Rasch model 於醫療領域之應用

謝清麟教授

臺灣大學醫學院職能治療學系

2018年年底

1970年代Rasch model即開始應用於醫療領域，但1970～1990年二十年期間應用很少。直到近10年，醫療領域才有較多學者採用。筆者於1994年接觸 Rasch model，第一次知道其功能強大，但沒信心未採納。直到2002年王文中教授至學系演講，才強化筆者信心並從此開啟長期採用Rasch model之大道。迄今（2018年年中）筆者以Rasch model為主要研究方法或分析方法，已發表25篇收錄於SCI/SSCI期刊之論文，其中多數期刊之學門排名於20%之內。因此Rasch model已被醫療領域學者積極運用。

筆者在王文中教授的帶領之下，應用Rasch model 於醫療領域之評估工具的驗證與發展，計四大方面：一、評估工具之心理計量特性驗證：1-9主要包含unidimensionality、信度、differential item functioning (DIF) 與個別層級反應性 (individual-level responsiveness) 等，以確認評估工具是否具有良好的心理計量特性。二、量尺轉換：2, 7, 10將評估所得原始分數（順序量尺），應用Rasch model 將轉換成等距量尺，以提升分數之統計分析與應用價值。三、簡化評估工具：11-13應用單向度或多向度Rasch model以簡化評估工具之項目數，但維持良好信度。簡化的評估工具有助於臨床或研究人員採用，並提升臨床效能。四、發展新評估工具：14, 15包含傳統紙筆評估工具與computerized adaptive testing (CAT)。

本章節包含諸多王文中教授指導或其學生（施慶麟博士、陳承德博士或周業太博士）協助完成之論文為例，說明上述Rasch model 於醫療領域之評估工具四大應用方面以及應用限制與後續發展等，主要內容如下：

第一節、評估工具之心理計量特性驗證

第二節、量尺轉換

第三節、簡化評估工具

第四節、發展新評估工具

第五節、應用限制與後續發展

筆者將以Rasch model 於醫療領域評估工具之實務應用（含題材選擇、資料分析、結果呈現及解讀、論文撰寫與發表歷程）與讀者分享，我們使用的Rasch相關統計軟體以 WINSTEP 或 CONQUEST為主。這些分享除了感念王文中教授的指導，亦期待讀者能應用Rasch model 於各自領域，以驗證、改良或發展評估工具，以期提升臨床專業與學術應用效能。

第一節、評估工具之心理計量特性驗證

評估工具之心理計特性（如信度、效度或反應性等）為其品質指標，也是使用者極其重視的特質。這些品質指標影響醫療領域常用問卷或量表（各具備多項目）之效能。如以評估工具評估個案/家屬之身心功能或醫療滿意度等是否精準有效或能反應個案之變化（以呈現療效）。醫學教育領域亦常用問卷或量表，藉以評估實習學生之學習狀況或標準病人之表現。醫學教育領域使用之問卷或量表的心理計量特性如何？也是使用者（含考官、學生與標準病人等）關心之特質，若評估工具之心理計特性不佳，則評估結果/數據將難以達成評估之目的。

Rasch model 可驗證評估工具之心理計量特性（含信度、效度或反應性等）。信度方面：Rasch分析可呈現個別評估分數之標準誤 (standard error, SE)，所以可以估算每位個案評估分數之SE，亦可換算成個案評估分數之信度。既有每位個案分數之信度，亦可計算所有個案（群體）分數之平均信度。這也是Rasch model跟古典測驗理論一大不同之處。依據古典測驗理論發展之評估工具僅有整體信度，而無個別分數之信度。所以Rasch model 所提供之信度資訊更為完整，用途更多（如判斷個別評估分數之SE/信度，這也是後續判斷個案層級反應性之基礎，詳下反應性之說明）。

效度方面，Rasch model 的主要功能之一為驗證評估工具之項目是否屬於unidimensionality（「單一向度」或「單向度」）。意即所有項目是否皆測量同一概念、dimension（向度）或latent trait（潛在特質）。此功能類似古典測驗理論之驗證性因素分析 (confirmatory factor analysis, CFA)。

反應性方面：Rasch model可估計每一個分數之SE，故可運用SE以驗證個案於前後測之變化量有無超過前後測差異之SE。這相對於傳統上的反應性驗證，帶來不同的概念與結果。傳統上，一般學者檢驗評估工具反應性之主要統計指標為 effect size，皆以團體層級為單位。而前述SE之運用，則可計算每位個案/分數之反應性，意即個別層級之反應性。相關實例詳後說明。

另外，differential item functioning (DIF) 之概念為：某項目對於相同能力但不同族群（如性別、地域等）之難度顯著不同。16具有DIF問題的項目之於不同族群具有系統偏誤，造成這些項目對不同族群的難度不同（如相同能力的男女性，女性於某項目之得分比男性高）。研究者若未驗證及處理評估工具項目之DIF問題，可能造成其評估結果偏誤。故完整的評估工具心理計量特性驗證須包含各項目之DIF檢驗。

