教育科學研究期刊　第六十二卷第二期
2017 年，62（2），31-60
doi:10.6209/JORIES.2017.62(2).02

數位學習實驗研究品質評估與現況分析：以行動學習為例

李漢岳
國立臺灣師範大學
心測中心

楊介銘
國立臺灣師範大學
心測中心

宋曜廷
國立臺灣師範大學
教育心理與輔導學系

摘要

由於近年來行動裝置與教育應用軟體的快速發展，探討行動裝置輔助學習效果的實驗研究大量增加。然而，目前尚缺乏以評估實驗研究品質為主題所進行的回顧性研究。因此，本研究的目的即為廣泛蒐集 ERIC 和 SSCI 資料庫自 2003 年至 2013 年間，197 篇以行動裝置輔助學習的實驗研究，並探討其在實驗設計嚴謹度、統計考驗之過程與結果的正確性，以及成效評估工具的品質等議題。研究結果顯示，實驗設計以準實驗研究占最多數（61%），但有四分之一的準實驗研究未考量實驗組與控制組之起點能力均等性問題。統計結果效度部分則發現，半數以上研究未符合統計基本假設，七成研究所採用之樣本數偏低，使其統計考驗力低落，效果值估計情形未臻精準。成效評估工具品質部分，則發現半數研究者未提供相關的測驗信度與效度資訊。最後，本研究根據研究結果與討論，對未來研究者提出相關建議。

關鍵詞：行動學習、研究品質、實驗設計

通訊作者：宋曜廷，E-mail: sungtc@bctest.ntnu.edu.tw
收稿日期：2016/07/11；修正日期：2017/01/23、2017/03/01；接受日期：2017/03/09。

壹、前言

因應科技的快速發展，數位學習的方式也歷經各種階段。包括早期的電腦輔助教學（Computer Assisted Instruction, CAI）、網際網路盛行後的遠距教學（distance learning）、加入行動載具輔助學習的行動學習（mobile learning）、到結合各種感測裝置的情境感知學習（context awareness learning）。然而，在各個數位學習的階段中，探討新工具對於教學效果的影響一直都是熱門議題（張菀珍、葉榮木，2014；Liu, Lin, & Paas, 2014; Sung, Chang, & Liu, 2016; Sung, Chang, & Yang, 2015）。

以行動學習為例，Wu 等（2012）回顧的 164 篇文獻中，即以「評估行動裝置輔助學習效果」為研究目的的文獻占最大宗（58%）。這些研究有些強調行動載具輔助學習者在教室中的創新教學（Kerawalla et al., 2007; Ketamo, 2003; Liu, Chou, Liu, & Yang, 2006; Zurita, Nussbaum, & Salinas, 2005），有些強調應用行動載具輔助學生探索博物館的展品（Hsi, 2003; Sung, Chang, Lee, & Yu, 2008），有些強調應用行動載具於戶外教學活動或實地觀察（Chen, Kao, & Sheu, 2005; Chu, Hwang, Huang, & Wu, 2008; Tan, Liu, & Chang, 2007），有些利用行動載具協助學習者在教室、實驗室以及戶外活動間進行無縫學習（seamless learning）（Liu & Hong, 2007）。由此可知，目前學界對於行動載具所能帶來的學習效果深感好奇與重視。

評估一項新的教學工具或教學方法的成效，最強而有力的方法即是實驗研究法。實驗研究可藉由操弄自變項，同時使其他變項保持恆定的情況下，來觀察依變項的變化結果。也因此，若期待實驗研究應用在教育上也可以證實預期的因果關係，就有賴對實驗研究的正確認識與實驗嚴謹度的精確講求。歷年來外界對很多教育研究的品質有不少批評，包括研究是否嚴謹、研究方法是否合宜以及研究結果的應用性等。Slavin（2002）就曾提醒教育研究在實驗法的運用與品質現況，已經落後醫療、農業、科學等領域一個世紀以上。此外，美國自 2002 年起一系列以實證本位的改革，也都在在強調實驗研究嚴謹度的重要性（Slavin, 2003; U.S. Department of Education, 2002）。而數位學習是個科技整合領域，許多研究者所受的實驗法訓練並不一致，且對實驗研究嚴謹度的要求亦不相同。因此，檢視行動學習實驗研究的品質，對於改進後續研究品質，並提供研究者參考有很大的價值。

Cheung 與 Slavin（2013）在回顧過去 30 年教育科技應用在數學學習的實驗研究時，即提醒讀者多數研究均面臨嚴重的方法學問題。主要的研究缺失包含實驗設計缺乏控制組、忽略實驗組與控制組起點能力不均等的問題、教學時間過於短暫、成效評量工具的品質不佳、採用的研究樣本數過少等。上述問題都可能對實驗結果的有效性形成嚴重威脅。Burston（2015）回顧過去 20 年行動裝置輔助語言學習的文獻也指出，儘管目前已累積上百篇的實驗研究成果，但半數以上的研究對於學習成果的測量並不可靠，研究者很少採取客觀與可信賴的測驗

工具，反而經常以教師主觀評估或學生的自我評估作為學習成效的工具。此外，許多研究也出現嚴重的實驗設計缺失，包含採用的樣本數過少、混淆變項控制不佳、缺乏控制組或對於控制組的描述相當模糊、未包含前測證據以說明實驗組與控制組在教學前的恆等性、不適當的統計分析程序等。

上述回顧性研究的結果，對於加深對實驗研究品質的重視具有很好的提醒作用：第一，實驗品質是累積正確證據的關鍵。數位學習與行動學習目前正處於快速發展的時期，我們是否更該在此階段檢視當前實驗研究方法的嚴謹度，以為未來此領域累積更多有意義的證據？Wu 等（2012）對於 2003-2010 年的行動學習研究回顧發現，86%的研究均具有正向的結果（positive outcome）。然而，缺乏研究方法嚴謹度的檢驗，我們如何確定行動學習的效果真有如此樂觀？這些研究中是否也包含一定程度有缺陷的實驗設計而誇大了實驗效果？第二，現有文獻缺乏對行動學習研究進行品質評估的回顧。過去有關行動學習的回顧性文獻大致包括探討筆記型電腦在學校學習的使用情形（Bebell & O'Dwyer, 2010; Fleischer, 2012; Penuel, 2006）、分析行動學習研究的發表趨勢（Frohberg, Goth, & Schwabe, 2009; Hung & Zhang, 2012; Hwang & Tsai, 2011; Wu et al., 2012）、分析行動載具對教學與學習的影響，例如對合作學習教學法的影響（Hsu & Ching, 2013）、分析行動載具對無縫學習的影響（Wong & Looi, 2011）。然而，目前尚缺乏以評估實驗研究品質為主題所進行的回顧性研究。

因此，本研究目的就是先介紹評估實驗嚴謹度的各種評估指標，以瞭解評閱或進行一份實驗研究時需要注意的重點。接著，透過搜尋與整理 ERIC 和 SSCI 資料庫中自 2003 年至 2013 年間，行動裝置輔助學習的實驗研究，以發現 10 年間數位學習實驗研究品質的概況與趨勢。

貳、實驗研究品質評估指標

本節統整適用於教育研究的兩份品質評估量表（WWC 與 DIAD）、相關學者的提醒，以及實驗方法教科書的建議作為介紹。其中，WWC 是美國教育部教育科學院（Institution of Educational Studies, IES）成立之「有效教育策略資料中心」（What Works Clearinghouse, WWC）所發布的教育研究評估標準（WWC, 2014）。該中心成立於 2002 年，主要任務為回顧各重要教育主題的實徵研究，以建立教育決策中可靠的證據來源。其發展的教育研究評估標準，用以辨識出哪些教育介入研究存有清楚的因果關係，得以提供該介入對學生學習是真正有效的證據（what works）。DIAD 則是由 Valentine 與 Cooper 於 2008 年所提出的研究設計與實施評估裝置（Study Design and Implementation Assessment Device, Study DIAD）。DIAD 由四個層面（內在效度、外在效度、統計結果效度、構念效度）與 32 個細節問題所組成，以系統性和全面性地評估個別實驗研究為其特色。有關 DIAD 詳細評估指標內容請參見 Valentine 與 Cooper（2008, p. 144）的研究。此外，近年來許多教育相關領域的研究皆使用 WWC 或 DIAD 的標

準進行研究品質的評估（e.g., Bernard, Borokhovski, Schmid, Tamim, & Abrami, 2014; Wolbers et al., 2015），顯見其在實徵研究品質評估方面的重要性及代表性。以下，本節將採用 DIAD 四個層面的架構，依序介紹相關的實驗研究品質評估指標。

一、內在效度

內在效度又可稱為「因果推論效度」，指的是實驗變項間之因果關係的推論有效性（Campbell & Stanley, 1966）。換句話說，內在效度著重在探討依變項的變化可歸因於自變項影響的程度。評估指標主要包含以下幾項：

（一）是否包含控制組

控制組的使用是實驗研究的必要特徵，缺乏控制組的研究又被稱為「有缺陷的實驗設計」，甚至不被認為是實驗研究（Campbell & Stanley, 1966; Fraenkel, Wallen, & Hyun, 2011; Slavin, 2003）。原因在於缺乏控制組的研究幾乎無法抵抗各種內在效度的威脅，例如無法區辨參與者的進步是來自教學介入或是個人心智的成長（mutation）、前測後的練習效果（practice effect）、霍桑效應（Hawthorne effect）等。此外，缺乏控制組也無法比較新的教學方式與傳統或其他教學方式的差異。甚至，缺乏控制組的研究由於僅比較教學前與教學後的差別，因此可能誇大了教學介入的真實效果。例如 Liao（2007）針對 52 篇電腦輔助教學對學業成就影響的後設分析即發現，未採用控制組的效果量為採用控制組的三倍（$g = 1.15$ vs. $g = 0.55$）。也因此 Slavin（2003）就曾指出一份以科學基礎為本位的實驗研究最基本的要求即是採用含有控制組的設計。

（二）是否有隨機分派程序

隨機分派參與者至實驗組和控制組被認為是實驗研究的「黃金標準」（Slavin, 2003），其原因在於隨機分派可以確保每位參與者皆有均等接受實驗處理的機會。且由於實驗組與控制組的樣本組成為隨機，因此也可確保兩組在前測階段於任一變項上均為均等（如學習能力、種族、性別、學習動機……），進而可完全排除選樣偏誤（selection bias）的威脅。由此可知，缺乏隨機分派程序的研究，其實驗組與控制組的起點能力可能在介入前就存有差異，使得介入結束之後測比較難以釐清是介入效果或起點能力差異所致。也因此 Valentine 與 Cooper（2008）和 WWC（2014）對於缺乏隨機分派程序的研究，均強調需要進一步檢驗兩組的起點能力均等性。

