特殊教育研究學刊
民 96，32 卷 2 期，1-15 頁

嬰幼兒綜合發展測驗之判定準確度及切截點分析

王天苗
臺灣師範大學特教系教授

廖華芳
臺灣大學物理治療系副教授

本研究主要目的在探究運用「嬰幼兒綜合發展測驗」（CDIIT）所得結果來判定發展遲緩的準確度及最佳切截點。依醫學診斷或鑑定老師研判為「發展遲緩」和「正常發展」之結果為效標，以 760 名 3 至 71 個月之幼兒（發展遲緩者 152 人，正常發展者 608 人）在 CDIIT 診斷和篩選測驗總分所得之發展商數和 Z 分數，進行「接受者操作特徵」（Receiver-Operating-Characteristic, ROC）分析，以探究該測驗之診斷與篩選準確率，並從敏銳度、特異度、陽性概率比、陰性概率比、擊中率、約登指數等指標，綜合研判測驗最佳切截點。結果顯示，CDIIT 篩檢測驗的 ROC 曲線下面積為.92、診斷測驗為.96，篩檢或診斷之準確率極高。根據上述指標綜合研判的結果，CDIIT 篩選測驗總分所得 Z 分數用以判斷疑似遲緩的最佳切截點為-1，而診斷測驗總分所得 DQ 用以判斷發展遲緩的最佳切截點為 77.5（即-1.5SD）。

關鍵詞：ROC 分析、切截點、嬰幼兒綜合發展測驗

本研究承國科會專題計畫補助（計畫編號 NSC91-2413-H-003-039 和 NSC92-2413-H-003-67），特此致謝。

緒論

針對特殊教育學生，專業人員最常運用評量的結果來做「資格認定」（entitlement decisions）（Salvia & Ysseldyke, 1998）。所謂「資格認定」，主要是先決定哪些個案需要進一步評估（即篩檢），接著再依照身心障礙或資優的特質或鑑定標準，來判定個案是否屬於特殊教育服務的對象（即鑑定或診斷）。無論是篩檢或鑑定的目的，評量者常使用標準化測驗來瞭解個案在團體比較下的表現或能力（Simeonsson, 1986）。

如果以標準化測驗結果做為教育上的「資格認定」或任何「特殊教育之決定」（decision making in special education）的參考，往往涉及測驗檢測「標準」（standards）、「切截點」（cut off points）或「切截分數」（cutoff scores）的問題（路君約，1989；Crocker & Algina, 1986; Harber, 2001）。測驗編製者如果能提供篩檢或診斷測驗的切截點資料，可以讓評量者從測驗結果更敏銳地找出需要服務的對象；或相對地，排除不需要者。

依 Crocker & Algina（1986）的說法，決定測量工具切截點或切截分數的方法是「標準設定」（standard setting）。標準設定的方法多，包括專家判斷或以受試者實際表現進行設定的平均數法、迴歸分析法、基準法、平均數以下一個標準差及總分之百分比等方法，此外，更有 Nedelsky（1954）的「臨界組法」（borderline-group method）、「對照組法」（contrasting-group method）、Berk（1976）的效標組法及吳裕益（1988）的對照組 M-SD 法。然而，這些方法被批評會因為選擇的樣本不同而有不同的設定結果（林惠芬，1993；鄭明長、余民寧，1994；Crocker & Algina, 1986）。

近年，有兩種新方法逐漸被運用在教育或心理測驗的標準設定上，一為「試題反應理論」（Item response theory, IRT）分析，另一為「接受者操作特徵」（Receiver-Operating- Characteristic, ROC）分析。其中，IRT 是從兩組受試者對試題反應之能力估計值來設定切截點（余民寧，1993）；ROC 分析則是考量測驗之預測效度來設定切截點。國外學者 Harber（2001）就建議，用 ROC 分析來探究不同切截點的診斷效果，並依分析結果做最佳的特殊教育決定。

至今，國內外在身心障礙領域中運用 ROC 分析的相關研究均少。其中，Watkins 等人（2002）以 ROC 分析探討 WISC-III 分測驗得分所得之「學習障礙指標」（Learning Disability Index）的診斷準確率，結果發現準確率僅 55%~64%，因而認為 LDI 並不是診斷學習障礙的有效指標。

檢視國內有關障礙兒童評量工具的文獻，發現有的評量工具之切截點是以兩組得分的平均值設定（林素貞，2000），有的是依智能障礙判定標準（-2、-3 和-4 標準差）來決定（林幸台、吳武典、吳鐵雄、楊坤堂，1992），有的以高低能力受試者表現來決定（黃惠玲，2000），有的則以 IRT 測驗訊息量分析結果決定（洪儷瑜、余曉珍，1998）。至於「自閉症兒童行為檢核表」，則是將切截點訂在能包含所有自閉症兒童的分數，以避免篩選階段就遺漏可能的個案（張正芬、王華沛，2005）。總之，國內用於篩檢障礙或遲緩的評量工具，除了「青少年社會行為簡式量表」之外，其他評量工具都未在編製報告中說明通過標準或切截點的設定方法與過程，似乎只由編製者自行決定而已。

研究者九年前編製的「嬰幼兒綜合發展測驗」（Comprehensive Developmental Inventory for Infants and Toddlers, 簡稱 CDIIT）（王天苗、蘇建文、廖華芳、林麗英、鄒國蘇、林世華，1998），至今是國內用於鑑定發展遲緩幼

兒最重要的一個標準化發展測驗。在常模建立及後續研究中，研究者已陸續探究該測驗的重測信度、內部一致性、觀察者信度、內容效度、建構效度、同時效度和預測效度等信效度資料（Liao & Pan, 2005; Liao, Wang, Yao, & Lee, 2005；王天苗，2005；王天苗等，1998；吳雪玉、廖華芳、姚開屏、李旺祚、王天苗、謝正宜，2005；廖華芳、王天苗、姚開屏，2002；廖華芳、王天苗、姚開屏、吳雪玉、李旺祚、鄭素芳，2002）。

