科學教育學刊
2008, 第十六卷第一期, 25-51

Chinese Journal of Science Education
2008, 16(1), 25-51

數理科在職教師評量素養培育模式之探究

王子華[1]　**范雅晴**[2]　**王國華**[2]

[1]國立新竹教育大學　教育學系

[2]國立彰化師範大學　科學教育研究所

（投稿日期：民國 96 年 3 月 27 日，修訂日期：96 年 12 月 17 日，接受日期：96 年 12 月 18 日）

摘要：教師評量素養會影響教學的成敗，許多針對「在職教師」或是「修過評量課程的在職教師」之評量素養的調查研究，都發現有缺陷。本研究的主要目的，是希望採用一個具備「個人化」與「情境化」特性的 e-Learning 型態之教師評量素養培育模式-「P2R-WATA（Practicing, Reflecting and Revising with Web-based Assessment and Test Analysis system）」，讓參與研究的在職教師，可以在 WATA 系統的 Triple-A Model（Assembling, Administering and Appraising）中進行個別化與情境化的評量知能學習。本研究採用準實驗設計，將參與研究的在職教師分為三組，分別為實驗組 A（「P2R-WATA」評量素養培育模式搭配面授活動，18 人）、實驗組 B（「P2R-WATA」評量素養培育模式，15 人）與對照組（「P2R」評量素養培育模式，16 人），研究發現，實驗組 A 與實驗組 B 的在職教師，在評量知識的改善情形上，均顯著優於對照組之在職教師，另外，也發現實驗組 A 的在職教師，在評量觀點的改善情形，顯著優於對照組之在職教師。

關鍵詞：WATA 系統、「P2R-WATA」評量素養培育模式、教師評量素養、師資培育、數位學習

壹、研究動機

評量與教學之間有著相當密切的關係，在一些研究中指出，評量在教師的整個教學活動中，可以佔用到 50% 以上的時間（Plake,1993; Stiggins, 1991a），基於此，許多研究者針對評量與學生學習效益之間的關係進行探究，均發現良好的評量可以提升學生學習效益（Brookhart, 1999; Mertler, 2004; Stiggins, 1999a），而評量除了對學生很重要外，對於教師也很重要，美國 National Science Education Standards [NSES]指出，「評量對於教師而言是一個重要的工具，可以用來進行好的教學探究。一個具有良好評量能力之教師，也可以藉由評量的技術來理解學生的想法、信念與推理（NRC, 1996, p. 63）」，由此可見，評量對於教師教學與學生學習而言，

均具有重要的意義。

自1990年代起至今，許多針對「在職教師」或是「修過評量課程的職前教師」所進行的評量素養研究均發現，教師的評量素養均不足夠，換言之，教師並沒有具備在課室中進行良好評量的能力（Arter, 2001; Brookhart, 2001; Gullickson, 1993; Marso & Pigge, 1993; Mertler, 2000, 2004; Plake, Impara, & Fager, 1993; Popham, 2006; Schafer, 1991, 1993; Stiggins, 1991a,b; Stiggins, 2001a），這問題由1990年代至今一直存在，國內，雖然沒有如同美國一樣進行大規模的調查(如: Plake et al., 1993)，但是一些區域性的調查，也反映出類似的問題（如: 王子華, 2005; 王子華、王國華、王瑋龍和黃世傑, 2004; Wang, Wang, Wang, Huang, & Chen, 2004; Wang, Wang, & Huang, in press)。由文獻中可知，在美國造成這種教師缺乏評量素養的因素，主要有兩個：「師資培育過程中與評量相關之訓練課程的缺失（Brookhart, 2001; Lukin, Bandalos, Eckhout, & Mickelson, 2004; Maclellan, 2004; McMillan, 2000, 2001; Mertler, 2004; Stiggins, 1991b, 1999a, b; Stiggins, 2004)」與「師資認證制度不重視教師評量能力或相關訓練（Stiggins, 1999b, 2004)」。王子華（2005）指出，上述兩問題也存在於臺灣的師資培育制度中，無論「師資培育課程」或「教師認證」方面，均不強調教師的評量能力：「師資培育課程」方面，其中的評量相關課程約只有四學分，且均屬於選修課，「教師認證」方面，教師評量能力的考核並沒有獨立出來，而是與其他科目包含在一個教育專業科目中，由此觀之，在臺灣亦可能存在與美國一樣輕忽教師評量素養的問題，以美國的經驗來看，臺灣教師的評量素養可能也有待改善（王子華, 2005）。

本研究之目的主要是針對「師資培育過程中與評量相關的訓練課程之缺失」這個面向提出改善策略，並主要針對在職教師在傳統課室評量方面的評量素養進行研究。王子華（2005）與 Wang 等人（in press）針對生物科職前教師建構了 e-Learning 型態之教師評量素養培育模式-「P2R-WATA（Practicing, Reflecting and Revising with Web-based Assessment and Test analysis system)」評量素養培育模式（圖1），研究發現，P2R-WATA模式對於生物科職前教師評量素養的提升有相當良好的效益，本研究將此模式延伸應用於數理科在職教師評量素養之培育，深入探究「P2R-WATA」評量素養培育模式對於數理科在職教師之評量素養培育的效益情形，本研究共有二個待答問題：

一、「P2R-WATA」評量素養培育模式對於改善數理科在職教師評量知識之效益為何？

二、「P2R-WATA」評量素養培育模式對於改善數理科在職教師評量觀點之效益為何？

貳、文獻探討

一、教師評量素養

Stiggins（1991b）在其所撰述的論文-As-

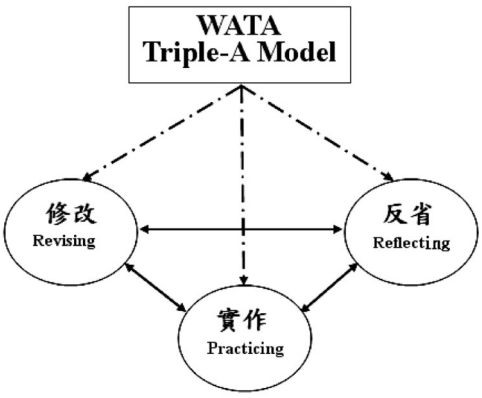


圖 1：「P2R-WATA」評量素養培育模式

sessment Literacy 中指出，要瞭解評量素養的定義，應該由不具有評量素養者（assessment illiterate）的特性瞭解起，Stiggins 指出，不具備有評量素養者「無法理解要採取哪些行動才可以獲得高品質的成績資料，因此無法以批判的角度評鑑自己所使用的資料，此外他們也接受成績資料的表面價值，而且容易被表面上較技術性的訊息或是複雜的考試成績的呈現所驚嚇（p. 535）」，相反的，一個具備評量素養者則應該有以下特徵：「知道好的評量與不好的評量的差異，而且不會被一些有時候看來神秘但總是讓人望之怯步的評量技術所驚嚇…並且知道自己在評量什麼？為什麼要這樣做？如何評量重要的知識與技能才是最好的？如何產生學生表現的好範例？什麼會使得評量出錯，以及如何避免錯誤的發生？…也會警覺到不精確的評量所潛在之負面後果…（Stiggins, 1995, p.240）」。除 Stiggins（1991b, 1995）外，亦有一些學術單位與研究者對於評量素養做了一些詮釋，如：「教育者準備好可以設計、執行與討論評量策略（North Central Regional Educational Laboratory, 2006）」、「與有效實施學生評量有關的知識（knowledge）與理解（understanding）（Phillips, 2003）」、「擁有執行良好評量的基本規則之相關知識，包含：術語、評量方法、技巧的使用與發展，並且對評量的品質標準熟悉，以及熟悉評量學習的另有與傳統方式（Mertler, 2004, p. 50-51）」等。除上述外，王子華（2005）將評量素養界定為包含：「評量知識」、「評量技巧」與「評量觀點」三向度，王子華進一步指出，該評量素養的三個向度，與「實作命題品質」均有顯著正相關（評量知識: $r = 0.265, p < 0.05$; 評量技巧: $r = 0.316, p < 0.05$; 評量觀點: $r = 0.283, p < 0.05$）。

本研究將採用王子華（2005）的定義，將評量素養界定為包含：「評量知識」、「評量技巧」與「評量觀點」三向度，另由於「評量技巧」與學科知識有關，然而，參與本研究的數理科在職教師人數較少，而且分屬於數學、物理、化學與生物等不同學科專長的教師，因此，在探究「評量技巧」方面會受到不少限制，因此，本研究將主要針對評量素養的「評量知識」與「評量觀點」兩向度進行探究。

二、教師評量素養的重要性

評量除了是教學活動中相當重要的步驟外（Popham, 2006），評量在教師教學過程中佔用的時間也有越來越多的趨勢，Plake（1993）與 Stiggins（1991a）指出，在未來，教師將花費50% 以上的教學時間進行課室評量，因為，一個好的評量與評分動作對於教師與學生而言，有相當正面的教育意義，除了可以協助教師改善教學與提升學生的學習動機外，更可以提升學生的學習效益（Brookhart, 1999; Campbell & Collins, 2007; Mertler, 2004），Stiggins（1999a）指出，改善教師課室評量的品質，可以提升學生在大範圍標準化測驗的成績表現約3/4 SD，綜合上述可知，一個好的教學一定要具備好的評量（Campbell & Collins, 2007; Stiggins, 1999a）。

近幾年，美國政府教育部（U.S. Department of Education）所推行的 No Child Left Behind [NCLB]計畫開始重視學校評量，因為很多證據顯示學校評量與標準化成就測驗成績表現是有關係的（Campbell, Murphy, & Holt, 2002; Mertler, 2004），Campbell 等人（2002）指出，在這樣一個趨勢下，教師開始被要求依據這些符合州標準之課程發展相關的評量能力，以提升其教學效益與學生的學習效益，因此，教師評量專業能力之提升便成為一個重要的議題（Mertler, 2004），同時也顯現出教師評量能力的重要性。除了 NCLB 的推行造

成教師評量能力的提升受到重視外，Lukin等人（2004）也指出，美國許多州都採用州評量來提升學生的學習效益，並藉此確保課程與教材的適當性，但內布拉斯加州（Nebraska State）於2000年啟動的教育改革計畫-Standards-based, Teacher-led Assessment and Reporting System [STARS]中，則改採地區性評量（district assessment）來精確管理課程教材與最佳化學生的學習效益（Bandalos, 2004; Plake, Impara, & Buckendahl, 2004），Lukin 等人更指出內布拉斯加州（Nebraska State）的 STARS 教育改革要成功，其中一個必要的因素就是，依據課程所執行之區域層次的評量要能夠用以改善學校的現況，並肩負起說明解釋的任務；另外一個影響 STARS 教育改革成敗的因素就是課室評量的改善，因為，STARS 是一種去中心化的教育改革模式，這種改革要成功，就是要許多老師願意參與，且教師將承擔更多的學生學習效益責任，在此狀況下，課室評量的品質以及教師的評量素養就是必須要重視的議題，因為課室評量可以協助教師的教學以及學生的學習，也因此在內布拉斯加州的州法律 LB 812就針對於教師評量能力以及培育進行規範（Buckendahl, Impara & Plake, 2004; Lukin et al., 2004; Plake et al., 2004）。

