教育科學研究期刊　第五十五卷第四期
2010 年，55（4），69-95

隨機化試驗在教育研究中的應用

譚克平
國立臺灣師範大學科學教育研究所
副教授

摘要

隨機化試驗是一種行之有年的研究方法，但近年來在不同社會因素的影響之下，隨機化試驗在教育界開始備受注意，並且引起多方的討論甚至爭議。本文旨在介紹隨機化試驗的內容，包括隨機分派與因果關係的建立；並介紹兩種隨機化試驗的形式，包括隨機化對照試驗與集群隨機化試驗，以及進行時考量的要點。本文並報導三個隨機化試驗的實例以供參考，從中可看出進行隨機化試驗時，因實務所需而會有或大或小的調整，最後則以綜合討論作結。本文希望能讓讀者對隨機化試驗有比較全面的初步瞭解，如日後要從事隨機化試驗時，在設計考量上能更加全面。

關鍵字：因果關係、集群隨機化試驗、隨機分派、隨機化對照試驗

通訊作者：譚克平，E-mail: t45003@ntnu.edu.tw
收稿日期：2010/05/31；修正日期：2010/11/25；接受日期：2010/12/25。

壹、緒論

教育政策研究常牽涉到方案（program），方案基本上可視爲是一組活動或計畫，其目的是要達成某些預先設定而且對公眾有益處的結果。通常與政策有關的方案大都是經由政府單位資助而執行，所以實質上它是依靠納稅人的稅捐才能進行的，因此，由政府資助的方案都必須接受評鑑，檢驗其實際效果是否值得接受政府的資金援助，以符合績效責任制度（accountability）的原則。方案評鑑在公共政策辯論以及監督政府所資助的計畫上扮演重要角色，它可以影響到有關計畫的設計、運作、資金運用等方面的決定。

方案評鑑有許多不同的技術，其中一項是隨機化對照試驗（Randomized Controlled Trials, RCT）。近年來，採用該方法來評鑑教育方案的效果，有愈來愈受政府決策單位以及教育研究人員重視的趨勢，此外，與其相似的集群隨機化試驗（Cluster Randomized Trials, CRT）研究方法亦愈來愈受到注意。事實上，該等研究方法在不少領域已行之有年，例如，Greenberg 及 Robins（1986）的論文中即曾指出，社會科學家應用隨機化試驗已有數十年的歷史，對於評量社會政策的效果方面已累積了不少經驗。其他研究領域如醫學及公共衛生（Borman, 2009; Donner & Klar, 2000），在過去 50 年間，透過大量採用隨機化臨床試驗的研究方式，成功撲滅了美國境內的麻疹及小兒痲痺症等疾病（Borman, 2009）。又例如美國的勞工局於上世紀 1980 年代期間，即已開始資助採用集群隨機化試驗的研究，以探討一些就業訓練方案的成效（Raudenbush, Martinez, & Spybrook, 2007）。教育研究應用隨機化對照試驗和集群隨機化試驗雖然起步較晚，但近年來也有一些呼喚教育改革的報告，文中亦提及在教育研究方面應進行隨機化對照試驗，例如，由美國國家數學諮詢委員會所提出之頗具爭議的《成功的根基》（*Foundation for success*）彙報，除了針對各級數學教學方面提出很多明確的建議，並鼓勵多執行高品質且能夠探討出因果關係的研究計畫，還特別強調進行隨機化對照試驗，可以對國家層級的政策提供有用的資訊（National Mathematics Advisory Panel [NMAP], 2008）。此外，美國聯邦教育部在 2002 年通過《教育科學革新法案》（Education Sciences Reform Act），正式成立了一個名爲教育科學院（Institution of Educational Studies, IES）的機構，專事教育方面的研究工作，並以推動隨機化試驗爲其首務之一（Education Sciences Reform Act, 2002），該機構自 2002 年至 2006 年這 5 年的期間，即資助了五十五個集群隨機化試驗的研究計畫，美國 IES 促進集群隨機化試驗的決心，於此可見一斑。

在臺灣方面，雖然一般研究方法的教科書都會對隨機分派及實驗法有所著墨，但教育研究期刊中採用隨機化試驗的論文爲數不多，筆者曾按一般途徑做初步的搜尋，只搜尋到 3 篇有用隨機分派的實驗研究法論文（方金雅、蘇姿云，2005；洪儷瑜、黃冠穎，2007；章勝傑、李冠蓉，2003），而且研究對象人數非常少。再者，究竟何謂隨機分派？何謂隨機化對照試驗與集群隨機化試驗？雖然從某一個角度而言，它們可以說是在教育方案評鑑與教育政策等領

域中較爲新興的方法，但爲何需要應用此類研究方法？它們有什麼優、缺點？在設計方面應該留意什麼事項？本文旨在對隨機化試驗做綜合的介紹，包括隨機分派與因果關係的建立；並介紹兩種隨機化試驗的形式，包括隨機化對照試驗與集群隨機化試驗，以及進行時考量的要點。然而，本文目的並非要對隨機化試驗做全面性的回顧，而是從較廣的層面，整理相關的背景知識，讓有興趣的研究者對此方法在概念上能有初步但較全面的瞭解，並能留意到在教育環境中執行隨機化試驗時，實務上並不簡單，需要有彈性做計畫上的調整，因此本文不列舉制式化的進行步驟，如讀者對隨機化試驗大致上的進行方式有興趣，可參考本文所報導的三個隨機化試驗實例。