身為編製者，至今仍有一項未竟之事，就是當初根據國外文獻在指導手冊初步建議之篩檢和診斷發展遲緩的切截點，依然有待驗證。基於發展測驗最需具備預測力（Meisels, 1985），而 ROC 分析又是考量測驗預測力最重要的標準設定方法，因此本研究的主要目的是採 ROC 分析及相關指標，探討 CDIIT 用來判定發展遲緩的準確度及設定最佳切截點。由本研究所得的結果，期待能使評量者在運用 CDIIT 來判定幼兒的發展問題時，能對此工具的準確性有更大的信心，而且有客觀的標準作為判定的參考。

文獻探討

一、ROC 分析與切截點設定

自 1950 年代以來，國外研究者會運用 ROC 分析來檢測篩檢或診斷工具的準確度及切截點設定，尤其在醫學界的運用最為普遍。近年來，有國外特殊教育研究者也建議運用 ROC 分析來探究身心障礙者的篩檢和診斷有效性的問題（Harber, 2001; Watkins, Kush, & Schaefer, 2002），不過，至今相關文獻仍有限。

由 ROC 分析結果，可得「接受者操作特徵曲線」（Receiver Operating Characteristic Curve，簡稱 ROC 曲線）。ROC 曲線是以「錯誤接受率」（false positive fraction，FPF，即 1-特異度，又稱偽陽率）為橫座標，「正確接受率」（true positive fraction，TPF，即敏銳度，又稱真陽率）為縱座標，由無數切截點值求出的無數對 FPF 和 TPF 作圖構成。換言之，由此可以分別計算出不同切截點（以 t 表示）的正確接受率和錯誤接受率交互而組成的曲線。其中，「敏銳度」（sensitivity，簡稱 SEN）係指正確檢測出問題個案之人數占所有問題個案數的比率；「錯誤接受率」是指錯誤接受為問題個案數占所有非問題個案數的比率；「特異度」（specificity，簡稱 SPE）則是指正確拒絕為非問題個案的人數占所有非問題個案數的比率（Harber, 2001）。

由 ROC 曲線下面積（area under the ROC curve, 簡稱 AUC）的大小，可以區別測驗的有用性（usefulness）或區辨值（discriminative value），面積愈大，測驗就愈有區辨效果，代表測量準確率愈高。當敏銳度和特異度均為 1 時，標記點在曲線的左上角，AUC 等於 1，ROC 曲線最完美（Hanley & McNeil, 1982; Rao, 2003）。一般來說，國外學者建議，AUC 大於.8 時，測驗才具篩檢或診斷的效果（Chong & Karlberg, 2004）。

此外，更可由不同切截點所得 TPF 和 FPF 而形成的 ROC 曲線，來判斷出最佳切截點，即指選擇 TPF（$y(t)$）較大但 FPF（$x(t)$）較小的某點（即在 ROC 曲線上離左上角直線距離最短之點）。由於測驗分數愈低代表愈有問題，因此設定的切截分數愈高，就會產生高敏銳度，但特異度就會低（錯誤接受率高）；反之，設定的切截分數愈低，會造成特異度高，但敏銳度低的結果（Harber, 2001）。

由於不同的切截點通常會使敏銳度愈大，則相對使特異度愈小，因此在綜合考量設定切截點時，還會檢視「概率」（likelihood）。概率可分「陽性概率比」（positive likelihood ratio，簡稱 LR+）、「陰性概率比」（negative

likelihood ratio，簡稱 LR-）及「兩點間差距概率」（LR（*x,y*））等三種（Choi, 1998）。「陽性概率比」，意指測驗結果能正確歸類遲緩幼兒的比率為不該歸為遲緩者（錯誤歸類一般幼兒為遲緩）比率的倍數，以敏銳度÷（1-特異度）（即 TPF/FPF）計算之。LR+愈高，測驗愈能有效地篩檢或診斷出問題。「陰性概率比」則相反，意指測驗結果誤將應為遲緩的幼兒歸為正常之比率為正確歸類一般幼兒比率的倍數，以（1-敏銳度）÷特異度（即錯誤拒絕/正確拒絕）計算之。LR-愈小，測驗愈具正確篩檢或診斷問題不存在的功效。至於「兩點間差距概率」，則是指測驗結果在兩切截點正確歸類遲緩幼兒之差距比率為不該歸為遲緩者之差距比率的倍數，通常用於測驗結果有多重分類之情形時用。

一般來說，LR+值大於 10 表示改變測量前後機率之程度屬「高度」，可幫助確定性診斷；5-10 為「中度」，2-5 為「輕度」，LR+值 1-2 之測量則對判斷陽性個案的幫助很少；相反地，LR-值小於 0.1 屬「高度」，有助確定性診斷，0.1-0.2 為「中度」，0.2-0.5 為「輕度」，LR-值 0.5-1 之測量則對判斷陰性個案的幫助很少（Hanley & McNeil, 1982；Jaeschke, Guyatt, & Sackett, 1994）。

除以敏銳度、特異度和三種概率比來判斷最佳切截點之外，國外學者還建議檢視不同切截點的「擊中率」（hit rate）、「約登指數」（Youden Index）及 Kappa 值。其中，「擊中率」即一致性（agreement），係指正確接受和正確拒絕個案數佔總篩檢或診斷人數的比率，若所得比率大於.8，表示測驗之預測力愈好。「約登指數」是指敏感度與特異度之總和減 1，數值所得最高者較佳，小於 0 代表是無用之測量（Chong & Karlberg, 2004; Straus, Richardson, Glasziou, & Haynes, 2005）。至於 Kappa 值，是用來檢測不同切截點所得測驗結果之分類與效標分類之間的關連性，若 Kappa 值介於.61~.80 間，兩者有「顯著」相關；若在.81 以上，則兩者有「接近完美」之相關（Domholdt, 2000）。

整體來說，如果由分析受試者實際作答的資料來設定標準或切截點時，Crocker & Algina（1986）認為應該使「錯誤分類」（misclassification）的比率或是錯誤分類可「預期的代價」（expected cost）減到最小。要使「錯誤分類」率小，設定的切截點應該使「錯誤接受」和「錯誤拒絕」的比率減為最小；要使錯誤分類可「預期的代價」減到最小，就是使效標分數大於切截點而能通過測驗（即被判定為精熟者），而效標分數小於此點者（即非精熟者）則無法通過該測驗（Crocker & Algina, 1986）。路君約（1989）提醒，在訂定切截點時，要特別注意「錯誤拒絕」的比率。尤其，發展篩選測驗更要有高「敏銳度」，使被「錯誤拒絕」的人數比例減到最少（Meisels & Provence, 1989; Scott & Hogan, 1982）。

