

測驗的建構： 因素分析還是 Rasch 分析？*

王文中**

摘要

測量始於有個被測的變項或建構，根據對該變項的理解，測量的工具於焉產生。在教育或心理測驗中，我們利用試題和受試者所引發的反應，來評估對該建構的測量是否良好。要分析測驗的建構，需從理論上去澄清建構的特性、建構的理論、建構的影響，然後據此設計測量工具。在編製過程中，要力圖測量工具與建構理論緊密配合。有了工具（試題）之後，加以施測，從實證資料中尋找回饋，以修正試題品質、或探索受試者的反應情形，甚至進一步反省建構理論。這種實證資料的分析，有賴於 Rasch（音譯為羅許）模式。當實證資料吻合 Rasch 模式時，我們可以宣稱測量已然進行。

基於這種測量的哲學，若用因素分析來探索測驗的建構，已經犯了邏輯上的錯誤。即使用驗證性因素分析也有困難，因為因素分析的基本單位就是等距的量尺。可是測驗的資料並沒等距的特性，頂多只有順序的特性而已。即使不是等距的資料，也都假設該資料已然來自變項。況且因素分析的結果取決於樣本，隨著樣本的變動，所得的因

* 本文的初稿發表於中央研究院第一屆調查研究方法與應用學術研討會。我感謝中研院統計所劉長萱博士以及兩位匿名評審的指正，以及林以正教授提供台大學生受測資料。本文所有的錯誤仍由作者負責。

** 作者現任職於中正大學心理系。

素結構會很大的變化。基於以上的種種缺失，因素分析不宜當作測驗建構分析的工具。在本文裡，我簡單比較因素分析和 Rasch 分析的理論，說明在測驗建構的釐清上，Rasch 分析更為簡單而直接，並佐以兩個實際的例子來說明兩種分析方法的差異。

測量的意義

心理測驗（包含問卷調查）的兩大功能是：測量(1)受試者間的個別差異，和(2)受試者內隨時間所產生的改變（Anastasi, 1988）。這裡包含兩個重要概念，一是被測量的變項，另一是測量本身。以下我先澄清變項的意義，然後說明測量的本質。透過此，我們將可更瞭解心理測驗的意義。

到底什麼是變項呢？性別是變項，因為有男女之分。身高也是變項，因為有高矮。如果所有的人都是女人，那麼就不存在性別這個變項（甚至也不該叫做女人）。同理，如果所有的人的身高都一樣，那麼就無所謂身高這個變項。因此變項的首要條件就是「變」。如果沒有變，就不是變項。變項的第二個要素就是單向度。身高是變項，體重也是變項，但身高和體重所形成的座標就不是變項。因為每個人在這座標上有兩個值，身高和體重，而不是單一個值。

變項依其性質又可分為質的變項和量的變項。性別、宗教信仰是質的變項。家庭收入、溫度則是量的變項。量的變項可分為間斷的和連續的。間斷的量變項如家庭收入、班級人數等。連續的量變項如身高、體重、溫度等。質的變項通常只用觀察或訪問即可得知，且理論上不會有誤差產生。間斷的量變項通常只需計數（count）即可得知，理論上也不會有測量誤差。連續的量變項一定會有測量誤差存在，不

過只要測量工具夠精確，就可以使測量誤差變得很小。

社會科學所處理的變項有質的變項也有量的變項。本文並不探討質的變項和間斷的量變項，因為它們並不是測量的範圍，也通常不是心理測驗所欲處理的變項，本文所強調的是連續的量變項。值得注意的是，並不是所有的變項都需透過直接測量，例如我們可以定義：

$$\text{體態} = \text{身高} / \text{體重}。$$

此時體態就是一個新形成的變項，因為每一個人都可以得到一個體態的值。我們並不直接測量體態，而是透過測量身高和體重，然後經由上述公式換算求得。至於這個體態的定義是否有意義和價值，這並不是測量所關心的主題，而是人們如何適切的使用這個變項，例如當作選美的一種標準。

密度也是這樣求得的，因為

$$\text{密度} = \text{重量} / \text{體積}，$$

$$\text{體積} = \text{長} \times \text{寬} \times \text{高}。$$

我們並不直接測量密度，而是測得長、寬、高後，換算成體積。同樣的測得重量後，再據以換算成密度。

這種新的變項，是基於理論所衍生出來的，並非從統計上產生。例如我們可以將聯考的科目：數學、國文、英文、自然、社會、三民主義等六科，利用人爲的共識，決定加總的權重，產生總分這個新變項，然後據以決定錄取與否。這個變項是否有意義，取決於人爲的共識（理論），而非取決於統計。

測量的先決條件就是存在一個連續的量變項。重量是個連續的量變項，所以可以測量，長度也是連續的量變項，所以也可以測量。體積則是先測得長度後，換算求得，因此體積並不是直接測得，而是間接求得。同理，密度也是間接求得。間接求得的變項主要仰賴換算的

公式，這個公式可能具有堅實的理論基礎，如體積、密度等。也可能缺乏很好的理論作後盾，如體態。

有了這個連續的量變項後，我們根據對這個變項的認識或理論，設計適當的工具，來進行測量。如果這個認識或理論越完整，測量工具就會越準確。反之，就越不準確。在心理測驗上，這個變項通常稱作建構 (construct)，測量工具通常是試題 (item)。測量源於假設存在某個我們關心的變項，如閱讀能力，而且受試者在這個變項上有高低之分 (否則根本不該叫做變項)，然後根據我們對這個變項的理解，設計了適當試題，將這些試題對受試者進行施測，以觀察其結果。

圖 1 闡述了測量的四個主要步驟。請注意，建構先於試題存在。我們是用試題和受試者的反應來評估對該建構的測量是否良好，而不是用測驗的資料來發現建構。圖 2 說明了受試者、變項、試題間的關係。測量必先存在一個連續的變項，如跳高能力或閱讀能力。然後我們將受試者置於這個變項上。同樣的，也將跳高用的橫桿或題目放在上面。

圖 1 測量的四大步驟

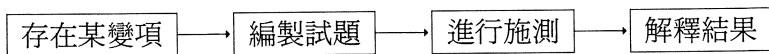
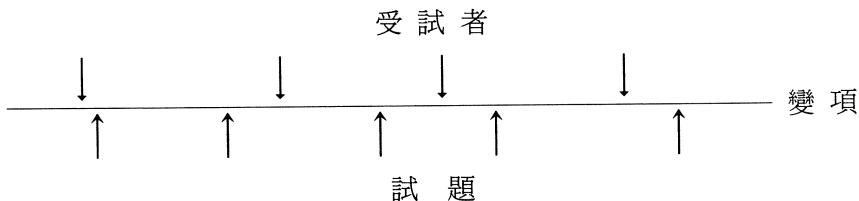


圖 2 受試者、變項、與試題間的關係



在圖2裡，我們發現有多道試題。因為如果只有一題的話，只能將受試者分為兩類：答對和答錯（如果是二分法的試題的話）。有鑑於此，我們又設計了多道試題，期能將受試者有效的區分開來，這才能達到測量個別差異的功能。當然前提是所加進來的試題仍然在同一變項，否則根本不能將受試者在這個變項上更進一步的區分。簡單的說，這些試題都在測同一個變項。

在測量時，無論要測的是外顯變項還是潛在建構（latent construct），它總是個事先既定的目標。要是沒有這個事先既定的目標，就無所謂測量的存在。測量是個有計畫的行動，而不是漫無目標的動作。如果我們對所欲測量的東西，一點都不清楚，那麼根本就發展不出測量的工具，也就毫無測量可言。如果對電流的特性還沒發展出令人滿意的理論，怎能發明好的測量工具，甚至我們怎知測到的是電流強度而不是別的。理論的建立緩慢而痛苦，在初期時，對電流的理論不甚理解，所用的工具也不甚精細。但依賴著這些不甚精細的工具，讓我們做實驗來修正理論，再據以發展更精細的工具。透過這樣的循環過程，理論得以清晰，工具得以精確，文明才得以進步。

就社會科學的測量而言，一直苦於缺乏完善的建構理論，以致無法發展出有效的測量。其實在自然科學界裡，有些建構也是很難測量的。就以「硬度」來說，現今對硬度的測量，只能令被測的物體，兩兩摩擦，看哪一個表面受損，其硬度就較小。所以對硬度的測量只停留在順序的尺度（嚴格說來並沒達到測量的水準），並無法像溫度測量一樣，具有等距的特性。如果有天，我們發現硬度可用某種公式來加以決定（就像密度一樣，由質量和體積來決定），那麼對硬度的測量才夠精確。

雖然社會科學中建構的理論，尚未清楚，但並不表示社會科學的

測量，無法具有等距的特性，以下我將會說明，若資料吻合 Rasch 模式 (Rasch, 1960/1980)，會有等距的特性。如此一來，我們對於社會科學的測量才有信心，才能勉強跟得上自然科學的測量水準——等距量尺。

