科學教育學刊
2012, 第二十卷第四期, 319-342

Chinese Journal of Science Education
2012, 20(4), 319-342

數學建模評量規準之研究

楊凱琳[1、*] **林福來**[1] **蕭志如**[2]

[1]國立臺灣師範大學 數學系
[2]東吳大學 數學系

(投稿日期：民國100年8月15日，修訂日期：民國101年4月30日，接受日期：民國101年6月28日)

摘要：數學建模問題是一種具有真實或逼真的情境問題，不同的解題者可能會有不同的考量和方法完成這樣複雜的任務。因此，數學建模問題也是一種能啟發學生主動學習的開放性任務。此外，在考量給予適當的評量規準有助於促進學生的數學學習之下，本研究想藉由實作評量與概化理論來探討評量數學建模表現的評量規準。首先，參考文獻上提出的建模歷程共同元素，再透過數學建模專家與國中數學教師的意見調整元素及其內涵，希望此規準能夠涵蓋數學建模中各種想法的共同元素，以完整地評量各種想法的長處與不足。形成數學建模表現評量規準後，基於概化理論檢驗此評量規準的變異來源並計算不同評分人數的概化係數和可靠性係數。最後，以準實驗研究檢驗其對學生數學建模表現的影響。結果顯示，在引模、探模和展模問題中，評分者在各個面向給分的一致性還不錯，而且如果有兩位評分者時，概化係數和可靠度係數皆能達至 .86的水準。另外，在學生尚未參與建構本評量規準的情況下，提供實驗組評量規準的學生數學建模表現仍顯著優於未提供本評量規準的對照組學生。

關鍵詞：概化理論、實作評量、數學建模

壹、研究緣起

九年一貫課程目標旨在發展國民所需的十大基本能力，其中包含主動探索和研究、獨立思考與解決問題等能力的激發與培養(教育部，2003)。主動性包含參與的意願和精神；探索的基本條件是好奇心與敏感度的表現；研究則是運用科學的方法探究問題，以發現新的事實、理論或法則。獨立思考意指學生進行問題解決的歷程，涵蓋批判性思考、創造性思考或反省性思考等高層次的認知運作。PISA (The Programme for International Student Assessment)將數學素養定義為個人辨認和瞭解數學在生活中扮演的角色、產生具有良好基礎的判斷、以及作為有建設性、富於社會關懷與勇於反省的公民運用和處理數學以滿足個人生活需求的能力(Organization for Economic Co-operation and Development [OECD], 2006)。無論是從課程綱要或國際評比的目標，都可看出培養處理

*通訊作者：楊凱琳

05-楊凱琳.indd 319 2012/9/5 上午 11:26:26

現實問題能力的重要性。

數學建模問題是一種非例行性的問題，它需要學生藉由形成數學模式並以數學觀點解釋現實情境以作為下決策的依據(Lesh & Zawojewski, 2007)。因此，不同國家的數學教育研究者都想在中小學的數學課堂中透過數學建模活動來培育學生在未來生活或職業上能使用所學的數學，即培養處理現實問題的能力，例如：德國的Blum (1993)，荷蘭的De Lange (1996)，澳洲的English (2006)，美國的Lesh與Doerr (2003)，丹麥的Niss (2001)。解決數學建模問題並非找關鍵字或直接轉譯成式子即可求出答案，它是一種解題(problem solving)活動(Schoenfeld, 1992)，特別是需要發展具有建設性數學方法來思考現實問題的解題活動(Lesh & Zawojewski)。由此可知，解決問題需要解題者進行推理、提出策略、連結不同概念或領域等能力，已不只是憶取既有知識的認知需求。而且，解決數學建模問題更加強調利用情境中的知識發展可解決問題的模式。相對地，傳統文字題或應用題通常是突顯或為了對應數學結構的情境題，其目的大都只是讓學生練習與應用學過的數學知識(Mousoulides, Sriraman, & Christou, 2007)。而數學建模問題需要學生將情境資訊數學化成數學問題或模式，也就是經由數學思考來結構化現實情境(Freudenthal, 1991)；並且也需要將所得的數學解再脈絡化，也就是解釋或效化此解在境境中的意義(Ferri, 2006)。

有些研究者利用數學建模問題中的情境來啟動學生的數學思考，以及作為促進數學學習的來源(Zbiek & Conner, 2006)；也有些研究者視數學建模歷程即為學習目標，讓學生從中學習如何應用調整所學的數學以適應現實的世界(Stillman, Brown, & Galbraith, 2010)。雖然在理論和實證研究上有不少探討數學建模學習與教學的相關文獻，但是能融入課室中的數學建模表現之評量工具仍不多見。Leahy, Lyon, Thompson與Wiliam (2005)認為在課堂中進行實作評量有利於追蹤學生的學習成長與設計教學介入；同樣地，如果發展能在課堂中實施且有效的數學建模表現之評量問題與評量規準，將來可用於追蹤學生的建模表現之成長，或設計以數學建模為學習目標的教學介入。再者，學習會因應評量的內容、方式與回饋而產生變化。如果學生常常接受只為了精熟練習數學技巧的測驗，其學習的方式和結果大都只是朝向快速的策略且短暫的保留。如果測驗的結果只是提供學生一個分數，他們從測驗所獲得的就不多。在臺灣考試領導教學的文化中，如此現象更是明顯。因此，本研究希望能發展非選擇或簡答題的評量數學建模表現的問題型式，以期更有效評量學生在較完整的建模歷程中的表現，而非片面性的建模歷程中的表現。而完整的建模過程通常是從現實情境問題出發，數學化成數學問題或模式後，進行數學解題，再脈絡化該數學解在現實情境中的意義為何，接著可以回到現實界重新評估情境中的資訊以形成新的假設或問題(Galbraith & Stillman, 2006; Lin & Yang, 2005)。「片面性的建模歷程表現」，意指相對於完整的建模歷程，只偏重需要某個或部分子歷程元素的表現。

本研究基於數學建模為培育學生提出假設、形成模式、詮釋與批判模式的一種教學材料(例如：Doerr & English, 2003; Sriraman & Lesh, 2006)；這些認知行為都需求學生主動思考(Mason & Johnston-Wilder, 2004)，因此參與數學建模活動是一種培養學生主動學習的進路。並且，本研究認同給予適當的評量有助於學生的學習(assessment for learning; Earl, 2003)。因而，意識到發展可融入課堂

的數學建模問題並提供教師與學生關於建模問題的評量規準是必要的。因此，為了將來能幫助學生瞭解自己的數學建模表現以及教師設計與介入提升學生的數學建模表現，發展一個具有信效度的評量規準是輔助學習與教學的重要議題。亦即，評量規準不只是可當成測驗、評鑑和報導學生成就表現的工具，將來也可以設計成有具有引導學生學習與老師教學的功能(Glickman-Bond & Rose, 2006)。據此本研究假設藉由評量規準中清楚描述學生各種可能的表現，有利於學生監控與反思自己的想法而提升其學習表現。本研究將在施測建模問題時，提供實驗組每位學生數學建模表現的評量規準，控制組則沒有此評量規準可以參考，並檢驗兩組學生在排除其起始的數學建模表現的影響後，是否在不同的數學建模問題表現上仍有顯著差異。所以，本研究預探討與檢驗所發展的評量規準之特性。依此目的所形成的研究問題如下：

一、建構一個評量學生在數學建模問題表現的評量規準，不同評分者使用此評量規準的一致性為何？評分變異是否主要來自與學生相關的變異來源？以概化理論推論多少評分者時的概化係數和可靠性係數可達 .80以上？

二、提供學生評量規準作參考，是否有助於提升學生數學建模的表現？

貳、文獻探討

一、評量數學建模能力的相關研究

在理論上或實務上致力於將數學建模融入學校教學前，我們必須先澄清何謂數學建模能力。我們先瞭解什麼是數學建模歷程，再從其過程和整體表現來思考數學建模能力。理論上，數學建模歷程需要在情境資訊中產生猜測與形成問題、發展數學的方法來解決問題、在數學解題的過程中會面臨提出假設或修改問題的情況、因而所求得的數學解答需要再透過現實脈絡的檢驗以釐清其限制條件或進一步修改猜測與數學模式(Blum, 1993; Lesh & Doerr, 2003)。實徵研究上，發現學生可能受困於情境資訊無法進一步發展數學模式(Galbraith & Stillman, 2006)，也可能直接將數學模式套用至情境問題而忽略其在情境中的有效性(Lin & Yang, 2005)。雖然文獻上與數學建模歷程的相關研究都提出不盡相同的建模歷程元素(Ferri, 2006)，數學建模主要是藉由模式來探究現象的規律，歷程中必定包含待建模的現實情境和其可能對應的數學想法，以及這兩類資訊間的變換(Mason & Davis, 1991)。也就是說，其中關鍵的元素為現實界(reality)和數學界(mathematics)，關鍵的歷程即是在這兩個元素之間的轉換：將現實界數學化(mathematizing)和將數學界脈絡化(interpretation)。就數學建模而言，這四個關鍵歷程元素的意義分別為：在現實界中瞭解、簡化或結構情境問題或模式(現實界)，將現實界的情境資訊轉化成數學界的數學問題或模式(數學化)，在數學界中解決數學問題或探討數學模式的性質(數學界)，重回現實情境中檢驗或解釋數學解或模式性質的意義(脈絡化)。