評估工具之效度驗證：確認評估工具項目之unidimensionality

unidimensionality屬於建構效度 (construct validity) 之一種，亦可稱為內在效度 (internal validity)。理論上，一般評估工具擬測量之概念應為單一向度（或單向度）或由多個單向度所組成。如生活品質問卷，即由多個單向度所組成。不同項目若具備單向度特質，因為這些項目皆測量同一概念（向度），故可加總這些項目的分數，以代表此向度（概念）之強度或能力高低。相對而言，若一評估工具之所有項目未具備單向度，則這些項目因為測量不同（多種）向度，故不宜將各項目之分數加總計分。否則加總之分數代表多種向度之強度/能力，其意義相當有限。如某一問卷擬測量A概念，但包含測量B概念或其它概念之項目。若將此問卷所有項目加總計分，則此分數難以代表A概念之強度，亦即未具備單向度特質，嚴重影響此問卷之效度。

資料準備/收案：200位個案為低標，300位為佳。且個案之能力需適當分散，非過度集中（過低、偏高或過度集中皆不宜）。另，若項目之評分量尺超過3，則宜確認每項目之各量尺是否皆有一定之個案數選擇之。個案數不足將造成分析結果不穩定/不肯定。若此，研究成果將不易發表於國際期刊。

實務分析時，Rasch model 之應用/分析皆以資料（個案之數據）是否符合Rasch model之假設起始。Rasch model之主要假設為 unidimensionality, local independence, sufficiency 與monotonicity。若資料未符合 Rasch model之假設，則後續之分析（如信度或反應性等）已無意義。上述假設，除了unidimensionality跟本章節內容較相關，於稍後說明，另3項假說於其它章節或基礎的Rasch model教科書皆有論述，故本章節未論述之。

Rasch model 假設之驗證而言，一般驗證與判斷標準為 infit mean square (MNSQ) 與 outfit MNSQ 同時介於 0.6~1.4 之間。不符合標準之項目，則須一一排除，再驗證之，直到所有項目皆符合標準為止。MNSQ可轉換為標準化Z分數 (ZSTD)，以進行統計顯著檢定。但ZSTD 如一般統計檢定，易受樣本數影響。樣本數若很大則容易獲得統計顯著不符合Rasch model 假設之結果，樣本數若小，則反之。故作者不建議後續使用者採用ZSTD作為驗證項目是否符合Rasch model 假設。

infit 與 outfit MNSQ 可考慮採用較嚴謹的標準，如0.7~1.3；或較寬鬆的標準，0.5~1.5，甚至有學者採用0.3~1.7。一般而言，較明確成熟之概念，或已有多種現有良好工具之驗證，宜採用較嚴謹之標準。反之，較新穎/抽象之概念，或現有工具之心理計量驗證證據有限，則可採用較寬鬆之標準。然而若採用較寬鬆的標準也代表該評估工具可能不甚良好，仍需改良。或是擬評估之概念較為抽象，被瞭解有限，不易被評估。

以 “A Rasch Analysis of the Frenchay Activities Index in Patients With Spinal Cord Injury” 為例，7此論文以 Rasch model驗證 Frenchay Activities Index (FAI) 應用於脊髓損傷病人之心理計量特性（單向度是主要特性之一）。FAI 共有15 項4分量尺之題目，欲評估「工具性日常生活功能 (instrumental activities of daily living, IADL)」。但量尺之類型不一，共有6種不同的4分量尺。作者依據文獻認為 IADL 為單向度，故應用 Rasch model分析之。因FAI項目之量尺類型不一，故作者採用 partial credit model（若量尺類型一致，可採用 rating scale model）。作者採用之model適配度判斷標準為：infit 與 outfit MNSQ (<1.3) 以及infit 與 outfit ZSTD (<1.96)。

研究對象為 233 位脊髓損傷病人，他們接受研究人員的面對面訪談15項 FAI。

研究者以partial credit model分析結果顯示：2項目（讀書與外出）不符合 model 假設（或預期），讀書：infit MNSQ/ZSTD 1.48/5.0, outfit MNSQ/ZSTD 1.91/6.3; 外出：infit MNSQ/ZSTD1.3/3.2, outfit MNSQ/ZSTD1.34/2.4，作者先刪除讀書項目（以較高的 infit MNSQ判斷）。刪除後結果顯示：外出項目仍不符合model 假設，故再刪除之。最後發現剩餘13項目皆符合原先設定之判斷標準，亦即此13項目皆符合 Rasch model 假設。因此刪除讀書與外出二項目，FAI 之13項目符合單向度。

再以 “A Rasch Analysis of a Self-perceived Change in Quality of Life Scale in Patients With Mild Stroke” 論文為例說明，9此論文以 Rasch model驗證self-perceived change in quality of life scale (CQOL) 之13個項目是否為單向度？

CQOL項目為4點量尺，擬評量中風病人自覺生活品質之變化（相對於中風之前）。

很少生活品質問卷直接評量「生病前後」之生活品質改變。CQOL之心理計量特性甚少被研究，作者以 Rasch model驗證 CQOL之心理計量特性（單向度是主要特性之一）。因CQOL項目之量尺類型一致，故作者採用 rating scale model。作者驗證單向度之方法有二：一、Rasch model 適配度（判斷標準為：infit 與 outfit MNSQ <1.3）；二、殘差 (residuals) 之主成分分析：殘差應無主要因素（向度）存在，若有主要因素即違反單向度之假設。