（三）起點能力均等性

對於缺乏隨機分派的準實驗研究，或是高樣本流失率的真實驗研究，實驗組與控制組的起點能力都有可能面臨不相等的問題，因此起點能力均等性的檢驗與相關的因應措施都是重要的議題。對於起點能力均等性的檢驗，有些研究者提供兩組的前測表現未達顯著差異作為

佐證（Edwards, Rule, & Boody, 2013; Kert, 2011），而 WWC（2014）則明確訂出兩組的起點能力差距的效果值以低於 0.25 作為起點能力均等的標準。若兩組的起點能力有所差異時，則有必要採取適當的等化方式以確保兩組的可比較性。常見方式包括比較實驗組與控制組的進步分數（後測－前測）差異（Basoglu & Akdemir, 2010）、以前測為共變量進行共變數分析等統計控制方式予以調整（Looi et al., 2011）。

過去幾篇行動學習的研究即顯示，有無考量實驗組與控制組的起點能力均等性，可能得出截然不同的研究結果。Kondo 等（2012）評估行動裝置對於學生學習英文的成效。若單純比較行動裝置組（實驗組）與一般教科書學習組（控制組）的後測成績（多益閱讀測驗），則實驗組與控制組無顯著差異。但考量前測時實驗組的起點能力即顯著低於控制組，因此研究者改以進步分數考驗兩組在實驗後的差異，結果即發現與上述相反的研究結果：採用行動裝置輔助英文學習具有較佳成效。相反的，Yen、Lee 與 Chen（2012）在比較影像化概念圖與文字化概念圖對電腦科學學習成就的影響時，未先行比較兩組在起點能力的差異，也未根據相關變項對兩組進行配對，因此當結果顯示兩種方式在成效上無顯著差異時，很難確認是此兩種方法的成效相似或是效果被各自的起點能力所抵銷。由此更可發現起點能力均等性的重要性。

（四）樣本流失率

樣本流失情形對於內在效度的威脅，是源自原先以隨機分派所建立的起點能力均等性，可能因嚴重的樣本流失而失衡。樣本流失率可分為整體流失比率（overall attrition rate）和差異流失比率（differential attrition rate）兩類。整體流失比率指的是總樣本的流失人數占全樣本人數的百分比；差異流失比率指的是實驗組內流失比率與控制組內流失比率的差值。當差異流失比率與整體流失比率兩者之比值過高時，即表示實驗組和控制組間的流失比率有明顯差異，此時原先經隨機分派所建立的起點能力均等性即可能遭到破壞。此外，樣本流失的類型亦可分為隨機性遺漏和非隨機性遺漏。隨機性遺漏指的是樣本流失的原因與介入情境無關（例如單純忘記填答、因搬家而離開實驗、實驗工具故障而被迫終止等），此種樣本流失類型對真實驗的威脅較低；而非隨機性遺漏指的是樣本流失的原因和實驗情境有特定關係（例如參與者不願意被分派到控制組而離開、在實驗組中表現不佳或缺乏興趣而不願繼續參與實驗），此種流失類型即有可能對實驗結果造成影響。

Valentine 與 McHugh（2007）曾回顧教育領域中樣本流失率對起點能力均等性的影響。在其所蒐集的 35 篇真實驗研究中，整體流失率介於 1.09%至 29.65%，差異流失率介於 0.45%至 36.67%，實驗組前測差異效果值介於-0.34 至 1.24。其中 22 篇效果值為正（實驗組大於控制組），12 篇為負、1 篇為 0。由此可見，樣本流失後一定程度的影響了兩組原先的均等性。此外，該研究也提及教育領域普遍未注意到流失率可能造成的影響，不僅在搜尋文獻時即發現有提供流失率資訊的研究付之闕如，且有報告流失率的研究中所提供的樣本流失資訊也不完整。例如缺少樣本為何流失的原因（此資訊可用來判別樣本流失的類型為隨機或非隨機）、差

異流失比率（此資訊可用於判別兩組流失的差異情形）、扣除流失樣本後的前測成績（此資訊可用於檢驗樣本流失後的兩組前測成績是否仍相等）等。

二、外在效度

外在效度是指實驗結果可以類推到不同母群與情境的廣度。意即，實驗結果的因果關係，在不同的參與者與不同的環境下是否仍能維持不變。外在效度包括母群效度（population validity）和生態效度（ecological validity），摘述如下：

（一）母群效度

母群效度指的是實驗結果可以類化到不同參與者的程度。例如能否將實驗結果類推到不同年齡、不同族群的參與者。因此在評估一項實驗研究的母群效度時，需檢驗參與者選樣的來源是否足夠涵蓋目標母群。舉例來說，該研究是否有明確說明研究結果欲推論的母群特徵？例如學生年齡、地區、成就水準（資優、一般、高風險族群、特殊需求學生）、社經地位、性別等。若已有明確需要推論的母群，於取樣時的代表性是否充足？

（二）生態效度

生態效度指的是實驗結果可以類化到不同情境的程度。例如，若想將研究結果推論到不同班級型態（小班教學、團體課程）、不同教學環境（室內、室外）、甚至不同應用情境（家庭、學校），研究者在實驗過程中所採取的情境廣度是否充足？另一個在數位學習研究中常見的生態效度問題是教學媒體的新奇效果（novelty effect）。亦即短時間的教學介入，參與者可能是基於對新的教學科技產品充滿新鮮感進而提高學習動機而產生的良好成效，但這樣的效果未必會出現在長期使用的情境下。例如 Kulik 與 Kulik（1991）以後設分析的方法整合 254 篇電腦化學習成效的研究，結果發現教學期程低於 4 週的效果（d＝0.42）顯著高於教學期程大於 4 週的效果（d＝0.26）。Cheung 與 Slavin（2013）進一步僅納入教學期程大於 12 週的研究進行後設分析，結果發現效果值為更低的 0.16。由此可知，若研究者進行較短期的教學實驗，則有可能受到新奇效果的影響而誇大了真實效果，使其不易將研究結果推論到一般的教學時間上。因此拉長教學介入時間並提供多次的測量結果（教學前、教學第 1 個月、教學第 2 個月、……、教學後），可作為檢驗新奇效果以及探討真實效果的良好方法。

三、統計結果效度

統計結果效度指的是統計推論的適當性與正確性。例如使用統計方法時有無違背該方法的基本假設？當實驗效果確實存在時，是否會因樣本數不足或相關因素導致統計考驗力過低而難以宣稱介入有效果？該研究有無提供效果值資訊？效果值之估計是否精準？將上述問題歸納為三點摘述如下：

（一）統計基本假設

在進行實驗研究的分析與推論時，多數研究者常採取母數統計的方式分析實驗資料（例如研究者常使用的 *t* 考驗、ANOVA、ANCOVA 等）。而統計基本假設指的即是這些統計方法的適用條件，例如「常態性」與「變異數同質性」兩項假設即為多數母數統計所需符合的假設，而「迴歸係數同質性」與「球型假設」則為共變數分析和相依混合設計之變異數分析所需符合的假設。由於這些統計方法的分析基礎源自各項基本假設的機率理論，若違反假設即意謂分析所得之機率推論是失效的，也因此資料分析前是否率先進行基本假設的檢驗，以及有無因應檢驗結果做出相應的調整均為重要議題。

過去一些行動學習研究者在資料分析前即有注意到此議題並採取相應的補救措施。例如，Rockinson-Szapkiw、Courduff、Carter 與 Bennett（2013）以 ANOVA 比較電子教科書與紙本教科書對學習效果的影響時，發現兩組的變異數並不同質，因此將 α 水準自 .05 修正為更嚴苛的 .025 以減低可能膨脹的第一類型錯誤率。Morris、Ramsay 與 Chauhan（2012）以相依樣本 *t* 考驗比較平板電腦的使用是否能改變大學生的學習行為時，發現過少的樣本數（24 人）導致違反常態性假設，而改以無母數統計以避免可能的推論錯誤。Chen、Hsieh 與 Kinshuk（2008）以重複量數設計，探討智慧型手機中四種教材呈現方式（僅有英文單字、英文單字附加中文註解、英文單字附加圖片註解、英文單字附加中文註解與圖片註解），對大學生英語學習的影響。由於每位學生均需重複測量四種情境的英語學習表現，因此 Chen 等即在資料分析前先行檢驗球型假設，並在檢驗出未符合假設後改採 Greenhouse-Geisser 校正公式以修正分析結果。Gentry（2008）在以個人數位助理（personal digital assistant, PDA）輔助多發性硬化症患者的職能治療訓練中，為正確分析參與者重複測量的資料，亦先檢驗重複量數設計所需符合之球型假設。同樣的，由於檢驗結果顯示資料不符合球型假設，因此改採 Huyn-Feldt 校正公式以修正分析結果。由此可知，統計基本假設的維持為確保統計分析結果正確性的基礎。

（二）效果值提供情形

在統計結果的呈現上，目前行動學習領域的研究者似乎還是偏重在呈現統計考驗顯著與否，而較沒有提供效果值以輔助統計考驗的習慣。然而，統計顯著性容易受到樣本大小的影響，可能形成「微小效果卻因大規模樣本而達到統計顯著性」，或是「有意義的效果但樣本數規模不足而未達統計顯著性」的歧異現象，因此已有多位學者主張以效果值輔助統計考驗的不足（Cohen, 1994; Kirk, 1996; Shadish, Cook, & Campbell, 2002）。

在行動學習的研究領域中，提供效果值至少有兩項功用：第一，檢驗研究結果具有統計顯著性的同時，在實務上是否仍有運用價值。例如 Solhaug（2009）以獨立樣本 *t* 考驗比較電腦使用環境對於批判思考練習的影響。結果發現，學生在「原教室使用一人一臺筆電」（實驗組）所進行的批判思考練習顯著高於「傳統電腦教室」（控制組）。然而，兩組在五點量表的