Niss (2004)將數學建模能力視為數學能力中提出並解決問題的子能力之一，數學建模能力又包含分析既有模式的基礎與性質、就現實情況轉譯和解讀既有模式、以及在給定脈絡下能主動建模。主動建模即經歷結構情境、數學化成數學問題與模式、解決問題與處理模式、分析批判模式的有效性並表達與模式相關的結果、並能監控整個建模歷程。而且，Niss也強調能力具有分析性

(analytical)和產出性(productive)等一體兩面的特點。分析性能力關注於理解、檢驗和評析數學現象或過程，產出性能力關注於主動的建構與執行數學想法。依據上述的觀點，數學建模能力不只需要數學概念知識與過程技能，還包含將這些知識與技能主動應用於解決情境問題的能力。也就是說，數學建模能力不只包含在分析既有數學模式時、或處理模式時需要數學概念知識與過程技能；還包含在主動建模時將這些知識與技能主動應用於瞭解與探討情境問題的能力。

評量數學建模能力的方式大致可分為兩種，一種是以若干選擇或簡答題來測驗某一子歷程的建模能力，一種是以單一題目來評量完成一個任務的建模能力。例如：Haines與Crouch (2005)的研究屬於前者；而Bell, Burkhardt與Swan (1992)以及Hjalmarson (2004)則提出以開放式的數學建模問題來評量學生的建模能力。此外，PISA (OECD, 2006)以素養(literacy)導向來建立評量架構，即重視將所學的知識應用在真實生活中的能力；而數學建模能力也是強調將數學知識與技能主動應用於解決現實情境的問題(Mason & Davis, 1991)。也就是說，兩者都強調數學知識在真實生活或現實情境中的效用(utility)。但是，從Turner (2007)挑出PISA 2000和2003與數學建模能力有關的問題並進行分析中，我們也可看出各問題通常只包含片面的數學建模歷程。例如：選擇題型的Rock Concert (Turner, p. 436)，偏重於數學化後的數學解。開放題型的Hartbeat (Turner, p. 438)，偏重於脈絡化數學式和數學解。而且，從PISA對學生在開放題型的給分編碼，也無法看出學生在不同建模歷程的表現。因此，我們認為PISA的理想目標與實際評量試題仍存有落差。這也顯示發展具有完整數學建模歷程的評量試題仍有待努力。PISA採用選舉題或簡答題，一方面便於評分，所以有利於大樣本資料的蒐集。另一方面便於增加題數以涵蓋充分的內容，所以利於增加一份測驗的內容效度(Baghaei & Amrahi, 2011)。在此，內容效度的定義為“the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” (Haynes, Richard, & Kubany, 1995, p. 238)。因此，通常被用來瞭解或比較某些地區或國家學生的表現。而開放性問題可以紀錄學生在數學建模歷程中的反應，也能深入分析其建模歷程的表現，因此有利於將來在學習與教學的互動下提供豐富的資訊。但是，以若干選擇或簡答題測驗子能力的缺點在於無法真正反應完成一個數學建模任務的表現，以開放式建模問題評量學生的不足在於難以概括課程中的數學內容。由於本研究的目的是以能融入教學的評量觀點來發展評量數學建模表現的問題與評量規準，因此將採用開放性問題作為本研究的評量問題之題型，而內容即先選定與規律相關的等差級數。

二、實作評量的原理原則

學習會因應評量的內容、方式與回饋而產生變化。如果學生常常接受只為了精熟練習數學技巧的測驗，其學習的方式和結果大都只是朝向快速的策略且短暫的保留。如果測驗的結果只是提供學生一個分數，他們從測驗所獲得的就不多。實作評量的方式不僅可用來克服上述的侷限，而且也可以檢驗學生是否能有效運用數學的語言來表達想法(Ellerton & Clarkson, 1996)。實作評量是希望藉由評量學生問題探究的表現以培養其實際解決問題的能力與知識，並不同於傳統測驗的目的在於證明學生的學業成就表現(Baron, 1991; Black & Wiliam, l998)。此

一另類評量方式也具有整合教學與評量的功能，但是在課堂上進行實作評量仍受限於時間和人力。當以紙筆作答進行數學建模問題的實作評量時，如何從學生紙筆作答結果給予客觀且有建設性的意見並不容易。既有的文獻少有針對開放性數學建模問題設計具有信效度的評量規準，如何應用實作評量的想法來評量數學建模表現仍是一個有待研究的問題。

Meyer (1992)指出真實評量和實作評量最大的區別在於，評量任務對真實世界的保真程度(fidelity)。真實評量比實作評量更強調評量任務和真實生活中所需知識、技能與態度間的關係。但是，設計評量任務時需要配合受評者的背景和受限於評量的時間和工具。因此，有時需要對真實情境加以簡化成部分類似真實或只是逼真的情境，而被視為一種實作評量。依照Meyer的觀點，真實評量也是一種實作評量，但是並非所有的實作評量都是真實評量。一般而言，真實評量需要受評者就任務選取、問題界定、資源取用與進程規劃加以思索，而實作評量有時候可能會對上述子任務給予若干限制而削弱其真實性。以數學建模評量為例，可將2010年全民節能減碳數學建模競賽視為真實評量的任務，而限制在一或兩節課內解決開放性且逼真的數學問題可視為實作評量的任務。也就是說，實作評量所提供的情境問題可以是簡化的或可想像的情境，而真實評量則特別強調提供與真實生活經驗非常接近的情境。數學建模始於某種情境，此情境不一定完全來自於真實生活，只要是可想像的真實即可(Lesh & Doerr, 2003)；此也符合實作評量的特徵。因此，藉實作評量的原理原則來評量學生在數學建模問題的表現應該是合適的。

除了提供一個較具挑戰性的問題任務外，實作評量不只提供分數或學生之間的相對排序結果，亦想提供質性的資訊以形成適切的回饋(Linn & Miller, 2005)。所以，一種多層次的評分規準來分析此多樣且複雜的解題過程則是必要的工具。Miller與Linn (2000)除了提出評定實作表現的十項原則，也進一步闡述Messick (1995, 1996)所提的效化實作評量(validity of performance-based assessment)觀點：問題內容(content)、思考歷程(substantive)、問題間結構(structural)、概化(generalizability)、外在效度(external)和結果(consequential)。從這些複雜的面向看來，都足以顯現實作評量的設計充滿挑戰。特別在訂定評分規準上，Stiggins (1994)也建議下列6個步驟：(1)列出重要特質或要素；(2)將特質或要素分類成幾個層面；(3)以簡單清楚的語言界定各層面；(4)分析實作表現；(5)敘述不同等級行為；(6)試用、修訂與改進。但是，訂出一個具有信效度的評分規準依然不是一件容易的任務(Miller & Linn)。

實作評量雖然有許多優點，例如：更能符合需要高階思考、推理、解題和溝通等教育目標(Resnick & Resnick, 1992)；讓老師更瞭解學生的學習情況與有利於提升學生的數學解題表現(Fuchs, Fuchs, Karns, Hamlett, & Katzaroff, 1999)；實作評量中的評量規準可作為教學介入的工具並改善學生的學習(Reddy, 2007)。但是，由於設計不易、施測時間較長、內容代表性較差、計分複雜主觀、評分者一致性不高、也難以進行團體比較等都限制(李坤崇，2006)，真正在課室中實行實作評量仍是困難重重。本研究希望評量工具可以在一堂課內完成，以利於往後將數學建模問題的評量融入課堂教學。基於國中一堂課是45分鐘，並且前5分鐘用來班級管理、後10分鐘可用來總結，所以估計合適的作答時間是30分鐘。另外，本研究也擬定

設計「引出學生想法」、「更深入探索相關想法」、「調整延伸既有想法」等一系列的數學建模問題(楊凱琳、林福來，2006)。如此，可以克服設計不易、施測時間較長、內容代表性較差等限制。之後，再發展可行且有效的評分規準以克服計分複雜主觀與評分者一致性不高等限制。如果內容代表性與評分規準的信度提高了，瞭解不同問題間、個體間或群體間的差異也就比較可行了。