研究對象為158位中風病人，個案自填完成問卷。

結果：3項目（說話能力之自覺變化、視力之自覺變化與思考能力之自覺變化）之infit and outfit MNSQs > 1.3，被逐一刪除後，剩餘10項目infit and outfit MNSQs皆符合標準。另，殘差 (residuals) 之主成分分析 (principal component analysis) 未發現主要因素（向度）存在（殘差之第一與第二因素各可解釋25% 與 16% 殘差之變異量）。這些結果顯示 CQOL 10項目符合 Rasch model 之假設且為單向度。

相關限制方面，Rasch model僅能驗證原設計為單一向度之評估工具，或多向度評估工具之個別向度內之項目是否屬於同一向度。亦即Rasch model只能逐一分析各向度之項目是否為單向度，而無法同時分析所有向度，亦無法驗證是否有高階向度/概念之存在。

評估工具之信度驗證：

資料準備/收案方面，200位個案為低標，300位為佳。且個案之能力需適當分散，勿過度集中（過低、偏高或過度集中皆不宜）。因為以Rasch model估計之分數為標準化分數，通常中間分數（靠近平均值）之 SE 較低（信度較高）；二端（較低或較高）分數之SE較高（信度較低）。故若樣本之能力分布未能分散，則信度估計易有偏誤（偏高或偏低）。

進行資料分析時，需先驗證資料是否符合 Rasch model 之假設，若資料不合 Rasch model。則不宜計算分數之SE/信度。Rasch model 可呈現每位個案分數之SE，故可計算每位個案之信度 (1-SE2)，亦可計算所有個案分數之平均SE或信度係數。

因為信度 (SE) 是評估工具的基本心理計量特性，故所有符合 Rasch model 之假設之評估工具，無論項目多寡，皆可計算樣本中每位個案分數之信度或所有個案平均之信度。以上或以下文內提到之Rasch model應用於驗證評估工具或發展評估工具，皆提到信度 (SE) 之驗證與數據呈現與解釋，請讀者自行參考之。

評估工具之個別層級反應性 (individual-level responsiveness) 驗證：

以 “Is the Long Form of the Fugl-Meyer Motor Scale More Responsive than the Short Form in Patients With Stroke?” 一文為例說明，長版Fugl-Meyer motor scale為二向度（上肢動作功能：26項題目與下肢動作功能：11項題目）皆符合 Rasch model 假設。短版Fugl-Meyer motor scale為從上述37題之中，再依據專家意見與難度分布挑選各6項題目，共12題。之前研究發現長版與短版Fugl-Meyer motor scale之團體層級反應性（effect size 數值）不分軒輊。此研究計算significance of change（「前後測差異值」與「差異之標準誤」二者之比值），其中「差異之標準誤」為（前測SE平方與後測SE平方二者之和的平方根），因為每位個案皆可計算significance of change數值，故可計算有多少個案此數值是否超過1.96，若超過即代表前後測之差異具有顯著差異。研究者可計算多少個案前後測之差異具有顯著差異，亦可比較不同評估工具之間，哪一個評估工具可以呈現較多的個案具有顯著差異，即是個別層級之反應性。

研究發現長版Fugl-Meyer motor scale比短版Fugl-Meyer motor scale可呈現較多的個案具有顯著差異，意即長版Fugl-Meyer motor scale之個別層級反應性比短版Fugl-Meyer motor scale為佳。長版Fugl-Meyer motor scale之個別層級反應性較佳的主因在於其SE較短版為低，故個案的前後測的差異較易超過significance of change數值。

此研究發現凸顯個別層級反應性與團體層級反應性之結果差異。我們已發表多篇論文，皆呈現類似結果。如“Comparison of the Responsiveness of the Long-Form and Simplified Stroke Rehabilitation Assessment of Movement: Group- and Individual-Level Analysis”4 以及 “Group- and Individual-Level Responsiveness of the 3-Point Berg Balance Scale and 3-Point Postural Assessment Scale for Stroke Patients”1 二論文皆呈現長版評估工具與短版評估工具於團體層級反應無顯著差異，但於個別層級反應性則長版優於短版。因此後續研究人員於驗證評估工具之反應性時，宜同時驗證個別層級反應性與團體層級反應性，以期全面呈現評估工具之反應性。

評估工具項目之DIF驗證：

以 “Development of a Computerized Adaptive Test for Assessing Activities of Daily Living in Outpatients With Stroke”17一文為例說明，此文作者所發展之日常生活活動CAT之題庫包含可能具有性別DIF之項目，作者使用IRT-based likelihood ratio test 驗證CAT題庫項目有無性別 uniform 或 nonuniform DIF。