二、判定發展遲緩之切截點

在美國，要研判幼兒有發展遲緩，常以發展測驗結果低於-1、-1.5 或-2 標準差為切截標準（Shonkoff & Meisels, 1991）。依據美國特殊兒童學會之「學前分會」（the Division for Early Childhood）的文章中提及，美國各州決斷發展遲緩的標準不一：全美有 34 州運用發展測驗分數決定發展遲緩，其中有 22 州以一項發展領域得分低於-2SD 或有兩項發展領域得分低於-1.5SD 為決定標準。在採用「遲緩比率」的州內，多數採取一項或兩項發展領域低於 25~30%的遲緩率為鑑定標準。只有九州以專業判斷、團隊共識或臨床的意見取代測驗結果來決定（DEC, 2001）。

對發展測驗的編製者來說，多數會在測驗指導手冊內提出判定發展遲緩之標準的建議，但是設定的標準不一；有些測驗則並不提

供任何判定標準的建議，全憑評量者自行做專業判斷。例如，在美國普遍使用的「貝萊嬰幼兒發展量表」（Bayley Scales of Infant Development）（Bayley, 1969），指導手冊內並未訂定判斷遲緩的標準。國內陳淑美、盧欽銘、蘇建文和鍾志從於 1991 年修訂該量表時，也依原版未設定標準。

此外，在美頗受歡迎的「伯特爾發展量表」（Battelle Developmental Inventory），包括診斷和篩選測驗兩題本，在診斷測驗記錄紙內之側面圖畫出陰影（即表示-1.5SD~1.5SD 之間的正常發展範圍），並建議診斷測驗結果落於-1.5SD 以下者為「確定劣勢」（definite weakness），大於 1.5SD 者則為「確定優勢」（definite strength）（Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984）。該測驗之篩選題本用以判定遲緩的標準則設定為：篩選測驗結果介於-1SD 至-1.5SD 之間，為「臨界」者；在-1.5SD 以下者，則確認有「缺損」（deficit）。至於用途為發展篩選測驗的「學前兒童學習能力測驗」（Developmental Indicators for the Assessment of Learning-Revised, 簡稱 DIAL-R），編製者設定-1.5SD 為檢測發展遲緩或疑似遲緩的切截點（Mardell-Czudnowski & Goldenberg, 1984）。

在台灣，除 CDIIT 外，現有的發展測驗都屬篩檢用途，包括早期編製的「嬰幼兒發展測驗」（徐澄清、廖佳鶯、余秀麗，1983）和「學齡前兒童行為發展量表」（簡稱 CCDI，徐澄清、蘇喜、蕭淑貞、林家青、宋維村、張珏，1978）及近年編製的「學前兒童發展檢核表」（台北市發展遲緩兒童早期療育通報轉介中心，1998）、「簡易兒童發展量表測定」（Wong, Chong, Wang, & Lin, 1997）和「零至六歲兒童發展篩檢量表」（黃惠玲，2000）。每一測驗的編製者都提出判定疑似發展遲緩或發展遲緩的標準，不過設定的標準不一，而且只在指導手冊內說明，並無研究報告說明設定的方法。

例如，「嬰幼兒發展測驗」篩檢發展遲緩的標準是：幼兒在比個案實齡簡單且同齡幼兒有 90%通過的項目不通過時，只有一項不通過是「有問題」，有兩項不通過則為「不正常」（徐澄清、廖佳鶯、余秀麗，1983）。CCDI 編製報告僅說明依原版的標準，即任何發展項目的得分如在少於該兒童實齡 20%至 30%此二條年齡線之間者，則判定該項目為「疑有發展遲緩」，若結果落在 30%年齡線以下者，則被判為「發展遲緩」。至於「零歲至六歲兒童發展篩檢量表」，編製者設定篩檢發展遲緩的標準是：先找出各發展項目的「基底水準」（即第一個出現「不會」或「不知道」之題目的前一題所標示的年齡）與「頂峰水準」（即最後一個出現「會」的題目所標示的年齡），任一大項的基底線低於實齡線，則視為「需追蹤對象」，任一項的基底線低於實齡線且頂峰線也低於實齡線，則視為疑有發展遲緩的對象（黃惠玲，2000）。

再就 CDIIT 而言，國內吳雪玉等人（2005）曾以 ROC 分析探究 CDIIT 動作分測驗與「皮巴迪動作發展量表第二版」對動作發展障礙兒童的診斷準確率。結果發現，CDIIT 動作分測驗的診斷準確度有.97，而其最佳分界點為動作發展商數 70，敏銳度為 87%，特異度 97%。雖然該研究對運用 CDIIT 判定發展遲緩的切截點有初步的探究，但是僅針對診斷測驗之動作分測驗進行檢驗而已。

為了設定 CDIIT 診斷測驗整體「一般水準」的區塊（即該測驗記錄紙內測驗結果側面圖上建議的陰影範圍。距平均數以下的陰影範圍之外者，建議判定有發展遲緩），故仍有必要探究出該測驗診斷測驗「總分」用來研判發展遲緩的最佳切截點。至於運用該篩選測驗用來決定疑似發展遲緩的最佳切截點，也需進一步分析。

研究方法

一、研究樣本

本研究樣本有 608 名一般幼兒和 152 名發展遲緩幼兒，共 760 名。發展遲緩幼兒的來源包括三部分：（1）84 至 85 年間建立 CDIIT 常模資料時收集就讀台北縣市公私立社會福利機構的「嚴重發展遲緩」效度樣本 26 名、（2）九十一學年度就讀台北市公私立幼稚園大班且經鑑定老師確認有發展遲緩的 47 名幼兒、及（3）衛生署民 90 年補助「嬰幼兒綜合發展測驗同時效度及反應度研究」案內經醫學診斷確認之 0-3 歲身心障礙幼兒 79 名。

