生之心理困擾，已連續進行五年，每年均施測SCL-90-R之四個分量表之部分題目和其他問卷。這是國內第一個關於SCL-90-R之縱貫性研究，因此頗具有示範作用。可是至今為止，對於SCL-90-R的分析，和絕大多數的量表一樣，基本上是屬於定性層次的次序量尺(ordinal scale)，尚未發展出屬於定量層次的等距量尺。例如絕大多數的分數計算是以原始總分或其線性轉換(如T分數)來表示受試者的程度。原始總分或其線性轉換分數，建立在古典測驗理論(classical test theory)的基礎上，充其量只有順序的特性，並無等距的意義。此外，縱貫性的研究，可能和吳齊殷的研究一樣，受試者

接受了數次的測驗，但每次測驗的題目不盡相同。題目不一樣，使得改變的測量變得十分複雜。由於原始分數並無等距的意義，連帶無法評量受試者的改變情形。我們需要新的測驗理論來獲致等距的量尺，並將不同時間點所施測的測驗放在同一量尺上，以衡量受試者的變化。

這種測量理論早在1960年代就由Georg Rasch (1901-1980)發展出來，通稱為Rasch測量模式(Rasch, 1960)。近年來，已經被廣泛使用在教育測驗、心理測驗、健康相關的測驗上。有關Rasch模式的介紹，可參見網站：<http://www.rasch.org>。本研究希望藉由吳齊殷(1997, 1998, 1999)的縱貫性資料，探討一些縱貫性研究上的測量議題。具體而言，本研究探討以下的測量議題：

1. 利用Rasch測量模式，探討測驗題目的適切性，以及將順序的量尺(總分)化為等距的量尺。
2. 探討量尺的點數是否可以縮減，以及縮減的方法。SCL-90-R的點數原為5點，但對於特定的樣本而言，是否需要5點？是否可以簡化，如簡化為2點？
3. 如何檢查測驗的建構(即題目的意涵)是否具有時間的不變性(invariance)。測驗的建構維持不變，才能評量受試者隨時間的改變情況。如果所代表的意義已經不同，就無法測量受試者的改變情形，因為此時所牽涉的問題不是量變，而是質變。
4. 如果測驗的建構沒有質變，或者即使有些微的質變，但小到可以忽略，接下來的任務就是如何將這多個時間點的資料，串連起來，放在同一個量尺上，以衡量受試者的變化。

在本文裡，我們透過SCL-90-R的縱貫性資料，展現具體方法和步驟，以獲致等距量尺，以及評量受試者隨時間的改變情形，可當作類似研究的參考，這算是本研究的附帶目的。

二、SCL-90-R之文獻與資料特性

近年來有相當多的論文在探討SCL-90-R的信度、效度、和應用價值等議題。例

如Derogatis與Savitz(1999)專章介紹SCL-90-R的應用。Schmitz, Kruse, Heckrath, Alberti與Tress(1999)探討SCL-90-R的臨床診斷價值。Woessner與Caplan(1995)探討該量表在中度和重度腦傷病人上的應用。Todd, Deane與McKenna(1997)探討正常人與臨床病人的差異。Woessner與Caplan(1996)探討中風病人的痛苦症狀。Woody, Steketee與Chambless(1995)探討SCL-90-R的某些分量表的功能。Carpenter與Hittner(1995), Cyr, McKenna-Foley與Peacock(1985), Rauter, Leonard與Swett, (1996), Vassend與Skrondal(1999)等人利用探索性或驗證性因素分析探討SCL-90-R的建構。Schmitz, Hartkamp, Brinschwitz與Michalek(1999)探討紙筆版本和電腦版本的差異。

國內亦有些許研究使用此量表，並獲致相當不錯的結果。顏永杰、鄭夙芬、楊明仁、何啓功、張明永(2000)利用華人健康量表(CHQ-12)及SCL-90-R，評估參與某工業意外災難之救難者，精神狀況之時序變化情形。柯慧貞、孫苑庭、林木芬、葉宗烈、陸汝斌(2000)由兒童期的行為抑制氣質探討婦女的焦慮與憂鬱是否具有相同的病因。林文香、夏萍緬、楊文山、洪志美(1999)討全身性紅斑狼瘡與類風溼性關節炎女性患者之身體功能、心理狀況與社會功能之影響程度及其影響因素。蕭淑貞、陳孝範、張珏(1999)採用多主題壓力調適工作坊，探討改善護理人員壓力症狀之成效。林文香、夏萍緬、楊文山(1997)以全身性紅斑狼瘡門診病患為樣本，檢定無助指標測量全身性紅斑狼瘡患者經驗無助的信效度。李毅達、吳晉祥、張智仁、陳純誠(1996)收集醫學中心健檢病房之1480名女性及1802名男性的基本資料及身體健康狀態。身體健康狀態分為無身體疾病或症狀、非特定身體症狀、慮病、及已確定之身體疾病四大類別，心理健康狀態之評量以SCL-90-R為之。陳映雪、葉紅秀、余鳳玉(1993)比較從未自殺過的青少年精神疾病住院患與企圖自殺者，發現在絕望與憂鬱的程度上並無顯著的差異。江麗珍、陳珠璋、彭素玲(1993)以四種評估工具(含SCL-90-R)及團體互動情形，評估精神分裂病患的團體發展過程及治療結果。李明濱、李宇宙、李蘭等(1990)將SCL-90-R由90題題目簡化為50題，發現其信度與效度在臨床診療使用上效果還算良好。李明濱、林憲、林信男(1988)取樣精神科門診病患，結果發現影響不良預後之預測因素包括：病人自覺治療者對病人之瞭解程

度較差，SCL-90-R量表中之身體化症狀較嚴重等。國內的一些學位論文也曾用此量表(如林木芬，1996；林玉慈，1999；張玉玲，2001；龔曉萍，2000)。

從以上的文獻回顧中，可以發現(1)SCL-90-R經常被使用於臨床上，且具有相當不錯的信度和效度。(2)SCL-90-R較少使用於正常人，也少使用於縱貫性之研究中，這一方面的信度與效度仍待進一步探究。(3)資料分析大多建立在古典測驗理論的基礎上，且已假設所獲致的量尺(如原始總分)是等距量尺，這個假設未必合理，因此其後續的分析是可議的。

本研究的資料來源為吳齊殷(1997, 1998, 1999)，他選用了SCL-90-R五個分量表共48道題目：體化症12題、憂鬱12題、焦慮10題、敵意6題、其他症狀(additional symptoms)8題，共48題，均為五點李克特氏量尺。1代表沒有這個症狀，2代表有點不舒服，3代表普通不舒服，4代表嚴重不舒服，5代表很嚴重不舒服。選擇這五個分量表施測的原因是：青少年較常反映有這些症狀，並且是比較容易辨認的。本研究的資料取共取用四個時間點，第一個時間點為1996年台北市國一的學生共1,434人。這些受試者於次年(國二)時又接受施測，人數仍為1,434人。再次年(國三)又被施測，不過人數略有流失，經補足後，共有1,449人。第四年(國中畢業後)施測，共完成1,182人。總計前後共有1,595人受測，其中1,101人連續接受了四次時間點的施測。在第一個時間點時，全部的48題均施測，第二和第三個時間點，只各施測47題，其中一題：覺得自己沒有價值(沒有用)被刪除，因為經初步分析，發現該題語意含糊並不恰當。在第四個時間點，為節省時間，只施測了15題，其中體化症6題，憂鬱、敵意、其他症狀各3題，並無焦慮的題目。由於每個時間點所施測的題目不盡相同，因此原始分數無法直接比較。在此我們並不分析「其他症狀」分量表，因為該分量表不像其他分量表一樣，具有明顯且一致的建構。