結果顯示：34題題庫之中有6項題目具有性別DIF (*p*<0.05)。17一般而言，若題庫題目充裕，則可刪除DIF項目。然而此題庫僅有34題，若刪除這6題，將使題庫與CAT表現變差，故作者保留之。17但估計個案分數時，則給予不同計分；意即男女性於此6項題目之計分不同，以免受到DIF項目影響結果分數。

再以「姿勢控制量表應用於中風病人：羅序分析」2一文說明之，此研究先針對整份量表的DIF狀況進行*X*2檢定，並計算DIF值與SE之比值。若DIF差異值與SE之比值大於3，即代表該項目存在DIF。另研究採Zieky 所提出之DIF效應值分類，18依據DIF之差異值分成A、B、C三等級：A （DIF估計 < 0.25）代表DIF很小，可忽略；B （0.25 ≦ DIF < 0.38） 代表題目有中等至嚴重的DIF；C （DIF ≧ 0.38），代表嚴重的DIF。當題目皆符合上述判斷DIF二條件 （「DIF差異值與SE之比值 > 3」以及「DIF差異值 ≧ 0.38」），則該題目具有DIF。

此研究作者檢驗中風病患之患側（左側與右側個案）與性別之於量表題目（共四個題目）有無DIF影響。2結果顯示四個題目於性別與患側DIF之卡方檢定未達統計顯著（*p* = 0.91與0.77），且DIF差異值與其標準誤的比值皆 < 3 (0.4 ~ 1.6)。在DIF效應值分類上，於性別方面，僅動態站姿平衡的DIF差異值為C級，其餘三個題目為A級。患側方面，靜態坐姿與動態坐姿平衡題目的DIF差異值為B級，而靜態站姿與動態站姿平衡題目的DIF差異值為A級。整體而言，此四個題目並無顯著之性別與患側DIF。2

以 Rasch model驗證評估工具其它心理計量特性：

其它心理計量特性包含「樣本與題目難度比較（或稱為 targeting）」以及「項目難度之延續性（有無 gaps）」等。研究者執行「樣本與題目難度比較」可顯示評估工具之所有項目對於研究對象（或驗證對象）是偏難還是偏簡單，意即targeting是否合適。研究者可藉此驗證評估工具之項目難易度與個案能力之一致程度。Rasch model設定所有項目之平均難度為0，所以若個案之平均能力在0左右，則顯示targeting 良好；若個案之平均能力為正值，則代表樣本之能力偏高（或題目偏簡單）；若個案之平均能力為負值，則代表樣本之能力偏低（或題目偏難）。

由於符合 Rasch model 假設之項目，Rasch分析可估計各項目之難度，故可瞭解評估工具所有項目難度之連續性如何（有無 gaps）。若連續性不佳（有一或多個 gaps），則該gap 所代表能力附近個案之估計分數之標準誤將增加（個別信度降低）。相關應用實例，讀者可參考第四節「發展新評估工具」所提供之論文實例說明。

第二節、「量尺轉換」：由順序量尺轉換為等距量尺

一般問卷或表現評估工具項目之量尺，大多如 Likert-like 量尺為「順序」或「次序」量尺 (ordinal level of measurement)。順序量尺所得分數 (0, 1, 2, 3, 4) 非真正量化/等距量尺 (interval level of measurement)。如其各項目之0～1距離（如很不滿意～不滿意）或3～4距離（如普通～滿意），並不一致。理論上，不同項目之順序量尺分數不能進行加總計分，即使加總計分，其分數或差異所代表之意義，並不符合其數值之意義（如5～8分之距離並不等於8~11分之距離）。因此一般問卷或表現評估工具項目之順序量尺，嚴謹而言，其量化/統計功能有限，甚至常被一般研究人員誤用。

符合Rasch model假設之評估工具分數可轉換成Rasch分數（logit或 log odds），此分數具備等距量尺之特性。19, 20等距量尺之Rasch分數，上述順序量尺之問題（項目分數不宜加總、同差距卻不等距等）皆已迎刃而解。因此，我們建議研究人員與評估工具使用者以轉換後的Rasch分數代表評估工具分數以及後續資料分析與呈現。

Rasch分數另一特點為標準化分數 (standardized score)。因此一般標準化分數之優點（如每一個分數距離平均值之距離等）Rasch分數亦同時具備。

因為符合Rasch model假設之評估工具分數即可轉換成Rasch分數，一般研究者以轉換表呈現單向度分數之原始分數以及Rasch分數與SE。如 “A Rasch Analysis of the Frenchay Activities Index in Patients With Spinal Cord Injury” 7與“Rasch Analysis of Combining Two Indices to Assess Comprehensive ADL Function in Stroke Patients”10 等論文。這SE亦是Rasch分數之特色（優點）之一，意即使用者立即可以瞭解該Rasch分數之誤差（不穩定性）有多少。然而，若是以多向度Rasch model驗證或發展之評估工具，則因個別向度分數皆可能考量其它向度分數（意即個案之分數可能受到個案於其它向度分數之影響），所以難以呈現如單向度簡單/直接的分數轉換表。一般研究者或以軟體程式提供後續使用者藉以轉換所需之Rasch分數，如“A Simplified Stroke Rehabilitation Assessment of Movement Instrument”13 與 “A Developmental Screening Tool for Toddlers With Multiple Domains Based on Rasch Analysis”15 等論文。因需使用特定軟體或轉換程式，這造成使用上一些不方便。