圖 1 和圖 2 的測量方式和現今有些心理或教育測驗的編製方式並不盡相同。因為在那些心理或教育測驗裡，並不存在著清楚的建構。題目的來源通常是主觀的喜好。其實在這種情況下，每一試題好像是在測不同的建構。這種作法當然並不好，因為在實質上不會有這麼多的建構存在。就算存在，每個建構測一題，也不能達到詳細區分受試者的目的。如果說整個測驗沒有建構，就好像整個測驗沒有在測量，這想必不是測驗編製者願意看到的吧！如果說有在測量，那麼就該有建構存在才是。合理的作法是，依照理論盡可能的將試題歸類，測相同建構的歸一類。也許整個測驗包含數個建構的分測驗。

測驗建構的實證分析：Rasch 分析

由圖 1 和圖 2 中，得知測量始於對某一變項感興趣，據我們對該變項的知識來設計工具，因此該變項早在實際測量之前，就被認知到，或已被假設存在著。如此一來，測驗的建構並不是被發現，而是假設它已存在，看測驗是否發揮測量它的功用。

從圖 2 中，若將受試者 n 在這個變項上的值定為 θ_n （某種能力或特質，為討論簡單起見，在此通稱為能力。），將試題 i 的特性（通常稱作難度，difficulty）定為 δ_i ，測量就是利用各種難度的試題，來將受試者的 θ 給找出來。受試者 n 在試題 i 上的反應取決於 θ_n 和 δ_i 的關係，其中最簡單的式子就是 $\theta_n - \delta_i$ 。 $\theta_n - \delta_i$ 的值介於正負無限大之

間。考慮到作答的隨機性，且以二分題（dichotomous items）為例，受試者 n 答對試題 i 的機率介於 0 和 1 之間。若 $\theta_n - \delta_i$ 大於 0 的話，答對的機率就大於 .5；若 $\theta_n - \delta_i$ 小於 0 的話，答對的機率就小於 .5； $\theta_n - \delta_i$ 等於 0 的話，答對的機率就等於 .5。爲了吻合這些條件，可將 $\theta_n - \delta_i$ 來個自然對數轉換，即 $\exp(\theta_n - \delta_i)$ ，使其值介在 0 和無限大之間。然後將這個值除於 $[1 + \exp(\theta_n - \delta_i)]$ ，就可以達到上述的要求。

換句話說，令 X_{ni} 表示受試者 n 在試題 i 上的反應，答對的話等於 1，答錯的話等於 0。那麼 X_{ni} 等於 1 和等於 0 機率分別是

$$P(X_{ni}=1|\theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad (1)$$

$$P(X_{ni}=0|\theta_n, \delta_i) = \frac{0}{1 + \exp(\theta_n - \delta_i)}. \quad (2)$$

這就是所謂的 Rasch 模式，以紀念其創始者 G. Rasch。

在這個模式裡， θ_n 和 δ_i 是在同一尺度上（要不然也無法相減）。它們具有什麼樣的特性呢？簡單的說，等距。說明如下。將公式(1)除以公式(2)就是所謂的勝算（odd），再將之取對數就是對數勝算子（logit，或直譯爲洛基）。即

$$\log \left[\frac{P(X_{ni}=1|\theta_n, \delta_i)}{P(X_{ni}=0|\theta_n, \delta_i)} \right] = \theta_n - \delta_i. \quad (3)$$

若有兩位受試者作答同一試題 i ，他們的對數勝算子分別是 $\theta_1 - \delta_i$ 和 $\theta_2 - \delta_i$ 。其差異爲 $(\theta_1 - \delta_i) - (\theta_2 - \delta_i) = \theta_1 - \theta_2$ 。若 θ_1 比 θ_2 多了一個單位，其對數勝算子的差異也相對的多了一個單位，無論 θ_1 的位置在何處（是 -5 還是 20 都好），均是相差一個單位，所以 θ 是等

距的尺度。既然 θ 是等距的尺度，那麼和它相同尺度的 δ 當然也是等距的尺度。

既然 Rasch 模式中所獲得的量尺 θ 具有等距的特性，所以可以直接用來進行統計分析，如 t 檢定，變異數分析，相關和迴歸等。可是二參數模式或三參數模式 (Birnbau, 1968) 內的量尺並沒有這種特性，原始分數也是如此。嚴格說起來若用它們來進行參數統計分析，已經違反了其對資料的基本要求。

Rasch 模式還有一個很重要的特性：明確客觀性 (specific objectivity)。簡單的說，題目的難度和受試者的能力彼此互為獨立，不相干擾。換句話說，所謂的明確客觀性必須達到以下兩個條件：(1) 如果受試者 1 比受試者 2 能力高的話，那麼受試者 1 答對任何試題的機率都要比受試者 2 來得高。(2) 如果試題 1 比試題 2 來得難的話，那麼對任何能力水平的受試者而言，答對試題 1 的機率一定要比試題 2 來得低。第一個條件在所有試題反應模式裡，都已吻合。因為試題特徵曲線 (item characteristic curve) 都是單調遞增 (monotonic increasing)。但第二個條件則只有 Rasch 模式吻合，其他二參數或三參數模式並不吻合。

以數學來說，明確的客觀性建立在充分統計量 (sufficient statistic) 上。在 Rasch 模式裡，受試者的原始總分就是其能力的充分統計量；就試題的難度而言，答對該題的總人數就是該試題難度的充分統計量。基於充分統計量，難度參數的估計不會受到樣本能力水平的影響，這就是所謂的樣本獨立的校準 (sample free calibration)；受試者能力水平的估計，不會受到試題難度的影響，也就是所謂的測驗獨立的測量 (test free measure)。因為 Rasch 模式具有優良的測量理念，因此 G. Rasch 本人原意要稱這種模式為「測量的模式」(model

for measurement)，但因恐這種稱呼過於籠統而作罷（Andersen, 1995）。

在此附帶強調的是，使用 Rasch 模式來檢驗資料並不是統計學上所謂的模式吻合（model fitting）的議題。因此，當資料不吻合 Rasch 模式時，並不是換個複雜的模式（如多參數模式）以增進資料的吻合度，而是檢驗資料為何不吻合 Rasch 模式，並改進資料。有關 Rasch 模式的特性，請參閱王文中（出版中）。

在以上的 Rasch 模式裡，只適用於二分題（對和錯）。晚近，有些學者承續明確客觀性的測量理念，發展出多種適用於複雜測驗情境的模式，如適用於評等表的評等量尺模式（rating scale model; Andrich, 1978），適用於部份得分的部份得分模式（partial credit model, Masters, 1982），適用於多種測量層面的多相模式（many-faceted model, Linacre; 1989），適用於同類別但不同計分法的順序分割模式（ordered partition model, Wilson, 1992），適用於複雜測驗情境的隨機係數多項洛基模式（random coefficients logit model; Adams and Wilson, 1996），和多向度隨機係數多項洛基模式（multidimensional random coefficients logit model; Adams, Wilson, & Wang, in press; Wang, 1994; Wang, Wilson, & Adams, in press）。由於這些模式基本上秉持著 Rasch 的精神，所以可通稱為 Rasch 家族（Rasch family）。

建構的驗證（construct validation）是個複雜的過程。在此僅提出重要的三大步驟：

1. 審視這個測驗所欲測量的建構。

誠如前述，測量是個有目標的行為，因此必存在一個所欲測量的建構。測驗的產生必須根據我們對這個建構的理解，如果在測驗裡，

看不到清楚的建構，那麼我們就要懷疑這些試題是否只是一些胡亂的組合而已，還是試題間並無交集，每個試題測的是不同的概念。此時，這種試題的組合嚴格說來並不夠資格說是測驗。要改善這種困境，就須回到最初的測驗編製的理念上。釐清為什麼要編製這個測驗，到底想測什麼等基本問題。

2. 檢查試題的編製是否根據對建構的理解。

由於對建構的好奇，引發我們編製測量工具。我們必須盡可能的使工具和建構間相互對應，這樣才能確保工具的建構效度。試題就是測量工具，因此它必須和建構緊密結合。因為我們對建構的認識通常不會完整，所以所編製出來的測驗也不會完美。雖然我們對建構的認識不夠完整，並不妨礙對它的測量。

以上兩個步驟著重在建構理論上的澄清，經過了這兩個步驟後，我們對所編製出來的測驗較具信心。在這兩大步驟裡，重點是建構的理論，任何的統計模式都幫不上忙。有了清楚完整的理論架構，以及工具與建構的緊密配合，才能有完善的測量。不過在社會科學上，建構的理論仍然未臻完善，當然就不可能有完善的工具，也就不會有精細的測量。在這種情況下，測量變得粗糙。不過仍然有可能達到測量的水準。我們仰賴的就是以下的 Rasch 分析。

