科學教育學刊
2022, 第三十卷第四期, 309-333
DOI:10.6173/CJSE.202212_30(4).0002

Chinese Journal of Science Education
2022, 30(4), 309-333

中學科學教師評量素養量表發展之信效度與恆等性分析

張仁誠[1]　洪菁穗[1、2]　吳心楷[1、3、*]

[1]國立臺灣師範大學　科學教育研究所
[2]臺北市立和平高中
[3]南非約翰尼斯堡大學　科學與科技教育學系

摘要

本研究目的在發展一份適用於臺灣國高中科學教師評量素養之量表，並進行量表信效度與恆等性之分析。此量表基於Abell與Siegel (2011)評量素養模型之構念與架構所編製，本研究檢驗量表之信效度，並檢驗跨性別與教育階段群體之測量恆等性，以確保在不同群體間皆能測量到相同的評量素養之構念組成。本研究的預試資料來自100位北部國高中科學教師，而正式施測收集到北部地區140所學校，872位中學科學教師之量表資料。本研究使用預試資料進行項目分析和探索性因素分析，正式施測資料則以驗證性因素分析與測量恆等性分析來處理。全部量表題項經分析刪減後，共包含35題。項目分析、探索性因素分析與驗證性因素分析結果支持Abell與Siegel評量素養模型架構。經刪題後，各構念題項呈現良好的信效度，且透過恆等性分析顯示量表具優良的跨性別群組及跨國高中階段測量的穩定性。

關鍵詞：信效度、恆等性分析、科學教師評量素養、量表發展

壹、緒論

在《十二年國民基本教育課程綱要》強調知識運用、高階思考、及問題解決的學習典範下，科學教室的評量，需同時針對學習過程和學習成果，結合學習檔案等非現行慣用的評量方法，讓評量不只是測驗，學習表現不只是分數，而是透過更多學習證據的設計、收集、和詮釋，以協助教師和學生進行科學的教與學。為瞭解中學科學教師對於評量是否具有足夠的理解來面對這些挑戰，需要發展相關研究工具來檢視臺灣國高中科學教師的評量素養。

評量素養通常是指教師是否熟悉課堂上的測量基礎知識(Popham, 2009)，所以在傳統上被定義為對教育評量和相關技能的基本理解，以將這些知識應用於學生表現的各種衡量標準，並被認為是教師專業素養的一部分(Xu & Brown, 2016)。許多教師評量素養的架構都是針對評量理論的應用，而且大多數是經由每個國家依其各自的政策和相關標準而制定(Pastore & Andrade, 2019)。然而由於教學

*通訊作者：吳心楷，hkwu@ntnu.edu.tw
(投稿日期：民國111年8月25日，修訂日期：民國111年11月14日，接受日期：民國111年11月14日)

科目和教學環境的不同，教師需要的評量知識和展現的評量實務也有差異，如：科學教學過程中除了課堂講述還有實驗操作，所以科學教師評量素養的主要核心，應以科學教師所知道和能夠做到的評量為主體，需要理解科學教師如何詮釋評量訊息和做出教學決定，且應以專業教師的觀點，運用針對科學教師的理論架構，因此如何將教師評量素養融合到科學教育領域，是本研究所關切的。

鑑於評量素養的重要性及科學教學的獨特性，有必要探索和測量科學教師的評量素養。目前評測科學教師評量素養的方法，大致上有量化與質化兩類(Demirdöğen & Korkut, 2021)，其中Ogan-Bekiroglu與Suzuk (2014)使用量化數據來確定職前科學教師的評量素養，探討的內涵包含評量的認知層次、評量類型、和評量標準等，涵蓋的層面少；而使用質性方法的研究探討了中學科學教師(Gottheiner & Siegel, 2012)、化學教師(Izci & Siegel, 2019)、職前物理教師(Ogan-Bekiroglu & Suzuk)和職前教師在探究課程如何規劃和教授科學的評量素養(Siegel & Wissehr, 2011)。相較之下，科學教師評量素養的質性研究數量較多，但專注於科學評量素養的量化研究卻很少。另外，這些量化研究所發展的評量素養量表，大多未針對科學教學設計而直接採用適用於一般教師的試題，如Ogan-Bekiroglu與Suzuk所使用的問卷來自McMillan (2001)；而且這些研究對於評養素養的內涵和相關構念，並未運用針對科學教育所發展的理論模型。

針對以上量化研究不足和未具科教理論基礎的研究缺口，本研究採用Abell與Siegel (2011)所提出的科學教師評量素養模型，來進行量表的發展。此模型的優勢(Demirdöğen & Korkut, 2021) 包括：一、針對科學教師所建立；二、奠基於有關教師評量知識和實務的實徵和理論文獻；三、已有質化研究文獻為該模型在職前和在職科學教師評量素養的適用性提供了實徵證據(Gottheiner & Siegel, 2012; Izci & Siegel, 2019; Siegel & Wissehr, 2011)。為發展量表和分析其信效度，本研究將Abell與Siegel模型中的構念給予操作型定義並依此編製題項，收集教師資料進行量化分析，以識別Abell與Siegel理論當中構成評量素養的組成架構，是否符合資料所呈現的國高中科學教師之評量素養結構。在檢驗理論架構時，各構念是否具跨組恆等性亦為重要考慮因素之一，即評量素養的構念組成、題項的解釋力和構念間的關係，能否在不同群體間都能達到相似的結果(Dyer, 2015)。當量表不具有群組測量恆等性時，群組平均數的差異比較是不合理的，因此要探究量表，在欲觀察或研究的現象是否測量到相同的屬性，缺乏測量恆等性的證據將無法明確說明個體或群體之間是否確實存在差異(李俊賢等，2016)。因此本研究近一步的使用恆等性分析進行跨性別及教育階段群體之穩定性檢驗。

綜上所述，本研究目的在發展一份基於Abell與Siegel (2011)評量素養的定義與架構且適用於國高中科學教師評量素養量表，檢驗量表之信效度，並進行跨性別及教育階段群體之測量恆等性的檢驗。

貳、文獻探討

本節將先回顧評量素養之文獻和理論模型，介紹Abell與Siegel (2011)評量素養模型及其構念，最後再描述現有評量素養的量表之設計和相關研究。

一、評量素養

(一)評量素養的意涵與架構

評量素養可定義為教師對於有效評量的知識和能力，包括評量的建構、執行、評分，和評量資料與結果的詮釋(Abell & Siegel, 2011; DeLuca et al., 2016b)。評量素養之所以受到重視，是由於教師所使用的評量方式和策略會影響到學生的學習成就和動機(Black & Wiliam, 1998)，應為教師專業知能的一部分，因此有些國家訂定評量素養的標準以提昇教育品質並達到教師認證的目的。

目前的評量素養模型架構可分為兩類，一類是一般性的架構，用於分析所有教師對評量的知識和技能；另一類是針對學科特色所發展的架構，強調學科評量的特殊性。一般性架構的範例為美國教師聯盟和相關單位(American Federation of Teachers [AFT] et al., 1990)率先在1990年提出針對學生教育評量的教師能力標準(standards for teacher competency in educational assessment of students)，這七項標準強調教師應能：1.選取適當評量方法；2.發展適當評量方法；3.執行、評分和詮釋不同評量方法所產生的評量結果；4.使用評量結果以做出關於特定學生、準備教學、發展課程、及學校改善等決定；5.發展具效度的評分程序；6.與學生、家庭和其他教師等溝通評量結果；7.知曉不道德、不合法或不適當的評量方法與評量訊息的使用方式。基於上述標準，教師評量素養可定義為教師在教育評量方面的知識和技術，且偏重對測量原理及評量設計的理解(Stiggins, 1991)。除此之外，Xu與Brown (2016)則提出了程序性的評量素養模型架構，包括六個組成部分：1.知識庫；2.教師的評量觀念；3.制度和社會文化背景；4.教師在實務中的評量素養；5.教師學習；6.教師身分(重新)建設為評估者。Pastore與Andrade (2019)針對一般教師提出了三維度概念架構：1.概念知識維度(conceptual knowledge dimension)；2.行動維度(praxeological dimension)；3.社會情感維度(socio-emotional dimension)。這些一般性的架構，為評量素養提出了適用於所有教師的定義和面向，其優勢在於適用於不同學科的教師，但較未考量教師任教學科的特性，為此類評量素養架構的限制。而且隨著學習和評量理論的翻新，教師評量素養的標準和內涵也有了變化(DeLuca et al., 2016b)，然而，這些標準和相關指標仍未考量教學科目和學習環境的特殊性。