三、概化理論

為了檢驗評分者使用多向度給分的評量規準時給分結果的可信度，除了以計算完全一致性比例和Cronbach's α係數外，也透過概化理論提供可特殊化為單一評分者與一般化為多位評分者給分結果的概化係數與可靠性係數。此外，本研究也利用概化理論同時檢驗分數變異來源是否主要為與受測者相關的變因，以及受測者在不同評分向度的表現是否具有一致性。

概化理論(generalizability thoery)可視為Cronbach's α的一般化，為了應用於非傳統單向度給分的測驗或評量。從根據一種誤差來源進行分數一致性的估計，一般化至同時估計不同的誤差來源，例如：同時估計評分向度與評分者所造成的誤差。在實作評量中，即需要此方法來估計比較各種誤差對測驗結果推論的影響程度(Brennan & Johnson, 1995)。概化理論藉由運用變異數分析的方法和隨機變異成份，估計總變異量以及相對與絕對的殘差變異，用以理解不同來源的測量誤差對分數概化一致性的影響(Brennan, 2001)。

概化理論可依研究目標分為兩種；第一種是找出各種的變異成份大小的概化研究(Generalizability Study, G-Study)，第二種即是利用已知的變異成份找出能使概化係數變大、測量成本變低的決策性研究(Decision Study, D-Study)。概化理論也以兩種係數來表示受試者反應的可信程度，一種用來作為相對性決策的概化係數(generalizability coefficient)，一種用來作為絕對性決策的可靠度係數(dependability coefficient)。相對性決策或絕對性決策之區分是來自於不同的評量目標，前者預計將受測者反應作出相對比較，而後者預計將受測者反應作成表現的層次(Brennan, 2001)。

概化理論中(Shavelson & Webb, 1991)，每一個可能的誤差來源可稱為一個面向，但是並不包含受測者。儘管變異數分析可提供最小誤差的不偏估計，但是概化理論在測量對象或面向的水準數過少時，有時會產生負數的變異數估計值。這違反變異數成份應該皆為正數，也是概化理論的最大缺點。不過，當估計值產生負值時，可藉由將負的估計值設定為0或是增加樣本數等方式解決。此外，變異來源中如果有固定效果或因素時，其處理方法有二：一是平均固定效果因素的各個層次，二是分別估算固定效果因素之各個層次。當使用第一種方法發現固定效果因素及其相關交互作用之變異成份過大時，則建議採用第二種方法。

概化理論優於古典測驗之處包含：可同時考量多種的誤差來源；提供一個統整的方法來處理不同的誤差並計算相對性或絕對性的測驗信度；並不需要假設各誤差來源是不相關的並且還可估計不同誤差來源交互影響所產生的誤差；也不需要假設每個問題的難度都是一樣的(Van Leeuwen, 1997)。而概化理論的缺點就是，其數學結構比古典測驗更加複雜，並且無法處理不完整的資料(Sun, Valiga, & Gao, 1997)。

05-楊凱琳.indd 324 2012/9/5 上午 11:26:27

參、研究方法

一、評量工具——數學建模問題

本研究參考Lesh與Doerr (2003)所提出的一系列建模活動：引模(model-eliciting)、探模(model-exploration)和展模(model-adaptation)活動之特徵，設計這三種類型的數學建模問題作為評量工具。依據Lesh與Doerr的想法，這三種建模活動是一系列依序出現的教學活動。引模活動的用意是透過合適的情境引出學生與教學設計預期發展的數學數學概念之初始模式，探模的用意是讓學生在各種表徵或媒介中更深入瞭解預期發展的數學概念，展模的用意是讓學生思考如何將在探模活動中發展的數學概念修改或推廣至不同的情境。但是，本研究將其應用於設計一系列的評量試題。此外，本研究設定各以30分鐘的評量時間與個別紙筆作答方式進行引模、探模和展模問題的施測。在用途、時間和實施方式的限制下，本研究的引模、探模和展模試題並非完全符合Lesh與Doerr在教學設計中所想的一系列建模活動。但是，本研究也是依序在學生未正式學習某數學概念時、在教師一開始引入數的規律和等差數列概念後、在等差級數概念教學後，使用引模、探模和展模問題評量學生的數學建模表現。引模問題的設計是讓學生應用既有的概念也能解決引模的情境問題，而且引出的既有概念是與等差規律相關的；探模問題的設計是可採用等差概念解決的情境問題；展模問題的設計是需要學生延伸或調整等差規律至等比規律的情境問題。這三種問題的用意均有所不同，所以本研究並不假設此三個建模問題是具有相同評量目標的任務。本研究也會藉由概化理論分析學生的表現以檢驗此假設是否成立。

圖1分別是引模、探模和展模問題：A、引模問題：老闆是騙人的；B、探模問題：珍珠有幾顆；C、展模問題：寶石比一比。解決引模問題主要想要引起學生以規律(比前一顆貴200)在特殊例(25顆)中檢驗資訊；解決探模問題需要學生利用未知數與規律(比前一顆貴1,000)來轉化與處理資訊；解決展模問題需要反思規律性問題的內容(總重量類比原來的價格)與結構(等差改為等比)。如果我們只提供一個開放性建模問題請學生作答，可能會面臨學生只就解決數學問題或模式來表達。為了避免他們因以前作答的習慣而沒有表現出情境或脈絡化的想法，所以將在各問題下設計四個子問題(1)在上述任務中，你考量哪些條件或線索？理由是什麼？(2)你將任務中的條件或線索轉成什麼樣的數學問題或方法？(3)請你將解題的過程寫下來。(4)請檢驗或解釋你的答案。

這三個問題設計是先從探模問題開始，以正在學規律與代數的國二學生為施測對象，蒐集分析其表現特徵。發現有國二學生能假設顆數算算看、判斷是奇數顆或嘗試要使用符號來代表未知數；依此設計引模問題以提供各種學生都能表現其想法。也發現有國二學生不能找到適合的未知數來符號化或未察覺自己得到的答案會使得某些珍珠的價值是負的，所以調整探模問題中的規律(相差1,000)，希望有利於學生主動察覺其答案是否合理。希望學完規律與代數的學生能夠將以數字或符號假設檢驗的思維擴展至等比的規律，依此設計展模問題。

巨觀分析上述三個問題時，發現這三個問題主要都是利用規律來檢驗、轉化或處理問題中的資訊，在解題的過程中，也都需要藉由自己重新組織或推論的資訊再回到問題情境中檢驗其合理性。但是微觀分析時，解決這三個問題所需的子能力又不盡相同。解

A、引模

美華要出席好友的婚禮，至珠寶店借了一串珍珠項鍊，不料歸還當天卻不小心將鍊子勾斷，最後將找回的 24 顆散落的珍珠拿回店裡歸還。老闆跟美華說：

這條項鍊售價 56,300 元，共有珍珠 25 顆，現在少掉的那一顆是最大顆的。由於從左右兩端開始，一直到中間最大顆為止的每一顆珍珠，都比前一顆貴 200 元。看在是老顧客的份上，壞掉的鍊子不用賠，只要賠遺失珍珠的價錢 3,800 元就可以了。

想想看，老闆要美華賠償 3,800 元，這樣合理嗎？

B、探模

在某次聚會中，王小姐的珍珠項鍊不小心被扯斷了。在場的來賓紛紛幫忙撿拾滿地的珍珠。王小姐說：

我記得項鍊是由 10 多顆珍珠串成的，珍珠都是經挑選後排列而成，中間的珍珠最大也最好。從左右兩側開始，每一顆連續的珍珠都比前一顆貴 1,000 元，直到大珍珠為止(含大珍珠)。項鍊的價格是 50,300 元。

如果王小姐的珍珠已撿回 13 顆，需要再請人幫忙找嗎？

C、展模

林太太有一條水晶顆數不超過 10 顆的水晶項鍊散落。只知從左右兩端開始一直到中間最大的水晶為止，每一顆水晶的重量都比前一顆多重 1 倍。若左右兩端最小的水晶重量是 6 克，撿回顆數的總重量是 210 克。請判斷水晶共有多少顆？

(1) 你認為問題(I)(引模A)有解嗎？ □ 有　□ 沒有　□ 不確定

(2) 如果你勾有，請你試著**解決**該問題；如果你勾沒有，請**修改**該問題成有解的問題。如果你勾不確定，請寫下你不確定的**想法**。

圖1：引模、探模和展模問題

決老闆是騙人的引模問題主要是能懷疑並檢驗問題中的數字資訊有所矛盾，相關的數學能力是有策略性地計算。解決珍珠有幾顆的探模問題主要是能符號化適當的未知資訊並檢驗價格不可能為負數，相關的數學能力是代數列式與運算。解決寶石比一比的展模問題主要是能依題意列舉倍數關係後並整合題意資訊推論水晶顆數，相關的數學能力是策略性地列舉與計算倍數關係。因此，研究中並不假設此三個建模問題在實作評量是屬於同一種作業。