3. 實證資料的驗證：Rasch 模式。

由於建構的理論仍然渾沌不清，所以測量的工具還需實證資料的檢驗。Rasch 模式就是用以檢驗測量工具的利器。如果實證資料吻合 Rasch 模式，那麼就知道這些測驗試題已經發揮了測量的功能。因為這些試題是在測同一建構，因此分數的加總才有意義，透過 Rasch 模式所得到的量尺（原始分數的非線性轉換）才具有等距的特性。至於測量的建構到底是什麼，那就是上述兩個步驟要回答的問題。因此配

合前述兩大步驟，我們就可以較有信心的宣稱我們已經對該建構進行了測量。

相反的，如果我們發現有些試題或有些受試者的反應不吻合 Rasch 模式，那麼這就是警訊。此時並不適合改採多參數模式，因為多參數模式所獲得的量尺並無測量的意義，而是要深入檢討試題的特性和受試者的反應。例如有些試題對能力高的人來說，比模式所預期的來的難，但對能力較低的人來說，卻比模式預期的來得簡單。這就表示這個試題對能力高的人來說，測到的是一種建構；但對能力低的人來說，測到的卻是另一種建構。總之，這個試題已包含了無關的難度 (irrelative difficulty)，使得試題無法發揮應有的品質，因此應該加以改良。

例如用英文出題來測數學能力（如美國研究所入學考試，Graduate Record Examination, GRE），對英文不是母語的人來講，這個試題恐怕不是在測數學，因為英文能力產生了干擾。假設台灣學子的數學能力高於美國，但英文能力則偏低。那麼對高數學能力的受試者而言（大都是台灣學子），這種英文的數學試題就偏難了；但對數學能力較低的美國學子來說，這個試題就偏簡單。補救之道，可能在降低該數學試題的英文閱讀難度，讓即使是英文能力不怎麼好的受試者也能瞭解題意。或者我們必須承認這種測驗並不適用於某些非英語系的考生，至少對這些考生而言，該測驗所測到的變項並不是當初我們所預期的數學能力。

同理，如果某受試者對某些試題答對（或答錯）的情形，與模式的預期相當不吻合的話，那麼有可能該受試者對試題的認知和我們的預期不相同，也就是說，這個測驗對他來講可能測到了不同的建構。同樣的用上述 GRE 的試題來說，對台灣的學子而言，如果該數學試題

所使用的英文字彙較簡單的話，該受試者就容易答對；反之，若字彙較難的話，則不易答對。我們以為這些數學題目是在測數學，但對這位台灣的學子而言，還包括了英文的能力。我們必須承認對這位台灣的學子的測量是有問題的，而應予以補救。

除了上述某些試題對不同受試者測到的是不同的建構外，還有可能是猜對、作弊、不小心、胡亂作答等。例如某受試者答對了遠比他能力高很多的試題，可能他猜對、作弊、或採用特殊的解題技巧。如果他答錯了遠比他能力低很多的試題，可能是粗心大意，或者隨意作答。如果我們認為測量對每位受試者都是嚴肅的事情，那麼就該進一步探究箇中的原因。

總而言之，透過 Rasch 模式的檢定，可以據此挑出有問題的試題和受試者，進一步去探索問題所在，加以解決。這就是 Rasch 模式所獨具的診斷功能。

測驗建構的實證分析：因素分析

因素分析被用來分析測驗的建構，已有一段很長的歷史。在因素分析裡，每個變項被化作少數幾個共同因素的線性組合，其目的就在資料的化約 (data reduction)。有人用因素分析來探討因素效度 (factorial validity)，其作法是將數種測驗放在一起，進行因素分析。某些測驗可能被歸為一類，如果這些測驗都有某種特質（如全部是測量語文能力），那麼這個因素就據此特質而命名。此因素和某測驗的相關（即此測驗在此因素上的負荷量，loading）就是因素效度 (Anastasi, 1988)。

因素分析也常被用來探索測驗的結構。研究者針對某一測驗的資

料進行因素分析，選定因素的個數，進行轉軸，然後將試題分為幾類（因素），據以宣稱該測驗含有數個因素，每一因素包含各哪些試題等。

讀者不難發現在量表的編製上（尤其在社會心理學或人格心理學的領域內），因素分析已被廣泛使用。就拿工作滿意度的量表編製來說，研究者可能編列了一些有關的試題，然後抽樣受試者進行施測，再以因素分析方法，對這一些試題進行歸類或刪題。接著研究者對所得到的因素結構，進行命名，據以宣稱該量表含有數個建構。這種作法和上述的測量理論背道而馳。因為在前述的測量理論中，建構先於測量工具而存在，用測量工具來探索建構，並不恰當。況且，當初研究者是想瞭解受試者的工作滿意度，有的人很滿意，有的人滿意度稍低，有人很不滿意，因此工作滿意度就是這個要測驗想測量的唯一建構。怎麼一進行因素分析，該測驗就變成了包含多個建構，這豈不矛盾！

如果研究者當初就認為工作滿意度實際上是多向度的概念，例如包括了實質利益與未來發展這兩個向度。那麼就該分別針對這兩個向度進行量表編製，就好像編製兩個測驗一樣。至於工作滿意度這個變項，可用人為的共識來加以定義，例如定義為實質利益乘以 2 加上未來發展。當然這種人為的定義必須有強而有力的理論作基礎，才會彰顯這種定義的價值。就像是經濟學上的各種指數一樣。以因素分析或者主成分分析等統計方法，來決定權重，並無多大意義，至少不能取代理論。

歸納言之，用因素分析來探索或驗證建構有著以下諸多缺點。第一，誠然在進行因素分析之前，也可進行上節中所述的建構驗證的前兩步驟。不過因素分析並不配合著如圖 1 所示的測量哲學。在圖 1 中，測量發生於有個我們想測量的變項，然後才根據對這個變項的認識編

製測量工具。因此建構已然先於工具存在，何來利用工具所獲得的資料來探索建構呢？這顯然犯了邏輯上的矛盾。

第二，除了在測量哲學上的矛盾外，因素分析在統計層面上也有缺失。例如大多數因素分析的方法都要求資料必須是等距的量尺，可是試題的分數或者測驗的總分並不見得是等距的資料，頂多只是順序的資料，因此用它們來當作因素分析的單位已經違背因素分析對資料的要求。

第三，如果我們用等距的資料來進行因素分析，那麼該等距的資料本身就來自外顯變項或潛在建構，又何來利用因素分析在找尋建構呢？即使我們利用四分相關 (tetrachoric correlation) 或多分相關 (polychoric correlation) 來進行因素分析，仍須假設這些資料已然來自不同的建構。所以進行因素分析只是對這些建構再進行一次歸類，這種歸類頂多只是資料的化約而已，不見得會有實質上的價值。

詳細的說，只用統計方法將變項加以組合而形成新的變項，這個新的變項並無多大意義。例如身高、手肘長度、下肢長度、體重、脖子大小、腰圍等六個變項都是屬於等距的變項，將這六個變項予以因素分析，可能得到兩大因素：第一個因素包括身高、手肘長度、下肢長度；第二個因素包括體重、脖子大小、腰圍。這兩個因素並無多大意義，充其量只是這六個變項的歸類而已。若要這兩個因素有意義，還需完整的理論和實證，光靠因素分析是不可以當作發現或驗證建構的直接證據。

第四，在探索性因素分析裡（事實上並無全然的“探索性”因素分析，因為我們仍然必須採用一些限制，如因素個數，才能求得因素結構和負荷量。）各種因素分析的技術仍莫衷一是。例如該用何種抽取因素的方法？主因素法 (principal factor)，最大概率法 (maximum

likelihood)，還是 alpha 法？要正交轉軸（orthogonal rotation）還是斜交轉軸（oblique rotation）？該用最大變異法（varimax），四方最大法（quartimax），還是均大法（equamax）？又該如何決定共同因素的數目？陡坡檢定（scree test）還是以特徵值（eigen value）？方法不同就會造成因素結構（factor structure）不同。到頭來取決的標準常常就是研究者主觀的判斷，容易解釋最重要。

第五，因素分析是樣本依賴（sample dependent）的統計模式（Wright, 1994）。即使我們用相同的方法進行因素分析，如果換做一個新的樣本，或者樣本人數有變動，或者題數不同（如刪掉或加入一些題目或測驗），通常因素的負荷量，甚至因素結構也會產生相當大的變化。此外，由於因素負荷量或因素結構的不穩定，使得我們通常不得不捨棄因素分數（factor score），而改用屬於該因素的題目的原始分數總和，來代表受試者在該因素的得分。