評分僅有 0.2 分的差距（Cohen's d＝0.24），而且在樣本數非常充足（712 人）的情況下，即顯示統計考驗在高樣本數時易達顯著的特性。若能在統計考驗之外，考量兩組在實務層面的低度效果，研究者即可對研究結果進行較為保守的解釋。第二，檢驗研究結果未達統計顯著性時，是否在實務上仍具有可運用的效果。例如 Chao 與 Chen（2009）比較智慧型手機（實驗組）與桌上型電腦（控制組）對於輔助大學生電腦科學學習的影響。雖然在統計考驗後未能發現兩組有顯著差異，但考量樣本數過少（僅 40 人）且效果值已達中度效果（百分制中實驗組高於控制組 7 分，Cohen's d＝0.5），顯示實際效果在實務上應具有一定的影響力，但受限於樣本數過少，使得統計考驗未能達到顯著差異。由此可知，若單純依賴統計考驗的結果，很可能忽略實務面的影響力。

（三）樣本數的適切性

過去多位學者對於實驗研究中的樣本數規劃，根據不同觀點提出相關的建議標準。例如從母數統計所應符合的常態性與變異數同質性假設出發，建議以細格人數 30 人作為樣本數規劃的參考（Pagano, 2007）；或從實驗組與控制組的起點能力均等性出發，建議隨機分派的單位（不論是學生、班級、或學校）應大於 30 人較佳（Cheung & Slavin, 2013）。或從效果值估計的精準度出發，指出研究者應規劃取得足夠精準的效果量估計值所需的最小樣本數，並以細格人數 50 人作為參考標準（Valentine & Cooper, 2008）；或從統計考驗力大小出發，建議在達到特定統計考驗力的條件下規劃所需的樣本數（Cohen, 1988; Dupont & Plummer, 1990; Faul, Erdfelder, Lang, & Buchner, 2007; Mayr, Erdfelder, Buchner, & Faul, 2007; Sink & Mvududu, 2010）。其中，統計考驗力以 0.8 以上為多數學者之共同建議標準（Aron & Aron, 1999; Cohen, 1988; Kirk, 1995）。由此可見，樣本數對於統計分析的各項環節均有重要影響。

一些行動學習的研究者有注意到此議題，並在研究方法中詳細交代樣本數規劃的過程。例如 Brooks、Miles、Torgerson 與 Torgerson（2006）在規劃電腦軟體輔助讀寫能力學習的樣本數時，採取統計考驗力至少大於 0.8 的標準。他們回顧教育心理學中教學介入研究，發現教學介入後的平均效果值約為 0.5 的中度效果。因此，以統計考驗力達到 0.8 的條件下，可偵測到中度效果的 128 位樣本數作為參考。

四、構念效度

構念效度層面主要是指結果變項之測量結果是否與該研究嘗試探討的構念一致，即「測驗工具之信度與效度」之檢核問題。

（一）測驗信度與效度

教育研究中關心的依變項多為學業成就、學習態度等抽象構念，因此檢驗研究工具能否測量到欲觀察的抽象構念（效度議題）和測量過程是否具一定的一致性與穩定性（信度議題）

即有其必要性。

現有行動學習的研究所探討的結果變項非常廣泛，不論傳統上與成就有關的語文（Oberg & Daniels, 2013; Sandberg, Maris, & de Geus, 2011）和數學（Ketamo, 2003; Roschelle et al., 2010）；或是教學活動後態度上的改變，如學習動機與興趣（Hwang, Shi, & Chu, 2011; Liu & Chu, 2010）；或是合作學習活動中的互動效果（Lin, Duh, Li, Wang, & Tsai, 2013; Zurita & Nussbaum, 2004）等。如果這些實驗研究在分析資料前缺乏測驗信度與效度的證據時，讀者實難判斷這些測量後的數據是否正確與值得信賴。尤其 Burston（2015）指出過去 20 年行動裝置輔助語言學習的研究者，經常以教師主觀評估或學生的自我評估取代客觀與可信賴的測驗工具作為學習成效的工具，更能提醒測驗品質對於研究結果的重要性。WWC（2014）曾訂出三種信度以及表面效度（face validity）的檢驗標準。其指出信度中的內部一致性信度至少需高於 0.5、再測信度需高於 0.4、評分者間信度需高於 0.5，而表面效度的判斷原則則是測量變項需要有清楚的定義與詳細的解釋。上述資訊亦可作為讀者評估文獻時的參考。

（二）介入固有的測量

「介入固有的測量」（treatment-inherent measure）指的是測驗內容與實驗組的教學素材過度的對應（over-alignment）。當測驗內容僅與實驗組的教學內容緊密連結，或測驗內容即是依據實驗組的教學內容而量身訂製時，測量結果將明顯地有利於實驗組而非控制組，此不僅失去了實驗組與控制組比較的公平性，也誇大了實驗組的教學效果。

Slavin 與 Madden（2011）曾回顧 17 篇同時採用「介入固有的測量」和「介入獨立的測量」兩項工具作為成效評量指標的研究，並比較同一篇研究使用此兩種不同工具所得到的介入效果是否有明顯的差異。結果發現，在 10 篇以數學為依變項的研究中，以「介入固有的測量」為工具的介入效果均一致的大於以「介入獨立的測量」為工具所得到的介入效果，且前者之平均效果值（0.45）也遠大於後者（-0.03）。同樣的情況也發生在另七篇以閱讀為依變項的研究中。因此，未來研究者若要削弱此誇大實驗效果的方式，除了在編製測驗時即避免測驗內容專為實驗組的教學介入而設計的現象，也可同時採用多種測量工具以完整捕捉研究者所感興趣的變項。

參、行動學習實驗研究品質的現況分析

一、分析方法

本文以教育文獻資料庫（Education Resources Information Center, ERIC）與社會科學引文索引資料庫（the Social Sciences Citation Index database of the Institute of Science Index）自 2003 至 2013 年所收錄有關行動裝置輔助學習實驗研究的文章為分析資料。以行動裝置有關之關鍵字（mobile, wireless, ubiquitous, wearable, portable, handheld, mobile phone, personal digital

assistant, palmtop, pad, web pad, tablet PC, tablet computer, laptop, e-book, digital pen, pocket dictionary, and classroom response system）和與學習有關之關鍵字（teaching, learning, training, and lecture）做組合搜尋，共計取得 4,121 篇研究。接續，研究者閱讀每篇文獻之摘要與全文，並依據以下納入排除準則篩選符合研究目的之文獻：（一）文獻主題必須與行動裝置輔助教學與學習有關。單純設計一個學習系統並測試使用性的文章將予以刪除。經由此篩選步驟，文獻數量由 4,121 篇縮減至 925 篇。（二）文獻之研究方法必須包含實驗研究法，包括前實驗設計（單組前後測設計）、準實驗設計（非等組後測設計、非等組前後測設計、對抗平衡設計、多因子設計）、真實驗設計（隨機化等組後測設計、隨機化等組前後測設計）。此外，若某篇研究同時採用實驗研究法及非實驗研究法（例如先進行調查研究再進行準實驗設計），則因該篇研究中有包含實驗研究設計，因此將會被納入於本研究分析範圍中。若僅採用非實驗研究方法，如相關研究、調查研究、單一受試研究與質性研究均於此階段排除。經由此篩選步驟，文獻數量由 925 篇縮減至 197 篇。本研究即以此 197 篇行動裝置輔助學習的實驗研究文獻進行現況與趨勢分析。

本研究主要參考 WWC 的評估架構，並補充 DIAD 中較基本、可客觀評估，且具有明確評斷標準的變項作為評估指標，共包含以下六點：（一）實驗設計類型；（二）起點能力均等性；（三）統計基本假設檢驗情形；（四）效果值提供情形；（五）樣本數之適切性；（六）測驗工具信度與效度。此外，尚有一些品質評估指標，例如內在效度之樣本流失率指標、外在效度之教學時間、採用之參與者特徵（如年齡級距、族群之多元性）、採用之教學情境（如教室、戶外、家庭）等指標，由於未有明確評斷標準，不易客觀評估（如需依照不同研究目的延請各領域專家討論後形成共識），因此未被列為本研究之評估指標。

二、實驗設計類型之現況與趨勢

依據 Campbell 與 Stanley（1966）之經典分類方式，將實驗設計類型分為：（一）前實驗設計，指沒有控制組，亦沒有隨機分派程序的設計。（二）準實驗設計，即有控制組（對照組），但缺乏隨機分派程序的設計。（三）真實驗設計，指的是既有控制組（對照組），又有隨機分派程序的設計。結果顯示，過去 10 年對於行動學習所進行的實驗研究，以準實驗研究為最大宗，計有 121 篇（61%）；其次為真實驗研究 58 篇（30%）。前實驗研究最少，計有 18 篇（9%）。

在研究趨勢方面，圖 1 顯示三種實驗設計在各年代中均以準實驗研究占最多數，其次為真實驗研究，並以前實驗研究最少。觀察年代間的趨勢則發現，準實驗與前實驗研究的比例有所下降，分別由 2003 年的 70%與 10%下降至 2013 年的 59%與 7.69%，而真實驗研究的比例則由 20%上升至 32.87%。顯示學界在 10 年間開始減少內在效度控制較差的前實驗研究比例，並更多嘗試以嚴謹的真實驗設計進行成效研究。