然而，非所有研究者皆會強調Rasch分數之等距特性，因此亦不會提供轉換表或轉換程式。若是如此，研究者可參考該研究所提供之參數（如各項目以及各量尺 threshold 之難度參數），以自行進行轉換。研究者亦可向原作者索取相關參數或是轉換表（程式），以利取得 Rasch分數。因為研究者有義務提供其發表論文之相關資訊，因此若可聯絡到原作者（如 e-mail），應即可取得相關資訊。

第三節、「簡化評估工具」

簡化評估工具意即減少施測時間，提升評估效能，無論使用者或研究人員皆相當期待評估工具被簡化。因此簡化評估工具之實務價值高。然而簡化評估工具的最高境界為：簡化後之評估工具仍須保有良好的心理計量特性（如信度與反應性），或心理計量特性與原評估工具相差不多。

運用 Rasch model 簡化評估工具之標的有二：簡化項目數或簡化量尺。簡化項目數可應用單向度或多向度Rasch model，一般而言，多向度簡化項目之機率較高與可減少之項目較高，意即運用多向度Rasch model可簡化多數評估工具且較可能保有良好的心理計量特性，或可達成簡化評估工具的最高境界。以下分別說明如何以單向度或多向度Rasch model簡化評估工具。

**單向度Rasch model簡化評估工具**

運用 Rasch model 達到上述目標的主要原理是：Rasch 分析可排列所有項目之難度，相同難度之項目即為多餘（無法提供更多的測量訊息或無法降低SE）可刪除之項目。或是項目之量尺多（如大於4），部分量尺可能多餘/累贅。適用對象為單向度評估工具，且項目數較多或評分量尺多之評估工具。

一般應用Rasch model簡化評估工具之過程如下：1. Data-model fitting 先刪除misfit項目；2. 驗證量尺是否合適，有無多餘之量尺點數；3. 挑選項目等。

以 “FIM-Minimum Data Set Motor Item Bank: Short Forms Development and Precision Comparison in Veterans” 一文為例說明之。21此研究探究復健領域二種常用評估工具FIM 與Minimum Data Set 有無簡化之可能性。FIM為13題，各題為7點量尺。Minimum Data Set亦具有13題，各題為6點量尺。研究者先確定項目之 data-model fitting，再排除 local independence 之項目，再選擇項目難度含括完整難度之項目以組成短版測驗。作者從二工具各別挑選出4題與8題短版版本（4題與8題FIM版本；4題與8題Minimum Data Set版本）。最後作者比較不同短版版本、原版以及所有符合data-model fitting項目之信度 (SE)，結果顯示：二工具8題短版之信度 (SE) 良好，且與原版以及所有項目之信度差不多，因此作者最後建議FIM 與Minimum Data Set 二種8題短版給後續使用者，以提升施測效能。

**多向度Rasch model簡化評估工具**

多向度Rasch model利用其它向度之測量資訊以推估各向度之分數，如三向度（A, B, C向度）評估工具，多向度Rasch model 於估計A向度分數時，除了利用個案於A向度之作答反應，亦從個案於B, C向度之作答資訊估計A向度之可能分數。因此如果A, B, C三向度之相關性高，則彼此推估之效能高。意即多向度Rasch model可用以提升多向度評估工具之信度，這優點特別適用於項目較少的多向度評估工具。提升多向度評估工具之信度是多向度Rasch model之最原始目的。對於項目較多的評估工具，則可運用此優點以簡化評估工具（減少項目數），但各向度仍保持良好的信度。

特別一提：王文中教授對於多向度Rasch model發表多篇相關理論與實務論文，有興趣之讀者可深入瞭解其相關著作。8, 13, 22-25

以 “A Simplified Stroke Rehabilitation Assessment of Movement Instrument” 一文為例說明之。13 Stroke rehabilitation assessment of movement instrument (STREAM) 評估三向度：上肢動作、下肢動作與行動能力等。三向度各10個項目，其中上肢動作與下肢動作為3點量尺，行動能力為4點量尺。

作者先驗證各向度項目之Rasch model適配度，接著驗證量尺設計是否合適，之後再挑選項目以簡化評估工具。

首先作者發現上肢動作與下肢動作各有2項目與1項目不符合Rasch model假設，個別刪除後，最後上肢動作8項目、下肢動作9項目與行動能力

10項目之Rasch model適配度良好。

 作者再使用 rating scale model 個別驗證三向度之量尺是否合適。他們發現三向度之量尺之階難度 (threshold) 皆可合適區隔個案之能力（數據詳原文941頁之表4），因此作者未調整各向度之量尺。

 最後作者挑選各向度之項目，挑選順序/原則主要有二：一、先從各向度中挑選最難與最簡單之項目，意即挑選 difficulty measure or logit 最大與最小的項目。二、挑選中間難度且難度平均分散之項目。若有類似難度之項目，則由作者與專家挑選之。最後各向度皆挑選5項目組成簡化版STREAM。

