以立即反饋系統搭配自動化短答分析提升課堂形成性評量之成效

簡佑達[1, 2, 3, *]、李育賢[4, 5]、李龍豪[6]、曾元顯[6]、李宗諺[7]、張俊彥[7, 8]

[1]國立臺灣海洋大學教育研究所
[2]國立臺灣海洋大學師資培育中心
[3]臺灣海洋教育中心
[4]國立金門高級中學
[5]國立臺灣師範大學物理研究所
[6]國立臺灣師範大學圖書資訊學研究所
[7]國立臺灣師範大學科學教育中心
[8]國立臺灣師範大學科學教育研究所
[*]通訊作者：ytchien@ntou.edu.tw

摘要

十二年國民基本教育課程總綱特別強調課程實施應兼重形成性評量，在課堂中關注學生的學習歷程，協助學生即時審視自身學習狀態並調整學習策略，以逐步接近學習目標。近年來，立即反饋系統（俗稱「按按按」）被推崇為實施形成性評量的利器。然而，為了易於蒐集與量化學生的課堂回答，傳統的立即反饋系統僅支持封閉式問題（如是非題與選擇題），教師事先設計的選項與學生實際想法總有落差，不易得知學生的思考歷程，也難以量測較高階的認知技能。因此，本研究運用自然語言處理技術結合行動裝置，開發次世代立即反饋系統，使教師得以在課堂上以開放式問答進行形成性評量，即時將全班學生的短文回答化約為數個具代表性的概念，期望能更精確評估學生的學習困難。但過去探討問題導向學習的相關研究指出，開放式問題很可能對低先備知識的學生遭成過多的認知負荷，反而不如封閉式問題有效。另一方面，高先備知識學生則具備解題相關的基模，封閉式問題的提示效益便驟減，甚或造成多餘的認知負荷，其效果反而比開放式問題來得差。因此，根據上述背景與前人研究成果，本研究在真實的課堂中探討立即反饋系統使用之問題形式（封閉式 vs. 開放式）、學生先備知識程度（高先備 vs. 低先備）以及學生概念學習成果之間的關聯。研究成果發現高先備知識學生的確有著較適合開放式問題的趨勢，而低先備知識學生則較適合封閉式問題。研究成果未來可轉化為適性化的問題驅動教學模組，根據學生先備知識程度提供不同的問題形式，更能落實形成性評量，進而最大化所有學生的學習成效。

關鍵字：形成性評量、科學教育、立即反饋系統、科技輔助教學

壹、研究背景與目的

我國教育部（2015）在《十二年國民基本教育課程綱要：國民中小學暨普通型高級中等學校自然科學領域（草案）》中明列，教師應「適時進行形成性評量和總結性評量，評估學生學習成就和診斷教學得失，並加以補救及調整，俾達成預期的教學目標」（頁 50）。根據 Black 與 Wiliam（1998）對於教育評量進行的文獻回顧，以及經濟合作暨發展組織（Organisation for Economic Co-operation and Development [OECD]，2005）出版之《Formative assessment: Improving learning in secondary classrooms》，雖然總結性評量與形成性評量均能評估學生階段性的學習精熟程度，兩者之實施時機與目的卻大相逕庭。總結性評量較偏重於成果與評等，如同一般大眾相當熟悉之段考、期末考與大考，總在教學或學習活動結束後才實施，針對學生的學習表現給予等第，據此進一步將不同等第的學生分流至不同的學習場域，或依總體學生等第來評定教師、學校以及教育體制的績效責任。形成性評量則側重於過程與回饋，在課堂中頻繁實施，協助教師較即時地針對學生的學習困難與需求來調整教學目標、方法、內容與步調；同時也促使學生省察自我的學習歷程，據此分析自身的學習困難，主動尋求協助並調整學習策略，甚而進一步提升後設認知能力（meta-cognition），學會如何學習（learning to learn），不單是精熟課程內容（Black & Wiliam, 1998; OECD, 2005）。