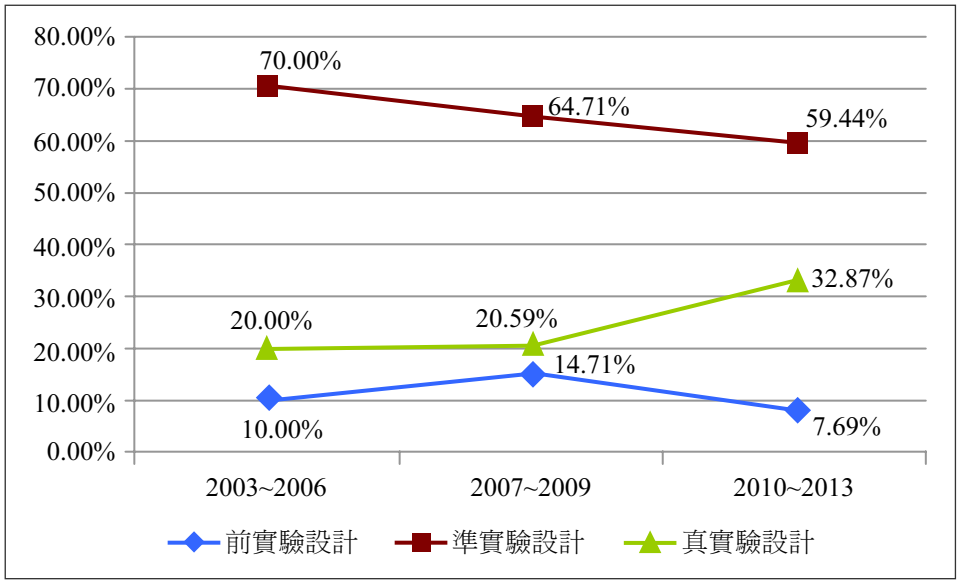


圖1.　實驗設計類型於三個年代之百分比折線圖

三、起點能力均等性之現況與趨勢

由於準實驗設計未隨機分派參與者，於前測時實驗組與控制組的起點能力即有可能出現不一致的現象，進而導致兩組在後測的差異難以分辨為實驗效果或是受到前測能力不均的影響。因此，將各研究者常使用之前測等化方式歸類如下：（一）未使用任何等化方法，意即該實驗無前測，或有前測但未進行任何調整；（二）針對前測成績進行 *t* 考驗並確認兩組無顯著差異；（三）採用共變異數分析（ANCOVA）作為統計控制；（四）採用進步分數為依變項。意即實驗組與控制組的比較，以進步分數（後測成績減去前測成績）取代後測，以消除前測不均等的現象；（五）對抗平衡設計，每一位參與者均需接受實驗情境與控制情境，且接受兩種情境的順序為隨機化，最後再比較實驗情境與控制情境的表現差異，由於兩種情境均為同一組參與者，因此可排除前測能力可能不均等之疑慮；（六）多因子設計，研究者將可能影響實驗效果的混淆變項，亦列為獨變項進行控制與探討，例如將學生能力視為實驗操弄以外的獨變項，探討實驗效果與高低能力學生的交互作用，或將前後測表現列為獨變項，探討實驗效果與前後測表現之交互作用。

接著進一步探討上一階段所得到的 121 篇準實驗研究對於前測等化方式的處理情形。由表 1 可以發現，仍有 25%的準實驗研究未採用任何等化的方式以確保實驗組與控制組於後測時之可比較性。另 75%有採取適當等化方式的研究中，則以針對前測進行 *t* 考驗以及使用 ANCOVA 調整後測成績為最常見之方法，分占準實驗研究中的 38%與 37%。其次為使用進步分數取代後測分數作為兩組比較的方式（13%），再其次為對抗平衡設計或多因子設計（各占 6%）。

表 1

準實驗設計所採用之前測等化方法一覽

準實驗設計	研究數	百分比（%）
1. 未使用任何等化方法	30	25
2. 以前測 *t* 考驗不顯著為證據	35	38
3. 採用 ANCOVA 調整	34	37
4. 採用進步分數調整	12	13
5. 採用對抗平衡設計調整	5	6
6. 採用多因子設計調整	5	6

在研究趨勢方面，圖 2 顯示各年代中採用準實驗設計的多數研究者，均有考量到起點能力均等性的重要性。且從年代間的趨勢亦可發現，有覺察到準實驗設計的選擇偏誤並進而調整的研究者逐年提升，比例自 2003 年的 64.29%提升至 78.82%；相反的，採用準實驗設計卻未針對實驗組與控制組的起點能力進行檢驗與調整的研究者則逐年下降，自 2003 年的 35.71%下降至 2013 年的 21.18%。顯示學界在 10 年間對於準實驗設計的嚴謹度要求有提高的趨勢。

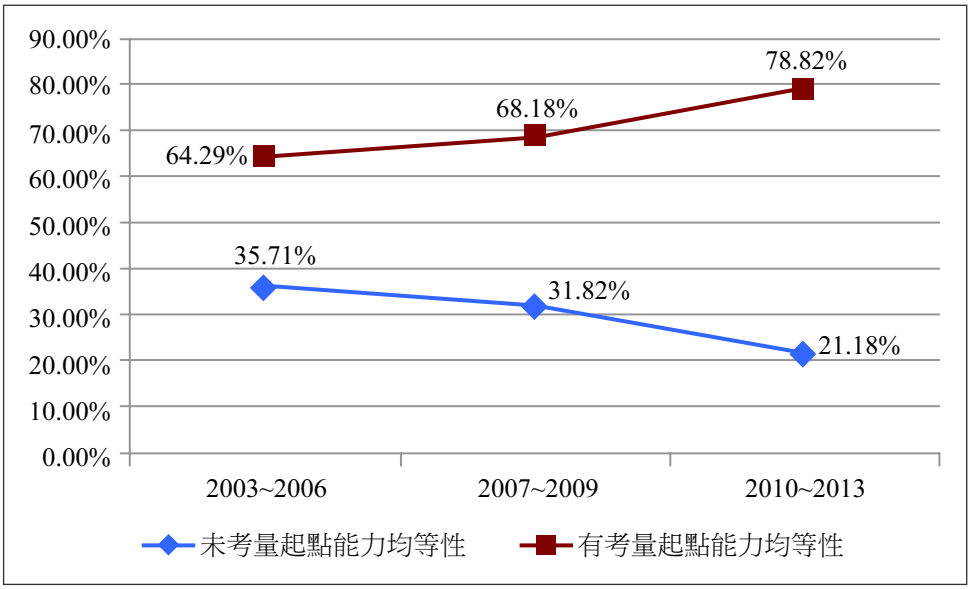


圖2. 實驗組與控制組之起點能力均等性於三個年代之百分比折線圖

四、統計基本假設檢驗情形之現況與趨勢

依據附錄之附圖 1 的判斷流程，將各研究分為「符合基本假設」與「不符合基本假設」兩類。結果發現，僅有 84 篇（45.65%）研究有符合統計方法之基本假設，其餘超過半數的研究（54.35%）均不符合假設。此結果顯示，過半數研究的統計推論結果存在錯誤決策的風險。

在研究趨勢方面，圖 3 顯示「2003~2006」與「2010~2013」兩階段，多數研究者仍在資料分析前忽略基本假設的檢驗，或檢驗結果發現資料特性不符合母數的基本假設。然而，在「2007~2009」階段則出現相反的現象。從年代間的趨勢則可以發現，「符合基本假設」的研究比例呈現先升後降的趨勢，自「2003~2006」的 36.84%，提升到「2007~2009」的 54.84%，再下降至「2010~2013」的 44.78%。

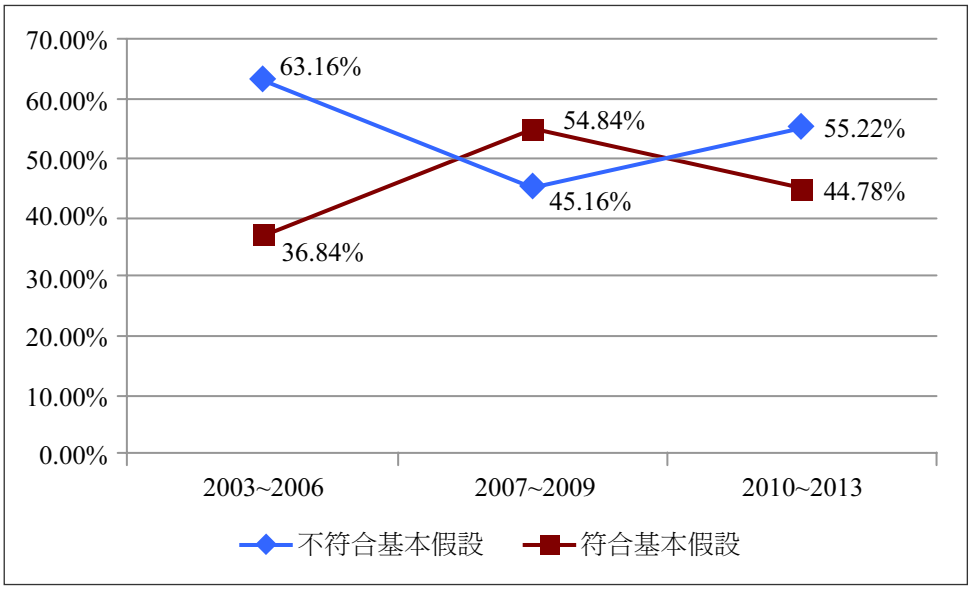


圖3.　各研究統計基本假設檢驗情形於三個年代之百分比折線圖

五、效果值提供情形之現況與趨勢

依據研究者在結果呈現效果值的狀況，分為「有呈現效果值」與「無呈現效果值」兩類。茲整理效果值提供情形之結果於表 2。由表 2 可以發現，僅有四分之一的研究者提供效果值資訊，而超過七成的研究者（75.54%）則未提供。若再依據統計方法分類為獨立設計與相依混合設計兩類，也發現有無提供效果值的比例相當一致（1：3）。

在研究趨勢方面，由圖 4 可以發現，無論在哪個年代，未提供效果值的研究比例均明顯大於有提供效果值的研究比例。從年代間的趨勢可以發現，自 2003 年至 2013 年兩者的差距有微幅的縮小，由接近七成的差距縮小到五成。儘管如此，「2010~2013」仍有高達 74.63%的

表 2

使用母數統計研究中之效果值提供情形

	有提供效果值	未提供效果值
獨立樣本設計	32（24.06%）	101（75.94%）
相依／混合樣本設計	13（25.49%）	38（74.51%）
總和	45（24.46%）	139（75.54%）