三、方 法

本研究使用Rasch模式及其延伸模式進行資料分析，而非一般常用的古典測驗理論，因為古典測驗理論有著嚴重的缺點。在古典測驗理論裡，受試者的特質(以

下通稱為能力)通常是以原始分數或其他線性轉換後的分數(如T分數)代表。題目難度(或閾值)則以通過百分比來定義,鑑別度則常以高低分組通過百分比的差異來表示。受試者的能力和試題特性並不在同一量尺上,因此兩者沒有關連,難以比較。此外,對受試者能力的估計會受到題目特性的影響,例如題目很簡單的話,受試者就容易得高分,因此我們就會認為受試者能力很高。反之,如果題目很難,受試者得分就低,我們就會認為受試者能力很低。同樣的,對題目特性的估計會受到受試者能力的影響。如果受試者能力很高,題目的答對率就很高,我們就會認為題目很簡單。反之,如果受試者能力很低,題目的答對率就低,我們就會認為題目很難。總之,在古典測驗理論裡,對受試者能力和題目特性的估計會互相干擾,因此不客觀。

為了克服以上的困境:受試者能力和試題特性不在同一量尺上,以及其估計會互相干擾,Rasch(1960)提出了所謂的Rasch測量模式。在這個模式裡,令 θ_n 表示受試者 n 的能力, δ_i 表示題目 i 的難度,則受試者答對該題的機率 p_{ni1} 為:

$$p_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \dots\dots\dots (1)$$

答錯的機率 p_{ni0} 為:

$$p_{ni0} = \frac{1}{1 + \exp(\theta_n - \delta_i)} \dots\dots\dots (2)$$

取勝算比(odds)再取對數,得對數勝算子(logit)為:

$$\text{logit} \equiv \log(\text{odds}) = \left(\frac{p_{ni1}}{p_{ni0}} \right) = \theta_n - \delta_i \dots\dots\dots (3)$$

在這個模式裡的 θ_n 和 δ_i 是在同一量尺上,因為其可以加減運算。此外,這個量尺具有等距特性。說明如下。現有二位受試者,其能力分別是 θ_1 , θ_2 。對第 i 題而言,他們的對數勝算子分別為 $\theta_1 - \delta_i$ 和 $\theta_2 - \delta_i$ 。兩者相距: $(\theta_1 - \delta_i) - (\theta_2 - \delta_i) = \theta_1 - \theta_2$ 。不管題目難度是多少,這兩位考生的對數勝算子永遠都是 $\theta_1 - \theta_2$,因此對考生能力差距的估計不受題目難度的干擾,即具有所謂特殊客觀性(specific objectivity)。

若 θ_1 比 θ_2 多了一個單位，其對數勝算子也相對的多了一個單位，無論 θ_2 是多大， θ_1 就是比 θ_2 多了一個單位，這就是等距的意義。同理， δ_i 也有和 θ_n 的特性，因為它們是同一個量尺。

上述的 Rasch 模式只適用於二元計分的題目（對和錯，或同意和不同意）。若是多元計分，則可將公式(3)擴大為：

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - \delta_{ij} \dots\dots\dots(4)$$

其中 p_{nij} 為受試者 n 在題目 i 中得 j 分的機率， $p_{ni(j-1)}$ 為受試者 n 在題目 i 中得 $j-1$ 分的機率， δ_{ij} 是題目 i 於分數 j 的難度，又稱梯級難度 (step difficulty)。這是由 Masters (1982) 提出的部份得分模式 (partial credit model)。這個模式適用於多元計分的題目，如計算題、問答題等。另位學者 Andrich (1978) 提出適用於李克特氏量尺的模式—評等量尺模式 (rating scale model) 如下：

$$\log\left(\frac{p_{nij}}{p_{ni(j-1)}}\right) = \theta_n - (\delta_i + \tau_j) \dots\dots\dots(5)$$

其中 δ_i 是題目 i 的整體難度 (overall difficulty)， τ_j 是得分 j 的閾難度 (threshold difficulty)。部份得分模式和評等量尺模式中的量尺和 Rasch 模式一樣，均具有特殊客觀性和等距特性。如果題目的反應真如 Rasch 模式或其延伸的部份得分模式、評等量尺模式所預期，則則所獲致的量尺分數 θ 具有客觀性和等距。反之，如果某些題目不吻合模式，這個題目對這些受試者而言，所代表的意義和其他的題目不同，也就無法和其他題目擺在同一個量尺上，因此值得進一步探究原因並加以修訂。

在上述的公式(3)、(4)、(5)中，反應試題特性的參數只有一種，就是難度。因此又稱為單參數模式。另有別的學者認為只用一種難度參數無法完全表達題目的特性，因此又增加了別的參數來反映題目的特性，如 Birnbaum (1968) 的二參數、三參數對數模式 (two-or three-parameter logistic model)。可惜這類多參數模式，並無 Rasch 模式的等距和客觀的良好特性。不論是單參數模式或多參數模式，共同點

就是在描述題目的作答反應，因此通稱為試題反應模式 (item response model)。由這些試題反應模式所建立出來的理論體系，稱為試題反應理論 (item response theory, IRT)。有關IRT的理論和實務，可參見Baker(1985, 1992)、Birnbaum(1968)、Embretson與Reise(2000)、Fischer與Molenaar(1995)、Hambleton與Swaminathan(1985)、Lord(1980)、Rasch(1960)、van der Linden與Hambleton(1997)、Wright與Stone(1979)等書籍。相關電腦軟體可參見相關網站：<http://www.assess.com>或<http://www.winsteps.com>。

本研究利用電腦軟體ACER ConQuest(Wu, Adams, & Wilson, 1998)來進行資料分析。該軟體是搭配多向度隨機係數多項洛基模式 (multidimensional random coefficients multinomial logit model, MRCML; Adams, Wilson, & Wang, 1997)。假設有 D 個特質決定了受試者於測驗上的表現，受試者 n 於這些特質上的程度置於向量 $\theta'_n(\theta_{n1}, \dots, \theta_{nD})$ 。假設受試者來自某個母群體，分佈為 $g(\theta; \alpha)$ ，其中 α 是該分佈的參數。如果 g 是常態分佈的話， α 就是平均數向量和變異數-共變數矩陣。在MRCML模式裡，受試者 n 於試題 i 的類別 k 的反應機率為：

$$f(X_{ik} = 1; \xi | \theta_n) = \frac{\exp(\mathbf{b}'_{ik} \theta_n + \mathbf{a}'_{ik} \xi)}{\sum_{u=1}^{K_i} \exp(\mathbf{b}'_{iu} \theta_n + \mathbf{a}'_{iu} \xi)} \dots\dots\dots (6)$$