然而，如同 OECD（2005）的報告指出，在目前的教學現場實施形成性評量常遭遇許多困難，其中一個嚴峻的挑戰即是形成性評量的進行相當耗時；即使以最簡便的單選題紙筆測驗方式進行，也難以在單一課堂上多次施測，並即時批改、統整再給予學生回饋。這樣的困境得以透過立即反饋系統（instant response system，俗稱「clicker」或「按按按」）的使用而有所緩解。如 Chien、Chang 與 Chang（2016）的回顧論文指出，傳統的立即反饋系統係由一組似電視遙控器般的訊號發射器、一個訊號接收器以及電腦自動統計軟體。而近年來隨著行動裝置的普及，已出現許多以智慧型手機與平板行動裝置作為載具，並運用網際網路進行資料傳輸的「次世代立即反饋系統」，例如 CloudClassRoom （https://ccr.tw）、Learning Catalytics（ https://learningcatalytics.com ）、 Socrative （ http://www.socrative.com ） 以 及 Zuvio（http://www.zuvio.com.tw）。上述的資訊設備得以讓課堂上的每個學生使用電視遙控器般的訊號發射器或自己行動裝置，立即針對教師的課堂提問進行回應。立即反饋系統則即時蒐集並統計全班學生的答案，形成答案分布報表，協助教師評估學生的學習表現，也刺激學生省察自我的學習狀態。根據 Beatty 與 William（2009）所進行的文獻回顧，過去 20 年間，已有越來越多的教師使用立即反饋系統執行科技輔助形成性評量（Technology-Enhanced Formative Assessment, TEFA）。許多針對立即反饋系統輔助教學的成效綜合分析研究亦指出（例如 Caldwell, 2007; Chien et al., 2016; Fies & Marshall, 2006; Kay & LeSage, 2009; MacArthur & Jones, 2008），立即反饋系統的確有助於形成性評量的實施。此外，相較於傳統講述教學，融入立即反饋系統輔助形成性評量的教學模式普遍而言伴隨著較好的教學成效（Chien et al., 2016; Fies & Marshall, 2006; Kay & LeSage, 2009; MacArthur & Jones, 2008）。

綜觀以立即反饋系統進行形成性評量的教學研究（例如 Beatty & William, 2009; Caldwell, 2007; Dufresne & Gerace, 2004; Smith et al., 2009），可發現其教學設計受限於上個世代的立即反饋系統硬體設計，課堂問答均使用是非題與選擇題。隨著次世代立即反饋系統的發展，例如前段所述之 CloudClassRoom、Learning Catalytics、Socrative 以及 Zuvio，教師得以使用開

放式問題來進行形成性評量，學生則得以使用智慧型手機與平板，輸入一段文字，以自己的話語來回答問題。本研究團隊（李龍豪、簡佑達、張俊彥、李宗諺、曾元顯，2016）更進一步將短文處理流程分析模組融入 CloudClassRoom 立即反饋系統，運用自然語言處理技術，在課堂上即時將學生的文字短答進行自動化關鍵詞彙擷取以及歸類分群，期望能讓老師得以快速掌握全班學生的文字短答內容，並據此給予回饋並調整教學。如同 Birenbaum 與 Tatsuoka（1987）、Dufresne 等人（2002）以及 Stanger-Hall（2012）建議，相較於是非題與選擇題等封閉式問題，開放式問題應可幫助教師更深入理解學生的思考歷程。然而，如 Beatty 與 William（2009）、Dufresne 與 Gerace（2004）以及 Smith 等人（2009）所述，若從建構主義（constructivism）的學習觀點來詮釋形成性評量如何影響學生的學習歷程與成果，教師提問不但是用來診斷學生的學習困難的工具，更是刺激學生運用個體經驗以及先備知識來建構知識的素材。換言之，問題的形式亦可能影響學生建構知識的歷程與成果。Kalyuga 以及其同事（Kalyuga, Chandler, & Sweller, 2001a; Kalyuga, Chandler, Tuovinen, & Sweller, 2001b）即曾比較不同形式的問題對於提升學生解題能力的影響。Kalyuga 等人（2001a & 2001b）的系列研究成果發現，對於低先備知識的學生而言，若提供越不完整的範例供其練習，學生越易耗費過多的認知資源於盲目試誤，不易建構出恰當的解題基模。另一方面，相較於低先備知識的學生，高先備知識學生本身即具備解題相關的基模，完整範例的提示效益便驟減，甚或造成多餘的認知負荷（cognitive load），其效果有時甚至比不完整的範例來得差。