就試題的變動對因素結構的影響而言，如果將上述的身高、手肘長度、下肢長度、體重、脖子大小、腰圍的六個變項中，將身高和體重換為血壓和心跳數，再進行一次因素分析，所得到的因素結構又會不同。再舉另一個例子，如果將快樂、平安、幸福、以及其他變項進行因素分析，那麼快樂、平安、幸福這三個變項極可能被放在同一因素，因此此時的快樂的建構就是「內在價值」這個概念。假如將平安、幸福換做財富、名聲，重新進行一次因素分析，此時快樂、財富、名聲這三個變項可能被放在同一因素。這時快樂的建構就變為「外在價值」。到底快樂是內在價值還是外在價值，視該因素分析中放的其他變項是什麼特性而定，因此快樂的建構取決於其他變項。從另一個角度來說，快樂可以有無數個建構，也就是毫無建構。如果測驗的建構是由這種方式產生，那麼豈不只是在創造新的詞彙，而這些詞彙可能缺

乏心理學的實質意義。

實例分析

在以下的兩個例子裡，我將分別檢查兩種測驗資料的建構。第一種是二分題的測驗，另一種是 Likert-type 的五點量表。我利用 Rasch 模式（或 Rasch 家族的模式）來分析這兩種測驗。雖然從測量哲學上，我並不建議進行所謂的因素分析，來探索或驗證測驗的建構，不過爲了說明它與 Rasch 分析的差異，以下仍然列出因素分析的摘要結果。

實例一：二分題

在筆者任教的教育心理學的期中考裡，有 15 道配合題，都是專有名詞和概念，考生必須指出這些專有名詞或概念是由何人提出，或和誰最有關連。試題裡共列出 13 位學者的姓名，考生將專有名詞或概念和姓名一一配對，姓名可以重複。答對者得 1 分，答錯得 0 分。受試者爲 88 位大學生。表 1 列出這 15 道試題和 13 位學者的姓名。

表 2 依照受試者得分大小，以及試題被答對的人數，重新排序，並將滿分者刪除，剩下 79 位受試者在 15 道試題的得分情形。我將答錯以空白表示，以利看清受試者與試題的交互作用。從表 2 可以約略看出，左上角的空白（答錯）較多，逐漸往右下角遞減。這是可喜的現象，因爲這代表著得分低的受試者很少答對較難的試題。在第 9 題裡，似乎得低分者也有一些人答對，得高分者也有一些人答錯，這個比例似乎嫌太多了一些。在第 10 題裡，似乎能力低的人太少人答對，能力高的人又似乎太少人答錯，因此也值得進一步分析。

表 1 教心測驗15道配合題與選項

專有名詞或概念	姓名
1. 效果律 (law of effect)	1. D. Ausubel
2. 前導組織 (advance organizer)	2. A. Bandura
3. 順應 (accommodation)、同化 (assimilation)	3. J. Bruner
4. 替代學習 (vicarious learning)	4. J. Dewey
5. 操作制約 (operant conditioning)	5. E. Erikson
6. 發展危機 (developmental crisis)	6. S. Freud
7. 前習俗 (preconventional) 道德期	7. L. Kohlberg
8. 接受學習 (reception learning)	8. A. Maslow
9. 古典制約 (classical conditioning)	9. I. Pavlov
10. 認知表徵 (cognitive representation)	10. J. Piaget
11. 本我 (id)	11. B. Skinner
12. 發現學習 (discovery learning)	12. E. Thorndike
13. 可能發展區 (zone of proximal development)	13. L. Vygotsky
14. 慾力 (libido)	
15. 知覺集中 (perceptual centration)	

Rasch 分析

這 15 道試題主要在測量考生對教育心理學中基本概念出處的理解，得分越高代表越理解其出處，得分越低代表越不理解。在此我假設這個建構存在而且試題也是根據我對這個建構的理解編製而成。當然這些試題只適用於選修的學生，而且也只適用於期中考，不適用於期末時才考，因為課程進度的關係。

由於這些題目是二分法：對或錯，因此可使用 Rasch 模式 (Rasch, 1960/1980) 來分析。就這 15 題而言，考驗其與 Rasch 模式的吻合度，發現只有第 9 和第 10 題的 INFIT t (Wright & Masters, 1982) 值略微超出 ± 2.0 ，分別為 2.2 和 -2.1。所以嚴格說來，這兩題

表2 排序後79位受試者在教心測驗15道配合題的作答情形
(空格爲答錯)

No.	8	4	1	10	2	9	5	12	15	6	7	3	14	11	13	總分
18							1					1	1		1	3
6									1	1		1			1	4
28										1			1		1	5
50										1	1			1	1	5
69					1				1		1	1			1	5
82										1	1		1	1		5
4					1	1				1			1	1	1	6
54		1			1				1						1	6
23									1	1		1	1	1	1	7
35			1				1				1		1	1	1	7
58					1			1	1			1	1	1	1	7
17	1					1	1			1		1	1	1	1	8
46				1		1	1			1			1	1	1	8
72	1							1	1		1	1	1	1	1	8
74					1	1	1	1			1	1	1	1	1	8
86					1	1		1			1	1	1	1	1	8
2		1			1			1		1	1	1	1	1	1	9
40		1				1	1			1	1	1	1	1	1	9
57			1		1	1		1	1			1	1	1	1	9
70			1	1		1		1		1	1	1	1	1	1	9
71	1	1			1	1		1		1	1	1	1	1	1	9
77					1	1		1		1	1	1	1	1	1	9
88		1			1	1			1	1	1	1	1	1	1	9
5			1	1		1	1			1	1	1	1	1	1	10
60					1	1	1			1		1	1	1	1	10
13		1	1		1			1		1	1	1	1	1	1	11
14		1	1		1				1	1	1	1	1	1	1	11
16				1	1	1			1	1	1	1	1	1	1	11
21		1	1		1			1	1	1	1	1	1	1	1	11
27		1			1	1	1	1		1	1	1	1	1	1	11
30			1	1	1				1	1	1	1	1	1	1	11
31			1	1	1				1	1	1	1	1	1	1	11
32		1				1	1	1	1	1	1	1	1	1	1	11
42		1			1	1		1	1	1	1	1	1	1	1	11
43	1		1	1	1		1		1	1	1	1	1	1	1	11
59		1	1	1	1			1	1	1	1	1	1	1	1	11
68			1	1	1	1		1		1	1	1	1	1	1	11
79			1		1	1			1	1	1	1	1	1	1	11
9		1			1		1	1	1	1	1	1	1	1	1	12
10		1			1	1	1	1	1	1	1	1	1	1	1	12
11		1			1		1	1	1	1	1	1	1	1	1	12
15		1				1	1	1	1	1	1	1	1	1	1	12
19	1		1	1	1	1		1	1	1	1	1	1	1	1	12
20	1				1	1		1	1	1	1	1	1	1	1	12
22			1	1	1		1	1	1	1	1	1	1	1	1	12
24		1			1		1	1	1	1	1	1	1	1	1	12
25	1		1	1	1	1		1	1	1	1	1	1	1	1	12
33		1	1		1		1	1	1	1	1	1	1	1	1	12
34		1	1	1	1	1		1	1	1	1	1	1	1	1	12
47	1	1				1	1		1	1	1	1	1	1	1	12
55	1	1	1	1	1			1	1	1	1	1	1	1	1	12
62		1			1	1	1	1	1	1	1	1	1	1	1	12
81			1	1	1	1	1	1	1	1	1	1	1	1	1	12
85			1	1	1	1	1	1	1	1	1	1	1	1	1	12
3	1	1	1	1		1		1	1	1	1	1	1	1	1	13
7		1	1					1	1	1	1	1	1	1	1	13
36		1	1	1	1	1		1	1	1	1	1	1	1	1	13
38		1	1	1	1	1		1	1	1	1	1	1	1	1	13
39	1	1	1	1	1	1		1	1	1	1	1	1	1	1	13
11		1	1	1	1		1	1	1	1	1	1	1	1	1	13
56		1	1	1	1			1	1	1	1	1	1	1	1	13
61	1	1	1		1	1	1	1	1	1	1	1	1	1	1	13
65	1	1	1			1	1	1	1	1	1	1	1	1	1	13
66	1	1	1	1	1		1	1	1	1	1	1	1	1	1	13
67		1			1	1	1	1	1	1	1	1	1	1	1	13
12	1		1	1	1	1	1	1	1	1	1	1	1	1	1	14
26		1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
29		1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
51	1		1	1	1	1	1	1	1	1	1	1	1	1	1	14
52	1	1	1	1	1	1		1	1	1	1	1	1	1	1	14
53	1		1	1	1	1	1	1	1	1	1	1	1	1	1	14
63	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
75	1		1	1	1	1	1	1	1	1	1	1	1	1	1	14
76	1		1	1	1	1	1	1	1	1	1	1	1	1	1	14
78	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
83	1	1	1	1	1		1	1	1	1	1	1	1	1	1	14
84	1	1	1	1	1	1		1	1	1	1	1	1	1	1	14
87	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	14
合計	28	42	44	45	48	48	54	55	56	66	66	73	73	74	77	