在國內方面，王子華（2005）指出，依據我國政府文獻資料網（http://twinfo.ncl.edu.tw）的政府公報查詢系統，可以發現我國最早可以追溯到西元1987年（76學年度）起，就開始進行「台灣省高級中等學校成績考查命題競賽」，由各年度的實施辦法可以發現，該活動的主要目的在於：「提高教師命題水準，增進評量效果並促進教學正常化，以加強教學成效」，競賽方式分為兩種，一種是「抽選」，另一種是「自由參加」，抽選是由教育廳抽選前一學年度（85學年度比賽就是抽選84學年度）上下學期的各競賽科目的期中考或期末考紙筆測驗試卷來參賽；「自由參加」則是由自行參賽之學校，自己選擇前一學年度上下學期的各競賽科目期中考或期末考紙筆測驗試卷來參賽，其中參賽的試卷必須是真正用來考試的原卷，且不可以是另外命題的試卷或是抄襲坊間的測驗試卷。教育廳會徵集評量方面的學者專家進行評分，經過評選優勝者會收錄於專書中提供各界觀摩，由上述的活動與措施可以發現，在國內也相當重視教師評量能力，並將良好的教師評量能力視為可以使「教學正常化」與「加強教學成效」的重要助力。

王子華（2005）進一步指出，除命題競賽外，依據「國民中小學九年一貫課程修訂綱要」（教育部, 2003）的內容可以發現，國內的教育改革趨向也如同內布拉斯加州的 STARS 教育改革一樣走向去中心化，在教科書的使用方面，揚棄統一教材的模式，改採一綱多本的模式，國民中小學教科用書依據部定課程綱要編輯，並經由審查機關審定通過後，由各校自己決定採用的教材，此外，特別強調「空白課程」與「學校本位課程」的理念，要求各校成立「課程發展委員會」，負責審查學校因應地區特性、學生特質與需求，選擇或自行編輯實施於全校或全年級的教材，這種「空白課程」與「學校本位課程」的理念，如果要能夠確實施行，教師除了需要具備有設計課程與教材的能力外，也必須要具備依據自己所設計的課程與教材，設計與執行評量的能力；在升學制度方面，由於教育部已經廢除聯考制度改採取高中職多元入學制度，雖然有辦理「國民中學基本學力測驗」統一性考試，但是，該多元入學制度中也會參考在校的評量成績，此設計迥異於過去只參考聯考成績，因此，無論由九年一貫課程精神的角度，或由升學的角度來看，

當前教師評量的能力均相當重要。

由上面的文獻分析發現，近十多年來，教學、評量與學習效益之間一直存在著密切的關係，而且在近幾年進行的教育改革計畫中，評量也扮演著輔助教學與學習以及提升效益的工具，但是令人憂心的是，許多研究者均指出，無論是在職教師或職前教師，都沒有足夠的能力執行或在未來執行課室評量，而這問題無論在1990年代（Gullickson, 1993; Marso & Pigge, 1993; Plake et al., 1993; Schafer, 1991, 1993; Stiggins, 1991a, b），或是在十多年後的現在(Arter, 2001; Brookhart, 2001; Mertler, 2000, 2004; Popham, 2006; Stiggins, 2001a）都一直存在著，而台灣也有相同的問題存在（王子華, 2005; 王子華等人, 2004; Wang et al., 2004; Wang et al., in press），也因此，這十多年來，許多研究者一直強調教師評量素養的重要性，並且認為有關教師評量素養的訓練課程應該融入在職教師與職前教師的師資培育課程中（王子華, 2005; Brookhart, 2001; Campbell et al., 2002; Lukin et al., 2004; Maclellan, 2004; McMillan, 2000, 2001; Mertler, 2004; O'Sullivan & Johnson, 1993; Plake, 1993; Wang et al., in press），本研究則將依循上述建議，將應用王子華（2005）與Wang 等人（in press）所發展出來的「P2R-WATA」評量素養培育模式，應用於數理科在職教師之評量素養培育，並深入探究該模式的效益情形。

三、教師評量素養師資培育模式

許多研究者均指出，師資培育過程中的評量訓練對於提升教師的評量素養是有助益的(Airasian, 1991; Arter, 2001; Brookhart, 2001; Johns & VanLeirsburg, 1993; Lukin et al., 2004; Maclellan, 2004; Plake et al., 1993; Plake & Impara, 1993; Schafer, 1991; Stiggins, 1991b），以下將深入介紹國外文獻中所提及之四套教師評量素養培育模式：

(一)Plake 和 Impara（1993）的評量素養訓練模式

Plake 和 Impara（1993）依據 Plake 等人（1993）的全國性調查結果設計評量課程，希望能夠改善教師的評量素養。Plake 和 Impara 所設計的訓練模式內容包含親師短文（parent teacher vignette, 一種 context-orienting vignette）、評量教學課程、實務活動、親身經驗、小組討論問題與已發展模式的回應（developed model responses）與家庭作業問題，其中，「親師短文」的設計主要用途在於營造一個與詮釋、傳達測驗結果有關的問題情境，用以發展一個親師論壇，實施方式為短文撰寫，而非一般的媒體演示方式。

(二)NAC（Lukin et al., 2004）

Nebraska Assessment Cohort [NAC]是首先由 University of Nebraska-Lincoln 於1999年起開始，歷時三年規劃完成，整個課程總共有18個學分，由 University of Nebraska-Lincoln 負責執行，執行過程中藉著對州政府教育部門的遊說，要求州內的教師參與 NAC 計畫。NAC 計畫正式開始於2001年，課程發展的主要目的，是為了提供有經驗、實際進行教學的教師與教育主管人員，進行評量素養訓練的機會，改善其評量素養並使其具備發展與執行課室評量和區域性評量的知識與技能，以符合內布拉斯加州（Nebraska State）之教育改革計畫-STARS（Bandalos, 2004; Plake et al., 2004）的需求，此外，也希望這些受訓完成之教師可以在評量過程中擔任領導者的角色。NAC 課程的18個學分包含連續兩個夏天的六小時課程，以及學年中的六小時實習課程（practicum），這課程設計的目的是希望參與教師可以在獲得評量知識與技能的同時，也可以應用於實務當中。在夏季學期，

由教師教授評量課程，而在春、秋學期，採用網路教學方式提供評量教學，該設計的目的，是希望能夠配合參與教師的學校課表。關於實習課程部分，參與的教師有機會在某一個環境中實作其所學到的知識與技能，而且此環境可以提供回饋與對話。

(三)ALLT（Arter, 2001; Stiggins, 2001b）

美國 Assessment Training Institute 的 Dr. Rick Stiggins 自1999年秋季起，發展了一套評量素養的學習課程-Assessment Literacy Learning Team [ALLT]（Arter, 2001），ALLT 的基本精神是邀請教師與校長等，自願參與小型專業團體，而以定期聚會與分享專業經驗的方式來培育教師評量素養。ALLT 的目標是希望參與者能理解好的與不好的評量活動之差異，堅持優良的評量，並學習使用課室評量來改善教學，並引起學生的學習動機。

(四)PALS 與 IPALS（Lukin et al., 2004）

Pre-service Assessment Learning Study Group [PALS]與 In-service and Pre-service Assessment Learning Study Group [IPALS]是由 ALLT 發展出來的。Lukin 等人（2004）指出，上述的 ALLT 對職前教師的助益不大，因為，原始 ALLT 設計中的參與人員，只有包含在職教育人員，於是針對職前教師發展出 PALS 與 IPALS。PALS 與 IPALS 均採用 ALLT 的設計，且均涵蓋職前教師與在職教師，而 PALS 與 IPALS 的差異主要在於，PALS 中的學習者是職前教師，在職教師的角色為學習團體中，其他成員的引導者或教導者，但是，在 IPALS 中，無論在職或職前教師都是學習者角色。PALS 與 IPALS 設計的共同特色就是，具有「教學實習學期（student teaching semester）」的設計，將職前教師與在職教師搭配在一起，然後實施一些與教室經驗結合的評量素養訓練，而這種真實的教室經驗，對於職前教師而言相當重要，也一直是許多師資培育課程設計中的挑戰（Lukin et al., 2004）。以評量素養訓練的角度來看，如果沒有這種真實的經驗，很難讓職前教師理解，良好的評量在學生學習過程中所扮演的角色，也很難讓職前教師評鑑各種不同的評量策略。

上述四個教師評量素養培育模式的主要特色，除了均採取傳統方式教授評量相關知識與技巧外，另有一個重要的共同特色，就是四種模式都提供將所學到的評量知識與技能，在教室中實作的機會，使課室經驗與評量素養訓練結合在一起，這種設計在許多相關文獻當中都受到肯定（Brookhart, 2001; Lukin et al., 2004; Mertler, 2004; Stiggins, 1999b; Tayler & Nolen, 1996），Tayler 和 Nolen（1996）甚至主張所有的評量概念，都應該在教學的實際情境中，教導給接受評量素養訓練的教師才會有效益，例如：Plake 和 Impara（1993）的模式中，包含有實務活動、親身經驗兩設計；在 NAC 模式中，於春秋兩學期進行的「實習課程」，該課程要求參與者實際將夏季學期所學的評量概念與技能應用於真實課室評量中；ALLT、PALS 與 IPALS 模式三者均類似，也都包含真實教室評量的部分，ALLT 將教育行政人員與教師配對在一起，協同於教室中執行評量，而 PALS 與 IPALS 設計中，亦包含「教學實習學期」，讓參與者可以將學習到的評量知識與技能，實際應用到教學情境中。

相較於國外，國內針對教師評量素養培育模式所進行之研究較少（王子華, 2005），王子華（2005）與 Wang 等人（in press）針對生物科職前教師發展出了「P2R-WATA」評量素養培育模式，該模式是以 e-Learning 環境為基礎所發展出來的，包含了上述四種模式的共同特色，詳細介紹請參考「四、