由於一般型架構的限制，開始有研究提出第二類的架構，深入探討符合學科類型的評量和教師應具有的評量素養，如：Starck等(2018)針對體育教師提出評量素養架構，該架構為四個階段：1.評量理解；2.評量理解和應用；3.評量應用與解釋；4.評量解釋和批判性參與評估。在科學教育領域，Abell與Siegel (2011)提出針對科學教師的評量素養架構，包含了教師對教學的價值觀和原則，與四類科學教師的評量知識：1.評量目的；2.評量內容；3.評量策略；4.評量解釋和由此產生的行動。此類評量素養的架構，有利於突顯學科特性(如：偏重實作技能和表現)以幫助學科教師發展對評量各面向的概念，並可整合到學科教學中的評量活動，藉以實現提高教師評量素養的目標。本研究即利用Abell與Siegel之評量素養架構，作為量表開發的基礎，此模型是以科學教師為基礎發展，其中提到的評量目的、評量內容與評量策略等，皆可搭配科學教育的學科特徵進行延伸，以下就Abell與Siegel之評量素養模型進行介紹，並搭配科學學科的評量範例以詳細說明。

(二)Abell與Siegel之評量素養模型

Abell與Siegel (2011)以Magnusson等(1999)提出的評量知識為基礎，試圖提出更全面的評量素養模型(圖1)。在此模型中，除了「評量什麼」(即評量內容)的知識外，有關「如何評量」的知識被拆解為評量目的、評量策略、評量詮釋和採取行動(action-taking)的知識。而這些知識和相關的教學決定，包括教師要用何種評量、教師為何／如何／何時評量、教師如何使用評量訊息，都會受到核心信念和價值的影響：即教師對學習的觀點，以及對評量持有的價值觀和原則。以下詳述模型的每項成分，即為本研究之量表開發所欲涵蓋的構念。

1.教師對學習的觀點

教師對教學和學習所持有的信念和觀點，對其教學、學習、評量的過程具重要的影響(Pajares, 1992)，因此教師所持的評量價值觀和對學習的觀點是該模型的核心，這些價值觀和原則是教師在科學課堂上做出評量決策的基本思想和信念(Abell & Siegel, 2011)。Chan與Elliott (2004)將教師對學習的觀點分為兩類：建構觀點與傳統觀點。採建構觀點的教師，傾向認為學習是學生與環境中的人事物互動而主動建構知識的過程；而具傳統觀點的教師，認為學習是學生從教師或教科書吸收知識的過程。不同的學習觀點可能會影響其他的評量知識。對學習持有建構觀點的教師，採用的評量策略和內容(如：使用實作評量來測得學生實驗表現)，與持有傳統知識傳輸觀點的教師可能會有所不同(如：使用紙筆測驗評量學生習得的科學知識、定律)。

2.評量目的的知識

教師評量學生的目的可以分為診斷性、形成性和總結性等類型(Abell & Siegel, 2011)。診斷性評量是在教學開始時進行的評量，用於瞭解學生對教學內容具有的先備概念、知識和信念，教師可利用診斷性評量的結果來設計教學(Abell & Siegel)。形成性評量是在整個教學過程中進行的評量，旨在向學生和教師提供關於學習和教學的回饋，以提高他們的學習能力(Abell & Siegel)。總結性評量是在課程結束時進行的評量，以記錄學生

圖1：科學教師的評量素養模型

資料來源：“Assessment literacy: What science teachers need to know and be able to do,” by S. K. Abell & M. A. Siegel, in D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221), 2011, Springer.

的學習情況並主要給出課程成績，並向教師提供有關其教學的回饋(Abell & Siegel)。

3.評量內容的知識

科學教師需要知道在他們的科學課堂上要評量什麼(Abell & Siegel, 2011)，即欲評量的科學內容或其他相關表現，如：科學過程、科學本質、概念理解、科學語彙、態度、實驗技能。評量內容與課程目標、重要學習內容和學習方式的價值觀有關，若持傳統學習觀點的教師，可能偏重概念和知識的獲得，若持較為建構學習觀點的教師，可能傾向評量學生學習過程的表現。

4.評量策略的知識

評量策略是指教師用來評量的各種方式，如：教師提問、測驗、討論、學習單與回家作業、報告、日誌等。而且評量策略可分為正式和非正式(Abell & Siegel, 2011)，正式評量通常會向學生宣布其分數和等級(如：測驗和期末報告)，而非正式評量是嵌在教學中的評量任務，可能無法被計分或評定等級(如：對課堂教師提問的回答、實驗室筆記等)。評量策略的知識是指教師對不同評量方式的理解，科學教師需要瞭解總結性評量中使用的正式評量策略(如：如何設計紙筆測驗的考題)和形成性評量中使用的非正式評量策略(如：如何在一堂課結束時，使用學生回答和筆記來衡量其學習情況)。這些策略可能與學習環境和教學科目息息相關，如：科學課和語文課的實作評量，其設計和施行方式就不盡相同。

5.評量詮釋與行動

評量詮釋和採取行動的知識是指教師如何處理評量資料和數據(Abell & Siegel, 2011)，並根據評量結果而採取的有關科學教學和學習的行動，如：引導出學生現有的知識、向學生提供回饋、提供學生更多的練習機會、監控學生的學習、改變教學方式、或重新編排教學內容等(Demirdöğen & Korkut, 2021)。在評量素養的研究中，詮釋經常被忽視，很少有研究探討教師如何詮釋課堂評量結果(Gottheiner & Siegel, 2012)，因此本研究將其納入量表中。

二、現有的評量素養量表工具及相關研究

過去已有文獻依據教師評量能力標準(AFT et al., 1990)和相關理論，開發評量素養的檢測工具。Gotch與French (2014)回顧了發表於1991到2012年間的評量素養研究所使用的34項測量工具(包括量表、問卷、測驗等)，以五項原則來檢視這些工具的品質包含：測驗內容、內在一致性、評分穩定度、內在結構、以及與學生成果之關連。Gotch與French發現使用這些工具的研究所提出的心理測量證據不足，而且工具內容缺乏當前評量視野的代表性和相關性。

比較了現有幾項評量素養工具所使用的信效度指標，這些工具所使用的信效度指標相當局限，提供的測量證據非常有限，呼應了前述Gotch與French (2014)的看法。如：Alkharusi (2011)同樣是基於AFT等(1990)的標準，來開發教師評量素養問卷，衡量259位職前教師對課堂評量實務基本原則的知識和理解。該量表由35項選擇題組成，使用難度、鑑別度、Cronbach α以及與測量課程分數的相關性作為量表分析的主要信效度指標。Jarr (2012)則是使用Means等(2011)的標準，探討220位在職教師詮釋和使用評量結果的能力，以及對評量實務的知識和技能的信心程度。向度一測量使用評量結果的能力，由11個選擇題組成，屬於認知測驗。使用難度、

鑑別度與Cronbach α作為量表分析的主要信效度指標；向度二測量信心程度有15題，使用Cronbach α與因素分析作為量表分析的主要信效度指標。而Latif (2021)同樣是基於AFT等機構的標準，僅使用Cronbach α作為量表分析的主要信效度指標。Vogt等(2020)同樣是以AFT等機構為基礎，雖然收集了大量資料，調查658名教師和1,788名學生語言學習課堂的評量實務、評量相關的回饋機制、教師的培訓需求、和學生的學習需求，但文章中缺乏信效度的檢測內容。

此外，近期用以檢測教師評量素養的工具，其涵蓋的面向，各有不同。但大多工具仍是以一般型的評量素養架構，即AFT等(1990)的教師評量能力標準作為基礎來開發，且注重教師對「評量內容」和「評量詮釋與行動」的知識，對於「評量目的」的探討付之闕如。再者，評量內容涉及學科特定的知識內容和能力展現，單一的教師標準可能無法通用，過去這些評量素養的量表並不一定符合科學教師評量素養測量的需求。所以為突顯科學教學和評量的特殊性，本研究應用Abell與Siegel (2011)的架構，且相比於過去的工具，多增加了教師對學習的觀點及評量目的等構念進行探討。

參、研究方法

一、研究對象

本研究為了瞭解科學教師的評量素養及其相關因素，預試樣本採取立意抽樣的方式，抽取願意配合學校高中部分7所共42位教師，國中部分5所共58位教師，合計100位教師進行預試分析。正式問卷部分採兩階段的分層抽樣方式進行資料收集。從臺灣北區(臺北市、新北市、桃園市及基隆市)的公立中學收集資料。在高中階段，依據各校入學考試的百分等級(PR值)分層，分別從各分層中按比例隨機抽取學校作為受試學校。在國中階段，則以各校所在區域和學校規模抽樣，依據各校學生人數分大中小型學校分層，分別從各分層中按比例隨機抽取學校作為受試學校。本研究在徵求學校及教師同意之後，以有意願配合學校之自然科教師人數為發放數，高中部分共抽取共有62校，發放問卷760份，回收353份，回收率46.45%；國中部分共抽取共有78校，發放問卷847份，回收519份，回收率61.28%。國高中科學教師總計共收集了872份量表資料。參與者之主要基本資料及各變項填答的分布情形，如表1所示。