貳、隨機化試驗備受注意的緣故

事實上，隨機化試驗並不是一個新的概念，隨機化的觀念可上溯至二十世紀二、三〇年代英國著名統計學家 Ronald A. Fisher（1925）所從事的研究，他認爲當要分派（assign；亦有翻譯爲分配或指派）研究對象到實驗處理（treatments）的不同組別時，分派這些研究對象的方式，必須是與研究對象的特質沒有任何關係的，例如，並不會因爲某研究對象比較高或者是比較敏捷而被分派至實驗組，因此分派的過程必須是隨機的，這樣的安排被認爲對實驗設計是既重要而且是相當可取的。Campbell 與 Stanley（1963）的名著 *Experimental and quasi-experimental designs for research* 一書中，除了向教育界介紹前實驗、準實驗與真實驗等設計的差別之外，並提供內在效度以及外在效度的各類威脅作爲一個架構，以比較不同研究設計的限制，尤其是在建立因果關係方面。該書所提出眾多研究方法的專業術語，時至今日，仍爲教科書所廣爲沿用，Campbell 與 Stanley 是忠實倡導教育研究應推行隨機化試驗的學者，他們認爲要面對教育實施（educational practices）方面的爭議，以及執行某項教育實施是否真的帶來進步，只有透過隨機化試驗才能得到解決。

教育研究雖然行之有年，但歷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方面。例如，美國教育部發表的 2002-2007 年政策計畫書中，就曾指出有別於醫學及農業學等領域，教育界長久以來依賴意識形態或者是專業人士所達成的共識來運作，很容易受到一時流行風尚的影響，因此不能像科學研究那樣，藉由科學方法的應用，以及系統化的蒐集與應用客觀的資料以制定政策，從而積累出進步的成果（U.S. Department of Education, 2002）。由此可見，至少當時美國教育部的官方立場，會認爲很多教育研究的可信度並不高。

另一方面，美國國會對於隨機化試驗之所以備受注意亦扮演一個重要角色，而且影響了不只一個面向。首先，Borman（2009）及 Slavin（2003）指出在 1998 年間，美國國會核准 1 億 5,000 萬美元作爲全面學校改革示範方案（Comprehensive School Reform Demonstration,

CSRD）的基金，給一般學校申請，以提供補助的方式，激勵學校發展全面改革，但前提是所擬實施改革的措施，必須具備確實有效的證據，而且應是經由採用實驗或準實驗研究法嚴謹評估其爲有效的措施，並要有家長的參與，目標是要讓學童能夠達到更高的學業標準。及至2001 年，該方案的經費更增加至每年 3 億美元之鉅。Borman 另外指出，美國國會亦開始留意到如果一些有瑕疵或有誤的研究成果被廣爲應用，可能會對兒童的學習產生不良後果，因此，國會的教育委員會在 2000 年時曾提出一個議案，要求教育研究必須要展現更多科學的嚴謹性，其中包括進行實驗時需要有妥善組成的對照組等要求，雖然，該議案最終並未通過，但國會關心教育研究的品質可見一斑。再者，美國近年來通過的聯邦教育法案－《不讓任何孩子落後法案》（No Child Left Behind），其中提及以科學爲基礎的研究（scientifically based research）一詞約一百一十次（Slavin, 2003），並且鼓勵經由實驗設計或準實驗設計（quasi-experimental design）而進行的研究方式。該法案要求所有接受資助的計畫，其教育措施應以高品質的研究成果爲依據，並鼓勵盡可能採用隨機化試驗來評估新教育實務措施的效果。最後，Borman 還提及了美國教育科學院（IES）所擔當的角色，該機構除了大力推動學術界採用隨機化試驗作爲做決策判斷的依據之外，並成立了著名的「有效教育策略資料中心」（What Works Clearinghouse），而該資料中心的任務，是要回顧已發表的教育研究，辨識出哪些教育介入或措施對學童學習存在清楚的因果關係，對於協助學童學習是真正有效的，透過肯定該等研究作爲一種推介的方式。該資料庫建立了一套高品質的標準，以檢視那些探討重要教育問題的研究，並將隨機化試驗設計視爲是其中一種最適合用來檢視教育方案或實務效益的判準，這反映出該機構對隨機化試驗重視的程度。

有鑑於過去許多教育研究或方案並非以隨機化試驗的方式進行，內在效度薄弱，故此無法提供教育介入是否有效的明確結論，因此在考量到包括學術、政策及社會等較大環境的觀點下，美國教育部開始推動教育研究要以證據爲依歸（evidence based）的整體路線（U.S. Department of Education, 2002）。在此大環境下，近年來有不少學者（例如 Borman, 2009; Cook, 2001, 2002; Mosteller & Boruch, 2002; National Research Council, 2002; Raudenbush, 1997; Shadish, Cook, & Campbell, 2002）皆鼓勵教育研究者遇到研究問題時，宜認真考慮以隨機化試驗設計從事研究的可行性。他們覺得過去教育研究以至於方案評鑑的文獻，皆過於忽略隨機化試驗的研究方式，已經達到不平衡的比例，他們深信提高應用隨機化試驗做研究的比例，將對教育研究與教育政策都有重要的貢獻。

事實上，在二十世紀末期，有一些專事教育相關研究的機構，其中包括 American Institutes for Research 與 National Research Council 等機構，亦致力於鼓勵教育研究者採用隨機化試驗進行研究（National Research Council, 2002），其他例如司法研究（U.S. Department of Justice）、疾病預防研究（Society for Prevention Research）等領域，亦有鼓勵採用隨機化試驗進行研究的情況，並且還有研究建議及建立嚴謹報告的規範。

參、何謂隨機化試驗？

一、名詞界定

在方案評鑑的領域，試驗（trial）基本上與實驗（experiment）同義。至於隨機化的意義則牽連較廣，暫可瞭解爲是意指需要透過機率處理，詳細須待下一節做說明。隨機化試驗是一個籠統的名稱，它可以衍生出一些其他相關的名稱。文獻中常見的如稱爲隨機化實地試驗（randomized field trials, RFT），由於質性研究學者習慣將 field 翻譯爲田野或現場，因此 RFT 或可稱爲「隨機化田野試驗」，或者是「隨機化現場試驗」；本文主要根據由杜祖貽與呂俊甫（2007）所編的《教育學詞彙》一書，而採用隨機化實地試驗的翻譯。field 這一個字最主要是要反映隨機化試驗是在真實世界的地點中進行。至於醫學領域的學者，則習慣將隨機化試驗稱之爲隨機化臨床試驗（randomized clinical trials），而一般領域的學者則習慣稱之爲隨機化對照試驗（randomized controlled trials），在文獻中兩者都有可能簡稱爲 RCT。對有些學者而言，隨機化實地試驗與隨機化對照試驗之間可能略有分別，前者 RFT 著重於真實世界的實際場所中進行，而後者 RCT 則著重在實驗室或類似實驗室的場所中進行，因此閱讀相關論文時宜留意上下文的意涵。