上述的性向—處理交互作用（aptitude-treatment interaction）或許可類推至以立即反饋系統進行形成性評量的教學情境：開放式問題由於缺乏解題的明確提示（例如限制選項），對於低先備知識的學生而言，反而不如封閉式問題來得有幫助；另一方面，封閉式問題的提示效果對於高先備知識學生可能幫助不大，甚或造成多餘的認知負荷，其效果反而比開放式問題來得差。然而，過去運用立即反饋系統輔助形成性評量的相關研究僅使用封閉式問題，上述假設仍有待驗證。因此，本研究嘗試在立即反饋系統輔助形成性評量的物理教學情境中，釐清問題形式（封閉式 vs. 開放式）、學生先備知識程度（低先備 vs. 高先備）以及學生概念學習成果之間的關聯。研究成果將有助於建立可對應學生先備知識的問題投遞模式，發展適性化的立即反饋系統，進一步提升形成性評量可帶來的學習成效。

貳、研究方法

一、研究對象與研究設計

本研究採用準實驗研究設計（quasi-experimental design），在立即反饋系統輔助形成性評量的情境中，探討問題形式（封閉式 vs. 開放式）以及學生先備知識程度（高先備 vs. 低先備）對於學生物理概念學習成果可能造成的影響。共有 47 位來自某國立高中的高三學生參與本研究，使用兩堂課（共計 100 分鐘）的時間來學習牛頓力學。共有兩種學習情境，其中 23 位學生以封閉式問題所組成之形成性評量，24 位學生則接受開放式問題所組成之形成性評量。在教學前，對兩種學習情境學生均施以牛頓力學概念前測，以評估各個學生的先備知識程度，並確認各學習情境中含有不同先備知識程度的學生，亦確認兩種學習情境中學生的先備知識程度分布並無顯著差異。在教學後，對兩種學習情境的學生再施以牛頓力學概念後測，以評估學生的概念學習成效。

二、教學模式

本研究採用 Beatty 與 William（2009）所提出之 TEFA 科技輔助形成性評量模式，以數個緊密相關的牛頓力學概念問題將兩堂課切割為數個小單元。在每個單元中，教師先拋出一個概念問題，並要求學生獨立思考，再使用立即反饋系統傳送自己的答案。當全班多數學生均傳送答案後，教師則使用立即反饋系統呈現全班之答案分布，但不直接公布正解，而是鼓勵學生透過小組討論來解釋自己的思考歷程，為個人的答案辯護。教師再邀請學生以立即反饋系統再次作答，若投票結果仍有爭議，教師才進行更深入的講解與澄清，否則僅作概要性的總結。整個課堂以上述流程重複進行，如 Beatty、Gerace、Leonard 與 Dufresne 建議（2006），教師的角色為提供合宜的試題以引導學生思考與揭露迷思概念，並且適時穿梭於小組間促進學生辯論。TEFA 模式與 Mazur（1997）提出之同儕教學法（peer instruction）非常類似。兩者之不同之處，乃是同儕教學法的實施通常在教師提出概念問題前，須先進行簡短的講述教學，而 TEFA 則否。由於本研究研究對象為高三學生，已於高二階段接受過牛頓力學的課堂教學，故使用 TEFA 進行概念的複習與補強，省略教師提問前的講述教學。

本研究參考 Hestenes、Wells 以及 Swackhamer（1992）研發之力學概念測驗（Force Concept Inventory, FCI），進而編製出與 FCI 試題概念相似，但情境相異之 10 題 TEFA 課堂單選題。課堂問題經由兩位高中物理專任老師獨立審查，確認各題目敘述與答案無誤，與牛頓力學概念高度相關，並且適合現行的高中物理課堂使用。隨後再將此 10 題的選項去除，成為 TEFA 開放式問題。其中一題單選題例題如下：

