立定跳遠發展階段觀察檢核表之多層面 Rasch 評分量尺模式分析

摘要

目的：利用多層面 Rasch 評分量尺模式分析立定跳遠發展階段觀察檢核表，Haywood(1993)以成份分析方法發展評估立定跳遠發展階段觀察檢核表，將立定跳遠動作分成起跳、和飛行/著陸兩個成份。Haywood 與 Getchell (2005)指出動作樣式、起跳高度與距離，將由起跳成份所決定，因此，將檢核表精簡成為起跳一個成份，只須觀察起跳時手臂和腿部動作。**方法**：採用觀察法，對象為國小一至三年級 36 位學生，以多層面 Rasch 模式來分析資料、篩選試題、受試者信度、失序問題，並建立具有等距特性的量尺分數。**結果**：以多層面 Rasch 評分量尺模式分析立定跳遠發展階段觀察檢核表，具有不錯的信度，題目適切性良好，符合單向度要求，接近有效測量的理想。**結論**：立定跳遠發展階段觀察檢核表具良好信、效度，未來將可提供實務檢核，以發展兒童動作學習之用。

關鍵詞：運動技能、Rasch 模式、評分量尺模式

壹、緒論

一、研究背景

跑步是反覆性動作，施作時並不會輕易感到疲憊，但評估立定跳遠動作每次卻只能實施一次，因為受試者會因反覆測試容易感受疲累，因此，如何以最精簡的工具去精確地評估運動技能，是非常重要的工作之一，由於古典測驗具有試題依賴和樣本依賴的特性，為追求客觀測量的體育測驗，必須選擇現代測驗理論的測量模式。

劉純忠(2005)以 Rasch 評分量尺模式評估立定跳遠發展階段動作檢核表(the assessing development level of the standing long jump observation checklist)，Chou 與 Yau (2006)以 Rasch 部分計分模式評估發展立定跳遠發展動作檢核表，前者之評估結果發現動作樣式(movement pattern）會影響成績表現，評分者間肯德爾和諧係數(Kendall's coefficient of concordances)為 0.97，檢核表重測信度(test-retest reliability coefficient)為 0.95，檢核表之內部一致性信度(Cronbach's α)為 0.63，檢核表概化信度係數(generalizability coefficient)為 0.98，受試者信度(person reliability) （劉純忠，2005)，相當於傳統測驗中的內部一致性信度(Cronbach's α)為 0.79，顯見檢核表內部可能是有問題的。

後者使用部分計分模式(Partial Credit Rasch Model)的研究中發現，檢核表第六題 OMNSQ（遠離受試者能力測量反應的非預期敏感度）高達 1.92，違反模式並且降低測量品質。第三題有類別失序的問題，可能要重新定義類別(Chou & Yau, 2006)。Haywood(1993)以成份分析方法發展評估立定跳遠發展階段觀察檢核表，將立定跳遠動作分成起跳、和飛行/著陸兩個成份。Haywood 與 Getchell (2005)以為動作樣式和起跳高度與距離，將由起跳成份所決定，因此，將檢核表修訂精簡成為起跳一個成份，只須觀察起跳時手臂和腿部動作，所以檢核表只有兩題。

姚漢濤(2002)指出現行標準中測驗的概念只有一個，效度是一個整體的概念，只是

效度需要許多證據來證明。效度最大的轉變，乃是從不同效度證明取代原有的效度型式，而 Rasch 測量可以提供良好效度證明。Baghaei(2008)指出建構效度的一個重要觀點，在於分數意義的可信賴度和它的解釋。科學探究目的在於建立這種效度觀點，也就是所謂的測驗效度證據的基礎，而 Rasch 模式則被視為一個建構效度的工具。

Wilson(2007)概述建構測量模式之架構（詳見圖 1），讓我們理解工具是以何種方法被建構出來的，也藉此架構去了解工具是如何地運作，建構測量模式的形成在根本上是具有架構的，事實上，提供藉由使用四個砌塊中每一個，去發展工具以供使用。這四個砌塊不但提供了關於推論建構的路徑，並且可以用為建構工具的指引，去測量那個建構。它們藉由定義這個建構的概念開始，然後開始去發展這個建構的工作和背景脈絡進而編製試題。這些試題產生反應，然後被分類和給分-這就是界定量尺。測量模式被應用去分析這些分數反應，然後這些測量可以被用以反思這些成功，以這個路徑可以繞回建構圖，並且已經測量到構念的方式。因此，透過這四個砌塊的順序，是一個真實地循環，當然，一個循環可能會被重複好幾次，推論的第二個步驟在於建立分數和建構的關連性。這將由第四個砌塊所完成，傳統上這被稱為測量模式，有時候被稱為心理測量（學）模式。