如果反應為題目 i 的類別 k ，則 X_{ik} 為1；否則 X_{ik} 為0， K_i 是題目 i 的類別數。 \mathbf{b}_{ik} 是題目 i 的類別 k 於 D 向量的計分向量， ξ 是題目參數的向量， \mathbf{a}_{ik} 是用以表示 ξ 元素間線性關係的設計向量。公式(6)的特性是使用了計分向量 \mathbf{b}_{ik} 和設計向量 \mathbf{a}_{ik} ，操弄這兩個向量，可以形成既有的各種Rasch模式及其延伸模式，如部份得分模式、評等量尺模式，多相模式 (facets model, Linacre, 1989)、等第分割模式 (ordered partitioned model, Wilson, 1992)、線性對數潛在特質模式 (linear logistic latent trait model, Fischer, 1973)、線性部份得分模式 (linear partial credit model, Fischer & Pononcy, 1994)，還有多向度的模式。在本研究裡，不同時間點的測驗被當作不同向度來一起處理，因此屬於多向度模式。透過多向度模式分析，可以將不同時間點的資料放在同一量尺上，儘管不同時間點施測的題目不盡相同，且可以直接估計四個時間點

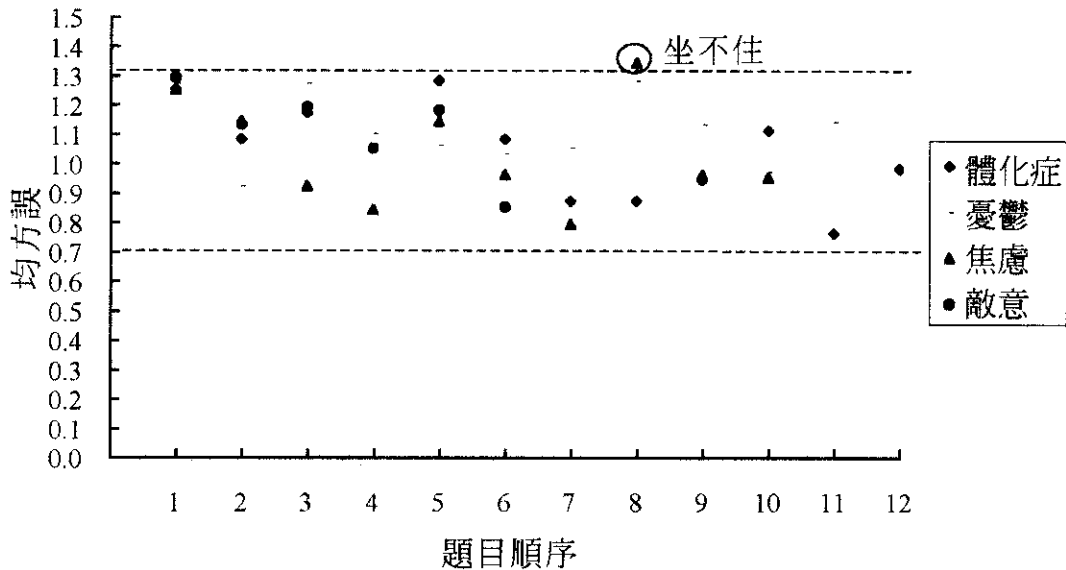
受試者母群體程度的變異數—共變數矩陣(相關矩陣)。關於此模式的理論和應用，可參閱上述文獻、軟體使用手冊、以及Wang, Wilson與Adams(1997, 2000)。

四、結 果

以下我們將說明如何利用Rasch模式及其延伸模式來進行分析。主要探討的議題包括：(1)資料是否吻合Rasch模式？如果有些題目不吻合，原因為何？應如何修改？(2)在維持測量精準度的前提下，五點量尺可否簡化為兩點？(3)未來修訂量表時，如何提高測驗的測量準確度？(4)題目的意義是否隨時間不同而不同？(5)如無不同，如何將不同時間點的題目放在同一量尺上，以衡量受試者的變化？

(一) 資料與Rasch模式的吻合度

由於本研究的資料為5點量表，因此可利用評等量尺模式(公式5)來進行分析。同樣的也可以將五點量尺簡化為兩點量尺(沒有不舒服，和有不舒服)，然後用Rasch模式(公式3)分析之。由於在以下的單元裡，我們發現五點量尺對這些受試者而言，並不實用，因此可簡化為兩點量尺。如前所述，如果資料吻合Rasch模式或其延伸模式的話，所得到的量尺分數具有等距和客觀的意義，因此第一個步驟就是要檢測資料是否吻合模式。圖一呈現四個分量表資料與Rasch模式的吻合情形。基本上如果資料吻合模式的話，其均方誤(mean square error)的期望值為1.0(Wright & Masters, 1982)。如果均方誤距離期望值1.0很遠的話，就表示該題不吻合模式。到底均方誤要距離1.0多遠時，才算不吻合模式，端看使用者對資料嚴謹程度的要求。有的人會使用 1.0 ± 0.2 ，即介於0.8到1.2之間。有的人則使用 1.0 ± 0.3 ，即介於0.7到1.3之間。本研究採用較為寬鬆的0.7到1.3的標準，這是因為人格測驗通常無法像能力測驗一樣，可以得到較為統一清晰的定義。如圖一所示，發現焦慮量表的「坐不住」不是很吻合Rasch模式。對國一學生而言，正值活潑的青春期的他們而言，「坐不住」可能代表著青春活力，而不是「焦慮」。未來可能考慮將此題目改為「坐立難安」，也許較能吻合焦慮的原意。在以下的分析裡，我們將這不當的題目加以刪除，重新進行分析。



圖一 四個量表吻合Rasch模式的情形

(二) 五點量尺的簡化

五點量尺的使用非常普遍，不過這並不代表所有的量表都適合使用五點量尺。如果受試者無法有效區分五點量尺的意義，或者作答反應不夠分散，五點量尺未必會比最簡單的兩點量尺來得適用。通常量尺點數的增加是爲了提高測量的準確度，如果點數增加，準確度並沒提高，那麼增加點數未必值得，畢竟點數增加可能會增加受試者的困擾和造成不易解釋測驗分數。在以下的分析裡，我們比較了SCL-90-R各個分量表在三個時間點(國一、國二、國三)時的五點量尺的信度和兩點量尺的信度。在Rasch模式下，信度的計算方法和傳統的計算(如Cronbach's alpha)略有不同。在試題反應理論裡，每道試題對於不同能力水平提供不同的訊息量(item information)，訊息量的倒數就是變異誤(error variance)，變異誤開跟號就是標準誤。換句話說，每道試題對於不同能力水平提供了不同的變異誤 $\hat{\sigma}_{\theta i}^2$ 。這和古典測驗理論中變異誤同質性的假設不同。將所有試題的訊息量加總，就是測驗訊息量(test information)。測驗訊息量的倒數就是該測驗對於不同能力水平的測驗變異誤。將所有受試者的測驗變異誤加以平均，就是平均變異誤 $\bar{\sigma}_{\theta i}^2$ 。令母體變異數的

估計值為 $\hat{\sigma}^2$ ，則信度為 $r = 1 - \frac{\bar{\sigma}_{\theta}^2}{\hat{\sigma}^2}$ ，這相當於古典測驗理論的信度，詳細作法可參見Lord(1980)。在五點量尺方面，我們使用評等量尺模式來分析。在兩點量尺方面，我們將五點量尺中的後四點(2, 3, 4, 5)合併為一點，因此五點量尺簡化為1(沒有不舒服)、2(有不舒服)，然後使用Rasch模式進行分析。除了信度之外，也計算這兩種量尺所得到能力估計值的積差相關，相關越高表示五點量尺和兩點量尺並無實際上的差別。信度和相關係數如表一。基本上，兩點量尺的信度略高於五點量尺(這是因為點數合併的隨機效果)，而且兩者的相關都高達.91至.97。這充分說明五點量尺對於這些受試者而言，並不見得比兩點量尺來得優越。在以下的分析裡，都是使用簡化後的兩點量尺。