二、科學教師評量素養量表的編製與發展

本研究在量表題項發展係以Abell與Siegel (2011)作為理論基礎，並選擇科學教育與教師教育相關之研究，作為題項發展之參考。由於部分試題為英文版，在題項翻譯方面，本研究邀請三位具有教育領域博士學位之學者，舉行多次專家會議，就中英文量表題項之內容差異性、語意正確性、文字流暢性，以及中學科學教師是否能理解題意等議題，進行討論和修改。評量素養量表各變項與填答說明詳述如下。

(一)對學習的觀點

教師「對學習的觀點」之題項發展係參酌Chan與Elliott (2004)，此問卷能貼合Abell與Siegel (2011)的理論，包含兩項構念，即建構觀點與傳統觀點。本研究選擇在Chan與Elliott問卷中因素負荷量高於.5的題目組成本研究的量表題項，以確保因素的效度。其中傳統觀點選擇了6題，示例為「講述的教學方法是最好的，因為它涵蓋了更多的資訊／知識」。建構觀點共有9題，例題為「好的課

表1：參與教師之基本資料和作答情況

基本變項	人數	百分比	傳統觀點		建構觀點		評量目的		評量內容		評量行動	
			平均數	標準差	平均數	標準差	平均數	標準差	平均數	標準差	平均數	標準差
性別												
男	462	53.0%	3.47	0.83	5.09	0.60	4.91	0.60	3.93	0.87	4.22	0.78
女	410	47.0%	3.22	0.80	5.12	0.49	4.87	0.54	3.86	0.78	4.13	0.77
教學階段												
國中	519	59.5%	3.45	0.81	5.13	0.55	4.92	0.56	3.94	0.83	4.23	0.77
高中	353	40.5%	3.21	0.81	5.08	0.54	4.84	0.58	3.83	0.83	4.10	0.78
教育程度												
大學	155	17.9%	3.28	0.86	5.16	0.54	4.92	0.55	3.83	0.92	4.14	0.85
40學分班	36	4.2%	3.35	0.76	5.12	0.48	4.87	0.49	3.81	0.77	4.02	0.91
教學碩士	49	5.7%	3.63	0.68	5.11	0.50	4.99	0.51	3.96	0.73	4.31	0.69
碩士	608	70.2%	3.35	0.82	5.09	0.56	4.87	0.59	3.92	0.82	4.19	0.75
博士	18	2.1%	3.28	0.84	4.98	0.43	4.76	0.48	3.94	0.79	4.13	0.72
年齡												
22～30	80	9.3%	2.99	0.80	5.13	0.54	4.93	0.56	3.87	0.85	4.22	0.74
31～40	260	30.3%	3.28	0.78	5.11	0.54	4.92	0.52	3.83	0.79	4.20	0.75
41～50	356	41.4%	3.43	0.84	5.08	0.58	4.89	0.54	3.90	0.85	4.14	0.79
51以上	163	19.0%	3.48	0.81	5.14	0.51	4.89	0.54	4.01	0.78	4.18	0.78
教學年資												
0～10	273	31.6%	3.17	0.79	5.13	0.52	4.91	0.54	3.84	0.85	4.22	0.74
11～20	327	37.9%	3.41	0.86	5.09	0.57	4.88	0.59	3.92	0.82	4.41	0.82
21～30	229	26.5%	3.45	0.81	5.12	0.54	4.88	0.59	3.94	0.80	4.22	0.73
31以上	34	3.9%	3.65	0.60	5.07	0.53	4.86	0.57	3.90	0.88	3.96	0.92

堂有民主自由的氛圍，以激發學生思考和互動」。每題採六點量尺，由「1 = 非常不同意」到「6 = 非常同意」，讓參與者依據自己對於教學與學習的看法對每個項目的敘述進行同意度的評價。

(二)評量目的的知識

教師對評量目的的知識(以下簡稱「評量目的」)題項是依據Abell與Siegel (2011)理論中關於評量目的的定義與其研究在科學教育課堂中的例子進行命題。針對診斷性評量，例題為：「評量幫助教師瞭解學生的科學背景知識」；而針對形成性評量，例題為：「在學習過程中，評量提供學生有關他們科學表現的回饋」。此分量表共13題，採六點量尺，由「1 = 非常不同意」到「6 = 非常同意」，由教師就自己執行科學評量的目的，選擇對題目敘述的同意程度。

(三)評量內容的知識

由於評量內容的知識(以下簡稱「評量內容」)涉及不同面向，並期望反映學科的脈絡特徵，因此題項需考量中學科學教師在科學課室中，欲評量的內容和項目。本研究參考相關文獻進行題項之設計，將Talanquer等(2013)、Grob等(2021)、Wang等(2010)的研究題項進行整合，這些研究分別探索職前或在職科學教師在科學課程、探究教學等不同脈絡下的評量內容，並能兼顧多元的學習表現。題項內容主要考量評量內容需包含：科學實務、互動溝通與自主學習、及態度與科學本質等評量內容，共設計16題。每題以六點量尺，由「1 = 不曾出現」到「6 = 總是出現」，由教師針對自己執行科學評量時，預期測得的學習表現之頻率進行勾選，如：「記憶科學知識與概念」、「計算或解題」和「認識科學本質」等。

(四)評量策略的知識

科學教師進行評量時，會以多樣化資料類型判斷學生的學習表現，因此，評量資料類型之題項參考Siegel與Wissehr (2011)、Grob等(2021)、Wang等(2010)對科學教育評量策略以及McMillan (2001)對中學教師的評量策略。這些研究涵蓋多元的科學教學脈絡，包含探究教學及一般科學課程、可能是總結或形成性評量，有傳統常見(如：試卷)及創新的資料類型(如：同儕、自我評量)來彙整科學教師常使用的評量型態，共17題，如：「教師自行命題或組題的試卷」、「學習單」、和「同學生上臺報告或辯論會的口語資料」等。每題為六點量尺，由「1 = 不曾使用」到「6 = 約每週二次或二次以上」，由教師就自己針對一個班級，使用不同類型資料來評量學生表現之頻率。

(五)評量詮釋與行動

評量詮釋與行動主要是測量除了計分之外，如何使用評量結果計畫教學以及幫助學生做出決定 (AFT et al., 1990)，題項是參考DeLuca等(2016a)的教室評量取徑問卷的第二部分，包括評量計分、評量回饋等，共設計9題。每題為六點量尺，由「1 = 不曾出現」到「6 = 總是出現」，由教師就自己執行科學評量後，對題目敘述的行動會採取的頻率選擇最適當的選項進行勾選。例題有：「使用評量證據促進學生的科學學習」、「提供充分、適當的訊息，讓學生與家長瞭解科學評量回饋和評分的含義」。

三、資料分析

為達到本研究目的，資料分析係針對教師評量素養量表的題項，進行信效度的檢驗與恆等性的分析。為確立量表信度，本研

究使用預試階段的資料，以項目分析進行極端組比較、相關性檢測與同質性檢驗，初步檢視量表各題項之適切性，以確認量表之題項是否具有基本品質。而效度的檢驗是透過探索性因素分析，檢視量表之因素結構，是否符合Abell與Siegel (2011)模型之五個理論構念。接著使用正式施測的資料進行效度分析，以驗證性因素分析進行構念信度、聚合效度與區辨效度等驗證，並搭配恆等性分析，來確認量表之因素結構在不同性別與教育階段的群體中是否穩定。

肆、研究結果

在本研究的量表信效度驗證上，考量「評量策略」一項並非詢問教師同意度或知覺的程度，而採教師在實務中特定策略的使用頻率來測量，因此使用「對學習的觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」作為主要的構念進行信效度驗證，並進行相關的信效度分析。最後針對相關架構進行性別與教育階段的恆等性分析。

一、信度與項目分析

本研究的信效度分析使用預試資料進行(N = 100)，採用項目分析之標準進行初期的量表題項分析，並參考吳明隆(2007)的觀點採用標準為極端決斷值應大於3。題項與總分相關以及效正題項與總分相關，應高於.4；每題項的題項刪除後α值應小於維度的α值；因素符合量高於.45為佳，而共同性係數大於.2之題項等標準進行項目分析，以下依各構念分別述明。