Duckor, Draney 與 Wilson (2009) 使用建構測量架構來接近精熟理論，提出了一種單向度的結構，用於理解 Wilson 所提出的建構測量架構中四個砌塊的使用，並且探究了關於和對照這個概念化的證據，研究中針對 72 位來自具有建構測量的經驗和專門知識者為受試者，在固定選擇反應試題的反應，資料藉由兩個評分者來給分，並且使用 ConQuest 程式進行 Rasch 部分計分模式的資料分析，藉由 1999 年所頒定的測驗標準所指導，所獲得的效度和信度分析的證據，提供了對於建構理論，和在試題設計的修改期間，工具使用的限定這兩者的支持。

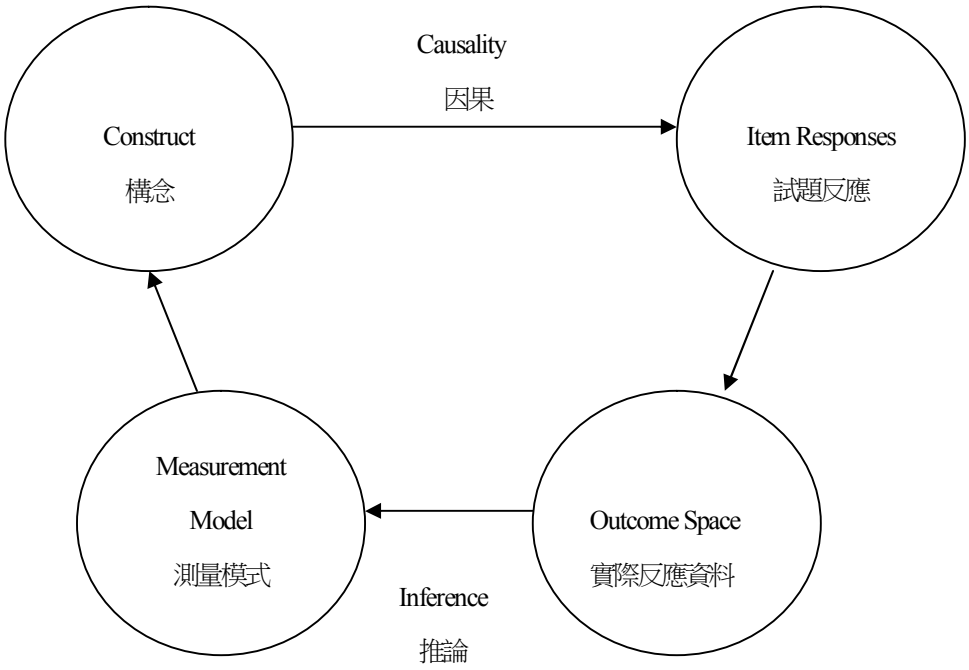

圖 1 四個砌塊展現了因果關係和推論的方向

取自 Mark Wilson(2007). *Constructing Measures.*

王文中（2004）指出應用 Rasch 模式測量，可以解決測驗依賴（test dependent）和樣本依賴（sample dependent）的問題，並且建立等距量尺(interval scale)，達到客觀測量（objective measurement）的效果。姚漢禱、紀世清、周嵩益、姚偉哲（2008）運用 Rasch 模式來修訂立跳遠發展階段觀察檢核表，並且加入結構方程模式的驗證模式架構。

本研究由於使用三位評分者進行動作檢核，檢核表量尺設定原為次序量尺，而且評分者給分方式為主觀給分，為求所獲得的量尺分數具有客觀性及等距性，本研究使用 Rasch 測量模式中的多層面 Rasch 評分量尺模式(Linacre, 1989)進行分析。

多層面評分量尺模式(Many-Facet Rasch Model)的數學公式如下：

$$\log (P_{nijk}/P_{nij(k-1)}) = B_n - D_{gi} - C_j - F_{gk}$$

簡單舉例來說，Cj 顯示了：裁判 j，當他給選手 n 在試題 i 上的施作表現，給予選手 n 評分給 k 時，評分的嚴謹或寬鬆度。若題目反應吻合模式預期，代表所得到的量尺分