由於這些受試者都是屬於正常的國中生，因此會有如SCL-90-R量表中的體化症、憂鬱、焦慮、敵意的情形並不多見。就算有，也很難有效區分其嚴重程度。因此五點量尺，反而不如兩點量尺來得方便。但這並不表示五點量尺就完全應該摒棄，而是對這樣的正常受試者而言，五點量尺並沒有比兩點量尺提供更多的訊息。如果受試者是有這種症狀的病人，那麼五點量尺就可以發揮區分嚴重程度的效果。此時，兩點量尺也許就會略嫌不足。

(三) 提高測驗的測量準確度

就試題反應模式而言，題目對受試者提供的信息，取決於題目與受試者能力

表一、五點量尺和兩點量尺的信度與相關

	第一年		第二年		第三年		第四年					
	信度	相關	信度	相關	信度	相關	信度	相關				
	五點	兩點	五點	兩點	五點	兩點	五點	兩點				
體化症	.64	.70	.96	.73	.75	.97	.76	.78	.96	.52	.53	.94
憂鬱	.67	.69	.96	.68	.76	.97	.73	.80	.97	.52	.54	.96
焦慮	.65	.67	.95	.67	.75	.96	.70	.77	.96			
敵意	.55	.56	.91	.60	.66	.93	.62	.68	.93	.05	.04	.94

的吻合度。吻合度越高，提供的訊息就越大，反之，則越少。訊息越大，就表示該題目可以越有效區分受試者的能力水準，因此測量誤差就越少。Rasch就模式而言，當題目的難度和受試者的能力一樣時，吻合度最大，兩者差距越大，吻合度就越小。如果受試者能力的平均數為0，那麼試題難度的平均值也最好接近0，才能提供大量的訊息。表二列出受試者於各分量表的三個時間點的能力估計值的平均數，試題難度的平均數都限制為0。由於這些分量表能力估計值的平均數都在-2左右，遠小於0，因此這些題目對這些受試者而言，都是偏難(在此意味著這些題目偏嚴重)。就以第一年的第一個分量表而言，一位平均程度的學生(-2.17)在作答一道平均難度的題目(難度等於0的題目：想吐或拉肚子)時，其答對(在此為回答不舒服)的機率為 $.10 \left(= \frac{\exp(-2.17-0)}{1+\exp(-2.17-0)} \right)$ 。換句話說，絕大多數的受試者在這些題目上的回答均傾向於答錯(在此為回答沒有不舒服)。

圖二呈現受試者在第一年的體化症的程度分佈和題目難度分佈的線性關係。其中難度最高的是：忽冷忽熱(難度值為1.72)，最簡單的是：頭痛(難度值為-1.54)。換句話說，受試者在回答是否頭痛時，遠比回答是否忽冷忽熱時，來得容易說：「有不舒服」。對一個程度在平均數(-2.17)的受試者而言，說會頭痛的機率為.35，說會忽冷忽熱的機率為.02。即：

$$p = \frac{\exp(-2.17 - (-1.54))}{1 + \exp(-2.17 - (-1.54))} = .35 ; p = \frac{\exp(-2.17 - 1.72)}{1 + \exp(-2.17 - 1.72)} = .02$$

從圖二可以看出，三種最嚴重的體化症狀依序為：忽冷忽熱、呼吸困難、感覺有東西卡在喉嚨。相反的，最輕微的四種症狀為依序為：頭痛、腰酸背痛、肌肉酸痛、頭暈。這和一般的臨床經驗相吻合。此外圖二還顯示出絕大多數的題目都位於受試者分佈的上方，這表示這些題目對這些受試者而言，太難了(即太嚴重了)。過度的難，將無法有效區分受試者的程度。例如，這12題對受試者而言，信度只有.70(見表一)。這樣的信度，無法當作臨床的充分依據。

同樣的，在其他三個量表上，也都出現題目難度過難以致信度不高的情形。圖三至圖五分別呈現受試者第一年於「憂鬱」、「焦慮」、「敵意」的程度分佈與題目難度

表二、受試者能力估計值的平均數與變異數

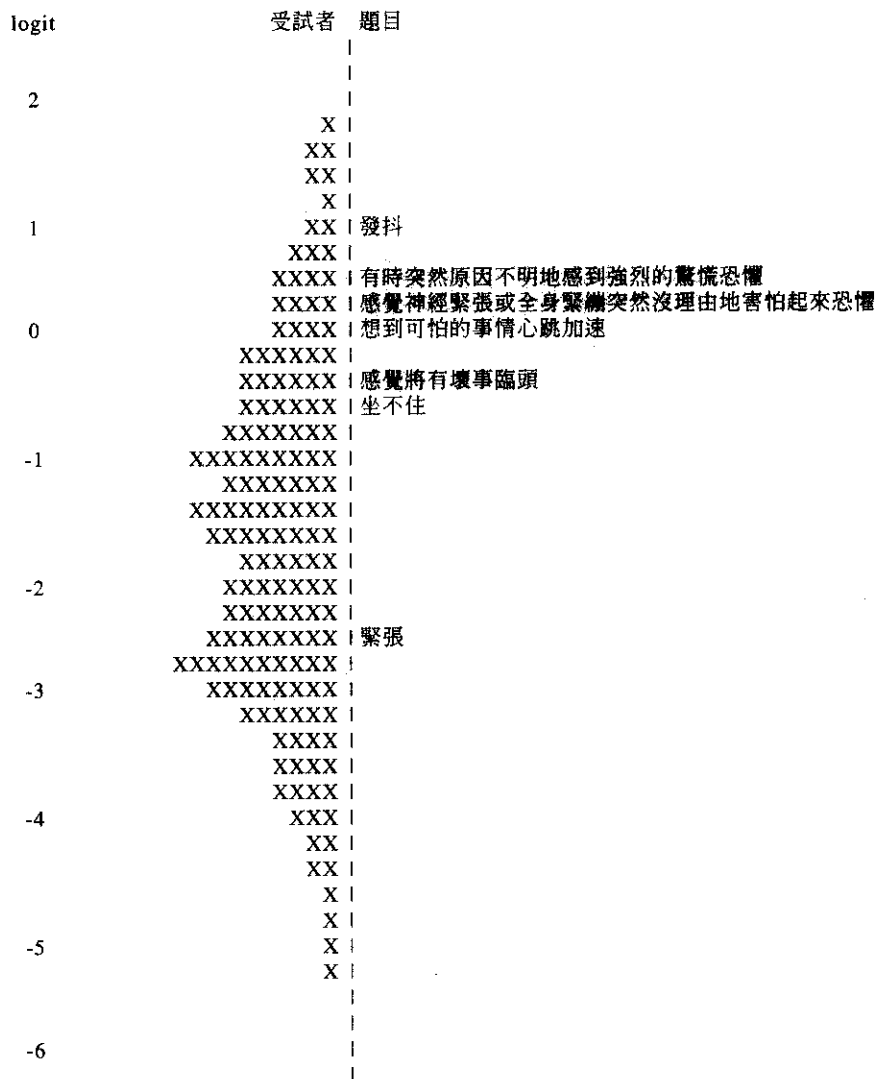
	第一年 (N=1434)		第二年 (N=1434)		第三年 (N=1449)		第四年 (N=1182)	
	平均數	變異數	平均數	變異數	平均數	變異數	平均數	變異數
體化症	-2.17	2.36	-2.38	3.93	-2.02	3.82	-2.56	1.98
憂鬱	-1.90	2.69	-1.92	4.58	-1.51	4.42	-2.22	3.78
焦慮	-1.71	2.41	-2.03	4.35	-1.82	4.82	NA	NA
敵意	-2.02	2.56	-1.83	3.44	-1.88	4.11	-2.96	1.09