針對教師「對學習的觀點」如表2所示，在傳統觀點的項目分析結果顯示，決斷值介於6.69 ~ 11.00之間，高於標準3.00。相關性檢測部分相關係數介於.67 ~ .81，校正項目總分相關係數介於.50 ~ .70，兩者所有題目均高於.40之標準，且皆達顯著水準。同質性檢測部分，題項刪除後的α值介於.79 ~ .83皆小於構念信度.83；共同性介於.41~ .66，高於標準.20；因素負荷量介於.64 ~ .81，高於標準.45，因此在傳統觀點的部分，項目分析6題全部保留。在建構觀點的項目分析結果顯示，決斷值介於4.35 ~ 10.89之間，皆高於標準3.00。相關性檢測部分相關係數介於.54 ~ .79，校正項目總分相關係數介於.39 ~ .72，只有建構觀點6略小，但是達顯著水準，因此在項目分析部分予以保留。同質性檢測部分，題項刪除後的α值介於.82 ~ .85皆小於構念信度.86；共同性介於.23~ .67，高於標準.20；因素負荷量介於.48 ~ .82，高於標準.45，因此在建構觀點的部分，項目分析9題全部保留。

在「評量目的」的項目分析結果顯示(表3)，決斷值介於6.44 ~ 11.55之間，相關性檢測部分相關係數介於.67 ~ .82，校正項目總分相關係數介於.60 ~ .78，皆高於標準值。同質性檢測部分，題項刪除後的α值介於.90 ~ .91皆小於構念信度.91；共同性介於.46 ~ .68，因素負荷量介於.68 ~ .83，皆高於標準值。因此在評量目的部分，13題全部保留。

表4呈現「評量內容」構念的項目分析結果。其中兩題，評量內容1 (t = -0.10) 與評量內容2 (t =1.17)，決斷值小於3.00，其餘題目決斷值介於3.54 ~ 13.25之間，皆高於標準3.00。相關性檢測部分，除了評量內容1 (r = .09)、評量內容2 (r = .21) 與評量內容5 (r = .39)相關係數小於.40，其餘相關係數介於.56 ~ .73，校正項目總分相關係數中，除了評量內容1 (r = -.01)、評量內容2 (r = .15)與評量內容5 (r = .30)相關係數小於.40，其餘題項介於.46 ~ .66，均高於.40之標準，且皆達顯著水準。同質性檢測部分，題項刪除後的α

表2：「對學習的觀點」項目分析摘要表

	極端組比較	相關性檢測		同質性檢測				
題項	決斷值	題項與總分相關	校正題項與總分相關	題項刪除後的α值	共同性	因素負荷量	未達標準指標數	備註
標準	≥ 3.00	≥ .40	≥ .40	< .83	≥ .20	≥ .45		
傳統觀點1	8.42	.67	.50	.83	.41	.64		
傳統觀點2	9.05	.72	.60	.81	.54	.74		
傳統觀點3	11.00	.81	.70	.79	.66	.81		
傳統觀點4	6.96	.70	.55	.82	.48	.70		
傳統觀點5	10.51	.77	.65	.80	.61	.78		
傳統觀點6	9.61	.78	.65	.80	.60	.78		
標準	≥ 3.00	≥ .40	≥ .40	< .86	≥ .20	≥ .45		
建構觀點1	5.16	.54	.40	.85	.23	.48		
建構觀點2	7.71	.74	.64	.82	.50	.71		
建構觀點3	9.42	.68	.58	.83	.49	.70		
建構觀點4	9.43	.76	.66	.82	.61	.78		
建構觀點5	10.89	.79	.72	.82	.67	.82		
建構觀點6	4.35	.54	.39	.85	.23	.48		
建構觀點7	5.30	.65	.55	.83	.44	.67		
建構觀點8	8.92	.76	.68	.82	.63	.79		
建構觀點9	5.19	.64	.54	.84	.44	.66		

表3：「評量目的」項目分析摘要表

	極端組比較	相關性檢測		同質性檢測				
題項	決斷值	題項與總分相關	校正題項與總分相關	題項刪除後的α值	共同性	因素負荷量	未達標準指標數	備註
標準	≥ 3.00	≥ .40	≥ .40	≥ .91	≥ .20	≥ .45		
評量目的1	7.89	.71	.64	.90	.54	.73		
評量目的2	7.49	.71	.66	.90	.54	.74		
評量目的3	6.44	.67	.60	.90	.46	.68		
評量目的4	7.34	.69	.63	.90	.47	.69		
評量目的5	9.69	.82	.78	.90	.68	.83		
評量目的6	7.21	.75	.68	.90	.55	.74		
評量目的7	11.55	.80	.76	.90	.66	.81		
評量目的8	6.64	.67	.60	.90	.43	.66		
評量目的9	5.07	.62	.55	.91	.35	.60		
評量目的10	7.44	.67	.61	.90	.44	.66		
評量目的11	4.68	.57	.46	.91	.27	.52		
評量目的12	10.09	.75	.70	.90	.58	.76		
評量目的13	9.04	.74	.68	.90	.54	.74		

表4：「評量內容」項目分析摘要表

	極端組比較	相關性檢測		同質性檢測				
題項	決斷值	題項與總分相關	校正題項與總分相關	題項刪除後的α值	共同性	因素負荷量	未達標準指標數	備註
標準	≥ 3.00	≥ .40	≥ .40	≥ .85	≥ .20	≥ .45		
評量內容1	-0.10	.09	-.01	.87	.00	-.07	6	刪除
評量內容2	1.17	.21	.15	.86	.01	.11	6	刪除
評量內容3	6.57	.56	.51	.85	.31	.56		
評量內容4	8.16	.61	.54	.84	.41	.64		
評量內容5	3.54	.39	.30	.86	.09	.30	5	刪除
評量內容6	7.03	.57	.49	.85	.34	.59		
評量內容7	8.46	.65	.58	.84	.52	.72		
評量內容8	6.15	.56	.48	.85	.30	.55		
評量內容9	13.25	.73	.66	.84	.59	.77		
評量內容10	7.40	.62	.54	.84	.41	.64		
評量內容11	8.68	.70	.63	.84	.51	.72		
評量內容12	8.33	.66	.58	.84	.46	.68		
評量內容13	9.76	.66	.58	.84	.46	.68		
評量內容14	7.69	.66	.58	.84	.41	.64		
評量內容15	5.06	.56	.46	.85	.26	.51		
評量內容16	6.91	.62	.53	.84	.40	.63		

值，除了評量內容1 (α = .87)、評量內容2 (α= .86)與評量內容5 (α = .86)刪除後的α值會高於構念信度.85之外，其餘題項介於.84 ~ .85皆小於構念信度.85；在共同性部分，除了評量內容1 (.00)、評量內容2 (.01)與評量內容5 (.09)共同性會小於標準.20之外，其餘題項介於.30 ~ .59之間高於標準.20；在因素負荷量部分，除了評量內容1 (λ = -.07)、評量內容2 (λ = .11)與評量內容5 (λ = .30)負荷量會小於標準.45之外，其餘題項介於.51 ~ .77，高於標準.45，因此在評量目的部分，考量刪除多項指標不合格的評量內容1、評量內容2與評量內容5，保留13題。

在「評量詮釋與行動」的項目分析結果顯示(表5)，所有題目的決斷值介於5.32 ~ 10.55之間，皆高於標準3.00。相關性檢測部分，除了評量詮釋與行動9 (*r* = .39)相關係數小於.40，其餘相關係數介於.56 ~ .75，校正項目總分相關係數中，除了評量詮釋與行動1 (*r* = -.35)、評量詮釋與行動7 (*r* = .37)、評量詮釋與行動8 (*r* = .39)與評量詮釋與行動9 (*r* = .17)相關係數小於.40，其餘題項介於.46 ~ .65，均高於.40之標準，且皆達顯著水準。同質性檢測部分，題項刪除後的α值，除了評量詮釋與行動9 (α = .83)刪除後的α值會高於構念信度.82之外，其餘題項介於.78 ~ .80皆小於構念信度.82；在共同性部分，除了評量詮釋與行動9 (.08)共同性會小於標準.20之外，其餘題項介於.34 ~ .60之間高於標準.20；在因素負荷量部分，除了評量詮釋與行動9 (λ = .28)負荷量會小於標準.45之外，其餘題項介於.58 ~ .78，高於標準.45，因此在評量目的部分，考量刪除多項指標不合格評量詮釋與行動9，保留8題。

表5：「評量詮釋與行動」項目分析摘要表

	極端組比較	相關性檢測		同質性檢測				
題項	決斷值	題項與總分相關	校正題項與總分相關	題項刪除後的α值	共同性	因素負荷量	未達標準指標數	備註
標準	≥ 3.00	≥ .40	≥ .40	< .82	≥ .20	≥ .45		
評量詮釋與行動1	5.32	.56	.35	.80	.34	.58	1	
評量詮釋與行動2	7.32	.66	.46	.80	.40	.63		
評量詮釋與行動3	10.55	.75	.51	.78	.54	.73		
評量詮釋與行動4	6.85	.66	.63	.79	.55	.74		
評量詮釋與行動5	7.50	.71	.65	.79	.60	.78		
評量詮釋與行動6	9.84	.73	.54	.78	.57	.75		
評量詮釋與行動7	6.63	.66	.37	.79	.42	.65	1	
評量詮釋與行動8	8.08	.66	.39	.80	.35	.60	1	
評量詮釋與行動9	3.56	.39	.17	.83	.08	.28	5	刪除