 以多向度Rasch model估計之簡化版STREAM三向度信度皆高於0.90 (0.91~0.95)，相當良好。只比多向度Rasch model估計之原版27項目STREAM（刪除3項目低適配度後）之信度 (0.93~0.98) 稍低。而且簡化版STREAM三向度分數與27項目STREAM三向度分數之相關性相當高（三向度Pearson *r* 皆等於0.99）。

 上述結果顯示：利用多向度Rasch model可減少 STREAM 50% 項目，但仍保有良好的信度與效度。因此此文支持多向度Rasch model簡化項目之成效良好。

 再以 “Refining 3 Measures to Construct an Efficient Functional Assessment of Stroke”11 一文為例說明，研究者將臨床上三種經常用一起使用的評估工具，應用單向度與多向度Rasch model組合成一個多向度評估工具。

 研究者先使用單向度 Rasch model 驗證三評估工具之單向度以及是否符合 Rasch model 假設，刪除適配度不佳之項目後，再以多向度Rasch model簡化項目數。簡化（挑選項目）之方法則以內容（專家挑選）與難度分布（平均分散）為原則。

結果原來三項評估工具共72項題目，刪除14項Rasch model 假設之項目後，再以專家挑選以及難度分布挑選符合假設的58題項目，最後只剩餘29項題目（只剩四成項目），而且心理計量特性（信度、效度與反應性）皆良好。

因此原本各自獨立發展/驗證之評估工具，可經由單向度與多向度Rasch model組合成一個嶄新且簡化的多向度評估工具。研究者使用多向度Rasch model故可大幅簡化項目（刪除約六成項目），然而這也是此新評估工具之缺點，意即三種評估工具必須一起使用，才能獲得上述好處以及 Rasch scores。

第四節、「發展新評估工具」

研究者若擬利用Rasch model發展新評估工具，需先確認該工具擬評量之概念屬於單向度，或是多向度之各個向度皆屬於單向度。且如前所提，一般項目不易符合Rasch model假設，因此研究者必須於一開始時多準備項目，且各題目需有難度差異，以利評估不同功能/能力之個案。一般而言，就一單向度新工具，且為5點量尺之項目，研究者可先提出約20題。以利後續可能排除高達50%之項目後，還有10題。

研究者亦可利用Rasch model 將幾種常用之評估工具發展成為一個新的評估工具。以 “Rasch Analysis of Combining Two Indices to Assess Comprehensive ADL Function in Stroke Patients”10一文為例，研究者將二個評估工具［10項目之Barthel index（巴氏量表，國內最常用的失能量表，10項目具備三種量尺：2～4點）與15項目之 Frenchay activities index（皆為4點量尺，但有多種選項設計）］彙整成一個新的評估工具。

研究對象為245位發病一年之中風病人，接受研究助理之面對面或電話訪談。資料分析時，研究者先以 partial credit model分析二工具量尺之合適性。若量尺不佳，則重組量尺後再分析之。之後研究者再驗證項目是否符合單向度（驗證項目之Rasch model適配度以及分析residuals有無存在明顯向度）。最後研究者再驗證剩餘題目是否充分（可評量不同功能程度之病人，意即項目難度間有無 gaps 存在）。

結果顯示：原本二工具中2項目之Rasch model適配度不佳，先被刪除。然剩餘25項目中，高達23項之量尺合適性差。這些項目之中間量尺僅有少數比例之病人適用，因此研究者將量尺合併成為2點量尺（研究者依據工具量尺選項之特質，將二評估工具之量尺以不同方式合併，詳原文）。10

這25項2點量尺項目再進行Rasch model適配度分析，又發現2項目之適配度差，研究者再刪除之。最後23項目之Rasch model適配度佳，而且因素分析residuals證實無明顯向度存在，因此這些結果支持這23題2點量尺項目符合單向度。且這23項目符合Rasch model之假設，故可轉換成 Rasch 等距分數。作者提供原始分數與Rasch分數轉換表，以供後續使用者轉換分數。

另外所有病人Rasch分數之平均信度高達0.94。然而有23項目之難度排序上有4個 gaps 存在，意即部分病人所得分數之信度（標準誤）可能較差。

整體而言，新的工具符合單向度且信度良好。項目可簡化（刪除4項目），且量尺較簡單，應可提升施測之方便性。

**發展電腦適性測驗 (computerized adaptive testing, CAT)**

 發展CAT之前提有二：一、項目符合Rasch model 或其它item response theory假設；2、項目夠多且難度分散佳（以適用不同功能高低之個案）。於此僅以Rasch model 為例說明。

 CAT之主要特性是從符合Rasch model 假設之項目（題庫）中挑選接近個案功能之項目施測（適性施測），因此只要題庫良好（符合上述二前提），則CAT可提供快速且信度良好（精準）之評估結果 (Rasch scores)。