二、形成評量規準的歷程

本研究除了設計合適的數學建模問題外，主要的目的是發展有用且有效的評量規準以評量學生在數學建模問題的表現。

Stiggins (1994)所提出的發展課室評量步驟是不分學科內容且一般性的步驟，其想法已被應用於大學的線上教學系統(例如：Liang & Creasy, 2004)、大學和研究所的統計教育(例如：Dickinson, Onwuegbuzie, Hines, & Hall, 2007)、大學的工程教育(例如：Trevisan, Davis, Calkins, & Gentili, 1999)、中學的數學和教學教育(例如：Brookhart, 1997)、小學科學教育(例如：Kilfeather, O'Leary, & Varley, 2006)。本研究進一步將Stiggins發展評量步驟的想法，推廣應用於建立數學建模的評量規準。本研究發展的歷程如下：

(一)列出數學建模的歷程元素與主要內涵

楊凱琳與林福來(2006)在融入數學建模教學活動的環境下探討參與教師的反思歷程時，曾以數學建模歷程和數學思維架構作為分析學生數學建模思維的架構。但是本研究是透過紙筆測驗評量學生的數學建模表現，所以並不能像個案研究中仔細探查學生建模歷程的想法。因此，既有的觀察學生思考的評量架構並不適用，而必需再依據數學建模問題的結構與內容調整成可行的評量規準。所以，我們捨棄此架構中數學思維向度，並以數學建模歷程元素為主。接著，我們將透過專家意見和試用過程重新修訂該架構。

(二)依據評量目的與目標分類

數學建模表現評量規準的評量目的在於，能幫助學生瞭解其數學建模表現的優缺點，能幫助教師知道學生之數學建模表現所需的教學介入目標為何。此規準的評量目標即是：評量學生數學建模的表現，包含現實界、將現實界數學化、數學界、將數學界脈絡化等四個數學建模歷程元素的表現。和Galbraith與Stillman (2006)將建模歷程區分成五階段不同之處在於，本研究其最後兩個階段：從數學解轉換成解的真實意義、從解的真實意義轉換成修改模式或接受解，併在現實界裡，因為本研究是以學生的書面資料作為評量的依據較難細分其過程和結果。

當然數學建模問題的難易深淺差異性很大，所以學生的數學建模表現除了數學內容外，也受到問題的複雜度與開放度的影響。假想有兩題複雜度與開放度均相當的建模問題，但解決問題所需的模式分別是學生較熟悉的和較不熟悉的，學生在較熟悉的建模問題上之表現理應會比較好。因此，數學建模表現評量規準雖然以數學建模歷程元素為主，但是在各元素的表現仍與建模者的知識、推理與策略等能力相關。此外，Galbraith與Stillman (2006)所發展的建模能力細目是需要透過訪談、討論或放聲思考蒐集學生思考的資料來診斷學生在建模歷程中可能遇到的阻礙。而本研究的研究目的並非瞭解學生的數學建模思考歷程，而是希望就學生的紙筆作答來評量學生。換言之，Galbraith與Stillman的建模能力細目就如同透過臨床訪談來診斷學生建模困難時所需的指標，不同於紙筆測驗的開放性試題之評量規準。由於本研究在設計數學建模問題時已考量數學建模歷程的重要元素，而針對四個歷程元素分別設計一個子問題。此外，學生在這些歷程元素的表現有何關聯性也是本研究想探討的問題之一；因此，這些歷程元素被視為評量規準的一個向度之一，另一個向度則是學生的解題表現。

(三)描述各層面的意義

接著以本研究的引模問題為例(見圖1)，描述現實界、將現實界數學化、數學界、將數學界脈絡化的意義。現實界即是思考老闆所說的資訊(例如：售價、顆數、最大顆的單價)是否為真，將現實界數學化即

是以老闆所提供的資訊列式(例如：以最大顆價值3,800為前提列出算式)，數學界即是運算所列式子的總和(例如：以最大顆價值3,800為前提算出所有珍珠的價值)，將數學界脈絡化即是利用數學解的結果來批判情境中老闆的論說(例如：以最大顆價值3,800的前提算出來的總和會大於56,300，表示老闆給的資訊有誤或高估最大顆的價值了。)。

(四)分析學生的建模實例

也以這四個建模歷程元素為主分析學生在賣魚建模問題的認知特徵，結果發現個人的現實與數學知識和建模表現有很大的關係，其中下決策、無限(多種)決策、想要最佳化與模式意義分別是啟動學生理解現實情境、將現實界數學化、解決數學模式和將數學界脈絡化的思維機制(楊凱琳、楊子錕，2008)。

(五)描述各層面的品質

將思考範圍或對象的真實性、系統性、完整性與結構性區分成低、中、高，如表1現實界中的提出關鍵因素或因素關係、提出關鍵因素的道理、提出因素關係的道理。只提出因素未有其理由則品質最差，不只論述單一因素的理由還論述因素間關係的理由則品質最佳。但是又考量有可能論述單一因素的理由之真實性又優於論述因素間關係的理由，所以將此區分視為各層面的質性分析。再以具有建模經驗的專家之觀點，檢驗規準的內涵和完整性(楊凱琳、林福來、蕭志如，2008)，並修改了質性分析(見表1第二欄)。第三欄表示品質等級，即則將各品質區分成高3、中2、低1或無(包含不知所云) 0等四種等級的分數。在實際的應用中，各高、中、低品質的區別是由評分者針對若干作品評析後再共同討論訂定的。

(六)以較一致的語言表達各層面的品質

以上述表1的評量規準，請2位國中數學老師、1位高中數學老師與3位研究生試著用來評量學生的建模結果，發現幾個問題：(1)

表1：第一版的數學建模表現評量規準

數學建模歷程的元素	質性分析	品質等級(0 ~ 3)
現實界(R)： 考量情境因素的質與量	提出關鍵因素或因素關係(R1) 修改後：任務的觀點和可行性(R1)	越貼近真實越高分
	提出關鍵因素的道理(R2) 修改後：提出考量的因素(R2)	越有系統越高分
	提出因素關係的道理(R3) 修改後：提出考量因素的道理(R3)	越有憑有據越高分
將現實界數學化(Ming)： 將情境轉成數學	量化或給出相關假設(Ming1)	越一致越高分
	表徵化情境因素(Ming2)	越完整越高分
	整合各個表徵(Ming3)	越有結構越高分
數學界(M)： 解決問題的數學方法	給出方法(M1)	越可行越高分
	執行數學方法(M2)	越流暢越高分
	給出數學結果(M3)	越完備越高分
將數學界脈絡化(Cing)： 數學方法與結果的意義	解釋數學方法的情境意義(Cing1)	越合情合理越高分
	解釋數學結果的情境意義(Cing2)	越有發展性越高分
	連結或延伸至其他的情境(Cing3)	越具變通性和獨特性越高分

不容易看懂評量規準的內容；(2)給分的一致性不高；(3)評量一份內容較多的結果平均需要約8分鐘的時間；(4)感覺評量規準有點複雜；(5)學生提出來的問題可能是以前看過的好問題，並非反思其建模歷程所連結或延伸的。因此，研究者試著重新結構原始的評量規準。首先將質性分析與品質等級合併，刪除連結或延伸至其他的情境這個子規準，並嘗試著在數學建模歷程中增加表達的面向。但是，初步分析結果發現會使得較不善於以文字表達想法但會以數學想法解題的學生之得分更加低於較善於以文字表達想法但不會以數學想法解題的學生之得分。因為在現實界和脈絡化已含有文字表達能力，如果又特別加上表達面向，會使得文字表達的權重太多。所以又將表達面向予以刪除，表2呈現修改後的評量規準。

綜言之，基於文獻上對數學建模歷程的瞭解，我們聚焦於現實界、將現實界數學化、數學界、將數學界脈絡化等四個主要的數學建模歷程元素，並從此四個元素分析臺灣學生在數學建模歷程中的思維機制，因而奠定了以此四個元素作為評分向度的效度。本研究的內容效度是奠基於將數學建模問題融入課程中有關規律的單元，所提供的建模問題包含正式學習前提供引模問題、學習中提供探模問題以及學習後的展模問題，而且各問題都列出與建模歷程相對應的子問題；此即為本研究工具的內容效度。其侷限性將在討論時補充說明。

三、評量對象與過程

參與對象為臺北縣某所國中二年級的學生，方便取樣其中兩個班級。這兩班不僅來自於同一所學校，其數學教師也是同一位老師；檢驗這兩班在國二一開學時的複習考數學成績也無顯著差異(t = -0.250, df = 67, p = .804)。實驗組一班34人，進行「引出學生想法」的引模問題前測、提供評量規準下解決「更深入探索相關想法」的探模問題、最後再進行「調整延伸既有想法」的展模問題後測。對照組一班35人，進行引模問題前測、未提供評量規準下解決探模問題、最後再進行展模問題後測。三次測驗各需30、30、30分鐘，測驗引模問題的時間是剛開始學習規律與代數的第一節課，隔天測驗探模問題，展模問題是已經學完規律與代數之後施測，與探模問題的施測時間相隔20天。如此，是為了減弱實驗組學生對評量規準的記憶效應。實驗組和對照組的差別只在於解探