此外，在進行隨機化試驗時，隨機分派的單位（unit）可以依據個人或群體來分類。前者爲常用的進行方式，而個人通常是指個別的學生，至於後者則屬於集群隨機化試驗所採用的方式，集群的單位可以包括如班級、學校或學區等，而且通常是指原封不動（intact）的群體。該研究設計在英文文獻中稱之爲 cluster randomized trials（例如 Bloom, Bos, & Lee, 1999），亦有學者（例如 Spybrook & Raudenbush, 2009）稱之爲 group randomized trials（GRT），以下本文一概稱之爲集群隨機化試驗，並以 CRT 的縮寫作爲代號。

在方案評鑑的領域常會遇到 treatments 以及 interventions 這兩個詞，本文將前者譯作「教育處理」，而後者則譯作「教育介入」或「教育干預」，而方案評鑑通常是要探討隨機化試驗所採用的教育介入方式是否真的會影響到學生學習等議題。

二、隨機化試驗的內涵

要介紹隨機化試驗這一個概念，本文會先從若干基本概念開始談起。教育研究者進行量化研究可能是出於不同的目的，但其中兩個重要目的是要建立因果關係，以及讓研究結果能推廣至研究情境外的更大範圍，這兩者皆與研究設計息息相關，研究宜採用隨機化的方式來進行，以增加研究的可辯護性。本節將簡要介紹因果關係（causation）、推廣性（generalizability）與隨機化（randomization）這三個觀念，進而介紹隨機抽樣（random sampling）與隨機分派（random assignment）的差別，以及它們與因果關係和推廣性的關係。由於方案評鑑最重要的

目的是要瞭解所採用的教育干預是否確實有效，值得政府運用經費繼續支持，這需要透過建立因果關係才能明確判斷，所以本節比較著重因果關係與隨機分派的說明。

要判斷兩個變項之間是否有因果關係，在研究方法上是相當困難的課題，歷來已有不少學說探討如何建立因果關係，包括早期休姆（Hume）及米爾（Mill）較爲哲學性的處理方式，到晚近 Pearls（2000）及 Rubin（1974）等現代技術性處理的模式。由於本文的目的並非要對因果關係做詳細的回顧，而是藉由因果關係扼要地介紹隨機化試驗和隨機分派的觀念，因此以下僅以傳統較簡化的方式討論因果關係，並以抽菸與肺癌爲例做說明。當然，抽菸是否會致癌亦有不少爭議，例如，統計學家 Fisher（1957）即強烈主張此議題必須考慮基因所扮演的角色，本文在此只是借用一個大眾頗爲熟悉的例子做引子，以方便說明。

要決定某變項 X 是否爲造成另一變項 Y 的原因，最少要考量以下三個要點：首先，X 必須要在時間上發生在 Y 之前；其次，X 與 Y 之間必須要有緊密的關係；第三、在其他變項皆被控制的條件下，X 與 Y 的關係必須依然不變（Knapp, 1998; Rosenthal & Rosnow, 1991; Williamson, Karp, Dalphin, & Gray, 1982）。茲以抽菸與肺癌爲例做一簡要說明，首先，假設某人有抽菸的習慣，後來發現有肺癌，很明顯抽菸是發生在發現肺癌之前，因此第一點成立。接著，很多研究顯示，抽菸與肺癌之間具統計上的相關，如抽菸的次數與是否得到肺癌的關係等，因此第二點表面上也成立。但要建立第三點卻並不簡單，雖然文獻中有幾份牽涉到模擬抽菸狀態的動物實驗研究，但像是遺傳問題與空氣污染等相關因素並未被列入控制的考慮，故仍然無法滿足上述第三個要點。需要留意的是，若某人欲探討某一關係是否爲因果關係，若該關係滿足前述第一與第二點，但是當其他因素都被控制後，第二點提及的關係就隨之消失，則此人探討的關係並非爲因果關係。不過，很多事物都是由多重原因所造成，因此若想找出造成某事物的唯一因素，那將會是艱鉅的任務。例如，即使其他相關因素都能加以控制，而兩者的相關仍然存在，在實務上還是只能說抽菸是造成肺癌的其中一個因素，而不是唯一的因素，因爲可能還有尚未瞭解的原因亦會影響致癌率。

若以建立研究結果的可推廣性而言，任何研究都需要擔心研究結果是否會以偏概全，亦即要擔心從研究的樣本而得的結果，是否可推廣至其所代表的目標母群體（target population），這是屬於研究結果的推廣性（generalizability）問題，該名詞或可翻譯爲通則化、概化或泛化等，本文譯爲推廣性。實務上很多研究的結論可能僅適用於該研究進行時的特定環境，若要推廣至其他環境，則可能會有風險或限制。

接著介紹很多量化研究會運用到的隨機化處理方式，而隨機化通常是透過隨機分派與隨機抽樣的方式來進行。爲免費時費力，研究者通常不會對整個母群體進行研究，而是從母群體抽取有限個個體組成樣本，並進行研究，然而該樣本必須能代表母群體，不然研究的結果只是針對特殊的樣本，無法與母群體做有意義的連結。而隨機抽樣是按照一些機率法則與抽樣設計來抽取樣本，從研究結果可推估所欲瞭解的母群體資訊，關於樣本抽樣有一些值得留