> 小明正在玩彈跳床，被彈跳床垂直上彈。考慮小明自彈開彈跳床後至再次碰到彈跳床的運動，假設空氣阻力微小至可忽略不計，則作用在小明身上的力有：
>
> 1. 僅受到一鉛直向下且大小固定的重力。
> 2. 一鉛直向下且大小固定的重力，以及一鉛直向上不斷減小的力。
> 3. 自彈起到達最高點的過程中，受一鉛直向上不斷減小的力；在下落過程中，受到鉛直向下不斷增大的重力。
> 4. 自彈起到達最高點的過程中，受一鉛直向下且大小固定的重力，以及一鉛直向上不斷減小的力；在下落過程中，僅受到一鉛直向下且大小固定的重力。

三、研究工具

（一）力學概念測驗

本研究之前測與後測均使用 Hestenes 等人（1992）編製之力學概念測驗（FCI），用以判斷學生對於牛頓力學的先備知識以及課後之概念理解程度。FCI 共計 30 題，每題均為五選項之單選題。FCI 在物理教育研究領域被廣泛使用（見 http://modeling.asu.edu/R&E/Research.html; 使用者必須對試題內容保密，若對 FCI 感興趣者請自行透過前述網址索取試題題本），是公認具好良好信效度的牛頓力學概念測驗。本研究前測結果的 KR20 為.73，後測結果之 KR20 則為.83，顯示 FCI 具備良好的內部一致性。

（二）*CloudClassRoom*

本研究使用 CloudClassRoom（https://ccr.tw）次世代立即反饋系統來執行前述之 TEFA 科

技輔助形成性評量模式。CloudClassRoom 不但支援封閉式問題，亦支援開放式問答的使用，並且具備短文處理流程分析模組（李龍豪等人，2016），可在課堂上即時將學生的文字短答進行自動化關鍵詞彙擷取以及歸類分群，讓老師得以快速掌握全班學生的文字短答內容。如李龍豪等人（2016）之實徵研究發現，使用 CloudClassRoom 進行課堂短答即時分析的結果雖然未臻理想，但仍能初步協助教師快速概覽全班學生的文字短答。

四、資料分析

由於 FCI 前測總分低於 12 分的學生佔樣本比例約半數（47%），本研究據此以 12 分做為先備知識高低之分界點。共有 11 位低先備知識學生與 12 位高先備知識學生使用封閉式形成性評量；另有 11 位低先備知識學生與 13 位高先備知識學生使用開放式形成性評量。兩種先備知識等級的學生在不同學習情境中的分布比例相當，卡方考驗亦指出學生先備知識高低之分布與問題形式之分派無顯著的關聯（$\chi^2 = .02, p = .89$）。此外，就概念前測的平均分數而言，使用封閉式問題驅動教學的學生與使用開放式問題驅動教學的學生表現並無顯著差異（$t = .03, p = .98$）。在確認兩種教學情境中皆含有不同先備知識程度的學生，亦確認不同教學情境的學生先備知識程度分布並無顯著差異後，本研究進一步採用雙因子變異數分析（two-way ANOVA），檢定「問題形式（封閉式 vs. 開放式）」、「先備知識（低先備 vs. 高先備）」、「問題形式*先備知識」對於學生「接受課程後之概念理解程度（FCI 後測分數）」的影響。本研究使用 IBM SPSS Statistics 22 版進行統計分析，顯著水準（p）設定為.05，並計算 Cohen（1988）的 d 係數以估算效果量。根據 Cohen（1988）的建議，$d = 0.20$、0.50 以及 0.80 可概略視為小、中等以及大效果。

參、研究結果

如表 1 所示，就低先備知識的學生而言，以開放式問題來進行形成性評量的組別，在概念理解後測的表現上，其平均分數比使用封閉式問題進行形成性評量的組別來的差。這個趨勢對於高先備知識的學生則反之；接受開放式問題來進行形成性評量的高先備知識學生，其概念理解後測的平均表現比以封閉式問題來進行形成性評量的高先備知識學生來的好。表 4 的二因子變異數分析結果更進一步指出，以學生的概念理解後測分數為依變數時，「先備知識*問題形式」交互作用接近顯著水準（$F = 3.53, p = .07$）。若進一步計算組間比較之效果量，可發現相較於常見的封閉式問題形成性評量，開放式問答對於低先備知識學生的學習成果有著負面且達中等的效果（$d = -0.5$），對於高先備知識學生的學習成果則有著正面且達中等的效果（$d = 0.6$）。