表2：第二版的數學建模表現評量規準

數學建模歷程	沒有反應(0)	模糊想法(1)	不完備想法(2)	完備想法(3)
現實界：考量情境的質與量	空白或不知所云	提出零星但不合理或不可行的條件、線索。	提出零星但有用的條件、線索。	提出完整或系統化的條件和線索。
將現實界數學化：將情境轉成數學	空白或不知所云	形成不清楚或有誤的數學算式、方法或問題。	形成清楚的數學算式、方法或問題，但與情境的連結不清楚或不合理。	形成清楚的數學算式、方法或問題，且與情境的連結清楚或合理。
數學界：解決問題的數學方法	空白或不知所云	解法不正確或不合理。 解法不清楚，未給出答案。	解法清楚，給出片面或有誤的答案。 解法不清楚，但是給出合理的答案。	解法清楚，而且給出合理的答案。(可包含只是計算錯誤)
將數學界脈絡化：解釋數學方法或解	空白或不知所云	不合理地檢驗或解釋其解的意義。	片面地檢驗或解釋解的意義。	具有正確且完整的解釋。

模問題時前者有評量規準而後者沒有，所有作答過程都在正式的數學課中由其數學老師監考，並鼓勵學生作答以取得神秘小禮物。於引模和展模問題作答前給予兩組學生相同的說明，於探模問題作答前在實驗組的說明加一句：提供一份評分標準讓大家參考，對照組則無；作答過程中數學老師並未提供其他資訊；作答後數學老師會讓學生提供想法後總結可能的答案，但兩班都未討論評分標準。

四、檢驗提供評量規準的方法

本研究一方面依據概化理論檢驗評量規準的信度，一方面藉由準實驗研究法來檢驗其對學生表現的影響。引模、探模和展模問題的作答者共有69位學生，三題都未有全空白的學生有37位，從中任意選取24位學生，作為分析數學建模表現評量規準信效度的資料。如此選取的理由是：(1)在三題都有作答的情況下檢驗評分者間的一致性，以避免評分者間的高一致性是由於學生未作答或不知所云的結果；(2)在學生人數不多的情況下檢驗本評量規準，以得知學生樣本較小時的概化係數和可靠度係數。除了描述各評分者間的評分一致性、各建模任務間的相關性與各評分向度間的相關性之外，本研究根據概化理論分析變異程度。造成評分變異的來源包含學生能力(p)、評分向度(d)、評分者的解讀(r)、任兩者間的交互作用(pd, pr, dr)以及殘差項。

依據本研究所形成的評量規準(表2)，24位學生的引模、探模和展模作答結果都各有三位評分者依各向度給予一個分數。此三位評分者都參與規準發展的調整與討論，並在評分前先就引模問題討論各面向給分可能對應的學生作答內容，其餘兩個建模問題只討論各面向可得最高分的解法或解答為何。然後，即使用表2的評量規準給予評分。在引模問題的給分標準是：學生在情境中若寫下「共56,300元，後比前貴200元，中間最大顆，老闆開價賠償3,800元」等完整主要線索則屬於「提出完整或系統化的條件或線索」，所以給予3分；若寫下部分主要線索則屬於「提出零星但有用的條件或線索」，所以給予2分；若只寫下「從左右兩端開始，掉的鍊子不用賠」等無關緊要線索則屬於「提出零星但不合理或不可行的條件、線索」，所以給予1分。學生在數學化中若寫下正確算式、解釋如何列式或假設未知數列式等都屬於「形成清楚的數學算式、方法或問題，且與情境的連結清楚或合理」，所以給予3分；若寫下以顆數當未知數列方程式則屬於「形成清楚的數學算式、方法或問題，但與情境的連結不清楚或不合理」，所以給予2分；若寫下錯誤的算式、代數式或只假設未知數則屬於「形成不清楚或有誤的數學算式、方法或問題」，所以給予1分。學生在數學中若算出大於56,300元或最大顆價錢少於3,800元都屬於「解法清楚，而且給出合理的答案(可包含只是計算錯誤)」，所以給予3分；若算出非整數顆數解則屬於「解法清楚，給出片面或有誤的答案」，所以給予2分；若代數運算有誤則屬於「解法不正確或不合理」，所以給予1分。學生在脈絡化中若提出總數不合或最大顆價錢不合等理由判斷老闆不合理則屬於「具有正確且完整的解釋」，所以給予3分；若認為總售價錯了或珍珠價錢錯了則屬於「片面地檢驗或解釋解的意義」；若認為沒有那麼多顆珍珠或應該加入鍊子的錢則屬於「不合理地檢驗或解釋其解的意義」，所以給予1分。

就評分的結果，除了分析學生在引模、探模和展模問題得分的相關性之外，也檢驗以此評量規準評量學生數學建模表現之各變

異來源的變異程度及其概化係數和可靠度係數。最後，以共變數分析檢驗實驗組和對照組在探模和展模的表現是否具有顯著差異。實驗組與對照組以不同的實驗處理(在解決探模問題時，分別提供與未提供評量規準)為自變項，兩組的「引模」問卷得分為共變量，「探模」和「展模」得分作為依變項，進行單因子共變數分析。如果以探模表現作為依變量時，發現實驗組的表現顯著優於對照組的表現，則表示在解決建模問題時提供評量規準是有助於提升學生當下的建模表現。如果以展模表現作為依變量時，發現實驗組的表現顯著優於對照組的表現，則表示曾經在解決建模問題時提供評量規準也是有助於提升學生之後的建模表現。也就是說，本研究只就提供規準是否對學生的表現所有影響來檢驗此工具的有效性，並非檢驗此影響來自於學習或教學。

肆、研究結果

一、本研究所發展之評量規準的特性

表3呈現三位評分者在24位學生回答引模、探模和展模問題中，所給各評分面向的分數平均數與標準差。在三個問題的各個面向中，分別以配對t檢定檢驗兩兩評分者的評分結果是否有差異，結果顯示只有在引模問題的現實界中，B和C評分者之評分結果有顯著差異(t = -2.145, df = 23, p = .043)；以及在展模問題的數學界中，C評分者和A (或B)評分者之評分結果有顯著差異(t = -2.563, df = 23, p = .017)。而且，每位評分者在各評

表3：三位評分者在各評分向度的平均數與標準差

問題	評分者		評分向度			
			現實界	將現實界數學化	數學界	將數學界脈絡化
引模	A	平均值	1.88	1.25	1.63	1.42
		標準差	0.74	0.90	0.92	1.06
	B	平均值	1.83	1.33	1.54	1.42
		標準差	0.70	1.01	0.88	1.02
	C	平均值	2.00	1.13	1.54	1.33
		標準差	0.78	0.95	0.88	1.05
探模	A	平均值	1.75	1.17	1.42	1.29
		標準差	0.79	0.64	0.58	0.62
	B	平均值	1.75	1.25	1.46	1.21
		標準差	0.68	0.68	0.59	0.59
	C	平均值	1.75	1.13	1.46	1.29
		標準差	0.74	0.74	0.59	0.62
展模	A	平均值	1.00	1.42	1.17	1.00
		標準差	0.93	1.06	0.96	0.66
	B	平均值	0.96	1.58	1.17	1.08
		標準差	0.86	1.18	0.92	0.65
	C	平均值	0.96	1.38	1.50	1.21
		標準差	0.91	1.01	1.14	0.88

註：n = 24。

分向度所呈現的高低起伏，並未出現特殊的評分趨勢(個人偏誤)。

就各個評分向度而言，表4是引模、探模和展模問題得分間的Spearman等級相關。結果顯示，學生在數學界評分向度的三種問題得分間的相關性都呈顯著相關。引模和探模都是從珍珠顆數和價錢衍生的問題情境，所以在現實界向度的引模和探模問題有顯著相關性；但是，引模和展模、探模和展模之間則無顯著相關。除了數學界評分向度外，各問題之間都至少有一個面向是沒有顯著相關，此符合本研究微觀分析解決此三題所需的能力是有所不同。