數，兼具等距和客觀的測量特性，相對地，如果題目不吻合 Rasch 模式所預期的，則代表該試題可能和檢核表中其他試題所測量的建構是不相同，則必須進一步修正或是探討是否刪除該試題。

二、具體研究目的：

（一）運用多層面 Rasch 模式去檢驗 Haywood 與 Getchell (2005)立定跳遠發展階段觀察檢核表題目的適切性。

（二）考驗檢核表信、效度，並建立檢核表的量尺。

貳、方法

一、研究對象

研究對象為彰化縣彰化市民生國小之 36 位學生，包括：一年級、二年級、三年級男性學生各 12 人，合計 108 筆觀察值（由 3 位評分者針對 36 位學生的評分資料）。九年一貫課程發展明訂「健康與體育領域」分成三個學習階段，每三年為一個學習階段，本研究以第一學習階段之一、二、三年級學生為研究對象。Rasch 模式的優勢，不同於一般傳統古典測驗為基礎的理論模式，Wright 與 Masters (1982)指出 Rasch 模式的重點，在於資料是否適合模式，只要資料能適合模式(data-model fit)，亦即實測資料表現符合 Rasch 模式預期，即可以推論母群，可謂測量有效，代表測量具有建構效度，因此，較大樣本數不是重點，而是要求資料要適合模式，換言之，Rasch 測量特別適合小樣本，這也是其他測量工具所做不到的。

二、測驗工具和方法

（一）測驗工具

本研究的測驗工具，採用 Haywood 與 Getchell (2005)所發展之立定跳遠發展階段觀察檢核表，使用二個試題(皆為 4 點計分)，只觀察起跳動作中的腿部和手臂動作，起跳部份將腿的動作分成四個水準、手臂的動作分成四個水準，每個水準皆設定簡明的特質標準，並且進一步詳細註記說明各個標準所在，以利評分者有客觀的評量標準。

1.起跳腿部動作要素

- 等級一：單腳起跳。從開始位置，起跳者單腳大步起跳（通常很少有準備性腿部彎曲動作）。
- 等級二：膝蓋首先伸展。起跳者在腳踝離開地面之前，就開始伸展膝關節，導致起跳太垂直向上，而不能達到最大的水平距離。
- 等級三：同時伸展。起跳者在腳踝離開地面之時，同時間彎屈膝蓋。
- 等級四：腳踝先向上。起跳由腳踝離開地面先開始，然後膝關節再伸展，起跳者顯示由向前傾斜開始起跳動作。

2.起跳手臂動作要素

- 等級一：沒有動作。手臂不動，或起跳後他們可能飛行（肩膀收縮纏繞）。
- 等級二：手臂向前擺動。從側邊起跳位置姿勢來看，手臂由肩膀向前擺動，或手臂也可能向外擺動，肩膀外展。
- 等級三：手臂伸展，然後部分彎曲。在腿部彎曲期間，手臂同部伸展，然後在起跳時，同時向前擺臂，或手臂擺動位置從不過頭。
- 等級四：手臂伸展，然後完全彎曲。在腿部彎曲期間，手臂一起向後伸展，然後向前擺臂至過頭位置。

（二）測驗方法

由在基層訓練站服務的專任體操教練 1 人，國小體育教師兼運動教育學博士生 1 人，大專體育教師 1 人(具備體育學碩士)，共 3 位擔任立定跳遠動作評分者，三位評分者藉由 X Video Converter 軟體從影片中，抽出最關鍵的圖片，以供三位評分者評分時的參

考。影響處理程式將立定跳遠動作轉換成分解動作後，依照檢核表項目進行觀察，並且評估立定跳遠動作，三位評分者在評分前，將進行評分者訓練，以提高評分者間之信度和效度。

測驗時架設數位攝影機拍攝整個立定跳遠過程，地上鋪設專用海綿墊，測驗前三位一組，只做簡單立定跳遠動作說明，然後逐一測試，每人先試跳一次，然後拍攝三次試做，取最佳立定跳遠成績的資料，分析影片中的立定跳遠動作。

三、資料處理

使用 Facets 測驗軟體，進行試題和受試者測量估計，包括：信度、向度性、題目閾值分布、量尺反應類別設定的適切性等，進行多層面 Rasch 評分量尺模式分析。

叁、結果

一、信度

受試者分類信度為 .92，顯示測驗的內部一致性非常好，本研究採用每位受試者試做三次，並且使用三位評分者的研究設計是很適切的。試題分類信度 .91，顯示試題本身內部的一致性非常的高。