較不吻合 Rasch 模式。INFIT t 值遠大於 2.0 者表示能力高的群體和能力低的群體，在該題的得分比例的差異比 Rasch 模式預期為小，因此其 INFIT t 值為正。相反的，若 INFIT t 值遠小於 -2.0 者，表示能力高的群體和能力低的群體，在該題的得分比例的差異比 Rasch 模式預期為大，即該題對能力高者太過簡單，但對能力低者卻過度的難。不過由於這兩個值只是約略超出 ± 2.0 ，所以並不明顯的不吻合 Rasch 模式。

第 9 題的題目是古典制約，正確答案是 I. Pavlov。可是由於答案中 B. Skinner，所以造成考生的混淆，因此使得能力高者和能力低者在這題的差異太小，鑑別力過低。第 10 題是認知表徵，正確答案是 J. Bruner。這一題的鑑別力過高，因為能力高者極易答對，但能力低者卻非常不易答對，主要的原因是因為能力低者很容易選擇 J. Piaget。

現將第 9 題和第 10 題刪除，然後再進行一次 Rasch 分析，結果如表 3，由該表的 INFIT t 可以得知，這剩下的 13 道試題的 t 值均介在 ± 2.0 ，這意味著這些題目頗吻合 Rasch 模式。再者，INFIT MNSQ 的平均數為 1.0，標準差為 .16；INFIT t 的平均數為 0.0，標準差為 1.0。這些也反映出這 13 題頗吻合 Rasch 模式。此外這 13 題難度的平均數為 0 logit（被限定），標準差為 1.46 logits。

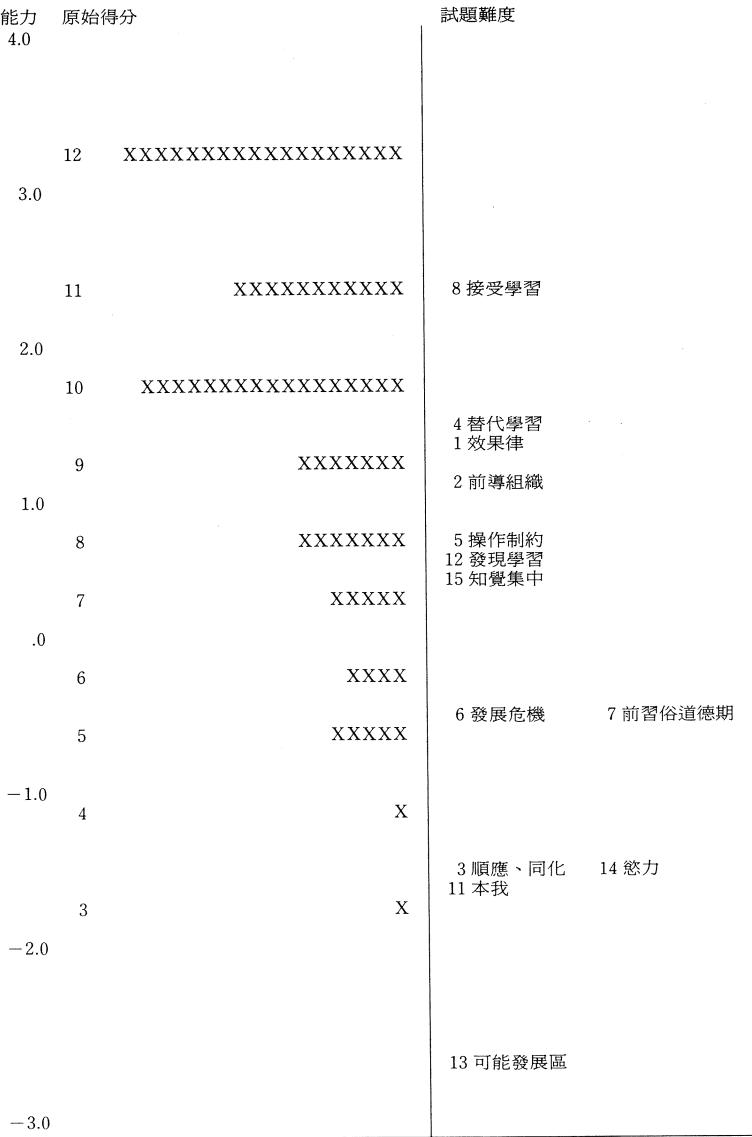
圖 3 說明了考生能力與 13 道試題難度的關係。在 88 位考生中，共有 12 位考生得滿分（對 13 題的測驗而言），因此他們的能力無法估計。剩下的 76 位考生的能力平均值為 1.55 logits，標準差是 1.28 logits。由考生能力和試題難度的分佈情形和平均數來看，考生能力的分佈略高於試題難度，這表示考生的平均能力高於試題難度。此外，共有 18 位考生的能力在 3.0 logits 以上，顯然對這些人來講，試題太過簡單，因此無法詳細區分他們的能力差異。改進之道在於編製困難

表 3 教心測驗刪除第9和10題後試題參數估計與模式吻合度分析

ITEM	SCORE	MAXSCR	THRSH	INFT MNSQ	INFT t
1	41	88	1.33 .27	1.03	.3
2	45	88	1.06 .27	1.22	1.7
3	70	88	-1.42 .46	1.25	.8
4	39	88	1.47 .27	.88	-1.0
5	51	88	.63 .29	1.11	.8
6	63	88	-.43 .34	.90	-.4
7	63	88	-.43 .34	.79	-1.1
8	25	88	2.41 .28	.90	-.8
11	71	88	-1.63 .50	.93	-.1
12	52	88	.56 .29	.79	-1.5
13	74	88	-2.61 .74	1.04	.3
14	70	88	-1.42 .46	.99	.1
15	53	88	.48 .29	1.21	1.4
Mean			0.00	1.00	0.0
SD			1.46	.16	1.0

註：MAXSCR：樣本數。
THRSH：試題難度和標準誤（置於試題難度之下）。

圖 3 受試者能力與試題難度參數的分佈



註：每個 X 代表 1 位考生。

較高的試題。

在 Rasch 分析裡，我們發現共有兩題約略不吻合 Rasch 模式，這是因為題目特性所引起，將之刪除後，發現剩下的 13 題均頗吻合 Rasch 模式。因此，這 13 題對這群考生而言，的確發揮了測量的功能。

探索性因素分析

計算這 15 題的積差相關矩陣後進行主軸因素分析，共可得到 5 個特徵值大於 1.0 的因素，第一個因素的特徵值為 3.80，佔總變異數的 25.3%。第二個因素只佔 10.1%，以此類推。（用四分相關求得相關矩陣後進行因素分析的結果，亦大同小異。）這 5 個因素是轉軸前的因素矩陣列於表 4。基本上，第一個因素的負荷量較其他因素為大，其中以第 10 題的負荷量最大。經直交最大變異（varimax）轉軸後，因素

表 4 教心測驗未轉軸前的因素矩陣

Item No.	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
10	0.70	-0.37	-0.08	0.09	-0.16
12	0.64	0.43	-0.13	0.19	-0.09
6	0.62	-0.17	0.23	-0.32	0.20
7	0.60	0.03	0.22	-0.09	-0.16
4	0.59	0.31	0.09	-0.04	0.33
8	0.49	-0.01	-0.11	0.09	-0.09
5	0.49	-0.30	0.22	-0.18	-0.10
1	0.49	-0.18	-0.30	0.11	-0.22
15	0.35	-0.27	-0.12	0.19	0.07
14	0.33	0.23	-0.08	-0.14	0.00
11	0.30	0.23	0.01	-0.19	-0.01
9	0.26	0.08	0.21	-0.03	0.08
3	0.12	0.08	0.56	0.55	-0.01
2	0.31	0.32	-0.33	0.07	0.00
13	0.22	-0.28	-0.26	0.21	0.48

負荷量矩陣如表 5 所示。

通常測驗分析者會根據表 5 的因素矩陣，而宣稱此 15 題共有 5 個因素，其中第一個因素是第 10，1，8，15 題；第二個因素是第 12，4，2，14，11 題；第三個因素是第 6，5，7，9 題；第四個因素只有第 3 題；第五個因素也只有第 13 題而已。也許分析者會將第四和第五個因素捨棄，因為各只有一題。也許他會將第 15，14，11，和第 9 題捨棄，因為負荷量未達 .4 的標準。

在上述的 Rasch 分析中，我們發現第 9 題的鑑別度過低，第 10 題

表 5 教心測驗直交轉軸後的因素矩陣

題號 題目	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
10 認知表徵	0.73	0.10	0.33	0.03	0.14
1 效果律	0.60	0.21	0.03	-0.08	0.07
8 接受學習	0.40	0.29	0.16	0.05	0.07
15 知覺集中	0.39	0.02	0.09	0.06	0.29
12 發現學習	0.30	0.72	0.11	0.21	0.00
4 替代學習	0.00	0.54	0.44	0.15	0.24
2 前導組織	0.14	0.53	-0.09	-0.06	0.07
14 慾力	0.07	0.38	0.18	-0.07	-0.04
11 本我	0.01	0.33	0.24	-0.06	-0.09
6 發展危機	0.20	0.12	0.73	-0.05	0.16
5 操作制約	0.38	-0.06	0.52	0.01	-0.04
7 前習俗道德期	0.33	0.26	0.48	0.15	-0.13
9 古典制約	0.01	0.13	0.29	0.16	0.02
3 順應、同化	0.02	-0.05	0.08	0.80	-0.01
13 可能發展區	0.17	0.01	0.02	-0.03	0.66