P2R-WATA 教師評量素養培育模式」的說明。

四、P2R-WATA 教師評量素養培育模式

(一)P2R-WATA 教師評量素養培育模式的設計理念

對於師資培育環境的設計而言，「個人化」與「情境化」設計是很重要的。在「個人化」設計方面，Fox（1995）指出，教師專業成長必須強調教師個人的專業發展，課程設計必須靠近教師本身的發展程度，也必須與教師本身的情感需求、興趣有關，換言之，課程本身就是要把「教師自己的需求」、「教學環境需求」與「學生的需求」連接起來，先以教師本身需求為起點，聚焦到工作需求，然後聚焦到學習者需求，最後達到專業能力的完善。在「情境化」方面，Clarke 和 Hollingsworth（2002）指出在師資培育課程當中，應該同時注重認知與情境面向的發展，亦即是同時重視教師知識與教師實務兩面向的發展，換言之，師資培育課程應該提供機會讓學習者於真實情境中進行理論的學習與驗證。Clarke 和 Hollingsworth（2002）主張「重視教師實務」的想法是來自於情境學習理論的觀點，情境學習理論主張知識是情境化的，學習應在真實的情境中進行，而且知識的建構是存在於人與過去情境間不斷的互動（Brown, Collins, & Duguid, 1989; Lave & Wenger, 1991; McLellan, 1993），換言之，重視教師實務的目的，就是希望學習者能夠在真實情境中學習與體會理論的意義，藉由親身的實踐，而能夠確實學習到與實務連結的理論知識，如此師資培育課程才能有機會達到預期的效益。

由上述文獻分析可知，一個良好的師資培育環境，應是一個「個人化」的學習環境，而且應融入 Clarke 和 Hollingsworth（2002）所主張之「師資培育課程應該重視教師實務」與「三、教師評量素養師資培育模式」中所提到的四種教師評量素養培育模式的共同特點-「課室經驗與評量素養訓練結合」(即情境化)，換言之，評量的概念應該在教學實際情境（instructional practice）中教導給接受評量素養訓練的教師。

「P2R-WATA」評量素養培育模式（王子華，2005; Wang et al., in press）是一種 e-Learning 型態的培育模式，該模式除將上述文獻中的建議均列入考量外，並採用「Web-based Assessment and Test Analysis system（WATA 系統，王子華等人，2004; Wang et al., 2004）」來建構 e-Learning 環境，以輔助教師評量素養的培育。「P2R-WATA」評量素養培育模式是延續 Wang 等人（2004）之研究所發展出來的，Wang 等人（2004）整理 Gronlund 和 Linn（1990, p. 109-141, p. 228）所歸納之「完整評量流程」:「確定測驗目的」、「建立雙向細目分析表」、「雙向細目表命題」、「選取適當的題型與適切的題目」、「組成試卷」、「執行考試」與「試卷之評定與結果的應用」，以及針對在職教師進行問卷調查與訪談後，指出一個完整的評量系統應該具備有 Triple-A Model（Assembling, Administering and Appraising），而這 Triple-A Model 分別為：

試題組成（Assembling）：教學者可以自建題庫並自編雙向細目分析表，且可依據此表進行命題。

考試管理（Administering）：教學者可透過該介面管理與執行多重考試行程。

試後評估（Appraising）：教學者可以分析出各種「試後分析（test analysis）」與「試題分析（item analysis）」的資訊。

下（表1）是以 Gronlund 和 Linn 的角度來比較 WATA 系統與其他六套網路評量系統

表 1：網路評量系統的比較（王子華, 2005; 王子華等人, 2004; Wang et al., 2004）

網路評量系統	Assembling		Administering		Appraising	
	確定測驗目的 & 建立雙向細目分析表（註一）	雙向細目表命題（註二）	選適當的題型與適切的題目（註三）	組成試卷（註四）	執行考試（註五）	試卷之評定與結果的應用（註六）
Gateway Testing System	X	X	◎	◎	◎	△
Lon-CAPA	X	X	◎	◎	◎	△
Mallard	X	X	◎	◎	◎	△
Question Mark	X	X	◎	◎	◎	△
Top-Class	X	X	◎	◎	◎	△
Web@ssessor	X	X	◎	◎	◎	△
WATA	◎	◎	◎	◎	◎	◎

X：代表完全不具備此功能

◎：具備此功能

△：具備部分功能

註一：具備「以建立雙向細目分析表的方式確定測驗的目的」之功能

註二：具備「根據雙向細目分析表搜尋題庫命題」之功能

註三：具備「建立題庫」以及「勾選、搜尋適當的題庫題目」之功能

註四：具備有「形成試卷」之功能

註五：具備有「執行考試」、「設定考試執行方式」以及「考試行程管理」功能

註六：除了學生成績分析（原始分數、T 分數、名次、成績分布圖等）之外，尚須包含「試卷分析（標準差、變異數、KR20 等）」與「試題分析（鑑別度、難易度、誘答力分析等）」。

的差異情形：

由（表1）可以發現，除了 WATA 系統之外，其他系統均不具備「雙向細目分析表」編輯與依據「雙向細目分析表」命題的功能，而且在「試卷之評定與結果的應用」方面也有很多差異，例如：Gateway Testing System、Lon-CAPA、Mallard、Question Mark、Top-Class、Web@ssessor 只呈現簡單的原始成績、統計圖表與試題對錯狀況，而沒有進行「試題分析」；而 Question Mark 則更進一步地紀錄了所有的學生作答過程資訊，並可藉此進行試題分析，但仍無法詳細地分析出學生的成績（如: T 分數）並進行試卷分析（如: KR20）。基於以上的分析可知，以師資培育的角度來看，WATA 系統因其具備有「完整評量流程鷹架」，而較符合教師評量素養培育的需求，因此，本研究採用 WATA 系統來營造 P2R-WATA 評量素養培育模式中的 e-Learning 環境。

除 Triple-A Model 外，「P2R-WATA」評量素養培育模式具有「情境化」與「個別化」的設計，「情境化」的設計是讓教師有實作的機會，使課室經驗與評量素養訓練二者可以相互結合，而「個別化」的設計，則讓教師依循個別的學習進度，隨時獲得各種個別化的回饋，另外，也強調了教師個人的主動參與學習，由教師「自己決定學習步驟」，並「提供更多的反省時間」，使其依據自己的學習進度與遭遇的學習問題，隨時搜尋與獲得系統的回饋，換言之，教師進行「資

源為基礎的學習而非教學為基礎的學習」（Egbert & Thomas, 2001）。

(二)P2R-WATA 教師評量素養培育模式的實施方式

「P2R-WATA」評量素養培育模式中之「個人化」與「情境化」的設計，具體實施方式如下：

「個人化」具體設計方式

WATA 系統提供每位使用者一個獨立的 Triple-A Model 環境，每位使用者可以在裡面進行個別化的「試題組成」、「考試管理」與「試後評估」，此外，在 Triple-A Model 的每一節點與頁面，均提供與該節點或頁面執行動作相關之「電子化評量教材」，提供使用者隨時點選參考。

「情境化」具體設計方式

WATA 系統具備有 Triple-A Model，提供「完整評量流程鷹架」，可以協助使用者在網路上模擬進行完整的評量流程，包含：「完整評量流程」-即「確定測驗目的」、「建立雙向細目分析表」、「雙向細目表命題」、「選取適當的題型與適切的題目」、「組成試卷」、「執行考試」與「試卷之評定與結果的應用」等步驟。

另外，本研究亦延續 Wang 等人（2004）的研究，採用 WATA 系統的 Triple-A Model 為基礎，發展一個具有「完整評量流程鷹架」之師資培育的 e-Learning 環境，其中亦包含適當的「電子化評量教材」，而這些「電子化評量教材」打散並融入 WATA 系統中 Triple-A Model 的每個節點，讓教師可以隨著自己的學習進度來模擬命題、執行考試與進行試後分析，且在過程中可隨時獲得系統的支援與回饋。此外，「P2R-WATA」評量素養培育模式亦包含「個別化」與「情境化」的精神，「個別化」方面，在「P2R-WATA」評量素養培育模式中，系統會依據教師的學習進度，隨時提供各種個別化的回饋給教師，並強調教師個人主動參與學習的精神，由教師「自己決定學習步驟」，使教師依據自己的學習進度，以及學習過程中所遭遇到的問題，隨時搜尋系統並獲得系統的回饋；「情境化」方面，在「P2R-WATA」評量素養培育模式中，教師將可以模擬真實的評量執行步驟，可以線上命題、施測與進行試後分析等。關於「P2R-WATA」評量素養培育模式，詳細的執行方式請參考「參、研究方法」的「三、研究設計」說明。

(三)P2R-WATA 教師評量素養培育模式的相關研究

在「P2R-WATA」評量素養培育模式的實證性研究方面，王子華（2005）研究發現，生物科職前教師對於該模式持有正向的態度，認為該模式可以改善其評量素養，也發現該模式對於改善生物科職前教師的評量技巧與命題品質有良好效益，另外，該模式對於改善生物科職前教師之評量知識與評量觀點而言，亦具有良好效益（王子華, 2005; Wang et al., in press），王子華指出，該模式的優勢在於，具備了 WATA 系統所提供的 Triple-A Model「完整評量流程鷹架」，生物科職前教師可以在其中實作命題、執行考試、進行自動化試後分析、分析學習缺失與修改命題缺失，這種優勢對於提升評量素養有良好的助益。

本研究將持續針對「P2R-WATA」評量素養培育模式進行深入探究，了解將該模式應用於數理科在職教師評量素養之培育的效益情形，此外，由於 Santally 和 Raverdy（2006）指出，e-Learning 環境如果可以與面對面（face-to-face）型態的傳統教學取得一個平衡點，並加以整合實施，將可以獲得更佳的效益，因此，本研究亦另外針對「P2R-WATA」評量素養培育模式搭配面授

與不搭配面授的效益情形進行探究，了解「P2R-WATA」評量素養培育模式搭配面授的效益情形。

參、研究方法

一、研究樣本

本研究的參與者為國中數理科在職教師共54位教師，分為三組：「實驗組 A」、「實驗組 B」與「對照組」，每組18人，但是，研究過程中，因為部分在職教師臨時無法配合，所以最後本研究之有效樣本為49人，分組情形請參考「三、研究設計」。

二、研究工具

(一)WATA 系統

本研究採用「WATA 系統」，來建構本研究之實驗組的評量素養培育模式-「P2R-WATA」。WATA 系統屬於常規性的 WBT 系統（Regular Web-based Testing）（Wang et al., 2004），該系統具有個人化的 Triple-A Model（圖2），擁有「試題組成引擎」可以讓教學者管理考試行程，並建立個人題庫與採用雙項細目分析表進行試卷的編製（圖3）；「考試管理引擎」則可讓教學者管理考試的進行（圖4）；「試後分析引擎」可以快速進行試卷分析（圖5）與試題分析（圖6），並產生各種教育統計資訊。

(二)評量教材

對於實驗組 A 與實驗組 B 的在職教師而言，評量教材融入 WATA 系統的每一個命題步驟與頁面中，讓在職教師在執行評量的過程中，可隨時獲得必要的協助以輔助學習。然而，對於對照組的在職教師而言，評量教材則採用紙本方式呈現，教材內容與實驗組 A 與實驗組 B 的相同，實驗組評量教材呈現方式請參考（圖7~圖9）：