二、效度分析

(一)探索性因素分析

考量探索性因素分析所需之樣本數，根據Kline (2014)若因素結構穩定則至少需100位樣本；再者，Arrindell與van der Ende (1985)建議每因素向度需有20個樣本，本研究架構共有「傳統觀點」、「建構觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」等5個因素向度，應具有100個樣本。綜合以上建議，因此本研究使用預試資料(N = 100)進行探索性因素分析作理論初步結構的探索，將項目分析後保留下來的49題進行因素分析。本研究採用最大變異法進行因素分析，抽取5個因素，以分析量表因素結構。其中教師「對學習的觀點」可分為兩項因素：建構觀點和傳統觀點，其他三項因素分別為評量目的、評量內容、和評量詮釋與行動。在各題項中，建構觀點6未落入預設的層面中，評量詮釋與行動1因素負荷量過低，而評量內容15、評量詮釋與行動2和8，有跨因素負荷的情況故予以刪除。刪題後將剩下的44題再進行一次因素分析，所有題項的因素負荷量皆在.45以上，且仍維持在原先所設定的因素層面中，5個因素共可解釋總變異量的53.42%，顯示此量表具有良好的構念效度，因此全部題項共44題皆予以保留下來，結果如表6所示。

(二)驗證性因素分析

正式施測的階段(N = 872)，本研究首先採用結構方程模式進行問卷的驗證性因素分析，並參考吳明隆(2009)的觀點，從基本適配度、整體模式適配度及模式內在品質檢定，來進行本量表一階驗證性因素分析模式的檢定，以確立本研究問卷的構念效度。在驗證性因素分析的初步分析當中，因建構觀點1 (λ = .48)、評量內容3 (λ = .44)與評量行動5 (λ = .34)等題項的因素負荷量過低先進行刪除。至此分析階段，整份量表為41題。

從驗證性因素分析結果來看，在適配指標方面卡方檢定值為2,269.63 (p < .001)，顯示樣本共變數矩陣與理論模式共變數矩陣之間未達適配程度，但是卡方值極易受樣本數影響；因此需進一步參照其他指標評鑑模式之

表6：因素分析摘要表(刪題後)

項目	因素負荷矩陣				
	評量目的	評量內容	建構觀點	傳統觀點	評量行動
傳統觀點1	-.17	.11	.15	**.61**	.07
傳統觀點2	.01	.20	.03	**.70**	.01
傳統觀點3	.01	.18	.04	**.79**	.02
傳統觀點4	.15	-.01	-.03	**.66**	.09
傳統觀點5	-.05	.22	-.29	**.70**	.08
傳統觀點6	.16	.17	-.17	**.71**	.03
建構觀點1	.18	.10	**.46**	.32	.00
建構觀點2	.21	.09	**.65**	.18	-.05
建構觀點3	.10	.06	**.67**	.11	.14
建構觀點4	.02	.08	**.78**	-.09	.21
建構觀點5	.26	.20	**.75**	-.08	.10
建構觀點7	.14	.03	**.66**	-.06	.02
建構觀點8	.25	.01	**.77**	-.25	.07
建構觀點9	.16	.04	**.63**	-.14	.05
評量目的1	**.72**	.01	.16	-.06	-.05
評量目的2	**.72**	-.05	.20	-.01	-.02
評量目的3	**.62**	.13	.28	-.32	.10
評量目的4	**.63**	.06	.22	-.23	.09
評量目的5	**.82**	.08	.16	-.01	.02
評量目的6	**.69**	.09	.10	.00	.20
評量目的7	**.77**	.04	.31	-.15	.04
評量目的8	**.67**	.18	-.01	.06	.06
評量目的9	**.56**	-.09	.11	.06	.27
評量目的10	**.68**	-.03	-.06	.27	.17
評量目的11	**.50**	-.12	.09	.20	.20
評量目的12	**.77**	-.06	.12	.06	.02
評量目的13	**.73**	.08	.07	.21	.12
評量內容3	.25	**.54**	.19	.04	-.10
評量內容4	-.08	**.67**	.01	.20	.09
評量內容6	-.01	**.57**	.08	.09	.09
評量內容7	-.03	**.71**	.08	.06	.17
評量內容8	.19	**.54**	-.04	-.09	.00
評量內容9	-.03	**.79**	-.01	.10	.11
評量內容10	.01	**.65**	.01	-.10	.24
評量內容11	-.02	**.68**	.17	.13	.15
評量內容12	-.02	**.68**	.14	.15	-.10
評量內容13	-.03	**.67**	.03	.19	-.11

表6：因素分析摘要表(刪題後) (續)

項目	因素負荷矩陣				
	評量目的	評量內容	建構觀點	傳統觀點	評量行動
評量內容14	.16	**.59**	.03	.13	.14
評量內容16	-.06	**.59**	-.01	.08	.16
評量詮釋與行動3	.09	.35	.14	.12	**.55**
評量詮釋與行動4	.24	.02	.05	.10	**.77**
評量詮釋與行動5	.13	.14	.19	.08	**.78**
評量詮釋與行動6	.13	.13	.05	-.04	**.81**
評量詮釋與行動7	.15	.25	.08	.07	**.64**
特徵值	6.76	5.52	4.36	3.74	3.12
變異量	15.37%	12.55%	9.92%	8.50%	7.09%
累積解釋變異量	53.42				
KMO取樣適切性量數	.72				
Bartlett球形檢定	2762.62 ($p < .001$)				

註：因素負荷量 > .45，以粗體表示。

適配程度，由於均方根近似誤差(Root Mean Square Error of Approximation, RMSEA) (.05)、標準化均方根殘差(Standardized Root Mean Square Residual, SRMR) (.05)、配適度指數(Goodness of Fit Index, GFI) (.87)與比較配適度指數(Comparative Fit Index, CFI) (.91)之結果，多落於適配標準之邊緣或低於標準(如：GFI = .87 < .90)，顯示模型適配度尚有修正空間，因此本研究進一步運用修正指標(Modification Index, MI)採題項刪除法，進行模式修飾，並依MI值的大小依序檢視各題與量表構念之MI值。模式修飾過程逐步刪除評量目的12、評量目的2、評量目的11、評量目的8、評量內容7及評量目的10等題項。最終量表共計35題獲得適配度，RMSEA為.05、SRMR為.05、GFI值為.91、CFI值為.92，顯示本模式之適配度已達良好，適配指標如表7所示。

表8列出最終版本之35題，且呈現量表之各潛在變項內「傳統觀點」、「建構觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」之觀察變項的因素負荷量介於.53 ~ .80之間，皆大於.50之指標信度良好標準，構念信度為.82、.84、.90、88與.76，數值皆高於.60構念信度良好之標準，故各潛在變項的觀察變項之解釋變異量有達理想值。平均變異抽取量為.42 ~ .52之間，其值並沒有達到判別標準之臨界值.50，顯示聚斂效度並不佳。然而，依據Fornell與Larcker (1981)的看法，即使平均抽取變異量未達50%，但若是單獨以構念信度為基礎，研究者也可以做出適當的構念聚合效度。由於本量表中每項因素的構念信度皆達到.60的標準值(詳見表8)，因此本研究認為此量表仍具有構念效度。換言之，科學教師評量素養量表之效度考驗，由各個觀察變項與其潛在變項間之標準化因素負荷量可知，所有觀察變項之因素負荷量，均達顯著水準，故這些觀察變項皆能有效地反應出所要測量的潛在變項特質。

在區別效度檢定方面(表9)，本研究先

表7：驗證性因素分析之結果摘要表

評鑑項目	適配的標準	檢定結果數據	模式適配判斷
基本適配度			
是否沒有負的誤差變異數	> 0	無	適配
因素負荷量介於0.50至0.95之間	.50 ~ .95	.53 ~ .80	適配
是否沒有很大的標準誤	不能太大	.04 ~ .09	適配
整體模式適配度			
χ^2值	$p > .05$	1516.22 ($p < .001$)	不適配
χ^2自由度比	< 4.00	2.76	適配
GFI值	> .90	.91	可接受
AGFI值	> .85	.89	可接受
RMR值	< .05	.05	適配
SRMR值	< .05	.05	適配
RMSEA值	< .08	.05	適配
NFI值	> .90	.92	適配
CFI值	> .90	.92	適配
CN值	> 200	348	適配
模式內在品質			
所有估計的參數均達顯著水準	t值 > 1.96	15.56 ~ 23.66	適配
潛在變項的組合信度	> 0.60	.92 ~ .76	適配
潛在變項的平均變異抽取量	> 0.50	.43 ~ .49	不適配