的鑑別度過高。但在因素分析的表 5 的第一個因素負荷量上，我們卻發現第 10 題的負荷量最大，因此是“最佳”的試題。第 9 題的因素負荷量很小，因此並不是很好的試題。第 11 和第 14 題的負荷量也不大，但在 Rasch 分析中，這兩道試題還蠻吻合 Rasch 模式。由此可見，這兩種分析方式會得到很不同的結論。

在因素分析裡，除了用特徵值大於 1.0 以上者來決定因素個數外，通常也會參考陡坡圖 (scree plot) 來檢定因素個數 (Cattell, 1966)。圖 4 中的陡坡圖可以發現特徵值從第一個因素陡降後，自第二個因素起即逐漸變小。因此依照這個標準，共同因素只有一個。這種陡坡檢定的方式顯然比用特徵值大於 1.0 的標準來取捨因素個數，較和 Rasch 分析的結果相近。

驗證性因素分析

我們將這 15 題視為一個因素，並限定同類 (congeneric) 模式，¹ 然後進行驗證性因素分析 (Jöreskog & Sörbom, 1988)，所得的結果如表 6 所示。其中第 3 題和第 13 題的 t 值未達顯著水準。資料與模式的吻合度由卡方檢定來看， p 值為 0.15，已達 05 顯著水準，故資料與模式並不十分吻合。基於以上的結果，此時研究者可能會將第 3 題和第 13 題刪除，然後重新進行一次因素分析。雖然此時所有的估計值均達顯著水準，可是在資料與模式的吻合度上，並沒有改善多少。

現在讓我們將由 Rasch 分析中得知較不吻合的第 9 和第 10 題刪除，再進行一次因素分析，結果如表 7。此時的資料吻合度略有改善，並未達 05 顯著水準，雖然第 3 題和第 13 題的估計值仍未達顯著水

1 若用複本 (τ -equivalent) 模式來進行因素分析，在資料的吻合度上比同類模式稍差。不過複本模式較類似 Rasch 模式，而同類模式較類似二參數模式。

圖 4 教心測驗特徵值的陡坡圖

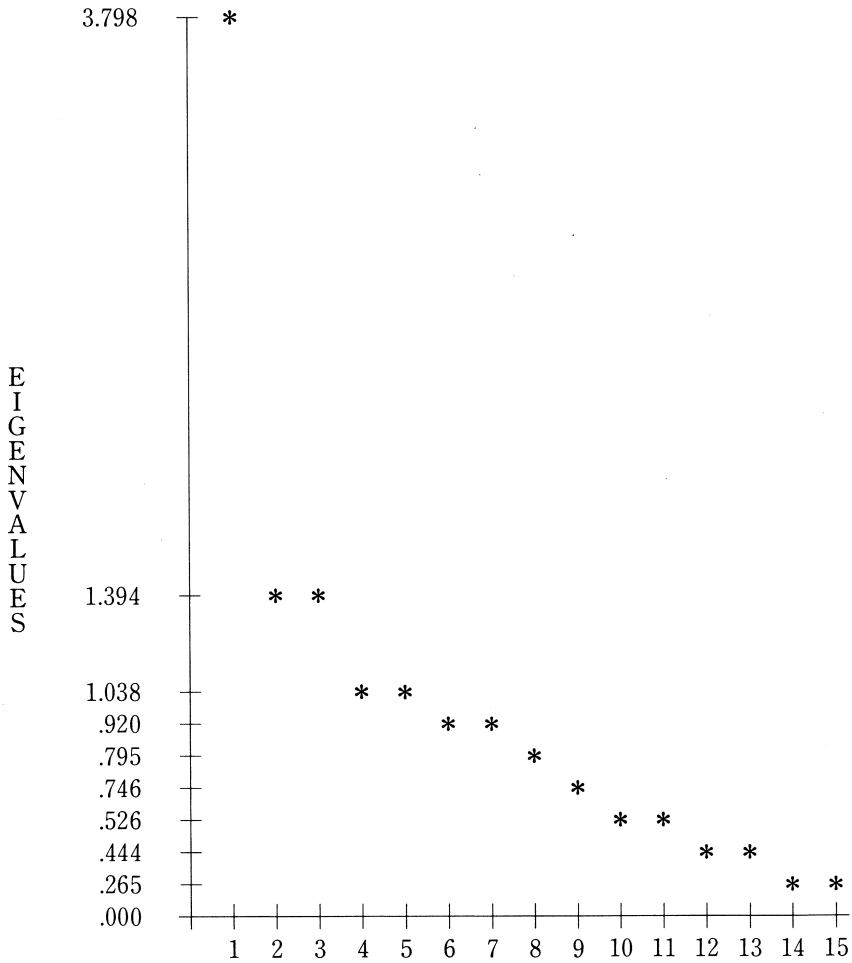


表6 教心測驗參數估計值、標準誤與 t 值與模式吻合度分析

Item No.	Estimates	S. E.	t-value
1	0.48	0.11	4.31
2	0.29	0.12	2.52
3	0.10	0.12	0.88
4	0.55	0.11	5.03
5	0.50	0.11	4.51
6	0.60	0.11	5.63
7	0.61	0.11	5.76
8	0.50	0.11	4.51
9	0.26	0.12	2.23
10	0.68	0.10	6.60
11	0.29	0.12	2.54
12	0.58	0.11	5.39
13	0.20	0.12	1.70
14	0.32	0.12	2.80
15	0.36	0.11	3.15

Fit Index

CHI-SQUARE WITH 90 DEGREES OF FREEDOM=121.43 (P=.015)

GOODNESS OF FIT INDEX=.844

ADJUSTED GOODNESS OF FIT INDEX=.792

ROOT MEAN SQUARE RESIDUAL=.082

準。如果我們事前並未經由 Rasch 分析，得知第 9 題和第 10 題並不十分恰當而應刪除，那麼這個驗證性的因素分析是無法進行的。

從上三種分析方式，獲得幾乎完全不同的結果。Rasch 分析能夠吻合理論的假設，而因素分析卻已經明顯違反線性資料的假設。而且 Rasch 分析是樣本獨立，而因素分析卻是樣本依賴。

表 7 教心測驗刪去第9和第10題後參數估計值、
標準誤與 t 值與模式吻合度分析

Item No.	Estimates	S. E.	t-value
1	0.42	0.12	3.63
2	0.34	0.12	2.91
3	0.10	0.12	0.86
4	0.64	0.11	5.92
5	0.42	0.12	3.63
6	0.55	0.11	4.96
7	0.61	0.11	5.58
8	0.49	0.11	4.30
11	0.34	0.12	2.94
12	0.64	0.11	5.94
13	0.17	0.12	1.39
14	0.36	0.12	3.05
15	0.28	0.12	2.33

Fit Index

CHI-SQUARE WITH 65 DEGREES OF FREEDOM = 84.48 (P.053)

GOODNESS OF FIT INDEX = .872

ADJUSTED GOODNESS OF FIT INDEX = .821

ROOT MEAN SQUARE RESIDUAL = .080

實例二：Likert-type 的多分題

寂寞 (loneliness) 的測量一直是社會心理學家所關心的主題之一。Russell, Peplau, 和 Cutrona (1980) 修訂了 UCLA 寂寞量表，該量表中共有 20 題，其中一半是正向題，一半是負向題，全部採用 Likert-type 的四點量尺。將負向題的計分方式更正後，得分越高表示

越寂寞。以 162 位大學新生為受試者，結果發現內部一致性 Cronbach α 係數高達 .94。此 UCLA 量表經翻譯成中文後，對 131 位台大學生進行施測，其中男生 36 人，女生 94 人。

Rasch 分析

利用 Masters (1982) 的部份得分模式進行分析，結果其中第 9 題、第 14 題、第 15 題較不吻合模式的預期。第 9 題（我是一個外向的人，反向計分）的 INFIT t 值為 2.6，表示此題在鑑別寂寞程度上的比 Rasch 模式預期來得低。簡單的說，能力高者（較寂寞者）在本題的得分上，和能力低者（較不寂寞者）的差異並不夠大。第 14 和第 15 題的 INFIT t 分別為 -2.2，-2.1，也約略超出顯著水準，不過由於幅度不大，因此在此並不將這兩題視為不吻合 Rasch 模式。第 9 題是關於內外向性格，可能和寂寞這個概念並不接近，因此導致寂寞者和較不寂寞者在這題的得分上的差異過小。