意的相關議題，可參考 Rosenthal 與 Rosnow（1991）做進一步的瞭解。至於隨機分派是指研究者將研究對象隨機地分派到實驗組與對照組，也就是說讓每一位研究對象都有同等的機率被分派至實驗組或對照組，而且控制各組人數大致相同，接著讓實驗組與對照組的介入運行，然後判斷實驗結果的差異是否可以用只是偶然出現來解釋，換句話說，所觀察到組別間結果的差異必須大於機率所可能預期的。但如果不透過隨機分派而用其他方式來分組，則各組的成員可能在某些特徵上是異質（heterogeneous）的，如此組別間結果的差距就不能只歸因於各組有不同的介入，而可能有其他解釋，例如是由於某些特別的研究對象被分派到某一組所引起。當然，即使是有執行隨機分派的過程，並不擔保各組成員就會完全相同或均質，此時實驗組與對照組的成員仍有可能在某些特徵上是不均質的，但這可能性會相對較低，隨機分派可以保證各組之間在實驗起步之前沒有系統性的差異，組別間若仍有差異，則僅止於是隨機性的差異，是偶然的現象。

隨機化通常是藉由使用一些公正沒有偏倚的機制或設備來進行，例如亂數表，藉以抽取出研究對象，或者是決定哪些研究對象被分派到實驗組與對照組。雖然隨機分派與隨機抽樣均可使用同一張亂數表，然而兩者的目的並不同。隨機分派研究對象到實驗組或對照組，是要從機率的角度確保各組的條件或特徵在實驗開始之前皆相同，若是實驗結束後，實驗組的結果有別於對照組，其差異則可以歸因於實驗處理，從而可以建立因果關係，這與研究的內在效度有關。另一方面，隨機選取樣本作爲研究對象，目的是要讓該樣本可以代表母群體，這與研究的外在效度有關。隨機化的觀念是應用在分派及抽樣的過程之上，而不是在結果之上，要求的是隨機化的機制要恰當（參 Knapp, 1998; Shadish et al., 2002）。

教育實驗因爲受教育制度及情境所限，通常無法進行隨機抽樣，在推廣性方面比較會有限制。因此，即使研究對象可能並不是母群體的典型代表，一般教育實驗仍會著重於隨機分派的執行，這是因爲隨機分派對於建立因果關係非常重要。然而調查研究（survey study）所遇到的問題剛好相反，一般的調查研究並無法操縱任何變項，嚴格來說，要建立因果關係十分困難，但仍可透過隨機抽樣來建立研究結果推廣至研究以外情境的可能性，隨機抽樣對於建立推廣性擔任十分重要的角色。近年來，有不少研究在探討如何從沒有用隨機化試驗的研究中整理出因果關係的資訊（如 Rosenbaum & Rubin, 1983, 1984），本文討論的部分會略做介紹。再者，在實務上要同時兼顧建立因果關係與推廣性並不容易，但實驗的目的本來就在建立教育介入與成效間的因果關係，因此，一般隨機化試驗比較著重於隨機分派的執行，而較少著墨於隨機抽樣。

肆、隨機化試驗的不同形式

隨機化試驗通常可透過兩種形式進行，分別爲隨機化對照試驗（RCT）與集群隨機化試驗

（CRT），後者由於近年來在技術上取得了不少成果，因此應用上也愈來愈多。本節將對兩者進行扼要介紹，由於前者已於上文大概勾畫出其重點，因此在此將著重介紹集群隨機化試驗，而更詳細的介紹可參看本文討論部分所推薦的專著。

一、隨機化對照試驗

在方案評鑑中，實驗組通常是接受新的教育處理或介入的那一組，而對照組通常是接受其他教育介入，或者是沒有接受任何教育介入的那一組。隨機分派是指隨機地將要研究的單位分派至實驗組或對照組，如果這些接受實驗或對照組教育介入的單位是指個體（例如個別學生），則稱此研究爲隨機化對照試驗。隨機分派的優點是在研究進行之前，除了兩組的教育介入此一因素之外，可以將已知或未知而會影響研究單位在依變項表現的各種干擾，亦按機率的原則，隨著研究單位的分派而平均分配到實驗組及對照組，從而讓各組成員在各方面的特質原則上相同，具可比較性，即控制了 Campbell 與 Stanley（1963）所謂的「研究對象選取偏差效應」（selection bias）。在試驗後，兩組依變項的平均若有差異，將可歸因於兩組教育介入的效果不同。如果缺乏隨機分派的機制，對於試驗的結果就不能做因果性的推論，這是因爲實驗組與對照組之間的成員有可能存在著系統性的差異。如果「各組皆能忠誠地執行預先設計的教育介入」此前提能夠成立，有不少學者會視 RCT 爲研究設計中的黃金標準（Shadish et al., 2002）。可是，亦有學者認爲，在教育研究的重重限制下，要各組皆能忠誠地執行預先設計的教育介入是十分不容易的事情，因此 RCT 在教育研究中較難視爲黃金標準（Cook & Payne, 2002; Shadish et al., 2002）。

二、集群隨機化試驗

一般的隨機化對照試驗，是隨機地將獨立的個別研究對象分派至實驗組或對照組，個體的隨機分派通常是最有效率且最便宜的方法。但在某些情況下，研究者隨機分派的單位不是個體，而是群組，此即本節所討論的集群隨機化試驗。被隨機分派到實驗及對照組的群組可以是班級或學校，也可以是其他原封不動的群體。爲何要隨機分派群體或集群？原因很多，例如，如果可以採用原封不動的班級，就可以不用爲了實驗的進行，從不同的班級抽取個別研究對象並重組成新的班級，在管理上比較方便，還可以增加研究對象參與研究的意願，而且很多時候實務上並不允許將不同班級的研究對象組成新的班級。另外一個原因，可能是出於教育介入將可以很自然地在原群組的形式中實施，或者是教育介入就是爲了現存的群組而設計的（Raudenbush et al., 2007）。而實施集群隨機化試驗通常最重要的原因是要減少在個體層級的隨機化對照試驗中，不同組別成員之間有溝通的機會，以致產生不必要的干擾。例如在一個班級內，部分學生被分派接受某一種教育介入，其餘學生被分派接受另一種教育介入。在此情況下，學生之間很可能會彼此交換意見或心得，從而讓實驗組及對照組各自採用的教育介入在作用上產生混淆（contamination），也因此無法判斷實驗組的教育介入是否確實有效