評量教材是採用王子華（2005）與 Wang 等人（in press）所發展的教材，該教材是參考「學生教育評量之教師能力標準（Standards



圖 2：WATA 系統個人化介面

CK-03 - 第 3 --> 3 章 雙向細目分析表

*部分請務必輸入

概念/布倫	知識	理解	應用	分析	綜合	評鑑
0304	4	1	0	0	0	0
0305	2	0	0	0	0	0
0306	0	1	1	0	0	0
0307	0	2	0	0	0	0

提交 重新設定

圖 3：依據雙向細目分析表命題

fish 的考試行程

更新考試行程

命題者	說明	範圍	時間
bio01 刪除 fish	> 設定考試參數 > (N/A) > 雙向分析表再命題 > 列出試卷內容 > 列印紙筆測驗 > 處理問題帳號成績 > 題目送分與答案變更 > 複製本行程	0 -- 1	2001 年 20 月 20 日 09 時 10 分 06 小時內可登入 登入密碼 30 修改 設定應考名單
bio02 刪除 fish	> 設定考試參數 > (N/A) > 雙向分析表再命題 > 列出試卷內容 > 列印紙筆測驗 > 處理問題帳號成績 > 題目送分與答案變更 > 複製本行程	2 -- 3	2001 年 12 月 12 日 07 時 20 分 1 小時內可登入 登入密碼 genbioeco 修改 設定應考名單
bio03 刪除 fish	> 設定考試參數 > (N/A) > 雙向分析表再命題 > 列出試卷內容 > 列印紙筆測驗 > 處理問題帳號成績 > 題目送分與答案變更 > 複製本行程	1 -- 6	2001 年 09 月 12 日 06 時 20 分 200 小時內可登入 登入密碼 fish 修改 設定應考名單

圖 4：多重考試行程管理畫面

for Teacher Competence in Educational Assessment of Students [STCEAS]」)（AFT et al., 1990）、Gronlund 和 Linn（1990）所提出之評量基本概念與三位數理科教育與評量專家的建議，並選擇對數理科教師較重要的教學評量概念所發展而成，其內容包含「試題的建構、試卷的形成、考試的執行」、「評量資料的後續分析與應用」兩大主題，其內涵分別為：

1.試題的建構、試卷的形成、考試的執行

該主題包含「選擇題的命題原則」、「選擇題特色」、「布倫分類」、「形成性評量與總結性評量的功能差異」、「總結性評量與形成性評量對教學的意義與實施的時間」、「一般命題流程」與「考試的執行」等內容。

2.評量資料的後續分析與應用

該主題包含「變異數」、「平均」、「標準差」、「KR20信度」、「試卷難度」、「一般命題流程」、「效度(內容效度-雙向細目分析表)」、「成績分布情形的分析」、「T 分數」、「z 分數」、「常態分布」、「選項誘答力」、「鑑別度」、「錯誤概念分析」與「試題難度」等內容。

試卷分析

人數	34
變異數	134.18
標準差	11.58
總平均	54.38
KR20	0.95
難度	48.92%

班級個人成績分析

學號/姓名	原始分數	z 分數[未常態]	T 分數[未常態]
01	57.78	0.29	52.9
02	62.22	0.68	56.8
03	35.56	-1.61	33.9
04	48.89	-0.46	45.4
05	55.56	0.1	51
06	48.89	-0.46	45.4

班級個人常態化成績分析

名次	學號/姓名	原始分數	z 分數[常態化]	T 分數[常態化]
1	20	73.33	1.35	63.5
1	24	73.33	1.35	63.5
2	29	71.11	1.18	61.8
3	33	68.89	0.93	59.3
3	28	68.89	0.93	59.3
4	09	66.67	0.82	58.2
5	23	64.44	0.63	56.3
5	08	64.44	0.63	56.3
6	02	62.22	0.45	54.5
6	27	62.22	0.45	54.5
7	25	60	0.30	53

圖 5：試卷分析

(三)評量知識測驗（Assessment Knowledge Test [AKT]）與評量觀點調查表（Survey of Assessment Perspectives [SAP]）

由於本研究採用王子華（2005）與 Wang 等人（in press）所發展評量教材，所以，本研究亦採用其所發展的 AKT 與 SAP 作為研究工具。AKT 主要用以評估在職教師的評量知識，AKT 發展前，先將評量教材所包含的概念加以編號，並且依據這些概念進行 AKT 初版試卷之命題，AKT 初版試卷共有50題，初版試卷經過某師範大學生物學系職前教師預試後，共保留40題作為 AKT 正式試卷，此測驗之試題均採四選一的單選題型式，每題為2.5分，總分為100分。AKT 試卷的內部一致性 Cronbach α 值為0.993，難度為0.505。

SAP 主要用以調查在職教師對於評量功能（assessment functions）與評量步驟（assessment procedures）的觀點，得分越高代表評量觀點越正向。SAP 主要根據 Gronlund 和 Linn（1990）對評量執行步驟、評量目的等觀點設計題目內容，再經由三位生物教育專家加以修改以建立專家效度。該調查表共十五題，其中有兩子調查表，分別為「評量功能子調查表」（七題）與「評量步驟子調查表」（八題），所得的資料均為二元性（binary）類別資料，SAP 針對某師範大學生物學系在職教師預試，得其內部一致性 Cronbach α 值為0.660。

在 AKT 與 SAP 的效度考驗方面，本研

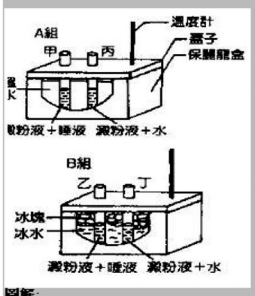
題目	A	B	C	D	E	N/A	Num	DP	ID
小英設計了下圖裝置，乙探討唾液所含的酵素特性。40分鐘後，甲、乙、丙、丁四支試管都加入本氏液，並隔水加熱，結果只有甲試管變色成正反應。根據此一實驗，請問小英想探討酵素的何種特性(！fish！357！3！*！普！應用！i0336-i0334----+！)  **圖解**: (A).本氏液對唾液所含酵素有何作用 (B).溫度對唾液所含酵素有何作用 (C).鹼度對唾液所含酵素有何作用 (D).澱粉對唾液所含酵素有何作用 (E).	25.21% H:0% L:21.43%	33.61% H:54.55% L:7.13%	4.2% H:9.09% L:14.29%	25.21% H:27.27% L:7.14%	0% H:0% L:0%	11.76% H:9.09% L:50%	119	0.47	34%
如果小傑穿白色褲子坐在草地上，褲子因為與草摩擦而沾染了綠色汁液，下列哪一種方法最容易把綠色汁液洗去(！fish！365！3！*！普！應用！0337-----+！) **圖解**: (A).用肥皂水洗 (B).用自來水洗 (C).用沸水洗 (D).用酒精洗 (E).	16.81% H:0% L:14.29%	3.36% H:0% L:7.14%	14.29% H:0% L:7.14%	57.14% H:100% L:21.43%	0% H:0% L:0%	8.4% H:0% L:50%	119	0.79	57%

圖 6：試題分析、誘答力分析

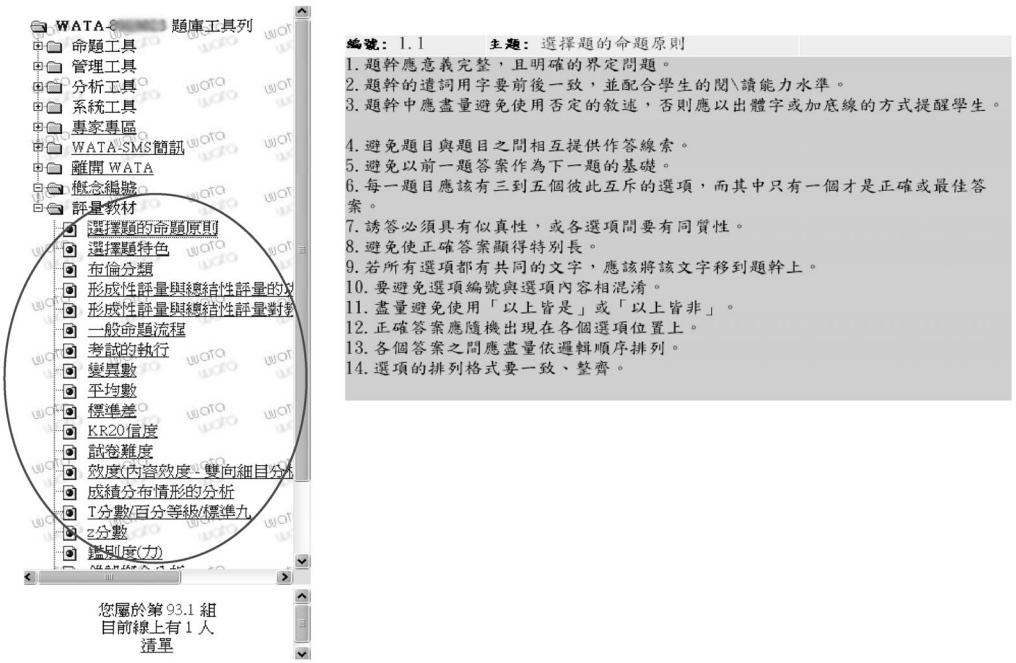
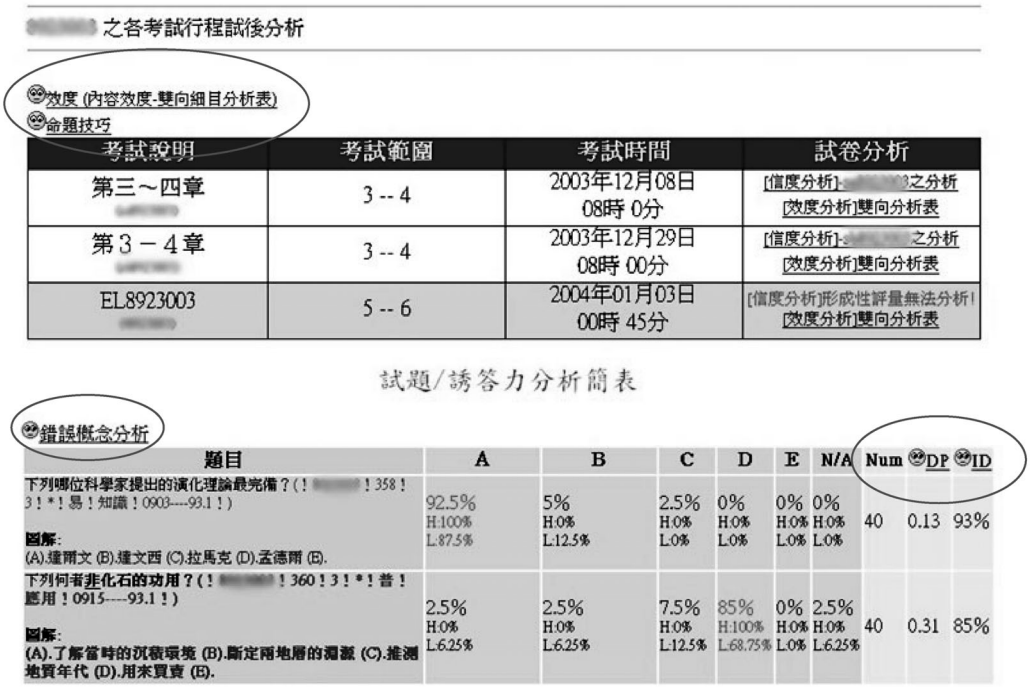