至於一般幼兒樣本，先選自 CDIIT 常模資料時的台北縣市常模樣本 1,325 名（CA=3～71 個月）；其次，再合併九十一學年度收集自就讀台北市 45 所公私立幼稚園普通幼兒樣本 173 名（CA 平均 67.5 個月），共有 1,498 名一般幼兒。雖然敏感度和特異度的分析不受研究樣本障礙出現率的影響，不過為顧及每一月齡都有約相同比例的樣本數，因此最後以 152 名發展遲緩幼兒為主，從上述 1,498 名一般幼兒中，經配對年齡結果，選取發展遲緩和一般幼兒約 1:4 比例的樣本數，共得 608 名一般幼兒樣本。

這 760 名研究樣本中，男 424 人、女 336 人，實際年齡（CA）介於 3~71 個月（平均 42.2 個月，SD=24.2）。由於經年齡配對，兩組幼兒的平均 CA 均為 42 個月。152 名遲緩幼兒中，男 98 人、女 54 人，CDIIT 發展商數（DQ）平均為 56.5（SD=24.9）；608 名一般幼兒中，男 326 人、女 282 人，DQ 平均為 101.8（SD=13.5）。

二、研究工具

「嬰幼兒綜合發展測驗」是一套評估嬰幼兒認知、語言、動作、社會和自理能力等五項發展能力及行為的個別發展測驗，適用年齡介於 3 個月至 71 個月的嬰幼兒。本套測驗有篩選和診斷兩題本，以全台地區 3,703 名嬰幼兒為標準化樣本，建立診斷測驗之整體發展及各發展領域之 30 個年齡組的年齡分數、百分等級和發展商數及篩選測驗之 7 個年齡組的 Z 分數等標準分數常模。全測驗是以直接施測和父母填寫問卷兩種評量方式收集嬰幼兒發展及行為資料。

該診斷測驗（共 343 題）的信度主要以重測信度、觀察者信度和內部一致性表示：重測信度係數在 .90~.99 間；觀察者信度係數在.19~.56 間，其中「情感與情緒」的觀察者信度較低，解釋應審慎；內部一致性係數介於.75~.99 之間。效度則包括表面內容效度、建構效度和預測效度：幼兒在全測驗測得各發展分數及總分均隨年齡增長而增加，顯示此測驗確為一可測量各發展領域的發展測驗；由各分測驗與總分間相關係數.96~.99 之間結果顯示，內部結構性高；本測驗對普通幼兒和嚴重發展遲緩幼兒的區別準確率較高（各占約 82.1%和 84.6%），對輕微發展遲緩幼兒的區辨率較低（約 55%）（王天苗等，1998）。至於 CDIIT 預測效度的檢測結果，學前測得幼兒在篩選或診斷測驗的結果，可以有效預測未來的特教需要和學業學習問題（王天苗，2005）。

CDIIT 發展篩選測驗是運用 Rasch Analysis 法由全測驗分析出各發展領域在各年齡組內最「適配」（即 Infit 指數在±2 之間）的題目且實施最簡易的 87 題，分測驗與總分得分間的相關在.96~.98 之間，各分測驗分數及總分均隨年齡增長而顯著增加，兩題本在分測驗和總分平均數的相關在.88~.99 之間（王天苗等，1998）。

在本研究，每名樣本都收集 CDIIT 篩選和診斷測驗結果的資料，而僅以兩項測驗結果的「總分」進行 ROC 分析。所以以測驗總分進行分析，理由是：研究者企圖在 CDIIT 診斷測驗的側面圖中，標示出可建議為「一般發展水

準」的陰影範圍，而這陰影範圍需涵蓋所有分測驗在內，因此研究者並不進行每一分測驗切截點的分析。至於，篩檢測驗的切截點分析也在整體考量下，以測驗總分進行 ROC 分析。

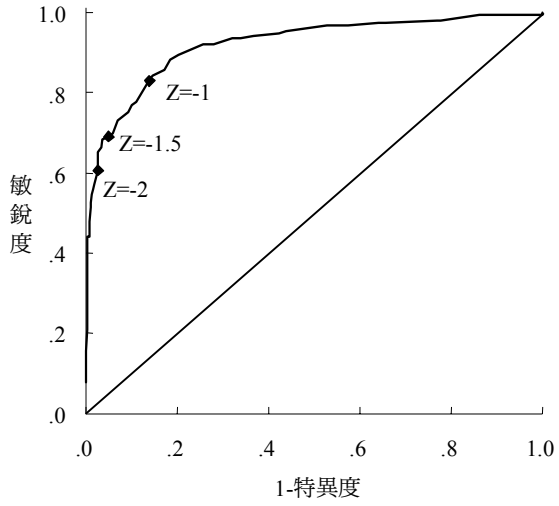
三、資料分析

本研究以 SPSS-PC 進行 ROC 分析。在本研究，「敏銳度」指切截分數足以正確歸類遲緩幼兒（即將遲緩幼兒歸為「遲緩」）的比率；所謂「錯誤接受率」（即 1-特異度），即指錯誤歸類一般幼兒為「遲緩」的比率；「特異度」是指，在被認定為發展無問題的個案中，由測驗結果判斷為發展無問題者的比例。

由於研究樣本來自醫院和幼兒園，故分別經由醫學診斷或鑑定老師綜合研判而獲知「發展遲緩」樣本，因此這些診斷及研判結果為效標。分析時，先從 ROC 分析獲得 AUC，以檢測 CDIIT 篩檢或診斷發展遲緩的準確度。其次，在分析最佳切截點時，考量發展測驗的切截點常建議在 z 值或標準差為-1（相當於 DQ85）、-1.5（相當於 DQ77.5）與-2（相當於 DQ70）處（Anatasi & Urbina, 1997），因此本研究以診斷測驗之發展商數 70、77.5、85 或篩選測驗之 Z 分數-2、-1.5、-1 為切截點，分別分析診斷和篩選測驗得分在不同切截點的敏銳度、特異度、FPF、LR+、LR-、擊中率、約登指數及 Kappa 值，最後綜合研判出用篩選測驗判定疑似發展遲緩或以診斷測驗發展遲緩的最佳切截點。由於本測驗僅有二元分類（遲緩、正常兩種），故不計算當測驗結果有多重分類之 LR（*x,y*）。