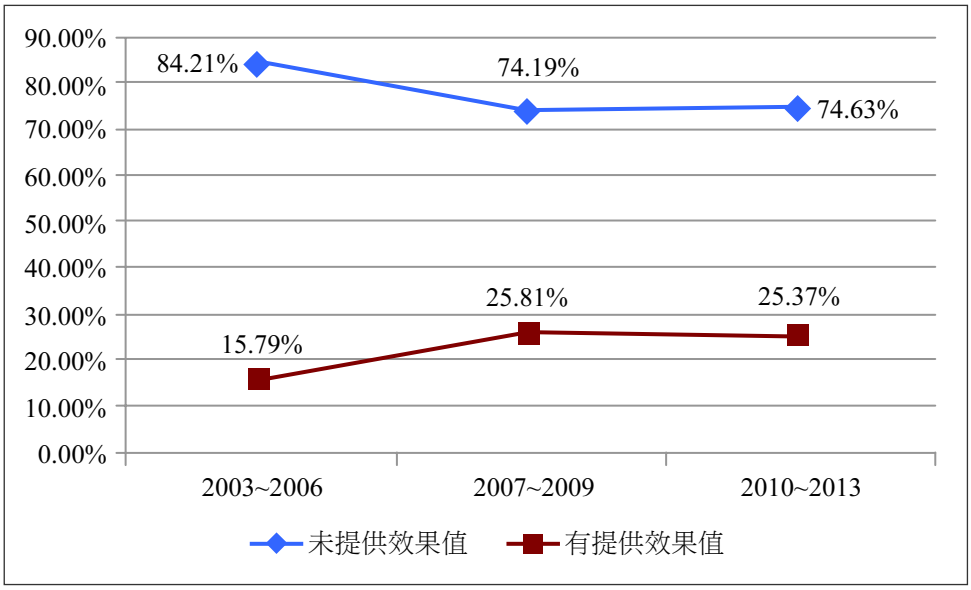


圖4. 各研究效果值提供情形於三個年代之百分比折線圖

研究者未提供效果值資訊。

六、樣本數適切性之現況與趨勢

統整過去學者對於樣本數規劃的看法，包含下列四種觀點：（一）樣本數對母數統計假設的影響（通常以各組人數相等，且每組人數大於 30 人為標準）；（二）樣本數對起點能力均等性的影響（通常以真實驗研究中，隨機分派之單位需大於 30 為標準）；（三）樣本數對效果值估計精準度的影響（以效果值估計誤差為 0.2 以內所需之樣本數為標準）；（四）樣本數對統計考驗力的影響（以統計考驗力大於 0.8 所需之樣本數為標準）。由於以上四種觀點之標準不盡相同，為完整檢驗樣本數對研究結果的影響，將以上述四種觀點之最大樣本數作為判斷個別研究樣本數是否充足之標準。結果顯示，樣本數規模得以同時符合上述四項指標之研究數量僅有 51 篇，占全部研究的 27.72%。換句話說，超過七成的研究所使用之樣本數並不充足。

在研究趨勢方面，由圖 5 可以發現，在三個年代中「樣本數不充足」的研究比例均明顯高於「樣本數充足」的研究比例。從年代間的趨勢可以發現，兩者的比例未有明確的上升或下降的趨勢。樣本數不充足的研究比例維持在68%到72%之間，樣本數充足的研究比例則維持在 26%到 32%之間。

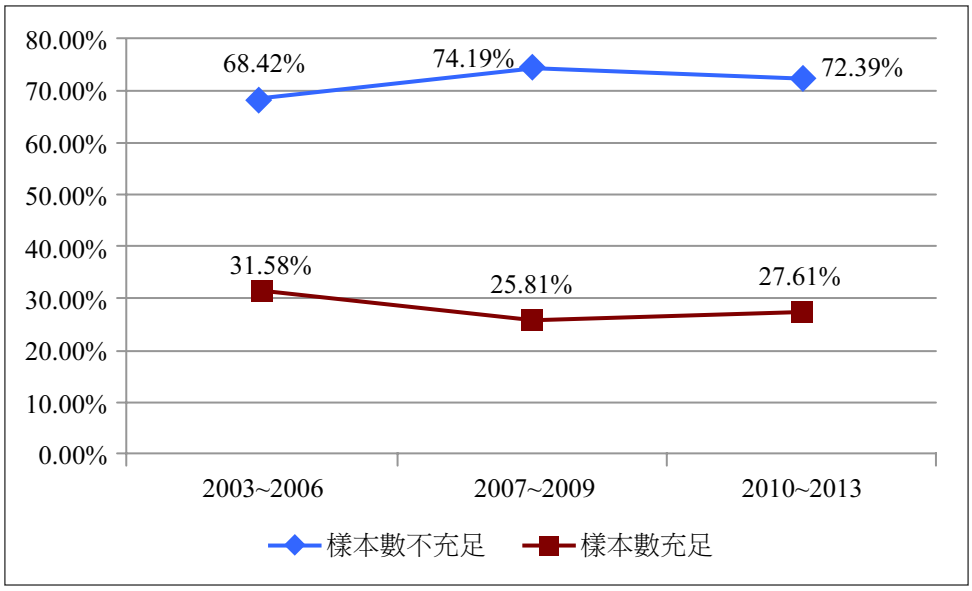


圖5.　各研究樣本數適切性於三個年代之百分比折線圖

七、測驗工具信度與效度之現況與趨勢

依據各研究報告中對測驗工具信度與效度提供之資訊可分為四類：（一）未提供信度與效度資訊；（二）僅提供信度資訊；（三）僅提供效度資訊；（四）信度與效度資訊均有提供。研究結果發現，以未提供信度與效度資訊的比例最高（104 篇研究，53%），其次為僅提供信度資訊（62 篇，31%），而信度與效度皆提供的研究僅有 29 篇（15%）。若就各研究信度考驗的類型與品質進行探討，則可發現內部一致性的 Cronbach's α 為最常用的考驗方式。其他考驗方式還包括折半信度、評分者間一致性、再測信度。在信度品質的判斷上，參考 WWC（2014）對於信度係數的最低標準進行檢驗：Cronbach's α 與評分者間信度需達 .5 以上，再測信度需達 .4 以上。若依此標準，則 91 篇有提供信度資料的研究中，有 88 篇（97%）符合最低標準。在各研究效度考驗的類型與品質方面，以內容效度為最常見的考驗方式（18 篇），其次為建構效度（九篇）、效標關聯效度（三篇）。在效度品質的判斷上，WWC 指出該研究所測量的構念需有明確的定義與解釋，且需提供該測量構念之相關證據。因此，審閱各研究中對效度考驗過程的相關描述，判斷是否符合此一最低標準。研究結果發現，在 31 篇有提供效度資料的研

究中，有 29 篇（94%）符合 WWC 之最低標準。

在研究趨勢方面，由圖 6 可以發現「無提供信度與效度」的研究比例有逐年下降的趨勢，自 2003 年的 75%下降至 2013 年的 46.85%；相反的，「有提供信度與效度」的研究比例則逐年上升，自 2003 年的 25%上升至 2013 年的 53.15%。由各年代中兩者的比較則可發現，2003 年至 2009 年，無提供測驗信度與效度的比例均大於有提供的比例，但兩者差距逐年縮小，至「2010~2013」年間甚至出現翻轉的結果，有提供測驗信度與效度的比例（53.15%）大於未提供的比例（46.85%）。

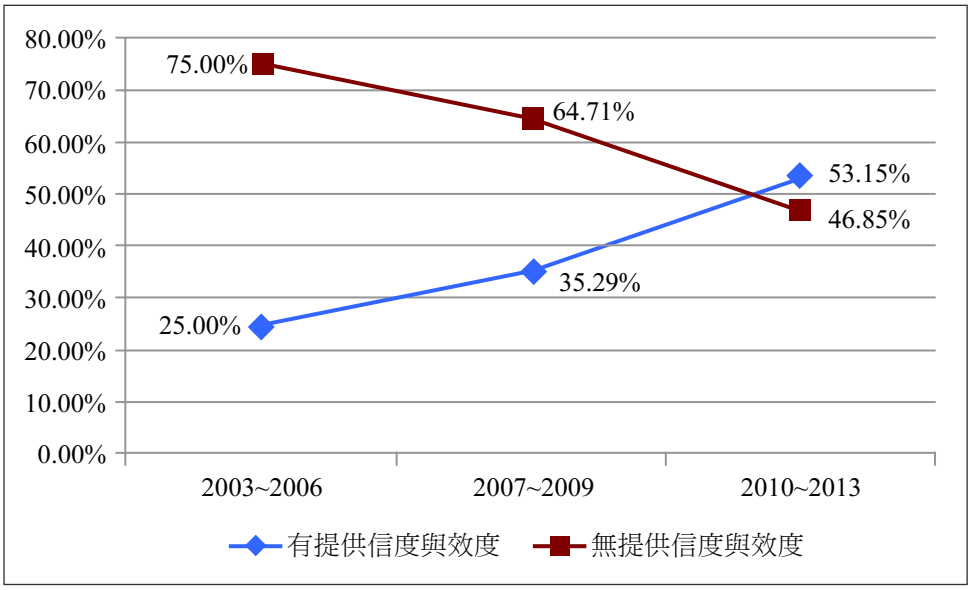


圖6. 各研究信度與效度提供情形於三個年代之百分比折線圖

肆、結論與建議

一、結論

本研究透過文獻整理，首先說明實驗研究常見的品質評估指標；其次整理 ERIC 和 SSCI 自 2003 年至 2013 年間與行動裝置輔助學習有關的實驗研究文章，並分析其在實驗設計類型、起點能力均等性、統計基本考驗檢驗情形、效果值提供情形、樣本數之適切性，以及測驗信度與效度的現況與趨勢。

結果發現，在實驗設計類型方面以準實驗研究占最多數（61%），而準實驗研究中，七成以上研究者有考量起點能力均等性。在統計結果效度方面，半數以上研究不符合統計基本假設、七成以上研究未提供效果值資訊，以及約七成研究所採用的樣本數偏低、使其統計考驗

力低落，效果值估計情形未臻精準。在測驗工具信度與效度部分，則發現半數以上研究者未提供任何的信度與效度資訊。

在研究趨勢方面則發現隨著年代演進，與內在效度有關的實驗設計類型與起點能力均等性均有逐漸嚴謹的趨勢。此外，各研究者對於與構念效度有關的測驗信度與效度的提供情形也有增加的趨勢。然而，與統計結果效度有關的統計基本假設檢驗情形、效果值提供情形以及樣本數的適切性，則沒有明顯的年代趨勢。而此三項指標在各年代中均顯示出較不佳的評估結果，可作為未來研究者在進行實驗資料分析時的參考。