 以 “Development of a Computerized Adaptive Test of Children's Gross Motor Skills”14一文為例說明。此研究之資料來自昔日研究，共有1738兒童接受醫療人員以嬰幼兒綜合發展測驗第II版 (Comprehensive Developmental Inventory for Infants and Toddlers II, CDIIT II) 施測。評估所得數據作為發展此CAT之用，並且藉以驗證此CAT篩檢能力與其它心理計量特性。

研究者先以 Rasch model 驗證後，仍有44項題目符合Rasch model 假設。研究者即以這44項組成題庫，以建置CAT。研究者經模擬CAT表現分析之後，決定 CAT 之終止施測條件：信度 ≧ 0.94 或是 limited reliability increase < 0.001（意即再施測所能增加的信度相當有限，故可終止施測）。此CAT所估計兒童分數之平均信度係數 ≧ 0.94，且與原始CDIIT II 評分之相關性極高 (Pearson *r* ≧ 0.97) 。這些結果顯示：此CAT之信度與同時效度極佳。另CAT所估計兒童分數之篩檢能力良好 (area under receiver operating characteristics curve ≧ 0.80)。

 另有500位兒童於6～14月之後，再接受CDIIT II 追蹤評估。追蹤資料之用途有二：一、驗證Rasch model 分析結果（各項目參數）之invariance；二、驗證此CAT之反應性。結果顯示：各項目參數於二時間點之invariance良好，因此研究者僅使用CDIIT II初測所得之項目參數作為後續建立與驗證CAT之數據。此CAT於500位兒童之6～14月後追蹤之 effect size ≧ 0.65，顯示反應性良好。

此CAT僅需施測約7～11項目，即可獲致良好的心理計量特性（含信度、效度與反應性）。因此，此新發展的CAT可提供臨床與研究人員快速且精準之評估結果 (Rasch scores)。

再以 “Development of a Computerized Adaptive Testing System of the Functional Assessment of Stroke”26 一文說明。此文源自前述 “Refining 3 Measures to Construct an Efficient Functional Assessment of Stroke”11 一文，研究者將臨床上三種經常用一起使用的評估工具，應用單向度與多向度Rasch model組合成一個多向度評估工具。該文中證實符合Rasch model假設之項目共有58題，研究者即以這些題目及其參數當成CAT之題庫。

研究者先驗證CAT之終止施測條件：研究者提出10種可能終止施測條件（達到最低信度標準或limited reliability increase之組合），經模擬CAT表現分析比較之後，決定 CAT 之終止施測條件：limited reliability increase < 0.010。

之後研究者再驗證以上述終止施測條件之CAT的心理計量特性（含信度、同時效度與反應性），結果發現此CAT平均只需8.5項目即可完成4向度之評估。而且此CAT所估計各向度分數之信度良好（信度係數為 0.88～0.93）、同時效度良好（與原始評估工具分數之Pearson *r* = 0.91~0.95），且反應性良好 (effect size = 0.65~0.76)。

以上結果顯示CAT可大幅提升施測效能（平均僅需施測8.5題，相對於原始評估工具之題目共72題或是以多向度 Rasch model 簡化後之項目數仍高達29題），且同時可保有良好的心理計量特性（含信度，效度與反應性等）。

第五節、「應用限制與後續發展」

Rasch model於醫療領域之應用限制或難題主要有二：一、一般評估工具之項目不易符合Rasch model 假設要求；二、驗證所需之樣本數大且特質分布需分散。

符合Rasch model 假設要求之評估工具，其心理計量特性良好，實用價值高。也因此Rasch model對於項目之要求標準高（含unidimensionality, local independence, sufficiency 與monotonicity等），因此一般傳統評估工具不易符合Rasch model 之假設要求，新發展之評估工具亦不易達到Rasch model 之要求。

以Rasch model驗證評估工具需要大樣本數，其原因為Rasch model建構於機率模型；且樣本之能力或功能分布需分散，以免造成偏誤（亦即需要不同能力之樣本以準確估計其相對應能力項目之參數，若樣本欠缺某些能力之個案，則估計相對難度項目參數之精準度下降）。如此樣本特質對於臨床研究人員，取得難度頗高。

另，如前所提Rasch model僅能驗證原設計為單一向度之評估工具。若為多向度工具，Rasch model只能逐一分析各向度之項目是否為單向度，而無法同時分析所有向度。Rasch model亦無法驗證是否有高階向度/概念之存在。

研究人員或專業人員需先瞭解上述限制或難題，以免應用後因不符預期而失望。另外，亦因上述限制或難題不易突破，因此少數研究者使用較低標準以使評估工具符合Rasch model假設要求，或使用於較低之樣本數或個案分數分布未能分散，上述這些狀況，將使研究結果不可靠，讀者宜留意。

最後，個人再度感謝王文中教授之指導，否則上述引用介紹 Rasch model之應用論文，就不會面世。個人及研究小組所發展的CAT亦不可能誕生。王教授之貢獻與學術風範永誌銘心！

參考文獻：

1. Huang YJ, Lin GH, Lee SC, Chen YM, Huang SL, Hsieh CL. Group- and individual-level responsiveness of the 3-point berg balance scale and 3-point postural assessment scale for stroke patients. *Arch Phys Med Rehabil*. 2018;99:529-533.