研究結果

本研究以發展遲緩和正常發展兩組幼兒在 CDIIT 篩選測驗總分所得之 Z 分數和在該診斷測驗總分所得之 DQ 進行 ROC 分析，結果見圖一。ROC 分析結果除可得 ROC 曲線的面積外，還可獲得由不同測驗切截點而得的 TPF 和 FPF，進而從曲線及其他指標來判斷最佳切截點。以下分別呈現 CDIIT 篩選和測驗總分所得 Z 分數和 DQ 進行的 ROC 分析結果：



圖一　CDIIT 篩選測驗總分所得 Z 分數之 ROC 分析結果

一、CDIIT 篩選測驗總分所得 Z 分數之 ROC 分析結果

（一）CDIIT 篩選測驗之篩檢準確率

圖一呈現 CDIIT 篩選測驗總分所得 Z 分數進行 ROC 分析所得之曲線。結果顯示，ROC 曲線下面積為.92，非常接近 1。由於 ROC 曲線下的面積即代表測驗的判定準確度，面積愈大，表示測驗判定的準確度越大，尤其最完美的曲線應是面積等於 1（當敏銳度和特異度均為 1 時），標記在曲線的左上角。由結果可知，CDIIT 篩選測驗的篩選準確率或區辨效果非常高。

（二）CDIIT 篩選測驗篩檢效果最佳之切截點

表一顯示 CDIIT 篩選測驗總分所得 Z 分數在三個切截點（-1、-1.5、-2）的敏銳度（SEN）、特異度（SPE）、錯誤接受率（FPF）、陽性概率比（LR+）、陰性概率比（LR-）、擊中率（Hit）、約登指數（YI）、Kappa 值的結果。結果顯示，這三切截點之中，Z 分數在-1 處能有較佳的敏銳度（.78）和約登指數（.67），LR-也最低，但特異度較低、錯誤接受率較高；三者的擊中率和 Kappa 值相近；Z 分數為-2 時的篩檢正確性最高（LR+=28.5）。

就敏銳度而言，三切截點的敏感度在 Z 分數為-1 處最高，換言之，若以該篩選測驗結果所得的 Z 分數小於-1 來判定疑似發展遲緩，則正確判定的比率較高。如果依 Barnes（1982）的建議，篩檢的敏銳度起碼應在.75 以上，那麼切截點設定在 Z 分數-1 處時，其敏銳度是唯一合標準者。再從三切截點所得之特異度結果（介於.89~.98 之間）來看，不但合乎 Carran & Scott（1992）及 Meisels（1989）建議不能低於.80 的標準，也能大約符合更保守的標準（.90 以上）（Glascoe, Martin, & Hunphrey, 1990）。此結果意指，以三切截點正確判定非遲緩個案的比率都高，相對地，將非遲緩個案錯誤接受的比率低。

表一　以醫學診斷與特教老師綜合研判結果為效標所得 CDIIT 篩選測驗之 Z 分數於三種不同切截點的預測估計比率及差異比較分析結果

切截點	綜合研判結果			SEN	SPE	FPF	LR+	LR-	Hit	YI	Kappa 值
	遲緩	無問題	合計								
Z=-1											
Z<-1	118（a）	67（c）	185	.78	.89	.11	7.09	0.25	.87	0.67	.63***
Z≧-1	34（b）	541（d）	575								
Z=-1.5											
Z<-1.5	105（a）	30（c）	135	.69	.95	.05	13.80	0.33	.90	0.64	.68***
Z≧-1.5	47（b）	578（d）	625								
Z=-2											
Z<-2	87（a）	11（c）	98	.57	.98	.02	28.50	0.44	.90	0.55	.64***
Z≧-2	65（b）	597（d）	662								
人數總計	152	608	760								

註 1：SEN=a/a+b; SPE=d/c+d; FPF=1-SPE; LR+=SEN/（1-SPE）; LR-=（1-SEN）/SPE; Hit=（a+d）/（a+b+c+d）; YI= SEN + SPE -1

註 2：$^{***}p<.001$

此外，三切截點所得的擊中率和 Kappa 值都相近，其中，擊中率介於.87~.90 之間，Kappa 值則介於.63~.68 之間。由這些結果顯示，無論採用哪一個切截點，都能有效地篩檢出疑似遲緩個案和拒絕非遲緩個案，而且由切截點判定的結果與診斷或鑑定的分類結果之間有「顯著」的關連。至於 Z 為-1 所得的 LR+值是三者中最低，LR-是三者最低。其中，LR+值（7.09）雖較最佳數值（大於 10）稍低，而 LR-值（.25）也較最佳數值（小於.1）稍高，但這些結果都仍接近 Jaeschke 等人（1994）訂出的中度標準。

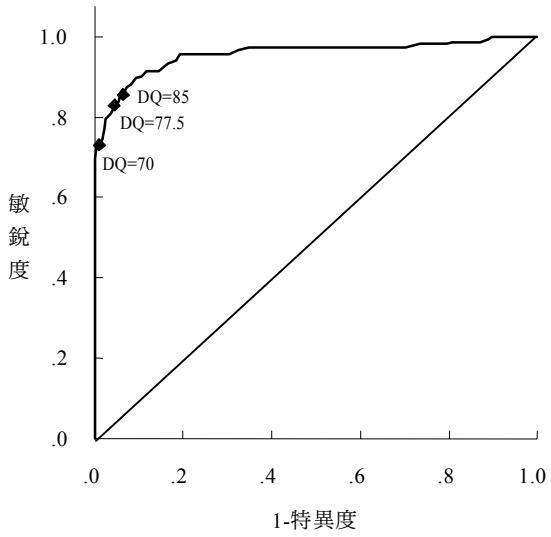
若從 ROC 曲線來選擇「SEN 較大但 FPF 較小的左上角某點」為最佳切截點，則從圖一也可以獲知 Z 分數=-1 最適合。

綜合以上數據資料，基於在選擇最佳切截點時，主要考量選取敏銳度、約登指數、擊中率及 Kappa 值最高且陰性概率比最小者，則 Z 分數為「-1」處是運用 CDIIT 篩選測驗總分來有效篩檢疑似發展遲緩個案的最佳切截點。

二、CDIIT 診斷測驗總分所得 DQ 之 ROC 分析結果

（一）CDIIT 診斷測驗之診斷準確率

在進行此項分析時，以臨床診斷或鑑定老師綜合研判之「遲緩」和「無問題」兩組幼兒在 CDIIT 診斷測驗總分所得之 DQ 結果進行 ROC 分析，結果見圖二。由結果得知，ROC 曲線下面積為.96，非常接近 1，可見運用 CDIIT 診斷測驗的診斷準確率或區辨效果非常高。