二、建議

由上述行動學習實驗研究的現況與趨勢結果可以發現，研究者在實驗設計的嚴謹度表現有進步的趨勢。隨著年代演進，不僅前實驗研究比例降低、真實驗研究比例提高，且準實驗研究中有考量起點能力均等性的研究也有所上升。儘管如此，以下兩點仍值得未來研究者注意：（一）前實驗設計較適宜作為初探性質的前導研究，在因果關係的解釋上需較為保留。前實驗設計採取「前測－介入－後測」的方式進行實驗，具有簡單、易於實施的優點；但同時，也因為缺乏控制組、無法避免成熟（maturation）和前測影響（testing）等內在效度的威脅，使得因果推論能力薄弱。由此可知，未來研究者可善用前實驗設計的簡單、便利的特性，將其用於大型計畫前的前導研究，或初步檢核新建構的教學系統或課程的初探性研究，但仍應瞭解前實驗設計於因果關係推論的限制，在解釋上較為保留。（二）準實驗設計仍需注意起點能力均等性議題。本研究發現近 10 年的準實驗研究中，每四篇即有一篇未考量起點能力均等性的議題。若實驗組與控制組在實驗介入之初即存有明顯的差異，那比較兩組在教學後的結果也變得失去意義。所幸本研究也發現，目前行動學習研究者已採取至少五種等化方式來確保實驗組與控制組的可比較性，均可作為未來研究者的參考。首先，在實驗教學前即先檢驗兩組在起點能力的差異，是最保險的作法。若兩組在前測無顯著差異，那研究者即可較為放心地比較兩組於教學後的差異情形，以作為教學效果的證據。然而，若兩組於前測即出現明顯的不同，研究者也可採取如進步分數（後測減前測）來比較實驗組與控制組的進步情形，或採取統計控制，以共變數分析將兩組的前測一併納入考量。另一方面，對抗平衡設計與多因子設計雖然使用的研究者較少，但所具有多項優點的特性仍值得未來研究者視研究目的採用。以對抗平衡設計而言，僅需使用單一組參與者，因此不會有實驗組與控制組起點能力均等性的問題。且單一組樣本也可大大減少樣本人數的負擔，所搭配的統計方法僅需低樣本數即可維持一定的統計考驗力也為其特色。另外，由於每位參與者接受實驗情境或控制情境的順序以隨機化的方式安排，因此前面情境可能影響後面情境之遷移效果（carry-over effect）的疑慮也可在綜合所有參與者資料時加以平衡。以多因子設計而言，除了實驗操弄與否為其一因子外，有些研究將前、後測視為第二因子，探討實驗組與控制組在前測與後測階段的差異

情形（Ardito, Costabile, de Angeli, & Lanzilotti, 2012; Hedman & Sharafi, 2004），有些研究直接將參與者依照能力分組，探討能力差異與實驗介入的交互作用（Billings & Mathison, 2012）。由此可知，使用多因子研究不僅可排除起點能力不均所造成的推論混淆，甚至透過交互作用的討論進而得出更細緻的研究發現。

另外，近 10 年的研究趨勢也發現，行動學習在統計結果效度部分並未特別重視。然而，若實驗資料不能符合統計基本假設，則分析結果就會有所偏誤。且樣本數不足所造成的統計考驗力低落、效果值估計誤差過大、影響資料完備性而違背基本假設所造成的後果，當可能導致推論結果的錯誤。因此，本研究針對統計結果效度部分提出以下兩點建議：（一）將統計基本假設的檢驗和效果值報告納入統計分析時的既定程序。例如以常用之母數統計方法，如 *t* 考驗和 ANOVA 所需符合之常態性與變異數同質性假設為例，若實驗處理之各細格人數不等，或採取之樣本數過少時，應在資料分析前先行檢驗此兩項假設。若資料已違反假設，可酌情採用資料轉換（例如平方根轉換、對數轉換、倒數轉換等）以減低資料違反常態性的程度（Ferguson & Takane, 1989; Kirk, 1995），或採用校正公式如 Welch procedure 以修正原始公式的分析結果（Howell, 2002），或轉而使用不需滿足母體嚴格假設的無母數統計法，例如以 Wilcoxon signed-rank test 取代相依樣本 *t* 考驗，以 Mann-Whitney U test 取代獨立樣本 *t* 考驗，或以 Kruskal-Wallis H test 取代 ANOVA 等（Aron & Aron, 1999）。此外，以 ANCOVA 作為分析方法時，應先行檢查其是否符合迴歸係數同質性假設。若資料已違反假設，可調整 α error 使其具有更嚴苛的顯著標準（Tabachnick & Fidell, 2007），或改而使用 Johnson-Neyman technique（D'Alonzo, 2004）。同樣的，當研究者採取相依樣本或混合樣本之 ANOVA 作為分析方法時，也應率先檢查其是否符合球型假設。若資料已違反假設，可採用校正公式如 Geisser-Greenhouse correction 或 Huynh-Feldt correction 等（Huck, 2008; Keppel, 1991）。在效果值的呈現部分，由於其計算公式較不受樣本數影響的特性，又方便用於實務上的理解與判斷，因此不論在統計考驗顯著與否的情況下均能提供充足的輔助資訊。此外，標準化後之效果值更因去除單位的限制，可方便於不同研究間的成效比較，並能作為後設分析時之整合應用，均為其特色（Lipsey & Wilson, 2001; Littell, Corcoran, & Pillai, 2008）。（二）綜合考量統計考驗力、母數統計的基本假設與效果值估計誤差，進而估計出適切的樣本數。本研究綜合各界觀點，統整常用的實驗設計類型與其搭配的統計方法所需的總樣本數如表 3 與表 4 所示。其中，各方格中所建議之樣本數，為下列三項指標集合之最大樣本數：1.達到 0.8 之統計考驗力所需之樣本數。2.確保常態性與變異數同質性假設所需之樣本數（各細格人數相等，且各細格人數均大於 30 人）。3.限制效果值之估計誤差在 0.2 以內所需之樣本數。由於統計考驗力除了受到樣本數影響外，亦會受到實驗效果的影響，因此本研究以表 3 表示以偵測中度效果值為依據所需的樣本數，表 4 為以偵測低度效果值為依據所需的樣本數。未來研究者若需採用更複雜的實驗設計與統計方法，可嘗試使用免費的樣本數規劃軟體 G*Power 3（Faul et al., 2007）。

表 3

在各種實驗設計下所採用的統計方法之總樣本數建議（在實驗效果為中度效果、α＝ .05 的條件下）

實驗設計 \ 統計方法	受試者內設計		混合設計	受試者間設計				
	相依樣本 *t* 考驗	多因子 ANOVA（2×2交互作用）	多因子 ANOVA（2×2交互作用）	獨立樣本 *t* 考驗	單因子 ANOVA（2水準）	單因子 ANOVA（3水準）	單因子 ANCOVA（2水準）	多因子 ANOVA（2×2交互作用）
前實驗設計	34	—	—	—	—	—	—	—
準實驗設計	—	—	—	—	—	—	—	—
以前測控制	—	—	—	128	128	159	—	—
以進步分數控制	—	—	—	128	128	159	—	—
以ANCOVA控制	—	—	—	—	—	—	128	—
以對抗平衡設計控制	34	30	—	—	—	—	—	—
以多因子設計控制	—	30	34	—	—	—	—	179
真實驗設計	—	—	34	128	128	159	128	179

表 4

在各種實驗設計下所採用的統計方法之總樣本數建議（在實驗效果為低度效果、α＝ .05 的條件下）

	受試者內設計		混合設計	受試者間設計				
統計方法 / 實驗設計	相依樣本 *t* 考驗	多因子 ANOVA（2×2交互作用）	多因子 ANOVA（2×2交互作用）	獨立樣本 *t* 考驗	單因子 ANOVA（2水準）	單因子 ANOVA（3水準）	單因子 ANCOVA（2水準）	多因子 ANOVA（2×2交互作用）
前實驗設計	199	—	—	—	—	—	—	—
準實驗設計	—	—	—	—	—	—	—	—
以前測控制	—	—	—	788	788	969	—	—
以進步分數控制	—	—	—	788	788	969	—	—
以ANCOVA控制	—	—	—	—	—	—	787	—
以對抗平衡設計控制	199	138	—	—	—	—	—	—
以多因子設計控制	—	138	200	—	—	—	—	1,095
真實驗設計	—	—	200	788	788	969	787	1,095

最後，本研究發現半數以上的研究者，沒有檢驗與報告測驗信度與效度的相關證據，顯示行動學習領域的研究者仍未十分重視測驗信度與效度的議題。因此，本研究針對行動學習實驗研究中，最常使用的成就測驗和態度測驗提出建議。首先，當研究者以成就測驗作為結果測量時，研究目的無非是想瞭解以行動裝置輔助教學對特定學業成績的幫助。因此，內容效度的證據將顯得格外重要。內容效度指的是試題內容所反映出教學目標與教材內容的程度。研究者除了在研究初期即依據教學目標命題，並盡可能使題目具有教材內容範圍的代表性外，也應在測驗編製完成後委請專家針對試題內容進行檢驗。Huck（2008）即進一步指出，委請專家審核的過程與結果描述更是關鍵所在。建議研究者應清楚報告專家組成是否多元且具有公信力（例如涵蓋學科專家、測驗專家與實務現場的教師等）、專家在審核的過程中被要求做些什麼（例如評估每道試題的描述方式是否適切、與教學目標是否契合等），以及專家評論的內容與研究者因應的調整等。另一方面，當研究者以態度測驗作為結果測量時，研究目的多著重於行動裝置輔助學習，對於學生在學習動機、自我效能、學習知覺、學習偏好等的影響。由於編製這些測驗工具是用以顯示參與者人格和心理的建構，因此提供編製這些測驗的構念效度證據即相當重要。構念效度指的是測驗能測量到研究者所關心的潛在構念的程度。研究者在建立此種工具的構念效度時，可嘗試對施測資料進行因素分析，以確認資料所反映的向度和當初定義與建構這份測驗時所持的看法是否相符。或採取 Campbell 與 Fiske（1959）所提出之 multitrait-multimethod matrix 的程序，以詳細檢驗所編製的測驗與相似和相異概念的測驗之間的關聯情形。

整體來說，或許有些行動學習領域研究者會認為，相對於研究中其他重要的問題而言（例如行動裝置的系統設計或行動學習教學法的內容規劃等），測量並沒有那麼重要，因此並未投入足夠的心力在編製一份有效的測驗或確保現有測驗的心理計量特性。然而，劣質的測量結果在本質上即限制了研究所能達成的結論效度。因此，對於一個較關心實質研究議題，而對測量本身不感興趣的研究者而言，更應從實驗開始前即盡可能使研究的測量正確有效，這也便於後續從事其他研究工作時不會受到太多測量問題的影響。