註：NA表示該時間點並無測量。

的線性關係。就圖三而言，發現幾乎所有的題目，除「不想活」外，都集中在一塊，不夠分散，而且難度都偏難，並無法有效區分這些受試者。圖四中題目略微分散，以「發抖」最難，「緊張」最易。也就是說，受試者比較不會有發抖的情形，但比較容易感到緊張。「緊張」的難度遠遠低於其他題目，難度值為-2.65，對焦慮程度為平均數-1.71的受試者而言，其感到緊張的機率為： $P = \frac{\exp(-1.71 - (-2.65))}{1 + \exp(-1.71 - (-2.65))} = .72$ 。換句話說，一般焦慮程度的國一學生在過去的一星期內約有72%的人感到緊張，這個比率並不算低。

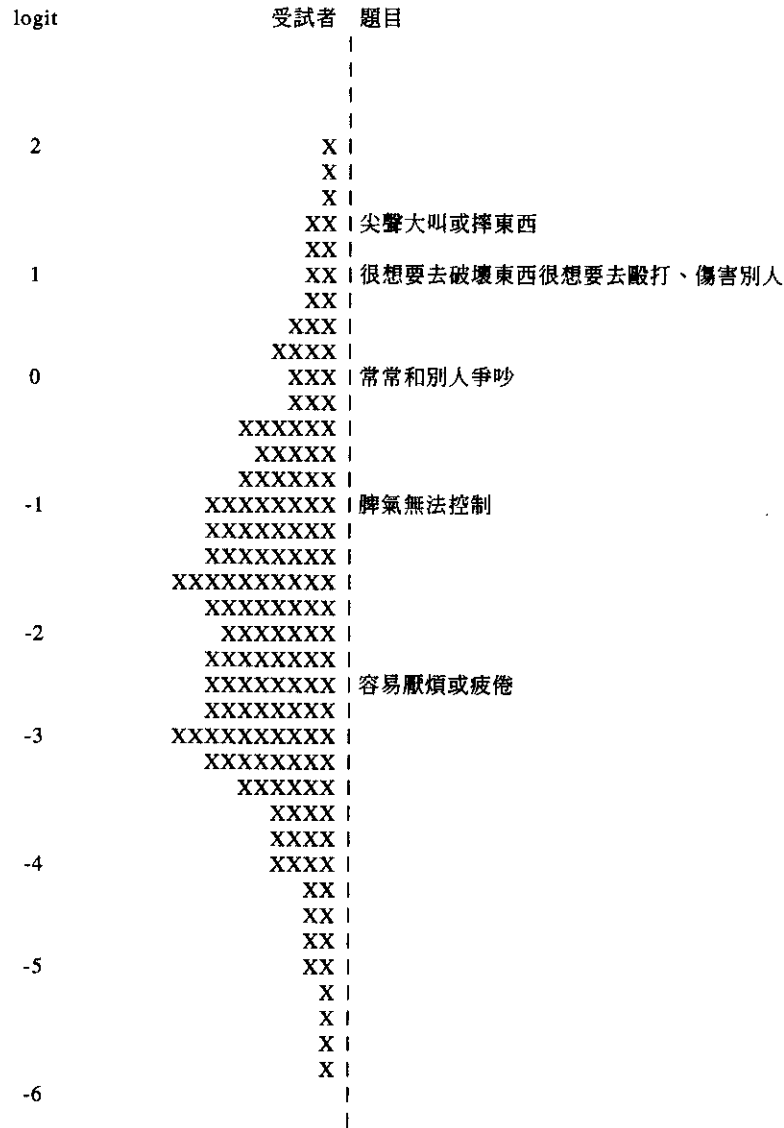
就敵意而言，如圖五所示，這6道題目的難度相當分散，比較容易區分受試者的敵意程度。不過整體而言，難度仍然偏高些。難度最難的為「尖聲大叫或摔東西」，難度為1.53。最簡單的為「容易厭煩或疲倦」，難度為-2.69。

圖二至圖五雖然只是列著受試者第一年的程度分佈，其實第二年、第三年、第四年的分佈和第一年分佈大同小異，只是受試者的平均數和變異數略有不同而已，詳見表二。難度的分佈則固定不變。從以上的分析中，可以發現這四個量表的題目，對這群國中生而言，實在太難了，而且難度又不夠分散，以致無法有效區分學生的程度。改進的方法，就是編寫適當難度的題目，而且每題的難度要足夠的分散。所謂適當難度的題目就是整體而言，平均難度應相當於受試者的平均程度。例如國一學生在「體化症」的-2.17平均數為，那麼適當難度就是在-2.17左右。所謂難度要足夠的分散，就是有的題目要難些，才能有效區分程度很高的學生。有些題目



註：每個X約代表9個受試者

圖四、受試者在第一年的「焦慮」的程度分佈與題目難度分佈的線性圖



註：每個X約代表9個受試者

■五、受試者在第一年的「敵意」的程度分佈與題目難度分佈的線性圖

難度要中等，才能有效區分中等程度的學生。同樣的，有些題目難度要低些，才能有效區分程度很低的學生。

如果題目的整平均難度相當於受試者的平均程度的話，信度可以提高多少呢？假設圖一中的12道試題重新編製，在受試者的程度維持不變的情況下，這新編製的12題的難度為原來的難度減2.17，也就是將圖一中的所有12道試題的難度往下移2.17個單位，使得難度的平均數由原來的0變為-2.17，剛好與受試者程度的平均值相等，然後重新計算信度，結果得到信度從原來的.70提高至.88。換句話說，適當難度的題目可以提高對受試者能力估計的準確度。表三列著各個分量表的新信度，假如題目難度的分佈恰巧對應於受試者能力分佈的話。顯然，題目的難度如果和受試者能力相當的話，信度可以大幅提高。

提高受試者的估計準確度的另一個方法，就是編製更多的題目，而且題目難度的分佈要和受試者的能力分佈相對應。由表二或圖二至圖五可知，絕大多數的受試者程度分佈遠低於題目難度。因此新的題目的難度應該降低，以圖二為例，新增加的題目難度應該盡量在-5到-2左右。如此一來，從-5到2難度都有了，將能很有效的區分受試者體化症的程度。要編製這類的題目可能並不容易，不過若能編寫出來，將能有效區分一般受試者體化症的程度。程度嚴重者應該入院治療，程度輕微者需要門診治療，再輕微者需要一些諮商輔導措施，更輕微者可以當作體化症早期的警訊。如果測驗的目的在篩選是否該門診或住院治療，那麼現有的12道題目的難度也

表三、原先兩點量尺的信度以及難度調整為適當時的信度

	第一年		第二年		第三年		第四年	
	原先題目	難度恰當	原先題目	難度恰當	原先題目	難度恰當	原先題目	難度恰當
體化症	.70	.88	.75	.85	.78	.88	.53	.75
憂鬱	.69	.86	.76	.84	.80	.85	.54	.62
焦慮	.67	.87	.75	.85	.77	.88	NA	NA
敵意	.56	.79	.66	.83	.68	.82	.04	.22