圖 7：評量教材總覽式呈現（點選左方工具列圈起來之處的「評量教材」選項，可以總覽系統中所有的評量教材）

究採用「多元特質多重方法分析」(multitrait-multimethod matrix, MTMM）來分析 AKT 與 SAP 二研究工具所蒐集到資料，分析結果如（表2）所示：

由（表2）可知，MTMM 矩陣中對角線上的值（AKT 與 SAP 的內部一致性信度 Cronbach α）大於其他值（AKT 與 SAP 間的 Pearson 相關係數），由此可知，AKT 與 SAP 所蒐集到的資料具有良好的區辨效度（discriminant validity）(Campbell & Fiske, 1959)。


之各考試行程試後分析

效度 (內容效度-雙向細目分析表)
命題技巧

考試說明	考試範圍	考試時間	試卷分析
第三～四章	3 -- 4	2003年12月08日 08時 0分	[信度分析]之分析 [效度分析]雙向分析表
第３－４章	3 -- 4	2003年12月29日 08時 00分	[信度分析]之分析 [效度分析]雙向分析表
EL8923003	5 -- 6	2004年01月03日 00時 45分	[信度分析]形成性評量無法分析! [效度分析]雙向分析表

試題/誘答力分析簡表

錯誤概念分析

題目	A	B	C	D	E	N/A	Num	DP	ID
下列哪位科學家提出的演化理論最完備？(! ! 358 ! 3 ! *! 易 ! 知識 ! 0903----93.1 !) 圖解: (A).達爾文 (B).達文西 (C).拉馬克 (D).孟德爾 (E).	92.5% H:100% L:87.5%	5% H:0% L:12.5%	2.5% H:0% L:0%	0% H:0% L:0%	0% H:0% L:0%	0% H:0% L:0%	40	0.13	93%
下列何者非化石的功用？(! ! 360 ! 3 ! *! 普 ! 應用 ! 0915----93.1 !) 圖解: (A).了解當時的沉積環境 (B).斷定兩地層的關係 (C).推測地質年代 (D).用來買賣 (E).	2.5% H:0% L:6.25%	2.5% H:0% L:6.25%	7.5% H:0% L:12.5%	85% H:100% L:68.75%	0% H:0% L:0%	2.5% H:0% L:6.25%	40	0.31	85%

圖 8：評量教材置入式呈現（評量教材隨著教師的個別學習需求，或是評量執行的步驟呈現，點選圈起來部分的超連結，就可以呈現對應之評量教材）

三、研究設計

本研究之時間從民國94年10月始至民國95年4月止，共進行約七個月，研究設計採準實驗研究法，將參與研究的在職教師分為三組：「實驗組 A」，採用「P2R-WATA」評量素養培育模式，並參與面授，共有18位教師；「實驗組 B」，採用「P2R-WATA」評量素養培育模式，但不參與面授，共有15位教師；「對照組」，僅採用「P2R」評量素養培育模式，完全在傳統環境中進行學習，不採用 WATA 系統輔助學習，也不參與面授，共有16位教師。三組教師所獲得的評量教材均相同（對照組採用紙本方式呈現），而且均經過相同的實際模擬命題與施測（practicing）、試後分析（reflecting）以及修改題目（revising）過程（對照組均在傳統課室環境進行，並沒有獲得來自 WATA 系統的任何資料與支援），三組實施的方式比較如（表3）所示：

實驗組 A 與實驗組 B 之在職教師所採用的「P2R-WATA」評量素養培育模式，由於是以 WATA 系統為基礎所發展出來的，其具備 WATA 系統的 Triple-A Model，而 Triple-A Model 恰可以提供「完整評量流程鷹架」，因此，在職教師學習過程中，Triple-A Model 將扮演核心的角色，作為整個評量素養培育過程的支撐鷹架，在整個培育過程中（Practicing, Reflecting, Revising）隨時提供「完整評量流程鷹架」、「電子化評量教材」、「個人化學習」與「情境化學習」的支援，其運作方式簡介如下：

「個別化」設計

所有參與的在職教師，都可以在個人化介面中進行個別化學習，系統可以隨著在職教師的學習進度，依循個別的需求，隨時提供各種個別化的回饋與「電子化評量教材」

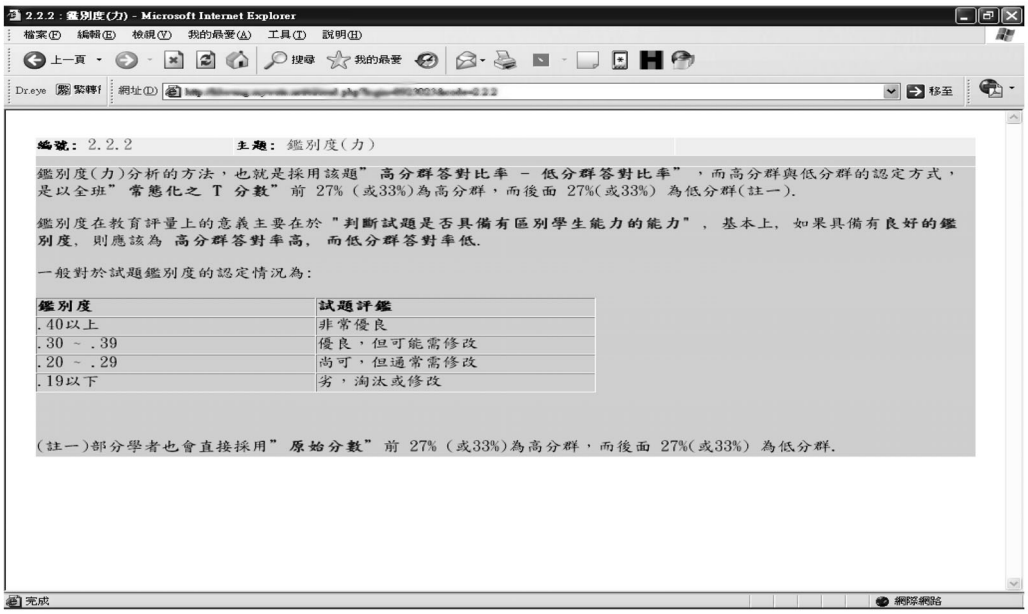


圖 9：評量教材內容畫面範例

表 2：AKT 與 SAP 之 MTMM 分析結果

	AKT	SAP
AKT	0.993[a]	
SAP	0.441[b]	0.660[a]

[a]為 AKT 與 SAP 內部一致性信度 Cronbach α
[b]為 AKT 與 SAP 間的 Pearson 相關係數

表 3：三組在職教師實施方式比較

分組	P2R	WATA[a]	面授[b]	評量教材[c]
實驗組 A	O	O	O	電子化
實驗組 B	O	O	X	電子化
對照組	O	X	X	紙本化

O：具備
X：不具備
[a]以 WATA 系統為基礎之 e-Learning 環境
[b]定期與教授進行面對面互動討論
[c]電子化與紙本化教材之內容均相同

給予在職教師。

「情境化」與「將課室經驗與評量素養訓練結合」設計

所有參與的在職教師，都可以在 WATA 系統中，依循 WATA 系統「完整評量流程鷹架」進行模擬命題，並且進行考試的「實作（Practicing）」，換言之，可以在 WATA 系統的輔助下，在網路上進行試卷模擬命題與實際執行考試，並由真實的學生在網路上進行考試，學生考試完畢後，WATA 系統可以立刻協助在職教師將學生考試結果進行試後分析，並且立刻回饋教育統計資訊給予在職教師，讓在職教師利用這些資訊「反省（Reflecting）」自己的命題缺失，並且找出教師的教學缺失與學生的學習缺失，並依據自己過去所學、「電子化評量教材」與 WATA 系統提供之試後分析資訊，針對自己的命題進行「修改（Revising）」，然後藉由 WATA 系統再次施測，來驗證自己的想法是否正確，以及自己的試題修改策略是否有效。

四、研究流程

首先，將54位國中數理科在職教師分為三組，每組18人，但完整參加整個研究流程的在職教師共49人，分別為：「實驗組 A」18

人、「實驗組 B」15人、「對照組」16人。三組在職教師均參與六次研習活動，分別說明如下：「實驗組 A」同時採用面授與網路自學的方式進行，其所參與的六次研習活動，均採面授方式進行，教授在面授過程中不對評量教材內容進行講解，而是與在職教師面對面互動討論一些與評量有關的議題；「實驗組 B」不參與面授，僅以網路自學方式進行，「對照組」則採傳統學習方式（即自行閱讀紙本教材），既不參與面授，也不進行網路自學。「實驗組 B」與「對照組」雖均不參與面授，但均會在「實驗組 A」參與面授研習的同一週，自行進行與「實驗組 A」相同的研習活動（除面授以外的活動，例如：填問卷、出考題與修改試題等）。六次研習活動的內容如下：第一次研習時，對在職教師說明研究目的，並對三組在職教師分別實施 AKT、SAP 前測；第二次研習時，請三組在職教師分別針對自己的教學科目與課本內容與雙向細目表命10題單選題，並形成一份總結性評量試卷，題目需包含題幹、選項、答案、章節、評量目標、預測難易度、出處等，之後將這些命好的題目，對任教班級學生（均為常態分班）實際進行施測。「實驗組 A」與「實驗組 B」之命題過程完全在 WATA 系統中操作，而且學生亦皆由線上測驗方式施測，另外，「對照組」教師方面，則都採用一般紙筆方式進行命題，而其學生則是採用傳統紙筆測驗方式進行施測。學生施測完畢後，第三次研習時，「實驗組 A」與「實驗組 B」教師需在 WATA 系統中依據學生的答題狀況進行試後分析，並依據分析結果對自己的命題撰寫試後分析報告，且需利用試後分析結果，將有問題的題目進行修改，「對照組」教師則是依據學生傳統紙筆測驗的考卷，依據自己的方式進行試後分析，並做試題修改；第四次研習時，三組教師則分別針對修改後的試題做第二次施測；第五次研習時，三組教師則分別針對修改後試題的施測結果進行分析（類似第三次研習的作法）。最後，第六次研習，對三組在職教師分別進行 AKT 與 SAP 後測，以瞭解其評量知識和評量觀點的改變情形。