圖二　CDIIT 診斷測驗總分所得 DQ 之 ROC 分析結果

（二）CDIIT 診斷測驗診斷效果最佳之切截點

從圖二的 ROC 曲線來看，若選擇 SEN 較大但 FPF 較小、最接近左上角的某點，則 DQ85 可設定為最佳切截點。

表二呈現 CDIIT 診斷測驗總分所得 DQ 在三個切截點（85、77.5、70）的敏銳度、特異度、錯誤接受率、陽性概率比、陰性概率比、擊中率、約登指數及 Kappa 值的結果。結果顯示，切截點在 DQ70（即-2SD）處所得的敏銳度低於一般可接受的.75，而且約登指數最低，因此排除將最佳切截點訂在 DQ70 的可能。

經比較其餘兩點的數值發現，DQ85（即

-1SD）處所得的敏銳度最大（.90），LR-最接近較佳值.1 的標準。然而，DQ77.5（即-1.5SD）處的 LR-（.18）與-1SD 者非常接近，約登指數的數值相同，因此最需考量的是兩者的敏銳度問題。依照 Carran & Scott（1992）的建議，若敏銳度超過.80，就表示測驗具有良好的預測力，因此雖然切截點在-1SD 時的敏銳度最大，但是-1.5SD 處所得的敏銳度(.83)也符合標準。

至於，切截點在 DQ77.5 處所得的特異度、FPF、擊中率和 Kappa 值，均較 DQ85 者為佳。其中，特異度和擊中率在.8 以上，FPF 在.20 以下，而 Kappa 值.78 則是在 Domholdt（2000）認定下屬「顯著」且近完美之相關。再從概率來看，若 LR+概率要大於 10，測驗才愈能正確診斷出發展遲緩問題，則 DQ77.5 的 LR+值才符合預測較佳標準，因此選擇 DQ77.5 為切截點較佳。

綜合研判以上各指標的結果，本研究建議運用 CDIIT 診斷測驗總分來有效診斷出遲緩個案的最佳切截點是「DQ77.5」或「-1.5SD」。

表 二 、以醫學診斷與特教老師綜合研判結果為效標所得 CDIIT 診斷測驗總分之 DQ 於三種不同切截點的預測估計比率及差異比較分析結果

切截點	綜合研判結果			SEN	SPE	FPF	LR+	LR-	Hit	YI	Kappa 值
	遲緩	無問題	合計								
-1SD											
DQ<85	137（a）	64（c）	201	.90	.89	.11	8.18	0.11	.90	0.79	.71***
DQ≧85	15（b）	544（d）	559								
-1.5SD											
DQ<77.5	126（a）	27（c）	153	.83	.96	.04	20.75	0.18	.93	0.79	.78***
DQ≧77.5	26（b）	581（d）	607								
-2SD											
DQ<70	110（a）	1（c）	111	.72	.99	.01	72.00	0.28	.94	0.71	.80***
DQ≧70	42（b）	607（d）	649								
人數總計	152	608	760								

註 1：SEN=a/a+b; SPE=d/c+d; FPF=1-SPE; LR+=SEN/（1-SPE）; LR-=（1-SEN）/SPE; Hit=（a+d）/（a+b+c+d）; YI= SEN + SPE -1

註 2：$^{***}p<.001$

討論與建議

在政府和民間都重視推動早期療育之際，專業人員如何及早篩檢或診斷出發展遲緩的嬰幼兒，往往是決定幼兒及其家庭是否能及時獲得所需的療育和支持性服務的關鍵第一步。為了使目前國內最重要的一項鑑定發展遲緩幼兒的標準化發展測驗—CDIIT 能有運用的具體證據，本研究採 ROC 分析來探究該測驗用來篩檢或判定發展遲緩的準確率及最佳切截點。首先，結果顯示，CDIIT 篩檢測驗的 ROC 曲線下面積為.92、診斷測驗為.96，可見本測驗的篩檢或診斷準確率均極高。如此可以肯定，無論為了篩檢疑似發展遲緩幼兒或確認幼兒的發展問題，運用 CDIIT 篩檢和診斷測驗題本都應該是最佳的選擇。

如果依照近年規畫的「幼托整合」計畫，未來兩歲至五歲的幼兒都將納入教育系統，此外，新住民子女的發展遲緩問題都逐漸浮現出來的現今社會，更凸顯幼兒發展問題篩檢和鑑定工作的重要。「工欲善其事，必先利其器」，

具有良好測量準確度的 CDIIT 正是可以擔當此任務的發展評估工具。

其次，本研究從敏銳度、特異度、概率比和擊中率等資料進行綜合研判可得，CDIIT 篩選測驗總分能有效判定疑似發展遲緩的最佳切截點為 Z 分數-1 處，而診斷測驗總分來有效診斷發展遲緩的最佳切截點為DQ77.5(即-1.5SD）處。後者之結果與吳雪玉等人（2005）發現 CDIIT「動作」分測驗對「動作發展障礙兒童」診斷的最佳分界點為動作 DQ70 的結果不同，這可能因該研究判定分界點的指標不同造成，也可能因該研究樣本多數為動作發展問題兒童有關，未來可進一步探究。由於本研究只企圖從 CDIIT 測驗總分整體探究足以包含各分測驗的切截點情況，故各分測驗的切截點分析並不在本研究探究的範圍。

總之，由本研究獲得有關切截點的結果，能提供評估者在運用 CDIIT 結果做資格認定或教育決定時的參考。當幼兒在 CDIIT 篩選測驗結果的 Z 分數落在-1 以下，則可預測該幼兒有「疑似遲緩」，而需進一步安排更深入的鑑定或診斷，以瞭解幼兒發展問題所在及嚴重程度。反之，當篩檢測驗結果所得之 Z 分數在-1 以上，則可初步研判幼兒沒有發展問題。至於以 CDIIT 診斷測驗結果所得的 DQ，若落在平均數-1.5SD 以上，可研判幼兒的發展屬一般範圍，但若落在平均數-1.5SD 以下，則可確認幼兒的發展有問題。