表5呈現就各個問題看來，每位學生的得分取三位評分者的平均後算出評分向度間的Spearman相關性。就引模問題而言；評分向度間的等級相關分析結果顯示，現實界和其他向度的相關性都最低，介於 .080 ~ .315，均無顯著相關；其他三個向度間都有顯著相關，介於 .666 ~ .904。在探模問題上，現實界和數學界、將數學界脈絡化兩向度的相關性最低，分別是 .117和 .111，均無顯著相關；其他向度間都有顯著相關，介於 .427 ~ .774。在展模問題上，現實界和其他向度的相關性都最低，介於 .162 ~ .302，均無顯著相關；其他三個向度間都有顯著相關，介於 .470 ~ .873。再由表3可知，現實界的得分不見得都是各向度間最高的。如此可以推論；並非學生只是從題目照抄一些資訊就可得分而其餘向度表現仍舊不佳，才導致此向度和其餘向度間的相關性最低。

就各個問題再來分析評分者間一致性時，使用的方法包括完全一致性比例(ratio of exact agreement)和Cronbach's α係數，依引模、探模和展模問題的分析結果分別列於表6、7和8。就三個問題的四個評分向度而言，有94.4%的百分比一致性比例都在17/24以上，最高可達24/24，最低也有15/24，都各出現一次。此結果顯示，評量規準是可以有效地讓不同的評分者對學生數學建模表現給予一致的評量結果。本研究在評分者訓練過程中，除了共同討論此評量規準的優缺點外，在評量前也先對各建模問題討論各面向得最高分的表現為何。如果實際上需要分配不同的受測學生給不同的評分者評分時，這樣的一致性比例仍需提高。本研究建議在評

表4：各評分向度上三種問題間的相關性

向度	引模和探模	引模和展模	探模和展模
現實界	.500*	.260	.049
將現實界數學化	.385	.639*	.508*
數學界	.426*	.683*	.573*
將數學界脈絡化	.456*	.262	.372

*相關的顯著水準為 .050 (雙尾)。

表5：各題的評分向度間之等級相關

問題	R & Ming	R & M	R & Cing	Ming & M	Ming & Cing	M & Cing
引模	.315	.137	.080	.783*	.666*	.904*
探模	.427*	.117	.111	.774*	.630*	.633*
展模	.167	.162	.302	.873*	.470*	.698*

註：R表示現實界、Ming表示將現實界數學化、M表示數學界、Cing表示將數學界脈絡化。
*相關的顯著水準為 .050 (雙尾)。

表6：在引模問題各評分向度中評分者間的一致性

一致性	現實界	將現實界數學化	數學界	將數學界脈絡化
完全一致性比例				
A & B	19/24	18/24	20/24	22/24
A & C	19/24	19/24	20/24	18/24
B & C	20/24	17/24	22/24	16/24
α係數				
A & B	.885	.925	.946	.979
A & C	.905	.937	.946	.940
B & C	.930	.927	.971	.913
A & B & C	.936	.952	.969	.963

註：$n = 24$。

表7：在探模問題各評分向度中評分者間的一致性

一致性	現實界	將現實界數學化	數學界	將數學界脈絡化
完全一致性比例				
A & B	20/24	20/24	23/24	22/24
A & C	22/24	17/24	23/24	20/24
B & C	22/24	19/24	24/24	22/24
α係數				
A & B	.913	.893	.969	.943
A & C	.962	.812	.969	.874
B & C	.955	.889	1.000	.943
A & B & C	.962	.906	.986	.945

註：$n = 24$。

表8：在展模問題各評分向度中評分者間的一致性

一致性	現實界	將現實界數學化	數學界	將數學界脈絡化
完全一致性比例				
A & B	23/24	17/24	19/24	20/24
A & C	18/24	17/24	18/24	17/24
B & C	19/24	17/24	15/24	18/24
α係數				
A & B	.987	.912	.891	.893
A & C	.870	.924	.900	.881
B & C	.875	.943	.896	.816
A & B & C	.940	.950	.928	.903

註：$n = 24$。

分者訓練的過程中，除了討論各面向最高分的表現外，也討論1分和2分的表現範例。如此，即可提升不同評分者使用本評量規準的一致性。當以Cronbach's α係數作為評分者間一致性的另一個指標時，就三個問題的四個評分向度而言，所有的一致性都大於.800，具有良好的一致性(Nunnally, 1978)。

接著，針對引模、探模和展模問題採用學生能力(p)、評分向度(d)和評分者的解讀(r)之雙層面的交叉設計。評分向度包含「現實界」、「將現實界數學化」、「數學界」和「將數學界脈絡化」等四面向，且評分者有三位。由於評量對象或是評分者在重作實驗時是可變的，屬於一種具有不確定性的因素；評分向度是從數學建模歷程所衍生的基本元素，重作實驗時是不變的，屬於一種確定性的因素。所以，將學生與評分者面向視為隨機因素，評分向度則視為固定因素。依混合模式以Restricted MLE估計各變異來源的變異數，分析結果呈現於表9。

無論是引模、探模還是展模問題，造成評分變異的主要來源皆為p × d的交互作用，其變異性分別佔總變異的47.6%、56.9%和45.6%；此結果顯示不同學生所善長的建模歷程面向可能有所差異。其次是p面向，分別佔總變異的40.9%、30.7%和36.4%。與評分者相關的變異來源之變異數成份佔總變異都近似0.0%，最多也只有1.1%；此結果顯示由評分者造成評分結果的不一致性，相較之下是可忽略的。

Shavelson與Webb (1991)建議固定效果的變異量大時，宜就固定效果的各層次(即各評分向度)分開估計變異成份。所以，本研究將依各評分向度重新計算學生和評分者所產生的變異成份，分析結果見表10。無論是哪一個評分向度，造成評分變異的主要來源皆為學生能力，其變異性分別佔總變異的76.4%至96.0%之間；此結果顯示分數變異性主要來自於該群學生的數學建模表現之間有所差異。評分者的變異數成份佔總變異都接近0.0%，最多也只有2.8%；此結果也再次顯示，由評分者造成評分結果的不一致性相較之下是可忽略的。所以，不同評分者可使用此評分規準而給出相近的評分結果。

由表11呈現決策性研究的概化係數得知，只有一位評分者時，各評分向度與建模問題的概化係數介於 .76至 .96之間。增加為兩個評分者時，概化係數都已提高至 .86以上。所以，增加評分者人數至兩人即可有效提升概化係數。由於評分者之間的變異很小，所以各評分向度與建模問題的可靠度係數和概化係數非常接近，只有一位評分者時，可靠度係數介於 .75至 .96之間。增加為兩個評分者時，可靠度係數都已提高至 .86

表9：建模歷程向度評分之變異數成分估計

問題	引模問題		探模問題		展模問題	
變異來源*	變異數成份	百分比(%)	變異數成份	百分比(%)	變異數成份	百分比(%)
學生(p)	0.362	40.9	0.145	30.7	0.332	36.4
評分者(r)	0.000	0.0	0.000	0.0	0.000	0.0
p × d	0.421	47.6	0.269	56.9	0.416	45.6
r × d	0.003	0.3	0.000	0.0	0.010	1.1
p × r	0.001	0.1	0.000	0.0	0.008	0.9
殘差	0.098	11.1	0.059	12.5	0.146	16.0

*評分面向(d)為固定因子。

表10：各建模歷程向度評分之變異數成分估計

問題	引模問題		探模問題		展模問題	
變異來源*	變異數成份	百分比(%)	變異數成份	百分比(%)	變異數成份	百分比(%)
情境向度						
學生(p)	0.457	82.4	0.486	89.7	0.682	84.5
評分者(r)	0.004	0.7	0.000	0.0	0.000	0.0
殘差	0.094	16.9	0.056	10.3	0.125	15.5
數學化向度						
學生(p)	0.786	86.3	0.359	76.4	1.017	85.9
評分者(r)	0.006	0.7	0.000	0.0	0.005	0.5
殘差	0.119	13.1	0.111	23.6	0.161	13.6
數學向度						
學生(p)	0.735	91.4	0.330	96.0	0.831	78.9
評分者(r)	0.000	0.0	0.000	0.0	0.029	2.8
殘差	0.069	8.6	0.014	4.0	0.193	18.3
脈絡化向度						
學生(p)	0.974	89.8	0.319	85.2	0.414	74.9
評分者(r)	0.000	0.0	0.000	0.0	0.005	1.0
殘差	0.111	10.2	0.056	14.8	0.133	24.1

表11：各建模歷程向度評分之決策性研究

問題	引模問題			探模問題			展模問題		
評分人數	1	2	3	1	2	3	1	2	3
現實界									
概化係數	.83	.91	.94	.90	.95	.96	.85	.92	.94
可靠性係數	.82	.90	.93	.90	.95	.96	.85	.92	.94
將現實界數學化									
概化係數	.87	.93	.95	.76	.87	.91	.86	.93	.95
可靠性係數	.86	.93	.95	.76	.87	.91	.86	.92	.95
數學界									
概化係數	.91	.95	.97	.96	.98	.99	.81	.90	.93
可靠性係數	.91	.95	.97	.96	.98	.99	.79	.88	.92
將數學界脈絡化									
概化係數	.90	.95	.96	.85	.92	.95	.76	.86	.90
可靠性係數	.90	.95	.96	.85	.92	.95	.75	.86	.90