五、資料蒐集與資料分析

本研究所蒐集到的資料均為量化資料，分別為 AKT 與 SAP 所調查之結果。AKT 量化資料方面，將採用 Cohen's *d* 估算各組間的效果量（effect size）（Cohen, 1988），估算方式如下：

$$\text{Cohen's } d = \frac{M_1 - M_2}{\sigma_{pooled}}$$

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

M_1、M_2：第一與第二組成績的平均值

σ_1、σ_2：第一與第二組成績的標準差

依據 Cohen（1988）的定義，Cohen's *d* 小於0.2屬於低效果量（small），介於0.2與0.5間屬於低到中效果量（small to medium），而介於0.5與0.8間屬於中到高效果量（medium to large），大於0.8屬於高效果量（large）。此外，亦採用共變數分析（Analysis of Covariance, ANCOVA），比較實驗組 A、實驗組 B 與對照組的在職教師，於評量知識方面的改善情形。SAP 量化資料方面，除採用 Cohen's *d* 估算各組間的效果量，以及採用 ANCOVA，比較實驗組 A、實驗組 B 與對照組教師在評量觀點方面的改善情形外，另採用 Cochran's Q 來分析比較各組教師在每一個題項填答的改變情形。由於本研究中，各組的樣本數較少，為避免違反 ANCOVA 分析的基本假設，進行 ANCOVA 分析前，均進行回歸係數同質性（homogeneity of regression coefficients）檢定與變異數同質性（homogeneity of vari-

ance）檢定。

肆、研究結果

一、評量知識的發展情形

以下針對「AKT 前測成績」、「AKT 後測成績」與「評量素養培育分組」進行共變數分析，以探討在職教師在不同「評量素養培育分組」下，其評量知識之學習效益的差異情形。ANCOVA 分析過程採用「AKT 前測成績」為共變量，以「評量素養培育分組」為固定因子，另外，由於各組樣本數較少，為避免違反 ANCOVA 分析基本假設，分析前分別進行回歸係數同質性與變異數同質性檢定，以考驗分析資料是否符合共變數分析的基本假定，結果發現，分析資料未違反回歸係數同質之基本假設（$F = 1.808$, $p = 0.176$），另外也發現，Levene 統計檢定量亦未達顯著（$F = 2.988$, $p = 0.060$），代表分析資料亦未違反變異數同質之基本假設。

三組在職教師 AKT 前後測成績之描述性統計請參考（表4），而共變數分析結果如（表5）所示，「評量素養培育分組」因子對於學習者的學習效益有顯著意義（$F = 5.103$, $p < 0.05$），另外，由 Post Hoc 分析可知（表6），實驗組 A 與實驗組 B，在評量知識的學習上，沒有顯著差異（$p > 0.05$），屬於低到中效果量（Cohen's $d = 0.372$），但是，實驗組 A 的效益顯著優於對照組（$p < 0.01$），屬於高效果量（Cohen's $d = 1.712$），而且，實驗組 B 亦顯著優於對照組（$p < 0.05$），屬於高效果量（Cohen's $d = 1.014$），由此結果可以發現，採用「P2R-WATA」評量素養培育模式，在評量知識的改善上，無論有沒有搭配面授的輔助，均有相同的效益，但是，如果僅採用「P2R」評量素養培育模式，而缺乏 WATA 系統的輔助，則評量知識的改善效益顯著較差。此研究結果與王子華（2005）採用「P2R-WATA」評量素養培育模式，針對職前教師之評量素養培育的研究結果相同，王子華指出，「P2R-WATA」評量素養培育模式比不採用 WATA 系統輔助培育之「P2R」評量素養培育模式，對職前教師之評量知識學習有顯著較佳的效益。綜合本研究與王子華

表 4：三組在職教師 AKT[a] 前後測成績之描述性統計

分組	人數	前測		後測	
		平均	標準差	平均	標準差
實驗組 A	18	60.556	18.502	73.056	15.567
實驗組 B	15	56.167	21.399	66.667	18.676
對照組	16	45.469	13.669	52.344	7.099

[a] 滿分 100 分

表 5：在職教師 AKT 後測成績之共變數分析摘要（n = 49）

變異來源	離均差平方和（SS）	自由度（df）	均方（MS）	F 值	顯著性
前測成績	4504.461	1	4504.461	38.579	0.000
評量素養培育分組	1191.532	2	595.766	5.103*	0.010
誤差	5254.176	45	116.759		
校正後總合	13509.694	48			

*$p < 0.05$

表 6：不同「評量素養培育分組」AKT 成績事後比較分析（n = 49）

處理模式	「後測成績」平均數差異	標準誤 [a]
實驗組 A–實驗組 B	3.986	3.797
實驗組 A–對照組	12.453**	3.944
實驗組 B–對照組	8.467*	3.996

**$p < 0.01$ *$p < 0.05$

[a] 多重比較調整: LSD

（2005）的研究結果可知，WATA 系統由於提供「個別化」的「完整評量流程鷹架（Triple-A Model）」與「電子化評量教材」功能，也提供「情境化」的功能，讓在職教師可以在「完整評量流程鷹架」下，實際實施評量，並利用 WATA 系統的統計分析功能，進行試後分析、反省與修改試題，這些設計提供了鷹架學習的效果，也提供了評量理論與評量實務交互映證的機會，因此，對於改善教師評量知識而言，有較佳的效益；反觀「P2R」評量素養培育模式，雖也有實際實施評量，並進行試後分析、反省與修改試題的機會，協助進行評量理論與評量實務交互映證的機會，但是，由於缺乏 WATA 系統所獨具的個別化「完整評量流程鷹架」，以及 WATA 系統協助進行試後分析的功能，將可能讓教師因為缺乏鷹架輔助學習，以及在進行試後分析時遭遇統計運算上的困難，而壓抑了評量理論與評量實務交互映證的效果，進而降低其評量知識的學習效益。

二、評量觀點的發展情形

以下針對「SAP 前測成績」、「SAP 後測成績」、「評量素養培育分組」進行共變數分析，用以探討在職教師在不同「評量素養培育分組」下，其評量觀點的發展情形。共變數分析過程採用「SAP 前測成績」為共變量，以「評量素養培育分組」為固定因子，另外，由於各組樣本數較少，為避免違反 ANCOVA 分析的基本假設，分析前亦分別進行回歸係數同質性與變異數同質性檢定，以考驗分析資料是否符合共變數分析的基本假定，結果發現，分析資料未違反回歸係數同質之基本假設（$F = 0.556$, $p = 0.578$），另外也發現，Levene 統計檢定量亦未達顯著（$F = 0.419$, $p = 0.660$），代表分析資料亦未違反變異數同值之基本假設。

三組在職教師 SAP 前後測成績之描述性統計請參考（表7），而共變數分析結果如（表8）所示，「評量素養培育分組」因子對於學習者評量觀點的發展有顯著意義（$F = 5.019$, $p < 0.05$），另外，由 Post Hoc 分析可知（表9），實驗組 A 與實驗組 B，在評量觀點的發展上，雖沒有顯著差異（$p > 0.05$），屬於中到高效果量（Cohen's $d = 0.641$），但二者在評量觀點的改變情形上，均較控制組佳（實驗組 A 與控制組：$p < 0.01$，Cohen's $d = 1.064$，高效果量；實驗組 B 與控制組：$p > 0.05$，Cohen's $d = 0.462$，低到中效果量），其中實驗組 A 的改善情形，則顯著較對照組正向。由此結果可以發現，採用「P2R-WATA」評量素養培育模式，無論有沒有搭配面授的輔助，在評量觀點的改善上，有相似的效益，但是，如果僅採用「P2R」評量素養培育模式（即缺乏 WATA 系統的輔助），則評量觀點的改善情形較差。此研究結果與王子華（2005）採用「P2R-WATA」評量素養培育模式，針對職前教師之評量素養培育的研究結果略有不同，王子華指出，「P2R-WATA」評量素養培育模式（不含面授）比不採用 WATA 系統輔助培育之「P2R」評量素養培育模式，在改善職前教師評量觀點上，有顯著較佳的效益，而本研究中，相較於採用「P2R」評量素養培育模式，採用「P2R-WATA」評量素養培育模式，無論是否搭配面授，均有較佳

表 7：三組在職教師 SAP[a] 前後測成績之描述性統計

		前測		後測	
		平均	標準差	平均	標準差
實驗組 A	18	12.722	1.320	14.444	0.922
實驗組 B	15	12.533	1.457	13.800	1.082
對照組	16	12.938	0.929	13.250	1.291

[a] 滿分 15 分

表 8：在職教師 SAP 後測成績之共變數分析摘要（n = 49）

變異來源	離均差平方和（SS）	自由度（df）	均方（MS）	F 值	顯著性
前測成績	0.356	1	0.356	0.289	0.594
評量素養培育分組	12.377	2	6.189	5.019*	0.011
誤差	55.489	45	1.233		
校正後總合	68.000	48			

*$p < 0.05$

表 9：不同「評量素養培育分組」SAP 成績事後比較分析（n = 49）

處理模式	「後測成績」平均數差異	標準誤 [a]
實驗組 A–實驗組 B	0.631	0.389
實驗組 A–對照組	1.210**	0.383
實驗組 B–對照組	0.578	0.403

**$p < 0.01$

[a] 多重比較調整: LSD

效益，但是，僅在搭配面授的狀況下才具有顯著較佳的效益，其原因可能是，面授的進行方式是授課教授與在職教師面對面討論評量相關議題，雖然不是針對評量教材進行教學，但是，在職教師可以在討論過程中，對評量功能與評量步驟有更正確的認識，因此，可以發展出較正確的評量觀點。

另外，針對參與研究之在職教師的「SAP 前測」與「SAP 後測」各題項的填答情形，進行 Cochran's Q 分析，用以探討在職教師評量觀點的細部改變情形，分析結果如（表10）所示。

由（表10）可以發現，實驗組 A 的在職教師在15個題項的填答中，有11個題項具正向的改變（73.3%），且其中有二個題項顯著改變，分別是「執行課室中的成就評量時，您是否會進行以下步驟中建立命題的雙向細目分析表」與「執行課室中的成就評量時, 您是否會進行以下步驟中依照雙向細目表命題」，而實驗組 B 之在職教師，則有10個題項有正向的改變（66.7%），而對照組的在職教師，僅有7個題項有正向的改變（46.7%）。