2. 吳佳樺, 黃千瑀, 李士捷, 陳佳琳, 尤菀薈, 謝清麟. 姿勢控制量表應用於中風病人：羅序分析. *職能治療學會雜誌*. 2016;34:250-264.

3. Huang CY, Song CY, Chen KL, Chen YM, Lu WS, Hsueh IP, Hsieh CL. Validation and establishment of an interval-level measure of the balance assessment in sitting and standing positions in patients with stroke. *Arch Phys Med Rehabil*. 2016;97:938-946.

4. Huang YJ, Chen KL, Chou YT, Hsueh IP, Hou CY, Hsieh CL. Comparison of the responsiveness of the long-form and simplified stroke rehabilitation assessment of movement: Group- and individual-level analysis. *Phys Ther*. 2015;95:1172-1183.

5. Hsiao YY, Shih CL, Yu WH, Hsieh CH, Hsieh CL. Examining unidimensionality and improving reliability for the eight subscales of the sf-36 in opioid-dependent patients using rasch analysis. *Qual Life Res*. 2015;24:279-285.

6. Chen KL, Chen CT, Chou YT, Shih CL, Koh CL, Hsieh CL. Is the long form of the fugl-meyer motor scale more responsive than the short form in patients with stroke? *Arch Phys Med Rehabil*. 2014;95:941-949.

7. Hsieh CL, Jang Y, Yu TY, Wang WC, Sheu CF, Wang YH. A rasch analysis of the frenchay activities index in patients with spinal cord injury. *Spine (Phila Pa 1976)*. 2007;32:437-442.

8. Wang WC, Yao G, Tsai YJ, Wang JD, Hsieh CL. Validating, improving reliability, and estimating correlation of the four subscales in the whoqol-bref using multidimensional rasch analysis. *Qual Life Res*. 2006;15:607-620.

9. Lin JH, Wang WC, Sheu CF, Lo SK, Hsueh IP, Hsieh CL. A rasch analysis of a self-perceived change in quality of life scale in patients with mild stroke. *Qual Life Res*. 2005;14:2259-2263.

10. Hsueh IP, Wang WC, Sheu CF, Hsieh CL. Rasch analysis of combining two indices to assess comprehensive adl function in stroke patients. *Stroke*. 2004;35:721-726.

11. Wang YL, Lin GH, Huang YJ, Chen MH, Hsieh CL. Refining 3 measures to construct an efficient functional assessment of stroke. *Stroke*. 2017;48:1630-1635.

12. Hsieh YW, Hsueh IP, Chou YT, Sheu CF, Hsieh CL, Kwakkel G. Development and validation of a short form of the fugl-meyer motor scale in patients with stroke. *Stroke*. 2007;38:3052-3054.

13. Hsueh IP, Wang WC, Wang CH, Sheu CF, Lo SK, Lin JH, Hsieh CL. A simplified stroke rehabilitation assessment of movement instrument. *Phys Ther*. 2006;86:936-943.

14. Huang CY, Tung LC, Chou YT, Wu HM, Chen KL, Hsieh CL. Development of a computerized adaptive test of children's gross motor skills. *Arch Phys Med Rehabil*. 2018;99:512-520.

15. Hwang AW, Chou YT, Hsieh CL, Hsieh WS, Liao HF, Wong AM. A developmental screening tool for toddlers with multiple domains based on rasch analysis. *J Formos Med Assoc*. 2015;114:23-34.

16. Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Med Care*. 2006;44:S39-49.

17. Hsueh IP, Chen JH, Wang CH, Hou WH, Hsieh CL. Development of a computerized adaptive test for assessing activities of daily living in outpatients with stroke. *Phys Ther*. 2013;93:681-693.

18. Zieky M. A dif primer. *Princeton, NJ: Educational Testing Service*. 2003.

19. Andrich D. Rating scales and rasch measurement. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11:571-585.

20. Tesio L, Simone A, Bernardinello M. Rehabilitation and outcome measurement: Where is rasch analysis-going? *Eura Medicophys*. 2007;43:417-426.

21. Li CY, Romero S, Simpson AN, Bonilha HS, Simpson KN, Hong I, Velozo CA. Fim-minimum data set motor item bank: Short forms development and precision comparison in veterans. *Arch Phys Med Rehabil*. 2018;99:534-541.e532.

22. Wang WC, Chen PH, Cheng YY. Improving measurement precision of test batteries using multidimensional item response models. *Psychol Methods*. 2004;9:116-136.

23. Wang WC. Direct estimation of correlation as a measure of association strength using multidimensional item response models. *Edu Psychol Measure*. 2004;64:937-955.

24. Adams RJ, Wilson M, Wang WC. The multidimensional random coefficients multinomial logit model. *Appl Psychol Measure*. 1997;21:1-23.

25. Wang WC, Chen PH. Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Appl Psychol Measure*. 2004;28:295-316.

26. Lin GH, Huang YJ, Lee SC, Huang SL, Hsieh CL. Development of a computerized adaptive testing system of the functional assessment of stroke. *Arch Phys Med Rehabil*. 2018;99:676-683.