註：NA表示該時間點並無測量。

許是恰當的。反之，如果測驗的目的在於對一般的(國中)受試者進行危險性評估，希望防微杜漸，那麼這樣的題目難度是不當的。如果我們的主要目的在於看出一般國中生在三年內體化症、憂鬱、焦慮、敵意的變化情形，那麼有必要編寫一些難度較低的題目，才能有足夠的敏感度，鑑別出學生在這三年內些微的變化。

(四) 時間引起的差異試題功能

量表的意義可能會隨時間變動而產生變化，例如原先很難的題目，後來變簡單了。在能力測驗上，最常見到這種情形。一旦題目被做過，就變得無比的簡單。一旦如此，這就意味著這量表在不同時間內，所測得的建構是不一樣的。當建構不一樣時，就無法探討不同時間所產生量的變化。就是表示試題在不同時間所發揮的功能是不同的，因此稱為差異試題功能(differential item functioning, DIF)。關於DIF的研究可以參見Holland與Wainer(1993)。為了檢定是否有DIF，我們利用概率比檢定(likelihood ratio test)兩個階層(nested)模式：一為允許題目有DIF的擴大模式(augmented model)，另一為限制題目無DIF的縮減模式(reduced model)。擴大模式為：

$$\log\left(\frac{p_{ni1t}}{p_{ni0t}}\right) = \theta_{nt} - \delta_{it} \dots\dots\dots(7)$$

其中 p_{ni1t} 和 p_{ni0t} 分別為受試者 n 在時間點 t 時答對(得1分)題目 i 和答錯(得0分)題目 i 的機率， θ_{nt} 為受試者於時間點 t 時的能力， δ_{it} 為試題 i 於時間點 t 的難度。在此模式裡，允許試題難度隨時間不同而不同，也就是允許DIF存在。縮減模式為：

$$\log\left(\frac{p_{ni1t}}{p_{ni0t}}\right) = \theta_{nt} - \delta_i \dots\dots\dots(8)$$

其中 δ_i 為試題 i 的難度，限制題目於不同時間的難度都相同，因此無DIF。如果統計上檢定發現這兩個階層模式達顯著差異，就宣稱題目有DIF，也就是題目於不同時間所代表的意義不同。表四列出四個分量表的擴大模式和縮減模式的概率離差(deviance = $-2 \times \log\text{likelihood}$)，自由度和概率比檢定的 p 值。結果發現在這四個

分量表裡，擴大模式和縮減模式都有統計上的顯著差異，因此沒有DIF的虛無假設被推翻。換句話說，題目的確因為時間產生了變化。

統計上的假設檢定，除了考慮顯著水準外，還要考慮實質的重要性。複雜的擴大模式雖然在統計上與簡單的縮減模式有顯著差異，但並不意味著就不能採用縮減模式。我們利用這兩種模式的信度和相關係數來說明模式的實質差異。從表五所列的信度來看，這兩種模式的信度是一樣的。這兩種模式對受試者能力的估計值的相關係數介於.967至.999之間。總而言之，雖然這兩種模式達到統計上的顯著差異(本研究的樣本數非常大)，但實質上並無差異，因此可以使用縮減模式即可。換句話說，題目的意義雖然隨時間產生了變化，但其影響性很小，可以忽略。

表四、擴大模式與縮減模式之概率比檢定

	縮減模式	擴大模式	差異	自由度	<i>p</i>
體化症	45702.04	45403.52	298.52	27	<.001
憂鬱	45586.85	45488.02	98.82	22	<.001
焦慮	35655.50	35552.36	103.14	18	<.001
敵意	22036.12	22192.88	156.76	12	<.001

表五、擴大模式與縮減模式的信度

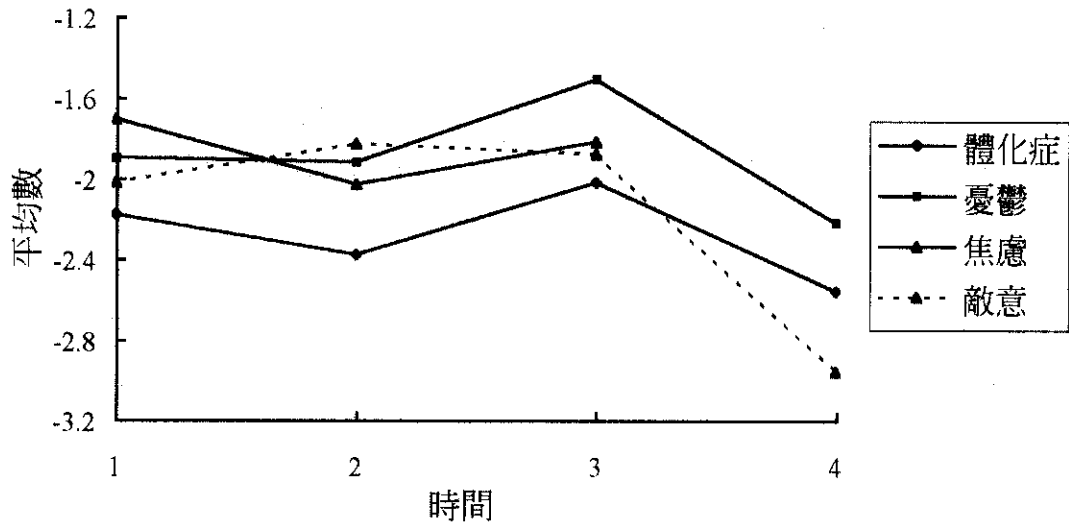
	第一年		第二年		第三年		第四年	
	縮減	擴大	縮減	擴大	縮減	擴大	縮減	擴大
體化症	.70	.70	.75	.75	.78	.78	.53	.52
憂鬱	.69	.69	.76	.76	.80	.80	.54	.53
焦慮	.67	.67	.75	.75	.77	.77	NA	NA
敵意	.56	.56	.66	.65	.68	.68	.04	.05

註：NA表示該時間點並無測量。

(五) 受試者的程度變化

由於題目的參數並沒隨著時間的不同而產生明顯的改變，也就是說沒有重大的DIF，那就可以將題目加以定錨(即固定題目的參數)，然後計算受試者在這四個時間點上的特質變化程度。受試者在這四個時間點(國一、國二、國三、國中畢業)上於四個分量表的平均數和變異數列於表二。由於樣本數很大，因此各時間點的平均數差異(表二)，大致都達到.05的統計顯著水準。在此用logit來表明增加或降低的程度的實質意義。例如就體化症而言，從國一到國二降低了0.21 logits，國二到國三則增加了0.36 logits，國三到畢業後(高一)降低了0.54 logits。換句話說，到了國二時，作任何體化症題目的勝算比是國一的0.81倍($=e^{-0.21}$)。國三的的勝算比是國二的1.43倍($=e^{0.36}$)。國中畢業後的勝算比是國三的0.58倍($=e^{-0.54}$)。表二內的其他平均數的解釋，可以仿照此法。最明顯的變化在於國三到國中畢業後的敵意降低程度：國中畢業後的勝算比只是國三的0.34倍($=e^{-1.08}$)。整體而言，這些受試者在體化症、焦慮上，國二比國一時來得低一些，但到國三時又增加起來。這可能是受試者到國二時，已經比較熟悉環境，能夠適應所致，因此略低於國一剛開學時。到了國三，可能因為聯考的壓力，使得體化症、焦慮又逐漸揚升。在憂鬱量表上，國一和國二沒有差別，但國三時則增加不少。敵意的變化在國中階段，頗為持平，並沒有因為步入國二而降低敵意，相反的，略微揚升。一旦國中畢業，可以發現受試者在這四種量表上，均大福的下降，尤其敵意降得最為明顯。