由（表8~表10）可以發現，在職教師採用「P2R-WATA」評量素養培育模式，對於改善教師評量觀點效果較佳，而其中以「P2R-WATA」評量素養培育模式搭配面授為最佳(實驗組 A)，另外也發現，該模式對於促進在職教師認同雙向細目分析表對於評量的意義而言，有較顯著的效益；實驗組 B 的在職教師雖採用「P2R-WATA」評量素養培育模式，但不參與面授，雖能改善在職教師之評量觀點，但是，其效益較參與面授者差；對照組的在職教師，不採用 WATA 系統輔助也不參與面授活動，雖然採用「P2R」評量素養培育模式，也進行模擬命題、施測

表 10：三組教師 SAP 各題項前後測填答情形之 Cochran's Q 分析

題項大意	實驗組 A（n = 18）	實驗組 B（n = 15）	對照（n = 16）
是否認為評量有鑑別學生學習成效的功能？	N/A	N/A	N/A
是否認為評量瞭解學生容易產生學習混淆教材概念的功能？	N/A	N/A	N/A
是否認為評量有分析個別學生所產生的不同迷思概念的功能？	N/A	N/A	N/A
是否認為評量有分析及改善教師教學策略的功能？	3.000	N/A	N/A
是否認為評量有了解題目命題的優劣的功能？	1.000	1.000	1.000
是否認為評量有了解題目中各選項命題的優劣的功能？	1.000	3.000	1.000
是否認為評量有了解整份試卷命題的優劣的功能？	1.000	3.000	N/A
執行課室中的成就評量時，您是否會進行「確定測驗目的」的步驟？	N/A	1.000	1.000
執行課室中的成就評量時，您是否會進行「建立命題的雙向細目分析表」的步驟？	6.400*	4.500	0.333
執行課室中的成就評量時，您是否會進行「依照雙向細目表命題」的步驟？	9.000**	3.571	1.000
執行課室中的成就評量時，您是否會進行「選取適當的題型與適切的題目」的步驟？	1.000	1.000	N/A
執行課室中的成就評量時，您是否會進行「組成試卷」的步驟？	N/A	N/A	N/A
執行課室中的成就評量時，您是否會進行「執行考試」的步驟？	N/A	N/A	N/A
執行課室中的成就評量時，您是否會進行「試卷成績的評定」的步驟？	N/A	1.000	1.000
執行課室中的成就評量時，您是否會進行「學習概念分析與結果的應用」的步驟？	4.500	N/A	0.667

**$p < 0.01$ *$p < 0.05$

N/A 前後測填答無任何差異，無法獲得統計量

與試後分析和修改試題，但改善其評量觀點的效果較為有限，由此足見，在職教師採用「P2R-WATA」評量素養培育模式搭配面授活動，對於改善在職教師評量觀點而言，有不錯的效果。

本研究將王子華（2005）與 Wang 等人（in press）所提出的「P2R-WATA」評量素養培育模式，應用於發展在職教師的評量素養，並進一步了解「P2R-WATA」評量素養培育模式搭配面授的效益情形，研究發現，相較於「P2R」評量素養培育模式，「P2R-WATA」評量素養培育模式無論是否搭配面授，對於提升在職教師評量知識均有顯著較佳的效益，然而，僅「P2R-WATA」評量素養培育模式搭配面授則對於改善在職教師評量觀點有顯著較佳的效益，該研究結果和王子華與 Wang 等人略有差異，王子華與 Wang 等人是採用本研究中實驗組 B 與對照組的研究設計進行研究，研究發現「P2R-WATA」評量素養培育模式（不搭配面授），對於生物科職前教師的評量知識、評量觀點、評量技巧與命題品質的改善效果均顯著優於「P2R」評量素養培育模式，王子華指出，在「P2R-WATA」評量素養培育模式下，由於具備 WATA 系統所提供的 Triple-A Model，可以提供「完整評量流程鷹架」，讓職前教師可以在其中實作命題、執行考試、進行自動化試後分析、分析學習缺失與修改命題缺失，因此具備良好的培養教師評量素養的效益。本研究在評量觀點上的研究結果，和王子華與 Wang 等人有差異的原因可能是，面授活動提供在職教師與教授面對面討論評量

相關議題，雖然不是針對評量教材進行教學，但是，在職教師可以有機會與教授討論課室評量實務，這種討論互動的機會，對於改善在職教師評量觀點而言，有更為良好的助益，反觀王子華與 Wang 等人研究中的職前教師，可能沒有豐富的課室教學與評量的經驗，因此，僅提供「P2R-WATA」評量素養培育模式（不搭配面授），職前教師就可以由這些訓練過程中，在 WATA 系統的輔助下，獲得評量實施的具體經驗，並藉由 WATA 系統在試後分析統計運算上的輔助，以及個人化的電子評量教材，逐漸改善其評量觀點，因此，相較於「P2R」評量素養培育模式，就可以有顯著較佳的效益。上述研究發現支持面授活動對於在職教師評量素養培育的價值，該研究結果可以由 Santally 和 Raverdy（2006）的研究結果獲得解釋，Santally 和 Raverdy 指出，如果 e-Learning 與面對面（face-to-face）型態的傳統教學取得某一個平衡點合併實施，將可以獲得更好的效益，因為，e-Learning 與傳統教學都各有其優點，e-Learning 可以藉由資訊技術，提供一個「個別化」與「情境化」的師資培育環境，但是，學習者在這環境中，將主要進行人機互動，而與教學者和同儕互動上，會比傳統教學環境少也缺乏即時性，雖然，e-Learning 環境可以藉由文字或是影音視訊同步互動與非同步互動技術，提供學習者與教學者和同儕互動的機會，但是，學習者在這種 e-Learning 環境中進行互動，除需要額外的訓練與學習外，亦須具備較高的資訊能力與投資較高的成本，然而，學習者與教學者和同儕的互動，在傳統教學環境則較容易達成，因此，兼取 e-Learning 與面對面型態傳統教學的優點，發展師資培育環境，將可以獲得更好的師資培育效益。

另外，本研究中的實驗組 A 與實驗組 B 均採用「P2R-WATA」評量素養培育模式，其差異只在於是否面授，而對照組則採用「P2R」評量素養培育模式，換言之，三組均已將 Tayler 和 Nolen（1996）、Clarke 和 Hollingsworth（2002）所主張之「師資培育課程應該重視教師實務」觀點融入其中，讓在職教師在能夠實際進行命題、對學生施測與試後分析的過程中，逐漸培養其評量素養，本研究發現，這樣的設計對教師評量素養之培育是可以獲得一定效益的，此研究結果與 Plake 和 Impara（1993）、ALLT、PALS 與 IPALS 等模式的發現類似，但是，本研究有更進一步的發現，如果「師資培育課程應該重視教師實務」的觀點，能夠應用資訊技術，並搭配上 Fox（1995）與 Wiske 等人（2001）所提出的「個別化」設計之觀點，建構出一個兼具「個別化」與「情境化」（即落實重視教師實務）的 e-Learning 環境，將更可以提升評量素養培育的效益。上述研究發現支持「教師實務」融入評量相關師資培育課程的重要性，雖然將「教師實務」融入評量課程，可以讓評量理論與評量實務交互印證，進而讓教師評量素養的師資培育課程有良好效益，但是，在實體環境中實施，卻有許多限制，而藉由本研究的設計發現，如果充分利用資訊技術，發展「個別化」與「情境化」教師評量素養培育環境，將可以有機會突破限制，讓評量素養培育課程可以有更好的效益。

伍、結論與建議

本研究發現，「P2R-WATA」評量素養培育模式對於改善在職教師的「評量知識」方面，無論是否搭配面授活動，其學習效益均沒有顯著差異，但是，均顯著優於不採用 WATA 系統輔助之「P2R」評量素養培育模

式組，換言之，「P2R-WATA」評量素養培育模式無論是否搭配面授活動，對於改善在職教師之「評量知識」有相同的效益，而 WATA 系統融入教師評量素養之培育，對於改善教師評量知識有顯著較佳的效益。而在「評量觀點」方面，在職教師採用「P2R-WATA」評量素養培育模式，無論是否搭配面授，在職教師的「評量觀點」改善情形雖無顯著差異，但是，「P2R-WATA」評量素養培育模式搭配面授，相較於不採用 WATA 系統輔助之「P2R」評量素養培育模式組，在「評量觀點」的改善上有顯著較佳的效益。此外，研究也發現，「P2R-WATA」評量素養培育模式搭配面授，對於提升在職教師認同雙向細目分析表對於評量的意義而言，效果最為顯著。綜合上述研究結果可知，採用 e-Learning 方式進行在職教師評量素養培育，對於在職教師的評量知識的學習而言，是有顯著效益的，但是，如果是要改善評量觀點，則需要搭配面授活動，才會有較顯著的效益。

由於「P2R-WATA」評量素養培育模式（實驗組 A 與實驗組 B）讓在職教師在整個訓練過程中，可以隨時使用 WATA 系統所提供之「個別化」的「完整評量流程鷹架（Triple-A Model）」與「電子化評量教材」來進行個人化學習，並提供「情境化」的功能讓在職教師可以依循 WATA 系統之「完整評量流程鷹架」進行「實作（Practicing）」命題且執行考試，考試完畢後，WATA 系統又可以立刻協助在職教師，將學生的考試結果進行自動化的試後分析，並且立刻回饋教育統計資訊給在職教師參考，讓在職教師可以進一步利用這些資訊來「反省（Reflecting）」自己的命題缺失，此外，在職教師還可藉由解析 WATA 系統所提供之試後分析結果，針對自己的命題進行調整與「修改（Revising）」後，再次以 WATA 系統對學生進行施測，以驗證自己的試題修改策略是否有效。但是，「P2R」評量素養培育模式（對照組），雖然有同樣內容的評量教材，並進行同樣的實作訓練過程（即上述的 Practicing, Reflecting, Revising），但是，由於缺乏 WATA 系統的輔助，在職教師無法獲得「完整評量流程鷹架（Triple-A Model）」與試後分析的支持與輔助，使得在職教師在進行試後分析的教育統計運算時，可能遭遇到較多的阻力，而降低評量素養培育的效果。綜合上述，「P2R-WATA」評量素養培育模式的主要優勢在於，WATA 系統具備 Triple-A Model，可以提供「完整評量流程鷹架」，使在職教師可以在反覆修改試題與進行實際施測與試後分析的循環過程中，來增強自己的評量素養，此外，WATA 系統也可以協助進行資料分析與統計運算，也將可以減輕在職教師在進行統計運算上的負擔。