以上。所以，增加評分者人數至兩人即可有效提升可靠度係數。由上述的分析可知，此評量規準是一個具有良好信度的評量工具，可用來評量學生的數學建模表現。

二、提供評量規準對數學建模表現的影響

在「引模」、「探模」和「展模」問卷的表現，對照組35人平均總得分為4.26、2.40和1.91分，實驗組34人平均總得分為4.53、3.73和2.79分。由獨立樣本*t*檢定的結果顯示，該兩組在引模問卷表現沒有顯著差異($t = 0.340$, $p = .735$)。接下來以「引模」得分作為共變量，分別以「探模」和「展模」得分作為依變項，進行組內迴歸係數同質性檢定考驗結果，組別*「引模」均符合組內迴歸係數同質性的檢定(探模：$F = 0.059$，$p = .445$；展模：$F = 0.069$，$p = .793$)，得以繼續進行共變數分析(ANCOVA)。在探模問題上排除共變量對依變項的影響後，來自實驗處理所造成的變異仍達到顯著水準，如表10所示。在展模問題上排除共變量對依變項的影響後，來自實驗處理所造成的變異則未達到顯著水準，如表12所示。也就是說，在解決建模問題當下提供評量規準時，可有效提升學生的建模表現；但是只提供一次評量規準時，並不能對往後解決建模問題產生有效的影響。

伍、結論與建議

本研究所發展的建模評量規準，包含現實界、將現實界數學化、數學界和將數學界脈絡化四個向度，各向度依其完備性又分成四個等級。往後的研究或教學若要以實作評量的方式評量數學建模能力，即可針對其使用的建模問題參考表2分析學生的表現並進行評分。本研究以此評量規準針對三個不同的建模問題進行評分者間的一致性與概化分析。結果顯示：無論是引模、探模和展模問題，評分者在四個評分向度給分的一致性都不錯，而且由其所造成的變異數成份之比例也很小。本評量規準的優點包含：具備數學建模的形式與內容，可用來評量各種建模問題，也可同時回饋學生在數學建模各面向的表現與整體的表現。但是，此評量規準仍有待改進的部分。例如：在Stiggins (1994)建議的第一個步驟：列出重要特質或要素中，仍要思考是否以及如何將文字表達能力和建模能力加以區分，在其建議的第六個步驟：以較一致的語言表達各層面的品質中，仍需進一步克服如何同時兼顧語言表達的一致性以及各個建模問題的特殊性，使其具有最一般化的特殊評量規準。

概化分析也顯示單一建模問題中，在24位學生且一位評分者的情況下，概化係數和可靠度係數皆能達至 .75的水準。這表示在一般的課室中，只要個別數學教師利用此評量規準與建模問題，即可進行有效的實作評量。當評分者增加為兩位時，即可將概化係數和可靠度係數皆能提升至.86的水準。這也意味著，此實作評量也可實際用來甄選在數學建模有特殊表現的人才。而評分向度

表12：兩組在「探模」和「展模」之共變數分析摘要表

依變項		探模		展模	
變異來源	自由度	*F*	Sig.	*F*	Sig.
引模	1	42.667	.000	24.151	.000
組別	1	5.843	.018	1.606	.210

與學生的交互作用佔分數變異的比例相較於評分者是較高的，顯示學生在建模歷程面向之能力各有不同。此現象也提醒教學者，每位學生優劣的面向會有所不同，可針對不同類型的學生再來設計教材與介入的時機。此外，學生在本研究所設計的引模和探模問題的現實界、數學界和將數學界脈絡化等三個評分向度的表現都有顯著相關，顯示本研究所設計的引模問題和探模問題過於相似。在如此相似的問題中，研究發現提供學生評量規準仍有助於其建模表現優於未有評量規準的情況。如果是在不相似的引模和建模問題中，其影響是否存在或是影響的程度是否不同仍有待進一步的探討。

在學生尚未參與建構本評量規準的情況下，提供實驗組評量規準的學生數學建模表現仍顯著優於未提供本評量規準的對照組學生。這顯示學生能夠從此建模規準擷取資訊以協助個人回答問題。為什麼學生有此評量規準可以參考時有利於提升其數學建模的表現呢？有可能是提供學生「有條理地說明分數會較高」，或許有利於激發學生的後設認知以監控其作答結果，並以能得高分的評量規準作為作答目標，而產生自我調整。除了後設認知的因素外，此評量規準是否對其認知上也有所影響或有何影響仍有待進一步的研究。但是此資訊的角色可能是協助目標設定或是形成策略仍未有定論，需要進一步地探討此評量規準如何影響學生的建模歷程，或是讓學生再參與評量規準的修改以提升其學習動機和自我調整能力(Stiggins, 2005)。另外，本研究尚未探討教師使用建模規準的方法及其可扮演的角色。

在教學中，我們可利用此評量規準追蹤記錄學生在不同建模問題之各種評分向度的表現。但是，如何從學生長期的數學建模表現資料推論其建模能力的發展仍有待進一步的探討。依據本研究的評量規準之特性，學生的表現會因不同建模問題和不同的評分向度而有不同層次的表現。也就是說，在引模問題的數學面得分較高的學生不見得在探模或展模問題的數學面得分也較高；或是在展模問題的現實面得分較高的學生不見得在展模問題的其他評分向度得分也較高。此特性顯示數學建模能力是相當複雜的構念，因此如何利用此評量規準來有效預測學生數學建模能力的發展仍需要克服許多挑戰。例如：探討由不同數學建模問題對學生數學建模能力的影響。

效化評量工具所涵蓋的範圍很廣，因此發展過程中很難顧全所有的效度問題，比較可行的是針對評量目的考量主要的效度問題(Shepard, 1993)。本研究的目的是要建構一個往後可在課室中用來評量數學建模表現的評量規準，而考量了思考歷程、概化和結果的觀點(Miller & Linn, 2000)來效化評量規準，即依據理論上數學建模歷程、學生建模認知分析與專家討論的過程檢驗評量規準中四個評分向度的效度，以概化理論檢驗評分者效果以及評分面向和學生交互效果的變異量，以評量規準作為介入工具檢驗對學生建模表現的影響。如果要擴展內容範圍和增加內容效度的話，則需要設計設計不同內容的建模問題，以及更多相似內容的建模問題。但是何謂相似內容也是需要進一步探討的議題，是所需的數學模式相似，還是建模情境相似，或是兩者兼顧才行。基於足夠多具有內容效度的建模問題下，也才能繼續探討問題和學生交互效果的變異量以及不同問題數的概化係數與可靠度係數。

往後的研究，也可依此評量規準設計適合中學或大學學生的評量單，接著給一個班級或是透過幾個個案試作該評量單來蒐集學生的想法。除了檢驗本評量規準的一般性

外，也可依照評量規準來分析學生建模表現的「優」與「劣」。以學生的「優」設計「引模」的學習單，以學生的「劣」設計「探模」的學習單，參考學生所提的相關問題設計「展模」學習單。如此發展一系列的教學實驗，檢驗依此設計的教學之學習效果，即可擴展本規準的實用性。

誌謝

感謝參與學生與老師的支持與配合，以及協助評分的人員以及協助資料處理的劉欣怡小姐，同時感謝審查委員的細心指正與批評使得本文更臻完備。本研究承蒙國家科學委員會的補助(NSC 95-2521-S-018-003-MY3)得以順利完成此論文，特此致謝。但是，文中論點為作者所有不代表國家科學委員會。

參考文獻

1. 李坤崇(2006)。**教學評量**。臺北市：心理。
2. 教育部(2003)。**國民中小學九年一貫課程之問題與檢討**。臺北市：作者。
3. 楊凱琳、林福來(2006)。探討高中數學教學融入建模活動的支撐策略及促進參與教師反思的潛在機制。**科學教育學刊**，**14**(5)，517-543。
4. 楊凱琳、林福來、蕭志如(2008，12月)。**概念化數學建模表現的評量規準**。發表於中華民國第二十四屆科學教育學術研討會。彰化市：國立彰化師範大學。
5. 楊凱琳、楊子錕(2008，12月)。**在數學建模問題脈絡下探討個案學生的思維特徵**。發表於中華民國第二十四屆科學教育學術研討會。彰化市：國立彰化師範大學。
6. Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling*, *53*(2), 192-211.
7. Baron, J. B. (1991). Performance assessment: Blurring the edges of assessment, curriculum, and instruction. In G. Kulm & S. M. Malcom (Eds.), *Science assessment in the service of reform* (pp. 247-266). Washington, DC: American Association for the Advancement of Science.
8. Bell, A., Burkhardt, H., & Swan, M. (1992). Balanced assessment of mathematical performance. In R. Lesh & S. J. Lamon (Eds.), *Assessment of authentic performance in school mathematics* (pp. 119-144). Washington, DC: American Association for the Advancement of Science.
9. Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan Online*, *80*(2), 139-148.
10. Blum, W. (1993). Mathematical modelling in mathematics education and instruction. In T. Breiteig, I. Huntley, & G. Kaiser-Messmer, (Eds.), *Teaching and learning mathematics in context* (pp. 3-14). London: Ellis Horwood.
11. Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
12. Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessment. *Educational Measurement: Issues and Practices*, *14*(4), 9-12.
13. Brookhart, S. M. (1997). Effects of the classroom assessment environment on mathematics and science achievement. *The Journal of Educational Research*, *90*(6),

323-330.