在四個時間點的相關方面，如表六所示的受試者母群體的變異數-共變數和相關係數估計值，可以發現這四個時間點的相關為中度相關，介於.41至.76之間。表六的相關係數反映出潛在特質的關連性，也就是說不受到測量誤差的弱化(attenuation)干擾。這是因為變異數-共變數矩陣是直接估計所得，而不是先算出受試者的程度估計值後，再利用積差相關求得相關係數。後者作法會受到測量誤差的影響，導致低估潛在特質的關連強度。讀者可參見Wang(1999)和Wang, Chen, & Cheng(in press)關於直接估計變異數-共變數矩陣的作法。



圖六、受試者在四種特質上的四個時間點上的程度變化情形

表六、四個分量表四個時間點的受試者母群體變異數—共變數矩陣與相關矩陣估計值

體化症	第一年	第二年	第三年	第四年	焦慮	第一年	第二年	第三年	
第一年	2.36	1.82	1.59	0.90	第一年	2.41	2.01	1.83	
第二年	.60	3.93	2.84	1.37	第二年	.62	4.35	3.28	
第三年	.53	.73	3.82	1.65	第三年	.54	.72	4.82	
第四年	.42	.49	.60	1.98	敵意	第一年	第二年	第三年	第四年
憂鬱	第一年	第二年	第三年	第四年	第一年	2.56	1.91	1.77	0.98
第一年	2.69	2.31	1.87	1.52	第二年	.64	3.44	2.83	1.48
第二年	.66	4.58	3.35	2.11	第三年	.54	.75	4.11	1.59
第三年	.54	.74	4.42	2.22	第四年	.59	.76	.75	1.09
第四年	.48	.51	.54	3.78					

註：對角線右上角共變數，左下角為相關係數。

五、討 論

在社會科學的縱貫性研究裡，由於處理的變項通常不是物理變項而是心理變項，因此如何確保所測量的分數是等距量尺，而且該變項的意義不會隨著時間不同而產生質變，是首先必須克服的問題。本研究利用Rasch模式分析了SCL-90-R之四種分量表的四個時間點資料。發現題目大致吻合Rasch模式，因此可以獲致等距的量尺分數。對於這些受試者而言，這些題目顯然「偏難」，以致無法有效區分受試者的程度，這是因為SCL-90-R原本的目的就在於鑑別心理症狀的患者。我們建議將來應該多設計一些適當難度的題目(我們現正致力中)，畢竟區分正常人的心理症狀強度，也是有預防勝於治療的功效。集結所有的題目就可以形成題庫。然後依照使用目的，編成多套版本的測驗，例如有專門為鑑別患者的版本，鑑別高危險群的版本，鑑別一般人的版本。不同版本的測驗，仍然可以透過試題反應理論加以連結，使得分數仍然可以比較。這種作法呼應了臨床上希望發展高信度的短式量表的需求。如果能夠進一步發展電腦適性測驗(computerized adaptive testing)，還會再提高測驗的效率。詳細作法可參見Wainer, Dorans, Flaugher, Green, Mislevey, Steinberg與Thissen(1990)。

在縱貫性研究裡，不見得每次時間點都會施測完全相同的量表。例如考慮到記憶的問題，不同時間點可能會施測不同的題目。在本研究裡，有時施測48題，有時47題，有時卻只有15題。如果利用原始分數的相減來代表改變的程度，並不恰當。本研究透過不同時間點的共同題目以及Rasch模式，將不同時間點的分數放在同一個量尺上，因此可以直接相減表示其改變的程度。現今的電腦軟體如ACER ConQuest，可以直接將不同時間點的分數加以連結。關於測驗等化或連結的的詳細作法，可參見Kolen與Brennan(1995)的測驗等化專書。

在提昇測驗的信度(測量準確度)的作法上，有兩種主要的觀點。一是增加題目的點數，例如由兩點量尺換為五點量尺。另一是維持兩點量尺，但增加題目。第一種取向能夠成立的前提是(1)受試者能夠適切區分各個點數所代表的意義(否則豈不是點數無限大最好)，(2)各點的作答反應要足夠分散，(3)難度要恰當。以本研究

分析的資料為例，對於一般的國中生而言，SCL-90-R中五點量尺並沒比兩點量尺提供更高的信度，因為這些題目所描述的情況，對一般的國中生而言，很少發生。即便偶而發生，也很難區辨其強度。難度不恰當，連帶使得受試者無法有效區分五點量尺的意義，因此第一種取向未必對所有的受試者都有效。反觀第二種取向，本研究發現在兩點量尺的信度並不比五點量尺低，而且只要將題目的難度加以調整，就可以大幅提高測驗的信度。如果再增加題數，當然還可以提高信度。

對測驗的編製者而言，增加題目的點數是輕而易舉之事，但增加題目可不容易，因此第一種取向廣受使用。就測驗分數的解釋而言，第二種取向可能比較實用。因為(1)受試者可能無法精確把握點數的語意(例如沒有不舒服、有點不舒服、普通不舒服、嚴重不舒服、很嚴重不舒服的區別見仁見智)，此時還不如單純的使用兩點(沒有不舒服和有不舒服)。有和沒有的語意比較清楚，較五點量尺不易產生混淆。(2)在臨床上，常會詢問個案有否出現某種症狀，尤其是一般的檢核表。測驗的使用者如醫師或臨床心理師，可以很清楚的看出患者有哪些症狀。例如以體化症的12題而言(如圖二)，如以兩點量尺而言，0和1計分，滿分為12分。如果以五點量尺而言，分別計分為0, 1, 2, 3, 4分，滿分為12的話，只需3題，假設這三題為腰酸背痛、想吐或拉肚子、呼吸困難。現有位受試者在這3題的反應分別得分為4分(很嚴重不舒服)、2分(普通不舒服)和0分(沒有不舒服)。另有位受試者是回答兩點量尺的12題，結果發現會頭痛、腰酸背痛、肌肉酸痛、頭暈、心臟或胸口痛、想吐或拉肚子，但沒有其他的六種症狀。對臨床心理師而言，顯然是兩點量尺的12題比五點量尺的3題來得實用。這是因為題目的描述，遠比點數的描述來得容易理解，而且具有臨床價值。因此我們建議測驗的編製者，不要一味的增加點數，而應該盡量朝增加適當難度的題數著手。

在縱貫性的研究裡，題目重複出現，可能會因為練習效果或時代的變遷，使得題目變質。如果題目沒有重複出現，就不會有這些問題，但卻有著更糟糕的問題，那就是無法評量改變。不同時間施測不同題目，將無法將兩次的分數放在同一量尺上進行比較。例如期初的考題很簡單，經過一學期上課後，期末考題非常難，結果每位考生的原始成績都是前高後低，這代表著學習退化！如果有些共同的題目出現

在這兩次考試裡，就可以利用這些共同題目進行連結，將兩此的考試成績放在同一量尺上，進行比較，就可以評量改變的情況。兩次時間點有著共同題的設計，雖然可以化解不同量尺的難題，但卻必須注意題目是否變質。本研究考驗了四個時間點的題目是否變質，結果發現並無明顯的變質，因此可以進行改變的評量。國中生在進入國二時，心理症狀的強度略微比國一時減緩，到了國三，又些微揚升。國中畢業後，則顯著下降。除此之外，四個時間點的心理症狀的改變呈現中度的正相關。