雖然本研究結果提供評量者在運用 CDIIT 時可參考的判定標準，但是，研究者仍須提醒，由於分析結果可能會受選取樣本特性而有不同，因此由本研究 ROC 分析所得之切截點結果，只能是「重要參考」，但非「絕對標準」。此外，研究者在探究 CDIIT 診斷測驗之最佳切截點時發現，如果考量測驗分數能正確篩選出發展遲緩個案之比率（敏銳度），那麼 DQ85 也是重要的切截點。因此，建議評量者在評估個案時，除了先檢視個案在該測驗分數所得 DQ 低於 77.5 的發展問題，還需要保守地檢視 DQ77.5 以上但低於 DQ85 之測驗結果，如此才能不忽略幼兒在發展上任何可能的問題。畢竟，評量者面對的是活生生的人，並非只是數據而已，因此需特別審慎解釋測驗結果。

尤其，國外臨床心理領域學者近年來提出「以實證為基礎之評量」（evidence-based assessment），期望從生態或環境脈絡下去掌握個案實際的狀況，使個案獲得「以實證為基礎的服務」（evidence-based services）（Mash & Hunsley, 2005）。Ysseldyke, Algozzine, & Thurlow（2000）也強調，在鑑定或評估身心障礙者時，該採「多元評量」模式進行。因此，測驗結果不過是資料中的一項而已，評量人員還應該收集各種與個案有關的資料（包括家庭、學校等生態資料），並且運用專業知識和經驗，做綜合的研判。如果只是參考測驗編製者設定的標準而無專業判斷或未考量生態因素，那麼可能會造成錯誤的認定。而錯誤的認定可能造成兒童就學機會的喪失，也可能造成錯誤的教育安置，因此鑑定人員不能不謹慎為之（Harber, 2001）。

此外，研究者在未來修訂 CDIIT 的指導手冊時，要將所得的切截點資料納入，使該測驗結果之解釋的說明更有依據。在目前使用的指導手冊內，研究者建議將篩選測驗結果判定「疑似遲緩」的切截點訂在 Z=-1，此建議在本研究獲得確認。不過，原本建議診斷測驗用以判定「發展遲緩」的切截點訂在-2SD（DQ70），這建議還需進一步修改，即改以-1.5SD（DQ77.5）為診斷發展遲緩的切截點。至於在 CDIIT 診斷測驗記錄紙內的側面圖上標示有「一般水準」的陰影範圍，未來也須由-2SD~2SD 之間更正為-1.5SD~1.5SD 之間，以方便評量人員在判斷測驗結果時參考。

至於未來研究，可以透過系列的個案研

究，探究 CDIIT 測驗結果由相同切截點或不同切截點判定出的幼兒發展情形。此外，也有必要檢測選取不同遲緩程度之樣本對切截點設定結果的影響。

最後，國內醫療界的論文偶見運用 ROC 分析及相關指標進行篩檢或診斷工具切截點的分析，但是心理與教育測驗領域則較少有關切截點的探討，多由編製者自訂。由本研究，ROC 分析及各種相關指標的分析確實可以為檢視測量工具之評量準確度及設定切截點的方法，可提供測驗編製者參考。

參考文獻

王天苗（2005）：嬰幼兒綜合發展測驗之預測效度研究。**特殊教育研究學刊，29**，1-24。

王天苗、蘇建文、廖華芳、林麗英、鄒國書、林世華（1998）：嬰幼兒綜合發展測驗之編製報告。**測驗年刊，45**（1），19-46。

台北市發展遲緩兒童早期療育通報轉介中心（1998）：**學前兒童發展檢核表**。台北：台北市發展遲緩兒童早期療育通報轉介中心。

余民寧（1993）：試題反應理論的介紹（八）--測驗編製。**研習資訊，10**（1），6-10。

吳雪玉、廖華芳、姚開屏、李旺祚、王天苗、謝正宜（2005）：「嬰幼兒綜合發展測驗」動作分測驗與「皮巴迪動作發展量表第二版」的診斷準確度。**台灣醫學，9**（3），312-322。

吳裕益（1988）：標準參照測驗通過分數設定方法之研究。**測驗年刊，35**，159-166。

林幸台、吳武典、吳鐵雄、楊坤堂（1992）：**學習行為檢核表**。台北市：國立台灣師範大學特殊教育研究所。

林素貞（2000）：國小中文讀寫障礙學生國語課學習行為特徵檢核之編製報告。**特殊教育研究學刊，20**，97-110。

林惠芬（1993）：通過分數設定方法在護理人員檢核筆試測驗之研究。**測驗年刊，40**，253-262。

洪儷瑜、余曉珍（1998）：青少年社會行為簡式量表在情緒障礙學生篩檢工作之運用。**師大學報：教育類，43**（2），43-52。

徐澄清、廖佳鶯、余秀麗主編（1983）：**嬰幼兒發展測驗**。台北市：杏文。

徐澄清、蘇喜、蕭淑貞、林家青、宋維村、張珏（1978）：學齡前兒童行為發展量表之修訂及初步常模之建立。**中兒醫誌，19**（2），142-157。

張正芬、王華沛（2005）：「自閉症兒童行為檢核表」之編製及相關研究。**特殊教育研究學刊，28**，145-166。

陳淑美、盧欽銘、蘇建文、鍾志從（1991）：**貝萊嬰兒發展量表常模的建立**。台北：國立臺灣師範大學教育心理與輔導學系。

黃惠玲（2000）：**零歲至六歲兒童發展篩檢量表指導手冊**。台北市：心理出版社。

路君約（1989）：**心理測驗（上冊）**。台北：中國行為科學社。

廖華芳、王天苗、姚開屏（2002）：早產兒於嬰幼兒綜合發展測驗與貝萊嬰幼兒發展評量第二版之同時效度。載於**第三屆全國早療相關服務發表大會暨國際研討會大會手冊&論文摘要**（63-64），台北市，12 月 12~13 日。

廖華芳、王天苗、姚開屏、吳雪玉、李旺祚、鄭素芳（2002）：**嬰幼兒綜合發展測驗的同時效度及反應度第一年期中報告**。行政院衛生署九十年度科技研究發展計畫。未發表。

鄭明長、余民寧（1994）：各種通過分數設定方法之比較。**測驗年刊，41**，19-40。

Anatasi, A., Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Pintice Hall.