註：GFI：配適度指數(Goodness of Fit Index)；AGFI：調整之配適度指數(Adjusted Goodness of Fit Index)；RMR：均方根殘差(Root Mean Square Residual)；SRMR：標準化均方根殘差(Standardized Root Mean Square Residual)；RMSEA：均方根近似誤差(Root Mean Square Error of Approximation)；NFI：標準配適度指數(Normed Fit Index)；CFI：比較配適度指數(Comparative Fit Index)；CN：臨界值(Critical Number)。

由每一個構念的平均變異萃取量(Average Variance Extracted, AVE)均方根大於各構念的相關係數之個數，且至少須佔整體的比較個數75%以上(Hair et al., 2019)。其分析結果顯示除了「建構觀點」與「評量目的」之相關高於其AVE平方根外，各構念皆滿足判別準則。近一步針對證明「建構觀點」與「評量目的」之相關，進行信賴區間分析，其信賴區間為[.67, .78]，顯示構念間雖具相關性但卻非相同之因素，亦即具有區別效度。

(三)恆等性分析

在進行恆等性分析前，本研究先進行4次的獨立驗證式因素分析，分別驗證獨立模型的適配度，再進行跨群組分析，以驗證科學教師的評量素養因素之間在跨群組上具有測量恆等的特性。分析結果發現各模式具有不錯的適配性，多數適配指標都有達到理想標準(表10)，RMSEA也在.05至.06之間，雖未達嚴格標準值.05但可接受，CFI值只有高中階段教師模型略低於.90之標準，但仍是合理適配，因此可進行下階段跨群組巢套模式恆等性的檢定程序。

在測量恆等性的分析上，Vandenberg與Lance (2000)認為欲檢測驗證性因素分析的測量不變性，必須經由一系列的巢套模型。本研究依下列巢套模式進行分析，分別檢證如

表8：觀察變項因素負荷量與潛在變項之信度與平均變異數抽取量摘要表

潛在變項	觀察變項	因素負荷量	構念信度	平均變異數抽取量
傳統觀點	傳統觀點1	.62	.82	.43
	傳統觀點2	.64		
	傳統觀點3	.67		
	傳統觀點4	.68		
	傳統觀點5	.66		
	傳統觀點6	.67		
建構觀點	建構觀點2	.61	.84	.43
	建構觀點3	.66		
	建構觀點4	.67		
	建構觀點5	.68		
	建構觀點7	.64		
	建構觀點8	.65		
	建構觀點9	.68		
評量目的	評量目的1	.76	.90	.52
	評量目的3	.69		
	評量目的4	.80		
	評量目的5	.78		
	評量目的6	.65		
	評量目的7	.77		
	評量目的9	.63		
	評量目的13	.68		
評量內容	評量內容4	.55	.88	.42
	評量內容6	.58		
	評量內容8	.60		
	評量內容9	.74		
	評量內容10	.72		
	評量內容11	.79		
	評量內容12	.65		
	評量內容13	.62		
	評量內容14	.53		
	評量內容16	.67		
評量詮釋與行動	評量行動3	.69	.76	.44
	評量行動4	.65		
	評量行動5	.73		
	評量行動6	.59		

表9：區別效度摘要表

變數名稱	傳統觀點	建構觀點	評量目的	評量內容	評量行動
傳統觀點	**.66**				
建構觀點	-.09*	**.66**			
評量目的	.07	.73***	**.72**		
評量內容	-.17***	.33***	.30***	**.66**	
評量詮釋與行動	.12**	.41***	.42***	.66***	**.66**

註：1.正對角線之數值(粗體部分)即代表平均變異萃取量之均方根。
2.*p < .05, **p < .01, ***p < .001。

表10：各模式驗證性摘要表

群組	χ^2	*df*	RMSEA	GFI	CFI	TLI	SRMR
全樣本(N = 872)	1516.22	550	.05	.91	.92	.92	.05
男性(n = 462)	1207.72	550	.05	.87	.91	.91	.06
女性(n = 410)	1000.08	550	.05	.88	.91	.90	.05
國中階段(n = 519)	1122.73	550	.05	.89	.93	.92	.05
高中階段(n = 353)	1132.06	550	.06	.84	.88	.87	.06

註：TLI：Tucker-Lewis適配度指數(Tucker-Lewis Index)。

下：模式1：結構相等；模式2：一階因素負荷量相等；模式3：潛在因素共變異相等；模式4：測量變項誤差相等。因卡方差異值很容易受到樣本數大小而波動，容易造成兩個原本沒有差異量的模式變得有顯著差異存在(Cheung & Rensvold, 2002)。因此，本研究為避免樣本數之影響，在恆等性分析的指標上使用Cheung與Rensvold所建議ΔCFI (< .01)與ΔTucker-Lewis Index (TLI, < .01)等兩項標準作為模式的取決依據。

表11為性別多群體巢套模式之比較(nested model comparisons)。此巢套模式之比較係由基底模式(MG1)來逐一加上限制「一階因素負荷量」相同(MG2)、限制「潛在變項共變異數」相同(MG3)以及限制「觀察變項測量誤差」(MG4)共4個模式之適合度以進行比較。

模式MG2為一階因素負荷量相等之模型，此模型將所有一階因素的負荷量皆限定為等同，MG2與MG1的$\Delta\chi^2$ = 48.53 (Δdf = 30)高於臨界值，顯示性別可能在因素負荷量上有所差異。近一步透過多群組雙參數檢定發現，主要的差異源自於評量內容的題項，如：評量內容4 (z = 2.61)、評量內容6 (z = 3.18)、評量內容8 (z = 2.20)、評量內容9 (z = 2.77)、評量內容10 (z = 2.44)、評量內容12 (z = 2.15)、評量內容13 (z = 3.04)與評量內容14 (z = 2.79)。但χ^2極易受到樣本數影響，容易造成兩個原本沒有差異量的模式變得有顯著差異存在(Cheung & Rensvold, 2002)。因此本研究採用Cheung與Rensvold所建議ΔCFI (< .01)與ΔTLI (< .01)等兩項標準作為模式的取決依據。本研究檢視MG2與MG1在ΔTLI = .001，ΔCFI = -.002顯示模型並沒有顯著的改變，所以依據Cheung與Rensvold的說法，可認定此模型並無顯著的差異。此結果顯示科學教師評量素養中，每項測量因素負荷量在性別因素上大致上具有不變性，即代表性別的結構

表11：性別恆等性巢套模式比較表

Model	χ^2	*df*	RMSEA	TLI	CFI
MG1	2207.78	1100	.034	.904	.911
MG2	2256.34	1130	.034	.904	.909
	$\Delta\chi^2 = 48.53$，$p = .017$	$\Delta df = 30$		ΔTLI = .001	ΔCFI = -.002
MG3	2305.45	1145	.034	.903	.907
	$\Delta\chi^2 = 49.10$，$p < .001$	$\Delta df = 15$		ΔTLI = .001	ΔCFI = -.002
MG4	2370.74	1180	.034	.903	.904
	$\Delta\chi^2 = 65.29$，$p < .001$	$\Delta df = 35$		ΔTLI < .001	ΔCFI = -.003

上個體在潛在變項上的每一題解釋力可視為恆等。將所有的因素負荷量皆設為恆等，和結構恆等的模式作比較，可發現恆等後，男女教師的科學素養評量因素之解釋力，並沒有太大的差異，但是仍不排除性別在評量內容的應用上可能有一定差異。

模式MG3為潛在變項共變異數之模型，此模型將所有潛在變項共變異數的相關皆限定為等同，此時一階因素負荷量與潛在變項共變異數皆限定等同，MG3與MG2的$\Delta\chi^2$ = 49.1 ($\Delta df = 15$，$p < .001$)高於臨界值，顯示性別可能在潛在變項共變異數上有些許的差異，此部分的差異主要源自於評量目的與評量內容($z = 2.34$)、評量內容與評量詮釋與行動(z = 4.47)、傳統觀點與評量詮釋與行動($z = 2.19$)、與評量目的與傳統觀點($z = 2.15$)。但近一步探究ΔTLI = .001，ΔCFI為-.002顯示在模式上並沒有產生很大的改變，潛在變項共變異數在性別上可視為是相同的，也就是代表五項潛在因素在男教師與女教師群組之間的相關係數是相同的，但是仍不可排除性別在相關性上可能存在一定差異。