將第 9 題刪除後，再重新進行分析，結果如表 8 所示，由該表的 INFIT t 可以得知，所有試題的 t 值均介在 -2.1 和 2.0 之間，這意味著這些題目頗吻合 Rasch 模式。INFIT MNSQ 的平均數為 1.0，標準差為 .14；INFIT t 的平均數為 -0.1，標準差為 1.2。這也反映出這 19 題頗吻合 Rasch 模式。此外這 19 題難度的平均數為 0 logits（被限定），標準差為 0.74 logits。

就試題難度與受試者能力水平來看，如圖 5 所示，大多數題目的「幾乎都是」選項顯得太難些（圖 5 內小數點後為 3 的題目）。負向題的第 7 題（我沒有與任何人親近）的「常常如此」選項也是非常的難。如果本測驗的目的是為了區分類似此一樣本（一般大學生）的受試者，那麼宜降低試題的難度，或增加一些低難度的試題。當然如果本測驗的目的是為了偵測寂寞的高危險群，那麼就無須降低試題的難度。

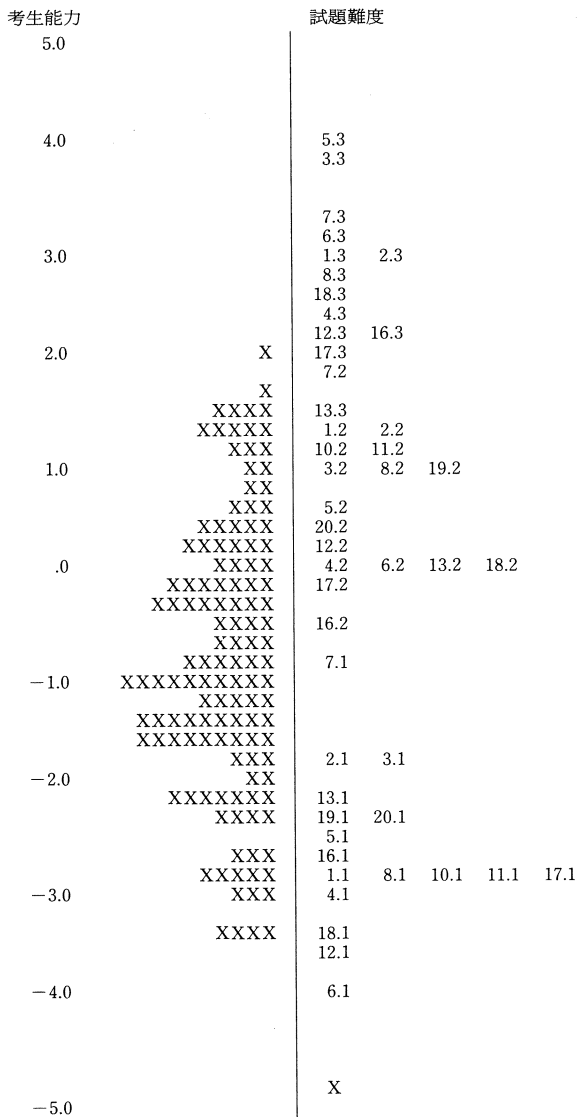
表 8 寂寞量表刪除第9題後試題參數估計與模式吻合度分析

ITEM	SCORE	MAXSCR	DELTA/S 1	2	3	INFT MNSQ	INFT t
1	128	390	-3.00 .33	1.58 .35	2.85 .77	1.06	.5
2	109	390	-1.86 .28	1.40 .34	2.93 .76	.89	-.9
3	114	390	-1.76 .28	.82 .31	3.91 1.04	.96	-.3
4	162	390	-3.18 .33	.05 .28	2.32 .48	1.08	.6
5	135	390	-2.54 .30	.55 .30	4.03 1.04	.98	-.1
6	167	390	-4.15 .43	.11 .29	3.07 .64	1.23	1.7
7	80	390	-1.01 .27	2.05 .40	3.20 1.05	.95	-.3
8	135	390	-2.90 .32	1.07 .32	2.69 .64	.99	0.0
10	133	260	-2.96 .32	.99 .32		1.04	.3
11	130	260	-2.98 .33	1.19 .32		.96	-.3
12	165	390	-3.66 .37	.26 .28	2.07 .46	1.01	.1
13	156	390	-2.21 .28	.01 .27	1.31 .37	1.14	1.2
14	128	390	-2.45 .30	.84 .31	3.92 1.04	.76	-2.1
15	133	390	-2.42 .30	.54 .29	4.04 .103	.79	-1.9
16	172	390	-2.76 .32	-.56 .27	2.14 .43	1.19	1.6
17	166	390	-2.91 .32	-.13 .27	1.88 .41	1.26	2.0
18	163	390	-3.45 .36	.10 .28	2.51 .52	.97	-.2
19	127	260	-2.44 .30	.82 .31		.82	-1.7
20	137	260	-2.45 .30	.30 .29		.84	-1.5
Mean			0.00			1.00	-.1
SD			.74			.14	1.2

註：MAXSCR：樣本數。

DELTA/S：試題梯級（step）難度和標準誤（置於試題梯級難度之下）。

圖5 寂寞量表受試者能力與試題難度參數的分佈



註：每個 X 代表 1 位考生。試題難度 i, j 表示第 i 題第 j 個梯級的難度（即得分在 j 以上和以下者的機率均是.5）的位置。

探索性因素分析

將這 20 題進行主軸因素分析，共有 4 個特徵值大於 1.0 的因素。第一個因素佔總變異數的 38.3%，第二個因素只佔 7.8%，逐次漸小。未轉軸前的因素結構如表 9 所示。如果依照特徵值大 1.0 的標準來選取這 4 個因素，並加以直交最大變異轉軸後的因素矩陣列於表 10。

在上述的 Rasch 分析中，發現第 9 題的鑑別度過低，第 14 和第 15 題題的鑑別度略微過高。在因素分析裡，從表 9 中，可發現第 15 和

表 9 寂寞量表未轉軸前的因素矩陣

Item No.	Factor 1	Factor 2	Factor 3	Factor 4
15	0.73	-0.01	0.13	-0.03
14	0.72	0.27	0.07	0.12
19	0.72	-0.50	-0.11	-0.28
20	0.71	-0.48	-0.10	-0.22
2	0.66	0.04	0.25	0.04
3	0.62	0.14	-0.11	-0.04
13	0.62	0.32	-0.50	0.14
5	0.62	-0.20	0.14	0.06
18	0.61	0.38	0.01	-0.34
7	0.60	0.13	-0.26	0.20
12	0.58	0.09	-0.11	-0.21
11	0.58	0.12	0.19	-0.05
4	0.58	-0.01	0.16	0.15
8	0.56	0.07	-0.08	0.31
16	0.54	-0.10	-0.32	0.12
17	0.54	0.36	0.22	-0.35
10	0.53	-0.28	-0.23	0.01
1	0.50	0.00	0.23	0.23
6	0.45	-0.23	0.25	0.25
9	0.43	-0.07	0.30	0.11

表 10 寂寞量表直交轉軸後的因素矩陣

Item No.	Factor 1	Factor 2	Factor 3	Factor 4
6	0.57	-0.02	0.09	0.22
1	0.54	0.15	0.19	0.08
2	0.54	0.37	0.18	0.20
4	0.50	0.21	0.24	0.18
5	0.49	0.18	0.17	0.38
9	0.49	0.16	0.03	0.13
15	0.47	0.39	0.25	0.32
14	0.46	0.44	0.45	0.07
17	0.21	0.73	0.07	0.06
18	0.12	0.73	0.26	0.14
12	0.13	0.44	0.30	0.32
11	0.40	0.42	0.17	0.15
13	0.04	0.29	0.80	0.13
7	0.24	0.19	0.60	0.17
16	0.16	0.07	0.50	0.37
8	0.40	0.11	0.49	0.10
3	0.23	0.38	0.41	0.23
19	0.26	0.22	0.14	0.85
20	0.29	0.19	0.16	0.80
10	0.19	0.05	0.34	0.52

第 14 題在第一個因素的負荷量最大，反而是“最佳”的試題。第 9 題在第一個因素上的因素負荷量最小，並不是很好的試題。由此看來，在 Rasch 分析中，鑑別度過高和過低都不是好的試題，因為它們和其他試題並不是在測同一特質。但在因素分析中，容易將鑑別度過高的試題，誤認為最佳的試題。

由圖 6 中的陡坡圖可以發現特徵從第一個因素陡降後，自第二個因素起即逐漸變小，因此共同因素只有一個。和例子一一樣，這種用陡坡檢定的方式較和 Rasch 分析的結果相近。