（Borman, Gamoran, & Bowdon, 2008）。此時，比較合理的安排是隨機分派部分班級接受某一種教育介入，另外一些班級接受另一種教育介入。雖然集群隨機化試驗是要將群組做隨機分派，但如果研究的推論是在獨立個體的層次，則隨機分派的單位（unit of randomization）將有別於分析的單位（unit of analysis）。

要進行集群隨機化試驗，常需要考量如何決定群組數以及群組大小，這兩者之間不但有關，而且還會涉及其他因素，特別是有效樣本數（effective sample size）與組內相關係數（intraclass correlation coefficient, ICC），因此決策上需要非常小心。茲以一簡化的例子說明，假設有研究者提出一新教學法，雖然經費比較高，但他相信此法會優於目前的教學法，因此希望能訓練教師，並進行實驗教學，以判斷教學法是否有效。爲了增加判斷的合理性，他希望接受實驗教學的學生能達 200 人。在集群隨機化試驗的架構下，研究者會讓實驗組接受新的教學法，而對照組接受原來的教學法，以做比較，他並打算各安排 200 位學生受教。如果願意參與計畫的班級是 20 人一班，則研究者需要安排十個實驗組的班級，並要訓練 10 位教師。但如果參與的班級是 40 人一班，研究者只需要安排五個實驗組的班級，並只需訓練 5 位教師。對照組方面亦如此類推。

20 人或 10 人兩種安排方式，選擇哪一種比較好？這須視研究目的是要瞭解教師還是要瞭解教學法而定。若研究目的在於瞭解教學法，在每班只有 20 人的情況下，因同群組內學生人數相對較少，能提供關於個別教師的資訊比較少，但因爲群組數比較多，對於瞭解教學法的效果比較有幫助，也較可能檢驗出與對照組之間是否達統計上顯著差異，不過費用亦會相對昂貴。而在每班 40 人的情況下，雖然費用比較節省，關於個別教師的資訊也會比較多，但對於瞭解教學法的效果幫助則比較少。因此，如果該研究的研究目的是要瞭解教學法之效果，應考量採用群組數比較多的安排。

然而，前述只屬初步的想法，另外還需要留意有效樣本數及 ICC 值這兩個議題。由於群組中的個體之間並非互相獨立，因此，研究的有效樣本數會與參與研究的實際人數不同。上述例子中，兩組參與人數共 400 人，但由於樣本人數並非來自簡單隨機抽樣，而是透過集群抽樣抽到一些班級而得到同樣大小的樣本，不過，因爲班級內的學生比較相似，不同班級的學生比較不相似，因此其有效樣本數會少於 400 人。關於此點可以這樣思考，先考慮一個極端情況，假設 400 位參與者皆互相獨立，彼此不影響別人的表現，則該研究有 400 筆獨立資料。接著再考慮另一個極端情況，假設有十個實驗組及十個對照組的班級參加，每班各有 20 位學生，再假設各班級內的學生彼此相似程度如同 1 人，表現完全相同，則該研究每一班級其實只提供 1 筆資料，整個研究基本上就只有 40 筆獨立的資料，而不是表面上的 400 筆。換句話說，研究的有效樣本數少於表面參與人數，所以當資料分析要進行統計考驗時，將會影響到統計考驗力。

而上述觀念又與 ICC 係數這個重要觀念有關，該係數可以用不同方式介紹，在只有兩個

層級的集群隨機化試驗中，例如只有班級及學生這兩個層級，ICC 可視爲是衡量組內同質性的一個指標，它被定義爲依變數（例如學生學業表現）的變異中，組別間（例如班級間）的變異占多少比率，即 ICC＝組間變異÷（組間變異＋組內變異）。當班級是隨機產生時，組別間沒有真正的差異，ICC 值爲零。但由於一般教育研究中所採用的班級並非隨機產生，而是採用原有班級做試驗，由於各班班風不同等緣故，班級之間通常會有差異，組內也會有同質性；只要有組間變異，ICC 值就不會是零。當各班級內的分數皆相同，但各班級的平均值皆不相同時，ICC 值應爲 1，達最大值，表示組內同質性最大。若組內有同質性，此時班級內個體之間並非互相獨立，如果忽略此點，直接以學生爲單位進行統計假設考驗，將會違反獨立性的條件（violation of independence assumption），而且組內變異很可能被低估，導致統計考驗的結果被高估。另一方面，ICC 值愈大，代表依變數的變異有較大部分來自班級，若能增加班級數，將比較能夠檢驗出教育介入效果的差異。若 ICC 值很小，則不需要太多的班級數，只需增加班級內學生人數，即可增加統計考驗力以檢驗出效果，而且當所需班級較少時，對學校的干擾也更少。因此，如果研究者有興趣想要進行集群隨機化試驗的話，在規劃時必須瞭解所需樣本數與 ICC 是有關的，如欲增加集群隨機化試驗的統計考驗力，在試驗前就應估算所需樣本數，其中一個重要考量點是要估計研究所牽涉到的 ICC 值，這通常需要參考過去類似研究所報導的 ICC 值以進行估計，而該等研究的規模需要夠大且群組數要夠多，才會增加其 ICC 值的參考價值。

其他相關資訊如有效程度或效果值（effect size）等，因篇幅所限在此不做介紹，建議有興趣者宜參考這方面的文獻（例如 Raudenbush, 1997），多加瞭解集群隨機化試驗之統計考驗力的原理。此外，還有一套名爲 Optimal Design 的免費軟體可協助研究者進行這方面的估計（Spybrook, Raudenbush, Congdon, & Martínez, 2009），建議研究者可以多加利用。