二、多層面模式分析

（一）在試題層面上，各題的均方值越接近 1.0，表示該試題非常穩合模式期望(Wright & Masters, 1982)，本研究在適合度考驗方面，適合度指標中，其中試題的 infit MnSq（訊息加權均方）介於.78~.79 和 outfit MnSq（偏離反應均方）介於 .70～.77 ，適配度皆在可接受範圍 0.5~1.5(Linacre, 2006)。兩種適合度考驗的 ZSTD 值也都介於 -2 ～ +2 之間（顯著水準約在 0.05 左右），都也在合理範圍（詳見表 1），表示題目吻合模式，檢核表的內容整體上反應出測量著相同的建構，大致為單一的向度，顯示出，測量是有效的。

（二）在評分者層面上，評分者 infit MnS 介於 .84～.91 和 outfit MnSq 介於 .69～.79，適配度在可接受範圍（詳見表 2）。ZSTD 值介於 -.3 ～ -1.15 間，也都在合理範圍中，表示評分者吻合模式，顯示本測量結果有效。

（三）在受試者層面上，受試者只有受試者編號 2、21、24、36 等少數 4 人之 Infit MnSq 和 Outfit MnSq 均超過 1.5 不符合模式預期（詳見表 3），然而當樣本數較小，卻只有少數人不吻合模式預期，尚可以宣稱接受，顯示出，此一量表的內容大致反映出測量同樣的建構，符合單向度要求。

表 1 試題適合度考驗統計表

試題	原始分數	整體難度	測量標準誤	Infit		Outfit	
				訊息加權均方	ZSTD	偏離反應均方	ZSTD
1	314	1.88	.24	.79	-1.4	.70	-1.4
2	243	-1.88	.25	.78	0	.77	-.8

表 2 評分者適合度考驗統計表

評分者	原始分數	嚴苛度	測量標準誤	Infit		Outfit	
				訊息加權均方	ZSTD	偏離反應均方	ZSTD
1	183	.23	.03	.91	-.3	.69	-1.1
2	186	-.03	.03	.84	-.8	.73	-.9
3	188	-.20	.03	.89	-.5	.79	-.6

表 3 受試者適合度考驗統計表

受試者	能力	模式測量標準誤	Infit		Outfit	
			訊息加權均方	ZSTD	偏離反應均方	ZSTD
2	5.38	.96	1.90	1.3	1.91	1.3
21	5.38	.93	2.78	2.1	2.79	2.0
36	5.38	.93	5.16	3..5	4..95	3.4
24	-7.85	1.20	4.33	2.4	4.58	2.4

三、整體題目難度分布與受試者特質的配合度

本檢核表目的在於有效區分受試者在立定跳遠動作發展表現上特質的高低，這兩個題目可以有效地配合受試者程度時，題目的難度不能夠太高或是太低。在題目整體表現方面，兩道題目的預設平均難度估計值為 0，而難度估計值標準誤的平均為 0.24，標準差平均為 0.01，說明難度估計的準確性不錯。受試者能力平均為 2.51，試題對於在試題

難度附近的受試者能力提供較高的測驗訊息。顯然這些題目對 1 到 3 年級男性受試者而言也是比較簡單的，未來應該再增加一些難度較高的題目，以有效區分高低分組。

四、量尺反應類別數的適當性

受試者在每個反應類別的平均特質（average measure）和閾值難度(threshold difficulty)，會隨著反應變項的增加而增加。反應類別的平均特質和閾值難度都呈現了單調性的遞增(monotonically)，而且反應類別間增加的範圍應該介於 1.4~5 logits 間，每個反應類別間閾值間距最好有 1.4 logits 以上，但是不要超過 5 logits，以避免在此變項的測量上有太大的間隙（gap）的發生，當超出此範圍時，必須修改反應類別數，比如：合併相鄰反應類別，使量表具有適當的反應類別數(Linacre & Wright, 1999)。