最後，在職教師的評量素養培育，建議可以參考本研究所建構之「P2R-WATA」評量素養培育模式，將網路評量系統應用於在職教師評量素養之培育，並搭配面授活動一起進行，如此，將可以兼顧傳統面對面教學（面授）與網路教學的優點，而獲得較佳的效益。師資培育者在面授活動中扮演的角色，並非在於評量教材內容的教學，而是與在職教師討論評量相關議題，本研究建議，這些議題可以包含：「課室評量實務」、「課室評量經驗分享」與「新興評量議題」等，在職教師將可以藉由與師資培育者和同儕的討論過程，獲得較佳的評量觀點發展機會。在後續研究方面，由於參與本研究的在職教師人數較少，且分為數學、理化與生物等不同領域專長的教師，因此，無法參照王子華（2005），同時針對評量知識、評量觀點與評量技巧進行探究，而僅能針對與學科知識較無關係之評量知識與評量觀點二向度進行研究。本研究建議，未來宜徵求更多不同學

科領域之在職教師參與研究，同時針對評量知識、評量觀點與評量技巧三向度，採取質化與量化研究並重的方式進行研究，以更精確了解「P2R-WATA」評量素養培育模式，對於不同學科背景的教師的效益情形。此外，本研究採用之研究工具-AKT 與 SAP 均針對數理科教師所發展出來的，而且本研究亦主要針對傳統課室評量方面的評量素養進行研究，因此，未來有必要朝向多元評量之評量素養之培育模式進行深入探究，並針對其他學科背景之在職教師與多元評量之評量素養發展適當的研究工具，以協助未來教師評量素養相關研究之進行。除發展量化研究工具外，由於該研究很難針對大量樣本進行研究，未來亦有必要發展質性研究工具，以利進行後續的質性分析，以蒐集到更為深入的佐證資料，例如：採用焦點團體訪談法或是參考 Plake 和 Impara（1993）的模式撰寫親師短文等。另外，本研究所調查的對象，僅屬於數理科在職教師，本研究建議，未來有必要擴大調查範圍，針對其他學科之教師，進行教師評量素養之調查，以瞭解國內教師評量素養的實況，並且歸納出國內教師評量素養的缺失，然後依照這些缺失發展更適當的評量素養培育模式，最後，由於教師評量素養改善的目的，主要是希望能夠提昇教師在課室環境中實施評量的品質，進而提升學生學習效益，因此本研究也建議，未來宜進行長期的縱貫性研究，除了進一步驗證「P2R-WATA」評量素養培育模式的效益外，也要探討經由該模式所培育出來的教師，是否可以將其評量素養展現在實際課室教學之中，進而提升學生的學習效益。

陸、致　謝

感謝國科會專題研究計畫 NSC 94-2511-S-018-009經費補助，讓本文順利完成。另外，亦感謝主編與審查者的寶貴建議，讓本文得以更加完善。

柒、參考文獻

1. 王子華、王國華、王瑋龍和黃世傑（2004）：網路評量系統融入師資培育之探究。**資訊管理學報**, **11(3)**, 121-154。
2. 王子華（2005）：**建構網路評量系統發展生物科職前教師評量素養**。彰化市：國立彰化師範大學科學教育研究所博士論文（未出版）。
3. 教育部（2003）：**國民中小學九年一貫課程修訂綱要**。2007 年 3 月 12 日，取自：http://teach.eje.edu.tw/9CC/index.php。
4. Airasian, P. W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10*(1), 13-16.
5. American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT, NCME, & NEA). (1990). The standards for teacher competence in the educational assessment of students. *Educational Measurement: Issues and Practice, 9* (4), 30-32.
6. Arter, J. (2001). Learning teams for classroom assessment literacy. *NASSP Bulletin, 85*, 53-65.
7. Bandalos, D. (2004). Introduction to the special issue on the Nebraska Standards-based, Teacher-led assessment and Reporting System (STARS). *Educational Assessment: Issues and Practice, 23*(2), 4-6.
8. Brookhart, S. M. (1999). *The art and science of classroom assessment: The missing part of pedagogy*. (ERIC Document Reproduction Service No. ED432938).

9. Brookhart, S. M. (2001). *The standards and classroom assessment research.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. (ERIC Document Reproduction Service No. ED451189).
10. Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.
11. Buckendahl, C. W., Impara, J. C., & Plake, B. S. (2004). A strategy for evaluating district developed assessments for state accountability. *Educational Measurement: Issues and Practice, 23*(2), 15-23.
12. Campbell, C. & Collins, V. L. (2007). Identifying essential topics in general and special education introductory assessment textbooks. *Educational Measurement, Issues and Practice, 26*(1), 9-18.
13. Campbell, D. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
14. Campbell, C., Murphy, J.A., & Holt, J.K. (2002). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.
15. Clarke, D. & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*(8), 913-1059.
16. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
17. Egbert, J. & Thomas, M. (2001). The new frontier: A case study in applying instructional design for distance teacher education. *Journal of Technology and Teacher Education, 9*(3), 391-405.
18. Fox, A. (1995). *Change as a way of life on education*. Tel Aviv (in Hebrew): Charikover.
19. Gronlund, N. E. & Linn, R. L. (1990). *Measurement and evaluation in teaching. (6th ed.)*. Macmillan, New York.
20. Gullickson, A. R.(1993). Matching measurement instruction to classroom-based evaluation: Perceived discrepancies, needs, and challenges. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 1-25). Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska-Lincoln.
21. Johns, J. L. & VanLeirsburg (1993). The impact of coursework in tests and measurements on assessment literacy. In Rasinski, T.V. & Padak, N. D. (Eds.), *Inquiries in literacy learning and instruction: The fifteenth yearbook of the college reading association* (pp. 193-199). Kent, OH: Kent State University.
22. Lave, J. & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.
23. Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23* (2), 26-32.
24. Maclellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualizations. *Teaching and Teacher Education, 20*, 523-535.
25. Marso, R. N. & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practice. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129-185). Lincoln,

NE: Buros Institute of Mental Measurement, University of Nebraska-Lincoln.

26. McLellan, H. (1993). Evaluation in a situated learning environment. *Educational Technology, 33*(3), 39-44.
27. McMillan, J. H. (2000). *Basic assessment concept for teachers and school administrators.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED447201).
28. McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32.
29. Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment reliability and validity. *Mid-western Educational Researcher, 13*(4), 29-35.
30. Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 32* (3), 49-64.
31. National Research Council [NRC]. (1996). *National science education standards*. Washington, DC: National Academy Press.
32. North Central Regional Educational Laboratory. (2006). *Indicator: Assessment*. Retrieved January 24, 2006, from http://www.ncrel.org/framewk/pro/literacy/prolitin.htm.
33. O'Sullivan, R. G. & Johnson, R. L. (1993). *Using performance assessment to measure teachers' competence in classroom assessment.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED358156).
34. Phillips, S. (2003). *Policy watch: Assessment literacy.* Canada: Society for the Advancement of Excellence in Education. Retrieved January 24, 2006, from http://www.saee.ca/policy/D_043_CCC_LON.php.
35. Plake, B.S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12, 39.
36. Plake, B. S. (1993). Teacher assessment literacy: Teacher's competencies in the educational assessment students. *Mid-Western Educational Researcher, 6*(1), 21-27.
37. Plake, B. S., Impara, J. C., & Buckendahl, C. W. (2004). Technical quality criteria for evaluating district assessment portfolios used in the Nebraska STARS. *Educational Measurement: Issues and Practice, 23*(2), 10-14.
38. Plake, B. S. & Impara, J. C. (1993). *Teacher assessment literacy: Development of training modules*. (ERIC Document Reproduction Service No. ED358131).
39. Popham, W. J. (2006). Needed: A dose of assessment literacy. *Educational Leadership*, *63*(6), 84-85.
40. Santally, M. & Raverdy, J. (2006). The master's program in computer-mediated computer communications: A comparative study of two cohorts of students. *Educational Technology Research & Development, 54,* 312-326.
41. Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice, 10*(1), 3-6.
42. Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice, 32* (2), 118-126.
43. Stiggins, R. J. (1991a). Relevant classroom

assessment training for teachers. *Educational Measurement: Issues and Practice, 10*, 7-12.

44. Stiggins, R. J. (1991b). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-539.
45. Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.
46. Stiggins, R. J. (1999a). Are you assessment literate? *The High School Journal, 6*, 20-23.
47. Stiggins, R. J. (1999b). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-27.
48. Stiggins, R. J. (1999c). Teams. *Journal of Staff Development, 20*(3), 17-21.
49. Stiggins, R. J. (2001a). The Principal's leadership role in assessment. *NASSP Bulletin, 85*(621), 13-26.
50. Stiggins, R. J. (2001b). *Student-involved classroom assessment (3rd ed.)*. New Jersey: Prentice Hall.
51. Stiggins, R. J. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 86*(1), 22-27.
52. Taylor. C. S. & Nolen, S. B. (1996). What does the psychometrician's classroom look like? Reframing assessment concepts in the context of learning. *Education Policy Analysis Archives, 4*(17). Retrieved January 24, 2006, from http://olam.ed.asu.edu/epaa/v4n17.html.
53. Wang, T. H., Wang, K. H., & Huang. S. C. (in press). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education*.
54. Wang, T. H., Wang, K. H., Wang, W. L., Huang. S. C., & Chen, Sherry Y. (2004). Web-based Assessment and Test Analyses (WATA) system: Development and evaluation. *Journal of Computer Assisted Learning, 20*(1), 59-71.
55. Wiske, M. S., Sick, M., & Wirsig, S. (2001). New technologies to support teaching for understanding. *International Journal of Educational Research, 35*, 483-501.

Implementation of Web-based Assessment System in Improving In-service Teachers' Assessment Literacy

Tzu-Hua Wang[1], Ya-Ching Fan[2] and Kuo-Hua Wang[2]

[1]Department of Education, National Hsinchu University of Education

[2]Graduate Institute of Science Education, National Changhua University of Education

Abstract

Teacher assessment literacy is a key factor in the success or failure of teaching. Studies investigating in-service and pre-service teachers' assessment literacy have revealed that even teachers who have taken assessment courses lack assessment literacy. The aim of this research is to propose an effective strategy for improving in-service teacher assessment literacy. The strategy is named ''P2R-WATA (Practicing, Reflecting and Revising with WATA system) assessment literacy development model'', and was developed on the basis of the WATA system (Web-based Assessment and Test Analysis system). The WATA system offers personalized and situated supporting resources and provides opportunities for in-service teachers to improve their own assessment literacy. In this research, quasi-experimental method was adopted. Participants were divided into three groups: experimental group A (P2R-WATA with face-to-face instruction, n = 18), experimental group B (P2R-WATA without face-to-face instruction, n = 15) and control group (P2R without face-to-face instruction and WATA system, n = 16). Results indicated that experimental groups' (A and B) in-service teachers' improvement of assessment knowledge was significantly better than control group's in-service teachers. In addition, it also showed that P2R-WATA with face-to-face instruction (experimental group A) had significantly better effectiveness in improving in-service teachers' assessment perspectives than control group's in-service teachers.

Key words: WATA System, P2R-WATA Assessment Development, Assessment Literacy, Teacher Education, e-Learning