至於資料分析，由於集群隨機化試驗牽涉到群組與群組內之個體的階層結構，因此分析上通常要採用階層線性模式的技術。

伍、著名的研究介紹

美國賓州大學（University of Pennsylvania）的 Boruch（1997）教授指出，自第二次世界大戰後，隨機化實地試驗即被用來測試沙克疫苗的有效性，這種試驗設計旋即被認爲是很多領域研究干預效果的黃金標準，它的運用具體地協助了小兒麻痺症（polio）及麻疹（measles）等傳染病從美國根除（Borman, 2009），影響的範圍甚廣。本節以下將依序介紹 3 份有份量的隨機化測試研究，以供讀者參考。第 1 篇先介紹小兒麻痺疫苗的試驗，雖然此研究是與醫學相關，但它可說是隨機化試驗的重要里程碑，極具歷史價值，亦可一窺隨機化試驗可應用於多大規模的範圍。第 2 篇介紹田納西州（Tennessce）檢驗班級大小對學習成就影響的隨機化

實地實驗，該研究爲一些努力推動教育研究多採用隨機化試驗的學者所重視。第 3 篇則介紹科學教師成長研究的集群隨機化實驗，該文作者在這方面研究十分活躍，從該文可參考即使隨機化試驗結果與預期的不太相同時，研究者有責任分析問題之所在，並思考下一個階段試驗是否應做調整。這 3 份研究皆有應用隨機分派，從中可看出進行隨機化試驗需要考量的層面非常廣泛，而且在研究進行期間，常有必要做計畫上的調整，從而讓讀者瞭解從事隨機化試驗可能遇到的困難，不宜固執於一個既定的流程，並應小心記錄所做的調整，仔細報導及解釋爲何要調整的原因。

一、小兒麻痺疫苗試驗

二十世紀前半葉，小兒麻痺症傳染病襲擊美國，其中以兒童受害尤深，導致不少 5 至 9 歲的兒童喪生。1954 年，美國執行了一項小兒麻痺疫苗試驗的大規模公共衛生實驗，它是由美國密西根大學（University of Michigan）Poliomyelitis Evaluation Center 的 Thomas Francis Jr. 博士所主持，此研究在當時共有將近 200 萬名兒童參與，耗費估計高達 500 萬美元，很可能是迄今美國史上最大規模的公共衛生實驗。該項研究的結果被廣爲報導，結案報告發表時備受關注，即使在今天仍有研究者會就該研究的合宜性進行討論，尤其是在隨機化試驗的設計以及倫理思維等方面。以下將根據 Anderson 與 Finn（1996）、Brownlee（1955）、Francis 等（1957）報導的資料，對該研究做一扼要的介紹。

1952 年，美國匹茲堡大學（University of Pittsburgh）的沙克（Salk）教授首度研發出相關疫苗，及至 1954 年，跡象顯示該疫苗很有希望能解決當時的處境，但疫苗必須經過更嚴謹的人體試驗，確認爲安全後才可以正式在社會大眾身上施用。爲了安全起見，疫苗試驗必須要小心的設計，並且需要有實驗組與對照組的對比機制，兩組除了有或沒有接受疫苗之外，其餘實驗狀況應控制爲一致，藉以排除其他解釋的因素，建立該疫苗是否有效的證據，並瞭解沙克疫苗在預防小兒麻痺上有多大的成效。因此在 1954 年間，Francis 博士領導了這個沙克疫苗的大型實地試驗。

根據 Anderson 與 Finn（1996）指出，在該試驗的初始計畫中，曾打算將沙克疫苗測試於小學二年級的兒童，即以二年級參與之兒童爲實驗組，而一及三年級的兒童則不予以施打疫苗，亦即以一及三年級學童爲對照組。初始的計畫是僅觀察各組小兒麻痺的感染情況，再將二年級的感染情況與一及三年級相比，而對照組與實驗組是在同一時期與同一地理環境下進行觀察。及後，有人提出了下列質疑：第一、在那個年代，對於施打疫苗的動作本身是否會增加或減少感染小兒麻痺的機率並不十分清楚，必須小心評估。第二、在發病率方面，二年級的兒童（約 7 歲）與一及三年級（約 6 歲及 8 歲）是否本來就有所不同？第三、當學童與他們的父母獲知他們有或沒有施打疫苗時，是否會影響到學童被感染的風險？第四、當醫生知道學童有或沒有施打疫苗時，其診斷是否會受到影響？其中第三及第四點牽涉到隨機試驗

中的雙盲（double blinded）設計的概念。

除此之外，該試驗亦牽涉到倫理層面的問題。由於實驗組的兒童有施打疫苗，控制組的兒童不施打疫苗，因此儘管暫時尚不清楚疫苗的效能，但可以確定的是，有些兒童將不會受利於疫苗，這是否合乎倫理的考量？再者，該試驗對社會整體的貢獻，是否比讓部分兒童在沒有疫苗保護下面對小兒麻痺症所承擔的風險來得重要？這在小兒麻痺症流行的時代，產生了倫理決策上的張力。

上述議題顯示出初始計畫有不周詳之處，因此後來又提出另外一套進行的方式，針對所有參與研究的學童，隨機讓其中一半施打一種看似疫苗卻是惰性溶劑的安慰劑（placebo），另一半則施打疫苗，使得一、二及三年級的學童皆會接受相同的對待。爲了增加試驗的嚴謹性，該計畫控制了各種會影響到如何挑選學童接種疫苗的偏倚因素，例如，有可能有些醫師會有意無意間傾向對較爲健康的學童施打安慰劑，從而導致疫苗的效果呈爲顯著。另外，也有可能因爲居住在不同地區學童的整體健康狀況或許有所不同，從而干擾到研究結果的解讀。爲了建立雙盲的機制，該計畫在每個箱子中裝有五十個玻璃的藥水瓶，其中二十五個裝安慰劑、二十五個裝疫苗，瓶子僅標示編號，只有特定少數服務於該評鑑中心的研究人員知道哪些瓶子裝有安慰劑，該等瓶子被隨機放置箱子中，因此，當醫生或護士注射時並不知道該疫苗爲何，藉此消除潛在的偏倚因素。