模式MG4檢測男教師與女教師的測量變項誤差相等是否相同，其模式是建立在MG3模型下，除了因素負荷量與潛在變項共變異數外，再加上測量變項誤差相同的假設，以確定在不同性別的結構下各題的誤差是否有所不同。MG4與MG3的$\Delta\chi^2 = 65.29$ ($\Delta df = 35$，$p < .001$)高於臨界值，顯示性別可能在誤差上有些許的差異，但考量ΔTLI < .001，ΔCFI = -.003顯示在模式上並沒有產生差異。透過MG1到MG4的比較可以發現，本量表在測量不同的性別結構上可視為一個具有穩定的測量模式，但是性別在評量內容的使用上有差異的可能。

表12為國高中教師多群體巢套模式之比較。此巢套模式設定與性別多群組相同。模式ME1主要是測量不同教育階段教師在評量素養上的結構不變性，顯示在不同教育階段模型中，國中教師與高中教師同樣具有五項主要的潛在因素，每項潛在因素下也負荷了相同的題項，模式的參數顯示模式ME1之結構不變性具有可接受的適配，因此可以繼續進行恆等性的測量。

接來的模式ME2為一階因素負荷量相等之模型，分析結果發現模型ME2與模型ME1在$\Delta\chi^2 = 40.97$ ($\Delta df = 30$，$p = .087$)，ΔTLI = .002，ΔCFI = -.001顯示模型並無顯著的改變。此結果代表在不同教育階段的教師，其評量素養在結構上每一題解釋力皆相等。若將所有的因素負荷量皆設為恆等，來和結構恆等的模式作比較，可以發現恆等後的各係數，大致上在不同教育階段教師之間，對於評量素養的因素構念解釋力，並沒有太大的差異。

表12：不同教育階段教師之恆等性巢套模式比較表

Model	χ^2	*df*	RMSEA	TLI	CFI
ME1	2255.00	1100	.035	.900	.908
ME2	2295.00	1130	.034	.902	.907
	$\Delta\chi^2 = 40.97$，$p = .087$	$\Delta df = 30$		ΔTLI = .002	ΔCFI = -.001
ME3	2313.39	1145	.034	.903	.907
	$\Delta\chi^2 = 19.42$，$p = .294$	$\Delta df = 15$		ΔTLI = .001	ΔCFI = -.000
ME4	2422.00	1180	.034	.900	.901
	$\Delta\chi^2 = 108.60$，$p < .001$	$\Delta df = 35$		ΔTLI = .003	ΔCFI = -.006

模式ME3為潛在變項共變異數之模型，此模型將一階因素負荷量與潛在變項共變異數皆限定等同，模型ME3與模型ME2在$\Delta\chi^2 = 19.42$ ($\Delta df = 15$，$p = .294$)，ΔTLI = .001，ΔCFI為.000顯示在模式上並無產生很大的改變。結果顯示潛在變項共變異數在性別的結構上是相同的，也就是代表潛在因素在國中教師與高中教師不同群組之間的相關係數是相同的。

模式ME4檢測國中教師與高中教師的測量變項誤差相等是否相同，其模式是建立在ME3模型下，再加上測量變項誤差相同的假設，已確定在不同教育階段模型結構下各題的誤差是否有所不同。ME4與ME3的$\Delta\chi^2 = 108.60$ ($\Delta df = 35$，$p < .001$)高於臨界值，顯示教育階段可能在誤差上有些許的差異，但檢視ΔTLI = .003，ΔCFI = -.006仍可視為教育階段在誤差模式上並沒有產生差異。透過以上ME1到ME4的比較可以發現，本量表在測量不同教育階段的科學教師，其評量素養在結構上同樣具有穩定的測量模式。

伍、討論與結論

本研究透過Abell與Siegel (2011)的理論架構，提出「對學習的觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」，嘗試建立科學教師的評量素養架構，接著經過題項的收集、設計、預試、項目分析、因素分析、信效度考驗與恆等性分析等步驟，編製出適用於中學科學教師評量素養的量表並考驗其穩定性。研究結果不僅能夠提供有效的科學教師評量素養之測量工具，亦擴展教育者和研究人員對科學教師評量素養架構和成分的理解，而且對於評量素養的未來研究和評量實務開闢若干途徑。

評量素養是整個教育系統的核心專業要求，因此衡量和支持教師的評量素養一直是過去二十年來相關研究的主要關注點(DeLuca et al., 2016b)。本研究與過去所發展的量表較大的不同在於，過去研究所發展的量表大多都是以AFT等(1990)所提出的教育政策評量能力標準，但是以政策導向出發的能力標準，不一定能夠符合學科需求，而且本研究之架構涵蓋「對學習的觀點」、「評量目的」等更多相關概念，較能貼近科學教師的需求。本研究同時提供Abell與Siegel (2011)的評量素養模型理論應用在臺灣地區科學教師的實證基礎，並驗證了該架構下評量素養量表具穩定的結構，而結構是理論模型的基礎。

根據跨群組恆等性檢驗之結果，可以發現Abell與Siegel (2011)的理論架構，與其衡量題項之間的關連強度穩定，不會因為測量高中與國中階段的男、女科學教師而有所差異，因此這些題項在測量國高中的男、女科

學教師生具有等同的效度。換言之，不同群體的科學教師是以相同的看法來評量自己在「對學習的觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」等行為上的表現，他們在所有題項的作答反應及其因素關係可以視為是相同的，但是在男、女科學教師在「評量內容」方面，可能的差異與變項間關聯性仍需進一步考量。但由區別效度中的相關矩陣可以發現，在學習的觀點中，傳統觀點與其他變項的相關性較低，這可能是對於學習、教學、教師及學生角色的概念會隨著時代的思潮(例：行為主義、社會文化理論)而改變(Abell & Siegel)，因此，隨著時代的改變，未來可能需進一步考量Abell與Siegel評量素養理論架構中傳統學習觀點的定位。

整體來說，本研究所發展之科學教師評量素養量表應可穩定有效測量高中與國中階段的男、女科學教師在評量素養當中的「對學習的觀點」、「評量目的」、「評量內容」與「評量詮釋與行動」，可作為測量科學教師評量素養之參考。雖然這項研究為科學教師的評量素養提供了重要的工具，但這項研究存在一些局限性。第一個限制是，雖然本研究抽樣方式是基於學校的PR值以及學校的大小，針對北部地區的中學教師做了多樣化的考量，但若在不同地區和教育階段，能夠招募更多樣化的教師樣本，可以為研究提供更豐富的資訊，並為科學教師的評量素養描繪出更精確的圖像。另外，為進一步檢驗本研究發展之量表之有效性和實用性，茲提出以下建議供未來研究之參考。

一、量表編製方面

在量表編制方面，可以思考是否增加效標參照指標，除了基於Abell與Siegel (2011)考量對於教師的「對學習的觀點」外，還有許多對於評量素養可能具有相關性的因素，是未來研究需要考量的，如：教師態度(Quilter & Gallini, 2000)。教師對評量和測驗的態度，最終會影響他們使用和詮釋各種測試結果的能力。如果教師對教育評量中的特定方法或策略產生消極態度，他們將來不太可能理解或使用這些方法 (Quilter & Gallini)。

另外，未來的研究可考慮招募來自不同教育階段(如：高等教育或國小階段)的教師，或比較不同科學學科(如：化學、生物和地球科學)的教師資料，以驗證量表。不同階段或學科的科學教師在執行評量素養上可能會有一部分的共性，這部分的測量是否穩定還值得探討。另外，不同階段或目在評量素養的特徵上可能也會有一部分的差距，如：相比於物理和化學，地球科學的教學涉及較少實驗操作的評量。如果有必要，它仍然可以在未來的研究中進行調整，以滿足基於學科的需求。

二、實務應用層面

在實務應用層面，應該重視科學教師評量素養的獨特性，科學教育領域的學科特徵會導致在評量上與其他學科有差異。所以在探討科學教師評量素養的研究，也應該探索「共同核心」和「特定學科」評量知識的關係，以期建立不同學科所特有的評量素養(Xu & Brown, 2016)。沒有一種單一的評量方法可以滿足所有評量需求的目的。不同的目的需要不同種類的訊息，因此需要不同種類的評量方式，教師有必要理解和運用多元評量。此外，課堂的評量需要關注個別學生，而政策級別的大型評量則需要總結大量學生的學習成就指標(Stiggins, 1995)，不同層次和尺度的評量也可能有賴教師不同的評量素養。更多的研究可以識別特定學科、特定層次的評

量素養之相關議題，並提供補充方法，以期更妥善地實施專業標準或政策。

另外，評量素養具動態和持續發展的特性。教師的評量素養是動態不斷發展的系統(Xu & Brown, 2016)。教師在應用評量進行指導學生、調整教學與開發課程的情境下為了與國家評量標準接軌(Brookhart, 2011)，使得教師常因應國家評量標準需求運用不同的評量知識與能力，所以多年來，評量素養在知識與能力的定義與標準不斷被學者們重新塑造和擴大(Yan & Pastore, 2022)。而且，除了評量素養的定義與標準之外，相關理論不斷發展，研究構念的數量也隨時間變化(Dyer, 2015)，這些因素使得評量素養的定義和架構不斷變化，因此對於評量素養的研究就必須時常更新，以便切合當下評量素養在國家政策與研究理論發展的需求。

誌謝

本研究成果承蒙科技部專題研究計畫「高中科學教師對『探究與實作』課程的評量素養(MOST 109-2511-H-003-015-MY3)」與「MOST 111-2811-H-003-010-MY2」經費補助，方能順利進行且完成，特此致上感謝之意。

參考文獻

李俊賢、黃芳銘、李孟芳(2016)。年輕族群男女在消費型態量表的測量恆等性研究。**管理與系統**，**23**(2)，273-302。

[Lee, C.-H., Hwang, F.-M., & Li, M.-F. (2016). Measurement invariance of consumer style inventory across gender on young. *Journal of Management & Systems*, *23*(2), 273-302.]