參考文獻

一、中文文獻

張菀珍、葉榮木（2014）。應用多媒體行動學習系統輔助大學生情緒管理學習成效與評量之探究。**教育科學研究期刊，59**（4），99-136。doi:10.6209/JORIES.2014.59(4).04

【Chang, W.-J., & Yeh, Z.-M. (2014). Exploration of learning effectiveness and assessment of emotion management for college students by using a multimedia mobile learning system. *Journal of Research in Education Sciences, 59*(4), 99-136. doi:10.6209/JORIES.2014.59(4).04】

二、外文文獻

Ardito, C., Costabile, M. F., de Angeli, A., & Lanzilotti, R. (2012). Enriching archaeological parks with contextual sounds and mobile technology. *ACM Transactions on Computer-Human Interaction, 19*(4), 1-30. doi:10.1145/2395131.2395136

Aron, A., & Aron, E. N. (1999). *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Basoglu, E. B., & Akdemir, O. (2010). A comparison of undergraduate students English vocabulary learning: Using mobile phones and flash cards. *Turkish Online Journal of Educational Technology, 9*(3), 1-7. Retrieved from http://www.tojet.net

Bebell, D., & O'Dwyer, L. M. (2010). Educational outcomes and research from 1:1 computing settings. *Journal of Technology, Learning, and Assessment, 9*(1), 5-15. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1606

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education, 26*(1), 87-122. doi:10.1007/s12528-013-9077-3

Billings, E. S., & Mathison, C. (2012). I get to use an iPod in school? Using technology-based advance organizers to support the academic success of English learners. *Journal of Science Education and Technology, 21*(4), 494-503. doi:10.1007/s10956-011-9341-0

Brooks, G., Miles, J. N. V., Torgerson, C. J., & Torgerson, D. J. (2006). Is an intervention using computer software effective in literacy learning? A randomised controlled trial. *Educational Studies, 32*(2), 133-143. doi:10.1080/03055690500416116

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. *ReCALL, 27*(1), 4-20. doi:10.1017/S0958344014000159

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multritrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105. doi:10.1037/h0046016

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chao, P.-Y., & Chen, G.-D. (2009). Augmenting paper-based learning with mobile phones. *Interacting with Computers, 21*(3), 173-185. doi:10.1016/j.intcom.2009.01.001

Chen, N.-S., Hsieh, S.-W., & Kinshuk. (2008). Effects of short-term memory and content representation type on mobile language learning. *Language Learning & Technology, 12*(3), 93-113. Retrieved from http://llt.msu.edu/vol12num3/chenetal.pdf

Chen, Y.-S., Kao, T.-C., & Sheu, J.-P. (2005). Realizing outdoor independent learning with a butterfly-watching mobile learning system. *Journal of Educational Computing Research, 33*(4), 395-417. doi:10.2190/0PAB-HRN9-PJ9K-DY0C

Cheung, A. C.-K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88-113. doi:10.1016/j.edurev.2013.01.001

Chu, H.-C., Hwang, G.-J., Huang, S.-X., & Wu, T.-T. (2008). A knowledge engineering approach to developing e-libraries for mobile learning. *Electronic Library, 26*(3), 303-317. doi:10.1108/02640470810879464

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003. doi:10.1037/0003-066X.49.12.997

D'Alonzo, K. T. (2004). The Johnson-Neyman procedure as an alternative to ANCOVA. *West Journal of Nursing Research, 26*(7), 804-812. doi:10.1177/0193945904266733

Dupont, W. D., & Plummer, W. D. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials, 11*(2), 116-128. doi:10.1016/0197-2456(90)90005-M

Edwards, C. M., Rule, A. C., & Boody, R. M. (2013). Comparison of face-to-face online mathematics learning of sixth graders. *Journal of Computers in Mathematics and Science Teaching, 32*(1), 25-47. Retrieved from https://www.learntechlib.org/p/39231

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Ferguson, G. A., & Takane, Y. (1989). *Statistical analysis in psychology and education* (6th ed.).

New York, NY: McGraw-Hill.

Fleischer, H. (2012). What is our current understanding of one-to-one computer projects: A systematic narrative research review. *Educational Research Review, 7*(2), 107-122. doi:10.1016/j.edurev.2011.11.004

Fraenkel, J. R., Wallen, N. E., & Hyun, H. (2011). *How to design and evaluate research in education*. New York, NY: McGraw-Hill Education.

Frohberg, D., Goth, C., & Schwabe, G. (2009). Mobile learning projects–A critical analysis of the state of the art. *Journal of Computer Assisted Learning, 25*(4), 307-331. doi:10.1111/j.1365-2729.2009.00315.x

Gentry, T. (2008). PDAs as cognitive aids for people with multiple sclerosis. *American Journal of Occupational Therapy, 62*(1), 18-27. doi:10.5014/ajot.62.1.18

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi:10.2307/1169991

Hedman, L., & Sharafi, P. (2004). Early use of internet-based educational resources: Effects on students' engagement modes and flow experience. *Behaviour & Information Technology, 23*(2), 137-146. doi:10.1080/01449290310001648251

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Hsi, S. (2003). A study of user experiences mediated by nomadic web content in a museum. *Journal of Computer Assisted Learning, 19*(3), 308-319. doi:10.1046/j.0266-4909.2003.jca_023.x

Hsu, Y.-C., & Ching, Y.-H. (2013). Mobile computer-supported collaborative learning: A review of experimental research. *British Journal of Educational Technology, 44*(5), E111-E114. doi:10.1111/bjet.12002

Huck, S. W. (2008). *Reading statistics and research* (5th ed.). Boston, MA: Pearson Education.

Hung, J.-L., & Zhang, K. (2012). Examining mobile learning trends 2003-2008: A categorical meta-trend analysis using text mining techniques. *Journal of Computer Higher Education, 24*(1), 1-17. doi:10.1007/s12528-011-9044-9

Hwang, G.-J., Shi, Y.-R., & Chu, H.-C. (2011). A concept map approach to developing collaborative mindtools for context-aware ubiquitous learning. *British Journal of Educational Technology, 42*(5), 778-789. doi:10.1111/j.1467-8535.2010.01102.x

Hwang, G.-J., & Tsai, C.-C. (2011). Research trends in mobile and ubiquitous learning: A review of publications in selected journals from 2001 to 2010. *British Journal of Educational Technology, 42*(4), E65-E70. doi:10.1111/j.1467-8535.2011.01183.x

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kerawalla, L., O'Connor, J., Underwood, J., duBoulay, B., Holmberg, J., Luckin, R., …Tunley, H. (2007). Exploring the potential of the homework system and tablet PCs to support continuity of numeracy practices between home and primary school. *Educational Media International, 44*(4), 289-303. doi:10.1080/09523980701680904

Kert, S. B. (2011). The use of SMS support in programming education. *Turkish Online Journal of Educational Technology, 10*(2), 268-273. Retrieved from http://files.eric.ed.gov/fulltext/EJ932245.pdf

Ketamo, H. (2003). An adaptive geometry game for handheld devices. *Educational Technology & Society, 6*(1), 83-95. Retrieved from http://ifets.info/journals/6_1/ketamo.pdf

Kirk, R. E. (1995). *Experimental design procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746-759. doi:10.1177/0013164496056005002

Kondo, M., Ishikawa, Y., Smith, C., Sakamoto, K., Shimomura, H., & Wada, N. (2012). Mobile assisted language learning in university EFL courses in Japan: Developing attitudes and skills for self-regulated learning. *ReCALL, 24*(2), 169-187. doi:10.1017/S0958344012000055

Kulik, C.-L. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior, 7*(1-2), 75-94. doi:10.1016/0747-5632(91)90030-5

Liao, C. Y.-K. (2007). Effects of computer-assisted instruction on students' achievement in Taiwan: A meta-analysis. *Computers & Education, 48*(2), 216-233. doi:10.1016/j.compedu.2004.12.005

Lin, T.-J., Duh, H. B.-L., Li, N., Wang, H.-Y., & Tsai, C.-C. (2013). An investigation of learners' collaborative knowledge construction performances and behavior patterns in an augmented reality simulation system. *Computers & Education, 68*, 314-321. doi:10.1016/j.compedu.2013.05.011

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University press.

Liu, C.-C., Chou, C.-C., Liu, B.-J., & Yang, J.-W. (2006). Improving mathematics teaching and learning experiences for hard of hearing students with wireless technology-enhanced classrooms. *American Annals of the Deaf, 151*(3), 345-355. doi:10.1353/aad.2006.0035

Liu, T.-Y., & Chu, Y.-L. (2010). Using ubiquitous games in an English listening and speaking course:

Impact on learning outcomes and motivation. *Computers & Education, 55*(2), 630-643. doi:10.1016/j.compedu.2010.02.023

Liu, C.-C., & Hong, Y.-C. (2007). Providing hearing-impaired students with learning care after classes through smart phones and the GPRS network. *British Journal of Educational Technology, 38*(4), 727-741. doi:10.1111/j.1467-8535.2006.00656.x

Liu, T.-C., Lin, Y.-C., & Paas, F. (2014). Effects of prior knowledge on learning from different compositions of representations in a mobile learning environment. *Computers & Education, 72*, 328-338. doi:10.1016/j.compedu.2013.10.019

Looi, C.-K., Zhang, B., Chen, W., Seow, P., Chia, G., Norris, C., & Soloway, E. (2011). 1:1 mobile inquiry learning experience for primary science students: A study of learning effectiveness. *Journal of Computer Assisted Learning, 27*(3), 269-287. doi:10.1111/j.1365-2729.2010.00390.x

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology, 3*(2), 51-59. doi:10.20982/tqmp.03.2.p051

Morris, N. P., Ramsay, L., & Chauhan, V. (2012). Can a tablet device alter undergraduate science students' study behavior and use of technology? *Advances in Physiology Education, 36*(2), 97-107. doi:10.1152/advan.00104.2011

Oberg, A., & Daniels, P. (2013). Analysis of the effect a student-centred mobile learning instructional method has on language acquisition. *Computer Assisted Language Learning, 26*(2), 177-196. doi:10.1080/09588221.2011.649484

Pagano, R. R. (2007). *Understanding statistics in the behavioral sciences* (8th ed.). Belmont, CA: Thomson Higher Education.