驗證性因素分析

將這 20 題視為一個因素，並限定同類模式，進行驗證性因素分析，所得的結果如表 11 所示，每一題的 t 值均達顯著水準。資料與模式的吻合度由卡方檢定來看， p 值為 .000，已達 .05 顯著水準，故資料與模式並不吻合。從上述 Rasch 分析中，我們發現第 9 題較不吻合，在此將第 9 題刪除，重新進行一次因素分析。此時所有的估計值仍均達顯著水準，可是在資料與模式的吻合度上，仍然沒有改善。

結 論

本文試圖從測量的觀點，來看心理測驗和教育測驗的建構。首先說明測量的先決條件就是存在一個可測的變項，根據我們對這個變項的認識，測量的工具才得以產生。在教育或心理測驗上，這個變項通稱為潛在的建構，測量的工具常是試題。由於一道試題無法詳細區分受試者，所以又多編了幾道試題。無論如何，這些多編出來的試題和原先的試題一樣，意欲針對這個建構將受試者加以區分。因此這些試題必須測試相同的特質或能力。

由於我們對建構的認識不夠完整，對所編出來的試題也不見得有完全的把握，因此必須經過實證資料的檢驗。Rasch 模式就是一個強有力的檢驗工具。當資料吻合 Rasch 模式時，就較有信心的宣稱測量已經發生，因為所得到的尺度具有等距的特性。反之如果不吻合

圖 6 寂寞量表特徵值的陡坡圖

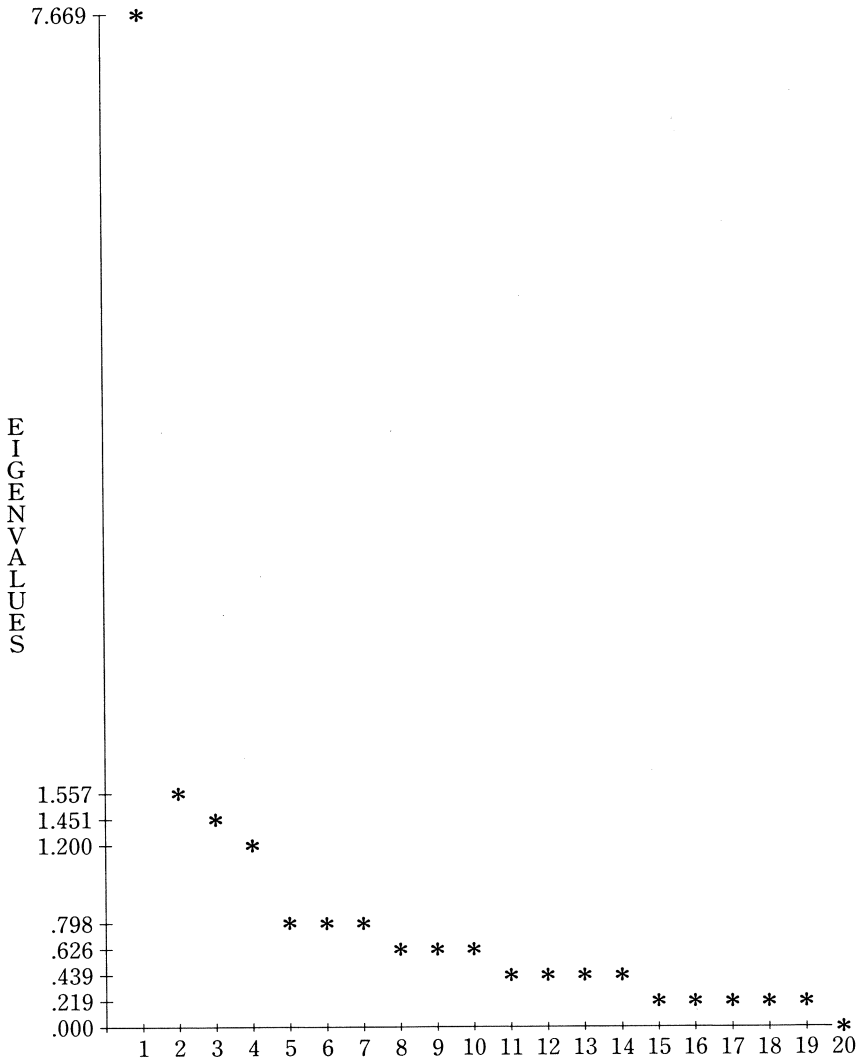


表 11 寂寞量表參數估計值、標準誤與 t 值與模式吻合度分析

Item No.	Estimates	S. E.	t-value
1	0.50	0.09	5.90
2	0.66	0.08	8.25
3	0.63	0.08	7.77
4	0.58	0.08	6.99
5	0.62	0.08	7.52
6	0.44	0.09	5.07
7	0.59	0.08	7.16
8	0.55	0.08	6.58
9	0.42	0.09	4.86
10	0.53	0.08	6.27
11	0.58	0.08	7.02
12	0.59	0.08	7.08
13	0.58	0.08	7.02
14	0.71	0.08	9.11
15	0.73	0.08	9.45
16	0.54	0.08	6.39
17	0.53	0.08	6.21
18	0.59	0.08	7.21
19	0.69	0.08	8.75

Fit Index

CHI-SQUARE WITH 170 DEGREES OFFREEDOM=446.67 (P=.000)

GOODNESS OF FIT INDEX=.743

ADJUSTED GOODNESS OF FIT INDEX=.683

ROOT MEAN SQUARE RESIDUAL=.082

Rasch 模式，那麼應對該試題或受試者深入分析，加以修正或放棄。

基於以上的理念，我們瞭解測驗的建構先於測量的工具。利用測量的工具來測量這個建構，而不是用工具來探索建構。因此，根本的問題在於受試者和試題所產生的行為是否吻合我們對該建構的預期，這就是 Rasch 分析的貢獻。

反觀探索性因素分析，研究者常用它來探索測驗的建構。這種作法和上述的測量哲學並不吻合。換句話說，建構反而存在於測量工具之後，這顯然犯了邏輯上的倒錯。此外就統計層面而言，大多數的因素分析，無論是探索性還是驗證性，其運算單位都必須是等距，但測驗的資料頂多只有順序的特性，因此就違反因素分析的假設。如果已經是等距的資料，那麼測驗的建構已然存在，何來用因素分析找另外建構呢？即使不是等距的資料，也假設背後已然存在著不同的建構。此外因素分析的結果仰賴於樣本，換了新的受試者或試題，所得的因素結構和負荷量可能截然不同。

在實例分析裡，我舉例說明如何用 Rasch 分析和因素分析來理解測驗的建構。我利用教育心理學的期中考的試題和寂寞量表，說明 Rasch 模式頗吻合這些資料，但因素分析卻得到不同的因素結構。如果我們的測驗要能吻合測量的哲學，那麼 Rasch 分析才是適當的檢驗工具。

參考資料

王文中

(出版中) 幾個有關 Rasch 測量模式的爭議。《教育與心理研究》，19。

Adams, R. J., & Wilson, M.

1996 Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard and M. Wilson (Eds.), *Objective measurement: Theory into practice*, Volume 3, Norwood, NJ: Ablex.

Adams, R. J., Wilson, M., & Wang, W.

(in press) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*.

Anastasi, A.

1988 *Psychological testing* (6th ed.). New York: Macmillan.

Andersen, E. B.

1995 What Georg Rasch would have thought about this book. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Andrich, D.

1978 A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.

Birnbaum, A.

1968 Some latent trait models and their use in inferring an examinees ability. In F. Lord and M. Novick, *Statistical theories of mental test scores*. Reading, Mass: Addison-Wesley.

Cattell, R. B.

1966 The meaning and strategic use of factor analysis. In R. B. Cattell (ed.), *Handbook of multivariate experimental psychology*. Chicago: Rand McNally.

Jöreskog, K. G., & Sörbom, D.

1988 *LISREL 7: A guide to the program and applications*. Chicago: SPSS, Inc.

Linacre, J. M.

1989 *Many-faceted Rasch measurement*. Chicago: MESA Press.

Masters, G. N.

1982 A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Rasch, G.

1960/1980 *Probabilistic models for some intelligence and attainment tests*.
Copenhagen: Danmarks Paedagogiske Institut.

Russell, D., Peplau, L. A., & Cutrona, C. E.

1980 The revised UCLA loneliness scale: Concurrent and discriminant evidence. *Journal of Personality and Social Psychology*, 39, 472-480.

Wang, W.

1994 *Implementation and application of the multidimensional random coefficients multinomial logit model*. Unpublished doctoral dissertation, University of California at Berkeley.

Wang, W., & Wilson, M. R., & Adams, R. J.

(in press) Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard & K. Draney (Eds.), *Objective measurement: Theory into practice*. Volume 4, Norwood, NJ: Ablex.

Wilson, M. R.

1992 The partial order model: An extension of the partial credit model. *Applied Psychological Measurement*, 16, 309-325.

Wright, B. D., & Masters, G. N.

1982 *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D.

1994 *Rasch factor analysis*. ERIC Document Service Reproduction No. ED 380476.