利用Rasch模式或其他的試題反應模式進行測驗資料的分析，已經廣被教育測驗所使用，例如SAT(Scholastic Assessment Test)、GRE(Graduate Record Examinations)、NAEP(National Assessment of Educational Progress)、TIMSS(Third International Mathematics and Science Study)、PISA(Programme for International Student Assessment)等，但在心理量表(如人格測驗、興趣測驗、態度測驗等)上，尚處於推廣階段。除了在教育心理的領域外，Rasch分析也逐漸被使用在醫學、公共衛生、管理、體育等學科的測量上。由於Rasch模式及其延伸模式具有等距量尺和客觀的特性，我們預見會有更多的量表進行Rasch分析。本研究的分析步驟希望能對這類分析的推廣略盡一份心力。

六、參考文獻

- 江麗珍、陳珠璋、彭素玲(1993)：〈日間留院精神分裂病患主題導向之人際互動團體心理治療〉。《職能治療學會雜誌》，11卷，頁51-63。
- 吳齊殷(1997)：〈青少年藥物濫用之起因：一個社會學習模型第一期〉。國家衛生研究院研究計畫DOH86-HR-621。
- 吳齊殷(1998)：〈青少年藥物濫用之起因：一個社會學習模型第二期〉。國家衛生研究院研究計畫DOH87-HR-621。
- 吳齊殷(1999)：〈青少年藥物濫用之起因：一個社會學習模型第三期〉。國家衛生研究院研究計畫DOH88-HR-621。

- 李明濱、李宇宙、李蘭等(1990)：〈簡式精神症狀量表在臨床醫療使用之信度與效度研究〉。《台灣醫學會雜誌》，89卷，頁1081-1087。
- 李明濱、林憲、林信男(1988)：〈門診精神官能性疾患之追蹤研究：結果與心理社會預測因素〉。《中華精神醫學》，2卷，頁105-121。
- 李毅達、吳晉祥、張智仁、陳純誠(1996)：〈健檢個案心理健康影響因子之性別差異〉。《中華精神醫學》，10卷，頁334-345。
- 林文香、夏萍緬、楊文山(1997)：〈無助指標的信效度檢定—以全身性紅斑狼瘡患者為例〉。《護理研究》，5卷，頁452-462。
- 林文香、夏萍緬、楊文山、洪志美(1999)：〈全身性紅斑狼瘡及類風濕性關節炎女病患的身體、心理、社會功能探討〉。《護理研究》，7卷，頁261-275。
- 林木芬(1996)：〈行為抑制、社會關係、人格與婦女焦慮之關係〉。中正大學心理學研究所未發表碩士論文。
- 林玉慈(1999)：〈親子溝通品質與青少年生活適應、偏差行為之相關研究〉。政治大學教育學研究所未發表碩士論文。
- 柯慧貞、孫苑庭、林木芬、葉宗烈、陸汝斌(2000)：〈由行為抑制探討婦女焦慮與憂鬱之共病機制〉。《台灣精神醫學》，14卷，頁13-21。
- 張玉玲(2001)：〈巴金森氏症患者之自我覺知功能〉。台灣大學心理學研究所未發表論文。
- 陳映雪、葉紅秀、余鳳玉(1993)：〈青少年的企圖自殺〉。《中華精神醫學》，7卷，頁234-245。
- 蕭淑貞、陳孝範、張珏(1999)：〈探討壓力調適工作坊改善護理人員壓力症狀之成效〉。《護理研究》7卷，頁90-98。
- 顏永杰、鄭夙芬、楊明仁、何啓功、張明永(2000)：〈某一工業災難救難人員的團體處遇經驗〉。《台灣精神醫學》，14卷，頁41-50。
- 龔曉萍(2000)：〈化學治療期間乳癌病患疲憊感及其相關因素探討〉。國防醫學院護理研究所未發表碩士論文。
- Adams, R. J., & Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random

- coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.
- Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Carpenter, K. M., & Hittner, J. B. (1995). Dimensional characteristics of the SCL-90-R: Evaluation of gender differences in dually diagnosed inpatients. *Journal of Clinical Psychology*, 51, 383-390.
- Cyr, J. J., McKenna-Foley, J. M., & Peacock, E. (1985). Factor structure of the SCL-90-R: Is there one? *Journal of Personality Assessment*, 49, 571-578.
- Derogatis, L. R. (1983). *SCL-90-R administration, scoring, and procedure manual-II*. Towson, MD: Clinical Psychometric Research.
- Derogatis, L. R., & Savitz, K. L. (1999). The SCL-90-R, brief symptom inventory, and matching clinical rating scales. In Mark E. Maruish (Ed). *The use of psychological testing for treatment planning and outcomes assessment* (2nd ed.) (pp. 679-724). Mahwah, NJ: Erlbaum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fischer, G. H. (1973). The linear logistic test model as instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Fischer, G. H., & Ponocy, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, 59, 177-192.

- Fischer, G., & Molenaar, I. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Kolen, M. J., & Brennan, R. J. (1995). *Test equating: Methods and practices*. New York: Springer.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests*. Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)
- Rauter, U. K., Leonard, C. E., & Swett, C. P. (1996). SCL-90-R factor structure in an acute, involuntary, adult psychiatric inpatient sample. *Journal of Clinical Psychology*, 52, 625-629.
- Schmitz, N., Hartkamp, N., Brinschwitz, C., & Michalek, S. (1999). Computerized administration of the Symptom Checklist (SCL-90-R) and the inventory of interpersonal Problems (IIP-C) in psychosomatic outpatients. *Psychiatry Research*, 87, 217-221.
- Schmitz, N., Kruse, J., Heckrath, C., Alberti, L., & Tress, W. (1999). Diagnosing mental disorders in primary care: The general health questionnaire (GHQ) and the symptom check list (SCL-90-R) as screening instruments. *Social Psychiatry and Psychiatric Epidemiology*, 34, 360-366.
- Todd, D. M., Deane, F. P., & McKenna, P. A. (1997). Appropriateness of SCL-90-R

- adolescent and adult norms for outpatient and nonpatient college students. *Journal of Counseling Psychology*, 44, 294-301.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Vassend, O., & Skrondal, A. (1999). The problem of structural indeterminacy in multidimensional symptom report instruments. The case of SCL-90-R. *Behaviour Research and Therapy*, 37, 685-701.
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislavey, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.
- Wang, W.-C. (1999). Direct estimation of correlations among latent traits within IRT framework. *Methods of Psychological Research Online*, 4, 2, 47-68.
- Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (in press). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*.
- Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139-155). Norwood, NJ: Ablex.
- Wang, W.-C., Wilson, M. R., & Adams, R. J. (2000). Interpreting the parameters of a multidimensional Rasch model. In M. Wilson, & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 219-242). Norwood, NJ: Ablex.
- Wilson, M. R. (1992). The partial order model: An extension of the partial credit model. *Applied Psychological Measurement*, 16, 309-325.
- Woessner, R., & Caplan, B. (1996). Emotional distress following stroke: Interpretive limitations of the SCL-90-R. *Assessment*, 3, 291-305.
- Woody, S. R., Steketee, G., & Chambless, D. L. (1995). The usefulness of the obsessive compulsive scale of the Symptom Checklist-90-Revised. *Behaviour Research and Therapy*, 33, 607-611.

- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA.
- Wu, M., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software*. Camberwell, Victoria: Australian Council for Educational Research.