本研究受試者在每個反應類別的閾值難度依序為-8.14、1.80、6.22，反應類別的平均特質，分別為-8.48、-1.28、1.83、3.93 來看，會隨著反應變項的增加而增加，由圖 2 可以發現這四個反應類別的平均特質和閾值難度都呈現了單調性的遞增，雖然符合單調性遞增原則，但是差異不在 1.4~5 logits 之間，小於兩個類別間閾值，增加最好至少有 1.4 logits，否則沒有區別性，超過 5 logits 則會有間隙(Linacre & Wright，1999)。

類別一次數特別少，可見此類別功能不大，建議修改反應類別數，建議合併類別一和二。此外，類別二和三的距離為 9.94，超過 5 個 logits。因此，建議增加新的類別在類別二和三間，以避免在此測量上有太大間隙。

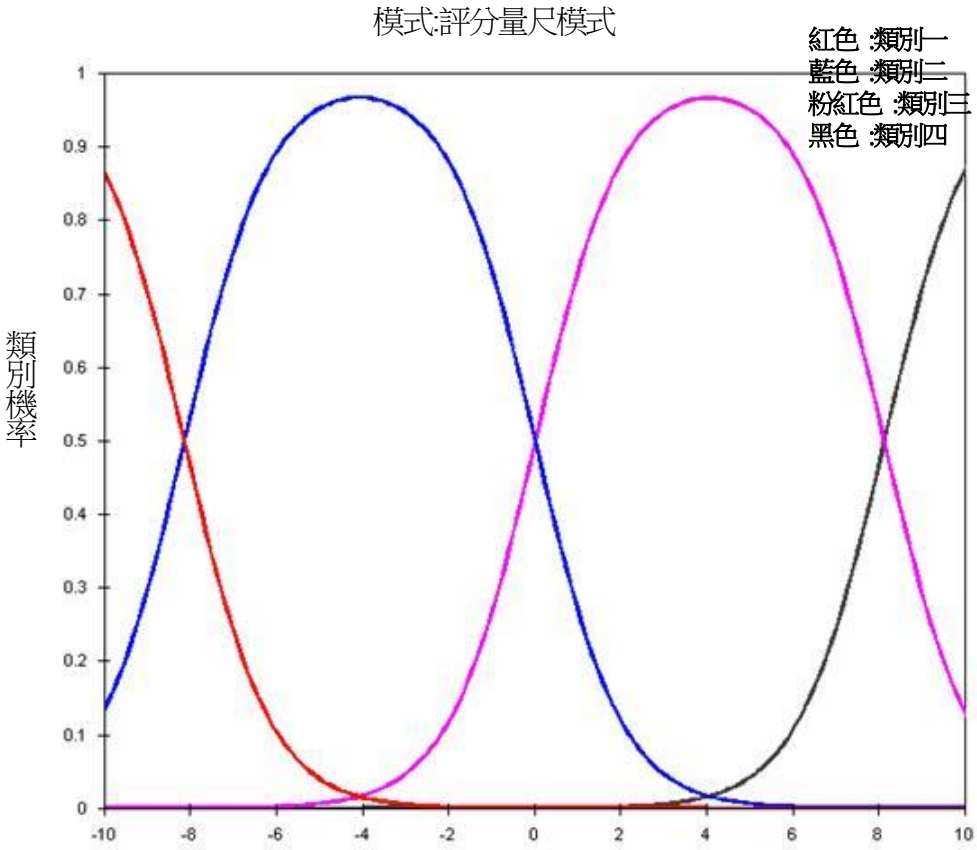

圖 2 類別反應機率曲線圖

肆、討論

一、討論

本研究運用多層面 Rasch 進行立定跳遠發展階段觀察檢核表試題分析，研究結論顯示題目反應吻合評分量尺模式所預期，代表所獲得的量尺分數具有客觀和等距的特性，本測驗具有建構效度。本研究建議應修改量尺反應類別數，因閾值難度差異不在 1.4~5 logits之間，小於兩個類別間閾值，增加最好至少有1.4 logits，否則沒有區別性，超過5 logits則會有間隙(Linacre & Wright, 1999)，其中類別一次數特別少，可見此類別功能不大，建議修改反應類別數，合併類別一和二，增加新的類別在類別二和三之間，以避免在此測量上有太大間隙(gap)，研究結果和姚漢禱等（2008）修訂立定跳遠發展階段觀察檢核表研究一致，藉由 Rasch 模式估計找出類別失序的問題。建議往後研究也應該在發現類別失序問題後，進行反應類別數的修改或合併，再經由 Rasch 模式分析來比較修改前後之差異，以取得最適當的量尺反應類別數。