14. De Lange, J. (1996). Using and applying mathematics in education. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 49-97). Boston, MA: Kluwer Academic.
15. Dickinson, W. B., Onwuegbuzie, A. J., Hines, C., & Hall, B. W. (2007). *Summative assessment strategies for statistical learning: Development, administration, and scoring of authentic and performance assessments*. Retrieved July 14, 2012, from http://www.stat.auckland.ac.nz/~iase/publications/sat07/Dickinson_et_al.pdf
16. Doerr, H. M., & English, L. D. (2003). A Modeling perspective on students' mathematical reasoning about data. *Journal of Research in Mathematics Education*, *34*(2), 110-136.
17. Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.
18. Ellerton, N. F., & Clarkson, P. C. (1996). Language factors in mathematics teaching and learning. In A. J. Bishop et al. (Eds.), *International handbook of mathematics education* (pp. 987-1033). Dordrecht, The Netherlands: Kluwer Academic.
19. English, L. D. (2006). Mathematical modeling in the primary school: Children's construction of a consumer guide. *Educational Studies in Mathematics*, *63*(3), 303-323.
20. Ferri, R. B. (2006). Theoretical and empirical differentiations of phases in the modelling process. *Zentralblatt für Didaktik der Mathematik*, *38*(2), 86-95.
21. Freudenthal, H. (1991). *Revisiting mathematics education*. Dordrecht, The Netherlands: Kluwer Academic.
22. Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., & Katzaroff, M. (1999). Mathematics performance assessment in the classroom: Effects on teacher planning and student problem solving. *American Educational Research Journal*, *36*(3), 609-646.
23. Galbraith, P., & Stillman, G. (2006). A framework for identifying student blockages during transitions in the modelling process. *Zentralblatt für Didaktik der Mathematik*, *38*(2), 143-162.
24. Glickman-Bond, J., & Rose, K. (2006). *Creating and using rubrics in today's classrooms*. Norwood, MA: Christopher-Gordon.
25. Haines, C., & Crouch, R. (2005). Applying mathematics: Making multiple-choice questions work. *Teaching Mathematics and its Applications*, *24*(2-3), 107-113.
26. Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247.
27. Hjalmarson, M. A. (2004). *Designing presentation tools: A window into mathematics teacher practice*. Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.
28. Kilfeather, P., O'Leary, M., & Varley, J. (2006). Adapting science performance tasks developed in different countries for use in Irish primary schools. *Irish Educational Studies*, *25*(1), 3-33.

29. Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, *63*(3), 18-24.
30. Lesh, R., & Doerr, H. M. (2003). *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching*. Mahwah, NJ: Lawrence Erlbaum Associates.
31. Lesh, R., & Zawojewski, J. S. (2007). Problem solving and modeling. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 763-804). Greenwich, CT: Information Age.
32. Liang, X., & Creasy, K. (2004). *Classroom assessment in web-based instructional environment: Instructors' experience*. Retrieved July 14, 2011, from http://www.pareonline.net/getvn.asp?v=9&n=7
33. Lin, F. L., & Yang, K. L. (2005). Distinctive characteristics of mathematical thinking in non-modelling friendly environment. *Teaching Mathematics and Its Applications*, *24*(2-3), 97-106.
34. Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, NJ: Pearson Education.
35. Mason, J., & Davis, J. (1991). *Modelling with mathematics in primary and secondary schools*. Geelong, Australia: Deakin University Press.
36. Mason, J., & Johnston-Wilder, S. (2004). *Fundamental constructs in mathematics education*. London: Routledge Falmer.
37. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.
38. Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington DC: National Center for Education Statistics.
39. Meyer, C. A. (1992). What's the difference between authentic and performance assessment? *Educational Leadership*, *49*(8), 39-40.
40. Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24*(4), 367-378.
41. Mousoulides, N., Sriraman, B., & Christou, C. (2007). From problem solving to modelling: The emergence of models and modelling perspectives. *Nordic Studies in Mathematics Education*, *12*(1), 23-47.
42. Niss, M. (2001). Issues and problems of research on the teaching and learning of applications and modelling. In J. F. Matos, S. K. Houston, W. Blum, & S. P. Carreira (Eds.), *Modelling and mathematics education: ICTMA 9: Applications in science and technology* (pp. 72-89). Chichester, UK: Horwood.
43. Niss, M. (2004). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagtsis & S. Papastavridis (Eds.), *3rd mediterranean conference on mathematical education* (pp. 115-124). Athens, Greece: The Hellenic Mathematical Society.
44. Nunnally, J. C. (1978). *Psychometric theo-*

ry. New York: McGraw-Hill.

45. Organization for Economic Co-operation and Development. (2006). *Assessing scientific, reading, and mathematical literacy: A framework for PISA 2006*. Paris, France: Author.
46. Reddy, V. (2007). *Mathematics and science achievement at South African schools in TIMSS 2003*. Cape Town, South Africa: Human Sciences Research Council Press.
47. Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston, MA: Kluwer Academic.
48. Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 334-370). New York: Macmillan.
49. Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
50. Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.
51. Sriraman, B., & Lesh, R. (2006). Modeling conceptions revisited. *Zentralblatt für Didaktik der Mathematik*, *38*(3), 247-254.
52. Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York: Merrill.
53. Stiggins, R. J. (2005). *Student-involved assessment for learning* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
54. Stillman, G., Brown, J., & Galbraith, P. (2010). Researching applications and mathematical modeling in mathematics learning and teaching. *Mathematics Education Research Journal*, *22*(2), 1-6.
55. Sun, A., Valiga, M. J., & Gao, X. (1997). Using generalizability theory to assess the reliability of student ratings of academic advising. *The Journal of Experimental Education*, *65*(4), 367-379.
56. Trevisan, M. S., Davis, D. C., Calkins, D. E., & Gentili, K. L. (1999). Designing sound scoring criteria for assessing student performance. *Journal of Engineering Education*, *88*(1), 79-85.
57. Turner, R. (2007). Modelling and applications in PISA. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education: The 14th ICMI study* (pp. 433-440). New York: Springer.
58. Van Leeuwen, D. M. (1997). Assessing reliability with generalizability theory: An application to inter-rater reliability. *Journal of Agricultural Education*, *38*(3), 36-42.
59. Zbiek, R. M., & Conner, A. (2006). Beyond motivation: Exploring mathematical modeling as a context for deepening students' understandings of curricular mathematics. *Educational Studies in Mathematics*, *63*(1), 89-112.

科學教育學刊
2012, 第二十卷第四期, 319-342

Chinese Journal of Science Education
2012, 20(4), 319-342

Study on a Rubric for Assessing Mathematical Modeling Performance

Kai-Lin Yang[1,*], Fou-Lai Lin[1] and Chih-Ru Hsiao[2]

[1]Department of Mathematics, National Taiwan Normal University
[2]Department of Mathematics, Soochow University

Abstract

Mathematical modeling problems are problems with real or realistic situations. Different problem solvers may use diverse methods to solve the same modeling problems, and thus we assumed that modeling problems are the type of tasks that can elicit students' active learning. Moreover, taking into account that providing appropriate assessment is helpful to facilitate students' learning of mathematics, this study aimed at constructing a rubric for assessing mathematical modeling performance based on generalizability theory and performance assessment. In order to construct a complete rubric which could be used to assess the strength and weakness of varied ideas for solving modeling problems, we firstly referred to the elements of modeling process in literature. Next, we modified the elements and their content according to a mathematics modeling expert and mathematics teachers' opinions. After formulating the rubric, its generalizability and dependability was evaluated by the generalizability theory, and a quasi-experimental study was implemented to assess its effect on students' performance in mathematical modeling. The results show that the agreement among the raters was good and both the generalizability coefficient and the dependability index with two raters were above .86 for each rating facet of each modeling task. Moreover, students who were provided this rubric performed better on the modeling task than students who were not provided this rubric.

Key words: Generalizability, Performance Assessment, Mathematical Modeling

* Corresponding author: Kai-Lin Yang

05-楊凱琳.indd 342 2012/9/5 上午 11:26:30