吳明隆(2007)。**SPSS操作與應用：問卷統計分析實務**。五南。

[Wu, M.-L. (2007). *SPSS operation and application: The practice of quantitative analysis of questionnaire data*. Wu-Nan.]

吳明隆(2009)。**結構方程模式：方法與實務運用**。麗文文化。

[Wu, M.-L. (2009). *Structural equation modeling method and practical application*. Liwen Cultural.]

Abell, S. K., & Siegel, M. A. (2011). Assessment literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The professional knowledge base of science teaching* (pp. 205-221). Springer. https://doi.org/10.1007/978-90-481-3927-9_12

Alkharusi, H. (2011). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia—Social and Behavioral Sciences*, *29*, 1614-1624. https://doi.org/10.1016/j.sbspro.2011.11.404

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice*, *9*(4), 30-32. https://doi.org/10.1111/j.1745-3992.1990.tb00391.x

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement*, *9*(2), 165-178. https://doi.org/10.1177/014662168500900205

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3-12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Chan, K.-W., & Elliott, R. G. (2004). Relational analysis of personal epistemology and conceptions about teaching and learning. *Teaching and Teacher Education*, *20*(8), 817-831. https://doi.org/10.1016/j.tate.2004.09.002

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016a). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment*, *21*(4), 248-266. https://doi.org/10.1080/10627197.2016.1236677

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016b). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251-272. https://doi.org/10.1007/s11092-015-9233-6

Demirdöğen, B., & Korkut, H. M. (2021). Does teacher education matter? Comparison of education and science major teachers' assessment literacy. *Eğitimde Nitel Araştırmalar Dergisi*, *26*, 23-52. https://doi.org/10.14689/enad.26.2

Dyer, W. J. (2015). The vital role of measurement equivalence in family research. *Journal of Family Theory & Review*, *7*(4), 415-431. https://doi.org/10.1111/jftr.12115

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, *18*(3), 382-388. https://doi.org/10.1177%2F002224378101800313

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2), 14-18. https://doi.org/10.1111/emip.12030

Gottheiner, D. M., & Siegel, M. A. (2012). Experienced middle school science teachers' assessment literacy: Investigating knowledge of students' conceptions in genetics and ways to shape instruction. *Journal of Science Teacher Education*, *23*(5), 531-557. https://doi.org/10.1007/s10972-012-9278-z

Grob, R., Holmeier, M., & Labudde, P. (2021). Analysing formal formative assessment activities

in the context of inquiry at primary and upper secondary school in Switzerland. *International Journal of Science Education*, *43*(3), 407-427. https://doi.org/10.1080/09500693.2019.1663453

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Izci, K., & Siegel, M. A. (2019). Investigation of an alternatively certified new high school chemistry teacher's assessment literacy. *International Journal of Education in Mathematics, Science and Technology*, *7*(1), 1-19. https://doi.org/10.18404/ijemst.473605

Jarr, K. A. (2012). *Education practitioners' interpretation and use of assessment results* [Unpublished doctoral dissertation]. University of Iowa.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge. https://doi.org/10.4324/9781315788135

Latif, M. W. (2021). Exploring tertiary EFL practitioners' knowledge base component of assessment literacy: Implications for teacher professional development. *Language Testing in Asia*, *11*(1), Article 19. https://doi.org/10.1186/s40468-021-00130-9

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, sources, and development of pedagogical content knowledge for science teaching. In J. Gess-Newsome & N. G. Lederman (Eds.), *Examining pedagogical content knowledge* (pp. 95-132). Springer. https://doi.org/10.1007/0-306-47217-1_4

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20-32. https://doi.org/10.1111/j.1745-3992.2001.tb00055.x

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. https://eric.ed.gov/?id=ED516494

Ogan-Bekiroglu, F., & Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *The Curriculum Journal*, *25*(3), 344-371. https://doi.org/10.1080/09585176.2014.899916

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, *62*(3), 307-332. https://doi.org/10.3102%2F00346543062003307

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, *84*, 128-138. https://doi.org/10.1016/j.tate.2019.05.003

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*(1), 4-11. https://doi.org/10.1080/00405840802577536

Quilter, S. M., & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, *36*(2), 115-131. https://doi.org/10.1080/08878730009555257

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, *22*(4), 371-391. https://doi.org/10.1007/s10972-

011-9231-6

Starck, J. R., Richards, K. A. R., & O'Neil, K. (2018). A conceptual framework for assessment literacy: Opportunities for physical education teacher education. *Quest*, *70*(4), 519-535. https://doi.org/10.1080/00336297.2018.1465830

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238-245.

Talanquer, V., Tomanek, D., & Novodvorsky, I. (2013). Assessing students' understanding of inquiry: What do prospective science teachers notice? *Journal of Research in Science Teaching*, *50*(2), 189-208. https://doi.org/10.1002/tea.21074

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. O*rganizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177%2F109442810031002

Vogt, K., Tsagari, D., Csépes, I., Green, A., & Sifakis, N. (2020). Linking learners' perspectives on language assessment practices to Teachers' Assessment Literacy Enhancement (TALE): Insights from four European countries. *Language Assessment Quarterly*, *17*(4), 410-433. https://doi.org/10.1080/15434303.2020.1776714

Wang, J.-R., Kao, H.-L., & Lin, S.-W. (2010). Preservice teachers' initial conceptions about assessment of science learning: The coherence with their views of learning science. *Teaching and Teacher Education*, *26*(3), 522-529. https://doi.org/10.1016/j.tate.2009.06.014

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Yan, Z., & Pastore, S. (2022). Are teachers literate in formative assessment? The development and validation of the Teacher Formative Assessment Literacy Scale. *Studies in Educational Evaluation*, *74*, Article 101183. https://doi.org/10.1016/j.stueduc.2022.101183

科學教育學刊
2022, 第三十卷第四期, 309-333
DOI:10.6173/CJSE.202212_30(4).0002

Chinese Journal of Science Education
2022, 30(4), 309-333

Examining the Validity and Measurement Invariance of an Assessment Literacy Inventory for Secondary Science Teachers

Ren-Cheng Zhang[1], Ching-Sui Hung[1,2] and Hsin-Kai Wu[1,3,*]

[1]Graduate Institute of Science Education, National Taiwan Normal University
[2]Taipei Municipal Heping High School
[3]Department of Science and Technology Education, University of Johannesburg

Abstract

The purpose of this study was to develop an inventory for measuring secondary science teachers' assessment literacy and to examine its validity and measurement invariance across gender and educational levels (i.e., junior and senior high schools). The inventory was developed based on a model for science teacher assessment literacy proposed by Abell and Siegel (2011). The inventory was piloted with 100 secondary science teachers in Northern Taiwan. The pilot data were used to conduct the item analysis and exploratory factor analysis. For the main study, data were collected from 872 science teachers teaching at 140 secondary schools in Northern Taiwan. The data collected from the main study were examined by confirmatory factor analysis, and the measurement invariance across gender and educational levels was tested. As a result of the analyses, the inventory was trimmed to 35 items. The results showed good reliability and validity of the items and confirmed the constructs and structure of Abell and Siegel's model. Additionally, the results of measurement invariance testing indicate that this inventory measured the same constructs in the same way across different groups. The factor structure of the model was stable across gender and educational levels.

Key words: Reliability and Validity, Measurement Invariance Analysis, Teacher Assessment Literacy, Instrument Development

* Corresponding author: Hsin-Kai Wu, hkwu@ntnu.edu.tw