Barnes, K.E. (1982). *Preschool screening: The*

measurement and prediction of children at-risk. Springfield, IL: Charles C. Thomas.

Bayley, N. (1969). *Manual of Bayley Scales of Infant Development*. New York: Psychological Co.

Berk, R. A. (1976). Determination of optiomal cutting scores in criterion-referenced measurement. *Journal of Experimental Education*, *45*, 4-9.

Berk, R. A. (1986). A consumer's guide to setting performance standards in criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172.

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics on Early Childhood Special Education, 12*(2), 196-211.

Choi, B.C. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology, 148*(11),1127-1132.

Chong, DSY, & Karlberg, J. (2004). Refining the Apgar score cut-off point for newborns at risk. *Acta Paediatr, 93*, 53-59.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: Holt, Rinehart & Winston.

Division for Early Childhood (2001). *DEC concept paper on developmental delay as an eligibility category*. Reston, VA: CEC.

Domholdt, E. (2000). *Physical therapy research: Principles and applications* (2nd ed.). Philadelphia: W. B.Saunders.

Glascoe, F.P., Martin, E.D., & Humphrey, S. (1990). A comparative review of developmental screening tests. *Pediatrics, 86*, 547-554.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology, 143*, 29-36.

Harber, J. R. (2001). Assessing the quality of decision making in special education. *The Journal of Special Education, 15*(1), 77-90.

Jaeschke, R, Guyatt, G.H., & Sackett, D.L. (1994). Users'guides to the medical literature, III: How to use an article about a diagnostic test, B: What are the results and will they help me in caring for my patients? *Journal of the American Medical Association, 271*, 703-707.

Liao, H.F., & Pan, Y.L. (2005). Test-retest and inter-rater reliability for the Comprehensive Developmental Inventory for Infants and Toddlers Diagnostic and Screening Tests. *Early Human Development, 81* (11), 927-937.

Liao, H.F., Wang, T.M., Yao, G., Lee, W.C. (2005). Concurrent validity of Comprehensive Developmental Inventory for Infants and Toddlers with Bayley Scales of Infant Development-II in preterm infants. *Journal of the Formosan Medical Association, 104* (10), 731-737.

Mardell-Czudnowski, C., & Goldenberg, D. (1984). *Developmental Indicators for the Assessment of Learning-Revised*. Edison, N.J.: Childcraft.

Mash, E.J. & Hunsley, J. (2005). Sepcial section: Developing guidelines for the evident-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 362-379.

Meisels, S.J. (1985). Predection, prevention, and developmental screening in the EPSDT program. *Child Development Research & Social Policy, 1*, 267-317.

Meisels, S.J. (1989). Can developmental screening test identify children who are developmentally at risk? *Pediatrics, 83*(4), 578-585.

Meisels, S.J., & Provence, S. (1989). *Screening and assessment: Guidelines for identifying young disabled and developmentally vulnerable children and their families*. Washington, D.C.: National Center for Clinical Infant Programs.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Newborg, J., Stock, J., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle Developmental Inventory-Examiner's manual*. Allen, TX: DLM Teaching Resources.

Rao, G. (2003). What is an ROC curve? *The Journal of Family Practice, 52*(9), 695.

Salvia, J., & Ysseldyke, J.E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.

Scott, G., & Hogan, A. (1982). Methods for the identification of high-risk and handicapped infants. In C.T.Ramey & P.L.Trohanis (Eds.), *Finding and educating high risk and handicapped infants* (pp.69-82). Austin, TX: Pro-Ed.

Shonkoff, J.P., & Meisels, S.J. (1991). Defining eligibility for services under PL99-457. *Journal of Early Intervention, 15*(1), 21-25.

Simeonsson, R.J. (1986). *Psychological and developmental assessment of special children*. Boston: Allyn & Bacon.

Straus, S.E., Richardson, W.S., Glasziou, P., Haynes, R.B. (2005). *Evidence-based medicine: How to practice and teach EBM* (3rd ed.). Edinburgh: Churchill Livingstone.

Watkins, M.W., Kush, J.C., & Schaefer, B.A. (2002). Diagnostic utility of the learning disability index. *Journal of Learning Disabilities, 35*(2), 98-103.

Wong M.K., Chong, C.K., Wang, C.M., Lin, H.T. (1997). Validation of a simplified Child Developmental Screening Test in Taiwan. *Formosan Journal of Medicine, 4*, 424-439.

Ysseldyke, J.E., Algozzine, B., & Thurlow, M.L. (2000). *Critical issues in special education* (3rd ed.). Boston: Houghton Mifflin.

Bulletin of Special Education
2007, 32(2), 1-15

Assessment Accuracy and Cut-off Points of Comprehensive Developmental Inventory for Infants and Toddlers (CDIIT)

Wang Tien-Miau
Professor, Dept. of Special Education
National Taiwan Normal University

Liao Hua-Fang
Associate Professor, Dept. of Physical Therapy
National Taiwan University

ABSTRACT

This study had two aims: first, to use Receiver Operating Characteristic (ROC) analysis to evaluate the ability of a Comprehensive Developmental Inventory for Infants and Toddlers (CDIIT) to discriminate developmentally delayed cases from normal cases; secondly, to determine the cut-off points which could assist assessors to identify cases with developmental delays. Total of 760 children aged from three to seventy-one months old, including 152 children with developmental delays and 608 normally developing children were involved. As the results showing from the area under the ROC curve (AUC), both CDIIT Screening Test (AUC=.92) and Diagnostic Test (AUC=.96) showed high accuracy for screening and diagnosis purposes. Judging from true positive fraction (TPF, also known as sensitivity), false positive fraction (FPF), positive likelihood ratio (LR+), negative likelihood ratio (LR-), hit rate, Youden Index, and Kappa, Z scores less than -1 and developmental quotients (DQ) less than 77.5 (-1.5SD) were identified as the best cut-off points for the CDIIT Screening and Diagnostic Test. That is, children whose total scores fall below the cut-off points can be considered as having “suspected developmental delay” in the Screening Test or “developmental delay” in the Diagnostic Test.

Keywords: Receiver Operating Characteristic (ROC), cut-off points, CDIIT