Penuel, W. R. (2006). Implementation and effects of one to one computing initiatives: A research synthesis. *Journal of Research on Technology in Education, 38*(3), 329-348. doi:10.1080/15391523.2006.10782463

Rockinson-Szapkiw, A. J., Courduff, J., Carter, K., & Bennett, D. (2013). Electronic versus traditional print textbooks: A comparison study on the influence of university students' learning. *Computers & Education, 63*, 259-266. doi:10.1016/j.compedu.2012.11.022

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuell, B., Nussbaum, M., & Claro, S. (2010). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathmatics learning. *Educational Technology Research and Development, 58*(4), 399-419. doi:10.1007/s11423-009-9142-9

Sandberg, J., Maris, M., & de Geus, K. (2011). Mobile English learning: An evidence-based study with fifth graders. *Computers & Education, 57*(1), 1334-1347. doi:10.1016/j.compedu.2011.01.015

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Sink, C. A., & Mvududu, N. H. (2010). Statistical power, sampling, and effect sizes: Three keys to research relevancy. *Counseling Outcome Research and Evaluation, 1*(2), 1-18. doi:10.1177/2150137810373613

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21. doi:10.3102/0013189X031007015

Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership, 60*(5), 12-16. Retrieved from http://www.ascd.org/publications/educational-leadership/feb03/vol60/num05/A-Reader's-Guide-to-Scientifically-Based-Research.aspx

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness, 4*(4), 370-380. doi:10.1080/19345747.2011.558986

Solhaug. (2009). Two configurations for accessing classroom computers: Differential impact on students' critical reflections and their empowerment. *Journal of Computer Assisted Learning, 25*(5), 411-422. doi:10.1111/j.1365-2729.2009.00318.x

Sung, Y.-T., Chang, K.-E., Lee, Y.-H., & Yu, W.-C. (2008). Effects of a mobile electronic guidebook on visitors' attention and visiting behaviors. *Educational Technology & Society, 11*(2), 67-80. Retrieved from http://www.ifets.info/journals/11_2/7.pdf

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education, 94*, 252-275. doi:10.1016/j.compedu.2015.11.008

Sung, Y.-T., Chang, K.-E., & Yang, J.-M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review, 16*, 68-84. doi:10.1016/j.edurev.2015.09.001

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics*. Boston, MA: Pearson/Allyn & Bacon.

Tan, T.-H., Liu, T.-Y., & Chang, C.-C. (2007). Development and evaluation of an RFID-based ubiquitous learning environment for outdoor learning. *Interactive Learning Environments, 15*(3), 253-269. doi:10.1080/10494820701281431

U.S. Department of Education. (2002). *Strategic plan 2002-2007*. Retrieved from http://www2.ed.gov/about/reports/strat/plan2002-07/index.html

Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the

methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13*(2), 130-149. doi: 10.1037/1082-989X.13.2.130

Valentine, J. C., & McHugh, C. M. (2007). The effects of attrition on baseline comparability in randomized experiments in education: A meta-analysis. *Psychological Methods, 12*(3), 268-282. doi:10.1037/1082-989X.12.3.268

What Works Clearinghouse. (2014). *WWC procedures and standards handbook version 3.0*. Retrieved from http://ies.ed.gov/ncee/wwc/Handbooks

Wolbers, K. A., Dostal, H. M., Graham, S., Cihak, D., Kilpatrick, J. R., & Saulsburry, R. (2015). The writing performance of elementary students receiving strategic and interactive writing instruction. *Journal of Deaf Studies and Deaf Education, 20*(4), 385-398. doi:10.1093/deafed/env022

Wong, L.-H., & Looi, C.-K. (2011). What seams do we remove in mobile assisted seamless learning? A critical review of the literature. *Computers & Education, 57*(4), 2364-2381. doi:10.1016/j.compedu.2011.06.007

Wu, W.-H., Wu, Y.-C., Chen, C.-Y., Kao, H.-Y., Lin, C.-H., & Huang, S.-H. (2012). Review of trends from mobile learning studies: A meta-analysis. *Computers & Education, 59*(2), 817-827. doi: 10.1016/j.compedu.2012.03.016

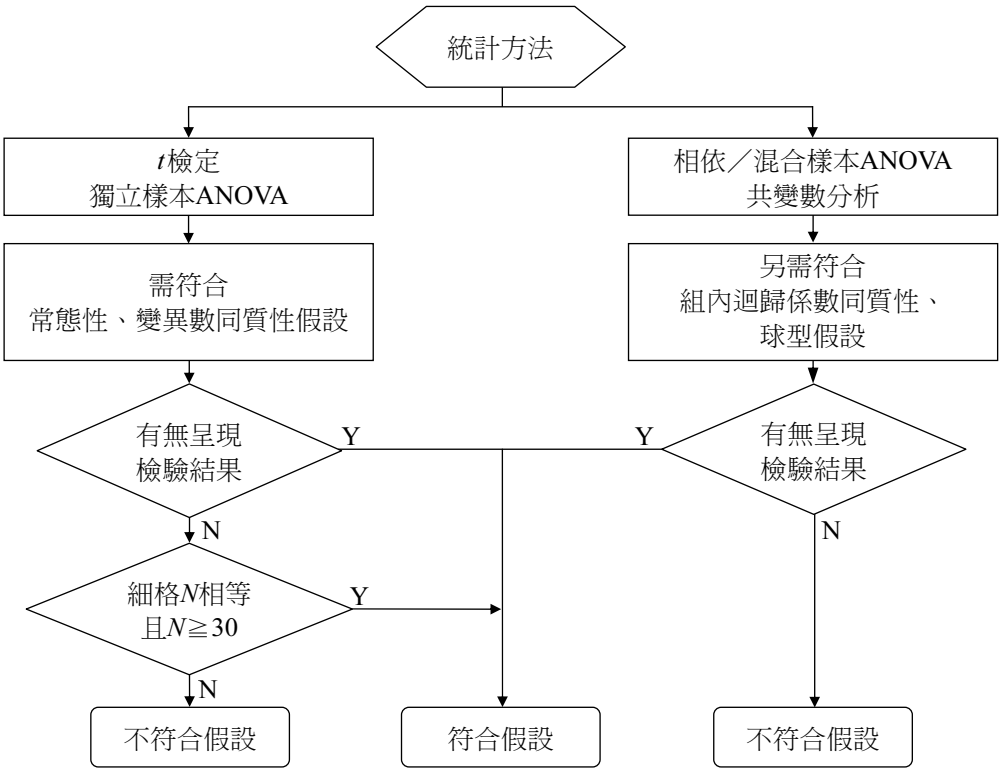
Yen, J.-C., Lee, C.-Y., & Chen, I-J. (2012). The effects of image-based concept mapping on the learning outcomes and cognitive processes of mobile learners. *British Journal of Educational Technology, 43*(2), 307-320. doi:10.1111/j.1467-8535.2011.01189.x

Zurita, G., & Nussbaum, M. (2004). Computer supported collaborative learning using wirelessly interconnected handheld computers. *Computers & Education, 42*(3), 289-314. doi:10.1016/j.compedu.2003.08.005

Zurita, G., Nussbaum, M., & Salinas, R. (2005). Dynamic grouping in collaborative learning supported by wireless handhelds. *Educational Technology & Society, 8*(3), 149-161. Retrieved from http://www.ifets.info/journals/8_3/14.pdf

附錄　統計基本假設判斷流程

本研究依照附圖 1 的判斷流程，將各研究在統計基本假設之符合情形進行分類。如附圖 1 所示，廣為研究者所使用之 *t*-test 與獨立樣本 ANOVA，需符合常態性與變異數同質性假設（Aron & Aron, 1999; Howell, 2002）。若該研究者有呈現檢驗結果，且在檢驗出違反假設後有採取因應措施，此研究將歸類為符合假設。此外，由於 *t* 檢定與獨立樣本 ANOVA 為強韌性（robust）之統計量，即在各組人數相等且均大於30人時，即使違反此兩項假設也不致對結果造成嚴重影響（Glass, Peckham, & Sanders, 1972; Pagano, 2007），因此，若該研究未對此假設進行檢驗，但各細格人數大於 30 人，也將其歸類為符合假設。另一方面，準實驗研究中常使用之 ANCOVA，因需依據各組依變項與共變項的迴歸線進行後測成績調整，因此另需符合迴歸係數同質性假設（Howell, 2002; Kirk, 1995）；相依混合 ANOVA 則因涉及多次測量之相關議題，另需符合球型假設（Keppel, 1991; Kirk, 1995）。由於相等之樣本數並不會對違反此兩項假設產生免疫力（Huck, 2008），因此本研究僅將有提供檢驗證據之研究歸類為符合假設，否則歸類為不符合假設。



附圖1.　統計基本假設符合情形判斷流程

Journal of Research in Education Sciences
2017, 62(2), 31-60
doi:10.6209/JORIES.2017.62(2).02

Quality Assessment and Situational Analysis of Experimental E-Learning Designs: A Case Study of Mobile Learning

Han-Yueh Lee
Research Center for Psychological and Educational Testing, National Taiwan Normal University

Je-Ming Yang
Research Center for Psychological and Educational Testing, National Taiwan Normal University

Yao-Ting Sung
Department of Educational Psychology and Counseling, National Taiwan Normal University

Abstract

Because of the substantial development of mobile devices and educational software in recent years, the results of mobile learning–based interventions represent a popular research topic for investigation. Although experiment quality is the basis of empirical research, few studies have explored this issue. Thus, the present study investigated the shortfalls in existing experimental research designs related to mobile learning over the past decade and offers suggestions in this paper. The researchers collected data from all 197 experimental studies on mobile learning published in the Education Resources Information Center and Institute of Science Index from 2003 to 2013. The findings of the present study are described as follows: (1) Quasi-experimental designs represent the most frequently used design type (61%); however, among the quasi-experimental studies, 25% did not consider baseline equivalence. (2) Over half of the studies may not have met basic statistical assumptions, and approximately 70% used insufficient sample sizes, leading to low statistical power and imprecise effect size estimation. (3) Half of the studies did not provide information on test reliability and validity. Finally, this paper discusses the results and their implications for future research and practice.

Keywords: experimental design, mobile learning, research quality

Corresponding Author: Yao-Ting Sung, E-mail: sungtc@bctest.ntnu.edu.tw
Manuscript received: Jul. 11, 2016; Revised: Jan. 23, 2017, Mar. 1, 2017; Accepted: Mar. 9, 2017.