肆、討論與建議

雖然問題形式與先備知識程度之交互作用項未達顯著（$F = 3.53, p = .07$），組間比較之效果量卻均達中等，意味著高先備知識學生有著較適合開放式問題的趨勢，而低先備知識學生則較適合封閉式問題。開放式問題可能由於缺乏選項限制，對低先備知識的學生遭成過多的

認知負荷，不利於提出自己對於題目的解釋；另一方面，封閉式問題的選項對於高先備知識的學生來說，可能是多餘的訊息，造成不必要的認知負荷。此外，後續研究應導入認知負荷的測量工具，並且針對學生的小組討論過程進行資料蒐集與分析，有助於進一步驗證上述解釋。此外，本系列研究成果之累積，可有助於適性化機制的研發與修正，使立即反饋系統得以根據學生當前的程度呈現合適的問題形式，以期同時提升課室內所有學生的學習成效。

表1、不同先備知識程度學生在接受不同問題形式之形成性評量後之概念理解後測表現

		先備知識	
		低	高
		M (*SD*)	*M* (*SD*)
問題形式	封閉式	11.36 (3.93)	15.42 (5.18)
	開放式	9.18 (4.75)	18.23 (4.23)

表2、先備知識程度與問題形式對於概念理解後測表現之二因子變異數分析摘要

變異來源	*SS*	*df*	*MS*	*F*
先備知識	501.83	1	501.83	24.26***
問題形式	1.17	1	1.17	.06
先備知識*問題形式	72.96	1	72.96	3.53†
總合	1475.87	46		

†$p < .10$; ***$p < .001$.

誌謝

本研究成果受科技部專題研究計畫補助（編號 MOST 106-2511-S-019-004 與 MOST 105-2511-S-003-012-MY3），特此誌謝。

參考文獻

Beatty, I. D., & Gerace, W. J. (2009). Technology-enhanced formative assessment: A research-based pedagogy for teaching science with classroom response technology. *Journal of Science Education and Technology, 18*(2), 146-162.

Beatty, I. D., Gerace, W. J., Leonard, W. J., & Dufresne, R. J. (2006). Designing effective questions for classroom response system teaching. *American Journal of Physics, 74*(1), 31-39.

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

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats: It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*(4), 385-395.

Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education, 6*(1), 9-20.

Chien, Y. T., Chang, Y. H., & Chang, C. Y. (2016). Do we click in the right way? A meta-analytic review of clicker-integrated instruction. *Educational Research Review, 17*, 1-18.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbam Associates.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*(6031), 862-864.

Dufresne, R. J., & Gerace, W. J. (2004). Assessing-to-learn: Formative assessment in physics instruction. *The Physics Teacher, 42*(7), 428-433.

Dufresne, R. J., Leonard, W. J., & Gerace, W. J. (2002). Marking sense of students' answers to multiple-choice questions. *The Physics Teacher, 40*(3), 174-180.

Fies, C., & Marshall, J. (2006). Classroom response systems: A review of the literature. *Journal of Science Education and Technology, 15*(1), 101-109.

Kay, R. H., & LeSage, A. (2009). Examining the benefits and challenges of using audience response systems: A review of the literature. *Computers & Education, 53*(3), 819-827.

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology, 21*, 5–23.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588.

MacArthur, J. R., & Jones, L. L. (2008). A review of literature reports of clickers applicable to college chemistry classrooms. *Chemistry Education Research and Practice, 9*(3), 187-195.

Mazur, E. (1997). *Peer instruction: A user's manual.* Upper Saddle River, NJ: Prentice Hall.

OECD. (2005). *Formative assessment: Improving learning in secondary classrooms.* Paris: OECD publishing.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*(1), 73-98.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science, 323*(5910), 122-124.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE-Life Sciences Education, 11*(3), 294-306.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4*, 295–312.

Sweller, J., van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

李龍豪、簡佑達、張俊彥、李宗諺、曾元顯（2016）。短文回應的主題自動歸類在行動教育活動上之應用初探。**圖書資訊學研究，11**（1），47-84。

教育部（2015）。**十二年國民基本教育課程綱要：國民中小學暨普通型高級中等學校自然科學領域（草案）**。臺北市：教育部。