該研究最後兩種設計的方式皆被採納，初始的設計施行於三十三個州，共約 57 萬學童參加，其中約 22 萬名二年級學童接受疫苗注射，其餘學童並未接受任何注射，僅接受觀察。另一方面，則以隨機化區組設計的方式於十一個州進行，共約 46 萬名一、二及三年級學童參加，其中約半數接受疫苗注射，其餘學童接受安慰劑注射（Francis et al., 1957）。基本上，Francis 等（1957）把初始設計視爲觀察性的研究（observational study），並對該方式所得的結果比較沒有信心，原因有四：第一個原因如前所述，二年級與一及三年級學童的年齡並不相同，發病率是否會有所不同？二爲依照 Francis 等人過去的經驗，勞動階級出身的兒童較易產生抗體，可能會因此影響觀察性研究的判斷；第三點，參與試驗的兒童包含不情願參加者，其合作態度與易受影響的程度難以估計；最後，初始的設計並非雙盲（double blind），讓每個人都知道誰有接種疫苗而誰沒有，即使不是有意產生偏見，偏見仍可能影響每個診斷階段或是對每個案例的觀察。

疫苗試驗的研究結果於 1955 年 4 月正式公布，Anderson 與 Finn（1996）將隨機化區組設計的研究結果簡化如表 1 所示。從表 1 中可見，接種疫苗的人數與接種安慰劑的人數大致相等，而未接種疫苗的孩童感染小兒麻痺症的人數爲有接種疫苗者的 3⅓倍（Anderson & Finn, 1996; Francis et al., 1957），由此可推論出該疫苗是有效的。該研究的結案報告厚達 500 多頁，雖然，以如此大規模的計畫而言，可以預期它必然會有不少缺失，例如有不少學童最終退出參與，另外爲數不少的同意書不翼而飛，有些學童搬家後無法追蹤等等；此外亦有學者對該

表 1　小兒麻痺症發生率的部分結果

接種與患病人數	接種疫苗	接種安慰劑
接種人數	200,745	201,229
小兒麻痺症案例數量	33	110

資料來源：Anderson 與 Finn（1996）

研究的資料分析有意見（Brownlee, 1955）。但整體而言，研究結果基本上認爲沙克疫苗是既安全又能有效防治小兒麻痺症的疫苗（Francis et al., 1957）。

有鑑於該試驗的統計結果，美國公共衛生機構開始著手於推動孩童注射疫苗的活動，而原本的沙克疫苗因仍有不少缺失，後來亦被沙賓（Sabin）疫苗的研發所取代。隨著小心執行隨機化的試驗，這令人生畏的小兒麻痺症終於從美國銷聲匿跡。

二、田納西州班級大小的實驗

此部分綜合 Finn 與 Achilles（1990）、Mosteller（1995）、Mosteller、Light 與 Sachs（1996）等人的研究做一綜合介紹。美國田納西州曾進行一個大型的班級大小研究，該研究旨在瞭解小班教學及上課時納入助教兩者在教學上的效益，研究由田納西州 4 所大學的人員協助設計與執行，共分三個階段進行。其中 1985-1989 年爲第一階段，由田納西州的教育機構執行稱爲 Student-Teacher Achievement Ratio（STAR）專案的 4 年實驗，旨在研究小班教學的效益，並與一般班級以及在一般班級中增加助教的情況做比較，探討是否會增加學生學習效果，研究對象爲幼稚園及一、二、三年級學童。第二階段爲 1989-1992 年 The Lasting Benefits Study（LBS）的事後追蹤研究，旨在研究參與第一階段實驗的小班級學生回到正常班級（四、五、六年級以後）的狀況，觀察這些學生的表現是否較佳。此階段的研究對象只限有參與過第一階段實驗的學生才能對資料有所貢獻。第三階段是 1989 年所執行的 Project Challenge 計畫，旨在幼稚園及一、二、三年級中執行小班級教學，主要分布在田納西州的十七個地區，這些區域的學生輟學機率很高，而且收入爲該州最低者。因篇幅所限，本文著重介紹第一階段的 STAR 計畫。

STAR 計畫每年約有 250 萬美元的經費用在額外支援的教師及助教上，而其餘經費則使用在資料分析上。在 STAR 試驗的規定中，作爲研究對象的班級必須是在內都市（inner city）、郊區（suburban）、城市和鄉村的學校。其中內都市與郊區學校被歸類爲大都市區域。內都市學校意指有一半以上的學生午餐費減免，郊區學校則是指位在大都市區域邊緣的學校。在非大都市區域裡，位於人口數超過 2,500 人的城鎮的學校則稱之爲城市（urban）學校，其餘則歸類爲鄉村學校。

1985 年，研究從幼稚園階段開始，一直進行到三年級，班級類型分爲三種：（一）小班級

－13 至 17 人；（二）正常班級－22 至 25 人；（3）正常班級，且配有助教。小班級人數平均有 15 位學生，比一班 22 或 23 人的正常班級少了 35%。在研究第 1 年中，有 6,400 位小學生參與研究，分派到一百零八個小班級、一百零一個正常班級，及九十九個配有助教的正常班級。老師與學生均隨機分派至班級。而助教並沒有任何特定的職務，大致上是協助教師的教學，且有薪水給付。參與研究的學校必須簽 4 年約，並且每一年級至少要有 57 位學生（以組成一班 13 人的小班級及兩班各 22 人的班級）參與，才足夠在同 1 所學校裡進行三種班級的比較。研究單位僅提供參與學校經費以補助額外支援的教師與助教，學校還需自行提供教室。每學年裡，實驗只會在一個年級裡進行以減少新教室的需求。儘管有 180 所學校表明願意參與研究，卻只有 100 所學校符合大規模學校的資格，其中 79 所在幼稚園階段參與研究。表 2 爲田納西州第 2 年實驗的一年級跨區域樣本數分布表。