本研究藉由適合度統計考驗，檢測出 4 位模式資料符合不良的受試者，Cole 與 Zhu(1994)認為，研究中只有少數的個人適合度統計考驗不適合模式期望，此測驗依舊可以被視為適合理論模式的期望。探究其原因可能在於當題數固定，樣本數太小時，這 4 人不符合模式預期，可能是因為拒絕區設定太大，樣本數又太少，容易掉入較為寬鬆的拒絕區（本研究設定為顯著水準在.05 左右），然而受試者反應並不是真正的不符合模式預期，這四個人應該還是來自同一個常態分配，因此，可以接受。

二、結論與建議

（一）本研究之「立定跳遠發展階段觀察檢核表」具良好信、效度，未來可提供體育教師發展兒童動作學習時之實務檢核之用。

（二）建議未來研究可以加大樣本數，以增加測量的穩定性。關於使用單一學校受試者研究，是否具有覆核效度的問題，因為並不在本研究提供效度證明之一，後續可再探討。此外，將檢核表試題精簡為只剩下兩題時，是否能測量到和原始的六道試題所測量的能力相同，應是後續研究可以探討的方向。

（三）本研究受試者為一至三年級男生，九年一貫課程發展明訂「健康與體育領域」分成三個學習階段，每三年為一個學習階段，本研究選擇第一學習階段之一、二、三年級學生，然而這兩道試題在性別和年級之間是否具有 DIF（Differential Item Function，簡稱 DIF）差異試題功能的問題，後續研究應該也要列入探討的方向。

引用文獻

王文中(2004)。Rasch 測量理論與其在教育和心理之應用。*教育與心理研究*,27,637-694。

姚漢禱(2002)。*體育測驗與評量*。臺北市:師大書苑。

姚漢禱、紀世清、周嵩益、姚偉哲(2008)。*修訂立定跳遠發展階段的觀察檢核表*。*國立臺灣體育大學論叢*,*19*(1),35-48。

劉純忠(2005)。*四至六歲男童立定跳遠動作與成績表現之研究*(未出版碩士論文)。國立體育大學,桃園縣。

Baghaei, P. (2008). The Rasch Model as a Construct Validation Tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Cole, E., & Zhu, W. (1994). IRT person fit statistics to diagnose motor function. *Research Quarterly for Exercise and Sport,* Supplement, A-16.

Chou, S. I. & Yau, H. D. (2006). An Evaluation of the Assessing the Development Level of the Standing Long Jump Observation Checklist. *Oral Presentation at 2nd Pacific Rim Objective Measurement Symposium of PROMS HK 2006,* Hong Kong.

Duckor, B., Draney, K., & Wilson, M. (2009). Measuring measuring: toward a theory of proficiency with the constructing: measures framework. *Journal of Applied Measurement, 10*(3), 296-31.

Haywood, K. M. (1993). *Laboratory activities for life span motor development.* (2th ed.). Champaign, IL: Human Kinetics Publishers.

Haywood, K. M. & Getchell, N. (2005). *Life span motor development* (4th ed.). Champaign, IL: Human Kinetics.

Linacre, J. M. (1989). *Many-Facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (2006). *A User's Guide to Winstep*. Chicago: MESA Press.

Linacre, J. M.., & Wright, B. D. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103-122.

Wilson, M. (2007). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associations.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA

Assessing the developmental level standing long jump checklist by many-facet Rasch model

abstract

The **pupose** of this study was to use the many-facet Rasch model to assess the developmental level standing long jump checklist. In the past, Haywood (2005) had used the motor analysis to construct the developmental level standing long jump checklist, and separated it into takeoff and flight components. In addition, Haywood and Getchell (2005) indicated that the motor style, height and length of the standing long jump based on the takeoff component. Therefore, the checklist was reduced in one component which was takeoff, and it needed to observe the leg and arm action of standing long jump. The subjects were 36 students who were from the 1st grade to 3rd grade of an elementary school. **The observation method** and the many-facet Rasch model were used to analyze the data, screen poor items, reliability, disordered quesitons, and to establish interval rating scale. The **result** which used the many-facet mdoel to analyze the developmental level standing long jump had the reliablility, item-fit, and single dimension to close the effective estimation. In **conclsuion,** the developmetnal level standing long jump checklist had the reliability and validity to check the motor skills of children.

Key words: motor skills, Rasch model, Rating scale model