表 2　田納西州第 2 年實驗的一年級跨區域樣本數

各參與類別之人數	地點			
	內都市	郊區	城市	鄉村
學校數	15	8	15	38
班級數				
全爲多數族群學生	0	18	28	119
全爲少數族群學生	65	0	13	0
綜合班級	5	23	21	39
總班級數	70	41	62	158
學生人數	1,495	804	1,214	3,059

資料來源：Mosteller（1995）

因篇幅有限，本文只介紹一年級的研究結果。該研究使用了標準化測驗與課程本位測驗來檢查學生的學習表現，一年級生接受了兩項標準化測驗的閱讀部分：（一）The Stanford Achievement Test（SAT）的字彙學習技巧及閱讀能力測驗；（二）The Tennessee Basic Skills First（BSF）的測驗，以及 1 份課程本位的閱讀測驗。在數學方面，一年級生參與了標準化的 SAT 及課程本位的 BSF 評比。

Mosteller（1995）指出，欲檢視實驗所施行新方法的效果，並與舊有方法的結果互相比較，通常可藉由將學生的成績以標準差爲單位表示，再將各組的差異以有效程度（effect size）來表示。有效程度指的是將組別平均成績的差除以沒有助教之正常班級的標準差。表 3 報導比較小班級與正常班級學生，以及有或沒有助教的正常班級學生在標準化測驗表現上之有效程

表 3　一年級小班級所可以獲得的有效程度

效果比較	SAT 閱讀	BSF 閱讀	SAT 數學	BSF 數學
比較小班級與（有與沒有助教的）正常班級學生表現的有效程度	.23	.21	.27	.13
比較有助教的正常班級與沒有助教的正常班級學生表現的有效程度	.14	.08	.10	.05

資料來源：Mosteller（1995）

度。由表 3 可看出，小班在數學與閱讀成績約優勝出四分之一的標準差，至於課程本位的 BSF，閱讀方面約超出五分之一的標準差，數學方面約超出十二分之一的標準差。Mosteller 進一步解釋超出四分之一的標準差約略所代表的意思，假設某學生並未接受小班教學，且假設其成績中等，即成績達到所有學生的第五十個百分位數，若他的成績增加四分之一個標準差，他將能從全部學生的 50%前進至 60%，比原本已經超越的 50%多越過 10%。這顯示了學生在表現上的確有向前躍進。

此研究原本的計畫是所有學生在實驗進行的 4 年中，均留在他們原本的班級類型中，但 1 年後有些家長對分派的方式有意見。經商討後，原本在有或沒有助教的正常班級中的學生，第 2 年重新分班，一半的學生隨機分派到有助教的班級，另一半則分派到沒有助教的班級，但小班級則不予以更動，之後亦沒有再改變。故此在研究的第 2 年尾聲，那些曾參與幼稚園與一年級階段的正常班級出現了以下四個情況：（一）2 年沒有助教；（二）2 年配有助教；（三）第 1 年沒有助教，第 2 年有助教；（四）第 1 年有助教，第 2 年沒有助教。由此可見，該研究開始演變得比較複雜，需要加入其他分析方法（如百分等級），有興趣的讀者可參考相關著作（Finn & Achilles, 1990; Mosteller, 1995; Mosteller et al., 1996）。由此研究可以觀察到，長期的隨機化試驗在執行上並不容易，有時會被迫做一些改變，但宜盡可能不牽涉到教育介入的內容，不然會影響到結果的詮釋。

綜合而言，STAR 計畫基本上應用了隨機化實地試驗作爲研究設計，Mosteller（1995）認爲其結果清楚顯示小班對幼童學習既有短程效益，也有長程效益，尤其是對於少數族群學生的學童。

三、科學教師成長研究

Borman 等（2008）曾進行一個與科學教育有關的集群隨機化試驗。他們指出，集群隨機化試驗將隨機化的單位層級提高至學校或教室，而在集群中蒐集個體（包含老師與學生）的資料。此外，因爲在學校的層級內，所有老師和學生均接受相同的處理或介入，所以能夠避免老師或學生因爲有機會溝通，得知另一組所採用之教育介入的內容，從而讓兩組產生混淆（contamination）。值得注意的是，集群隨機化實驗需要足夠的集群，才能提供足夠的統計考

驗力來估計集群層級的介入效果，以及進行適合分析集群資料的統計方法。

Borman 等（2008）研究所針對的，是在 System-wide Change for All Learners（SCALE）這個地方政府推動的科學教學與學習方案中，爲小學四及五年級教師提供的暑期專業成長研習，在地方政府持續提供輔導的環境下，讓教師學習如何運用該方案所發展以科學探索爲依歸的課程單元。Borman 等人的研究目的是要瞭解該專業成長研習及持續輔導的效果，爲此，研究者需要進行相關的師資訓練，而該研究主要報導在洛杉磯地區所進行的第 1 年情況與實驗結果。

（一）研究背景

美國輿論認爲，要追求科技的創新與經濟的成長，應回歸到年輕人的科學素養上，然而近年來，一些關於學童科學概念理解與知識的全國性評量卻出現令人擔憂的結果。2000 年所舉行的 National Assessment of Educational Progress（NAEP）測驗中，美國四年級學生的科學成績達到精熟程度者僅有 29%，與 1996 年之評量結果如出一轍。而近年 NAEP 的成績也顯示，在加州（California）地區甚至只有 17%達到精熟程度或以上，而全美國的平均則爲 29%。這些偏低的比例與大幅的比例落差，充分反映出增進基礎科學的教與學實爲全國性的需求，尤其是在那些擁有高比例弱勢學生之地區。

然而，何以學生在科學上的表現如此弱勢？教育學者和政策制定者又該如何修正這個情況？Borman 等（2008）指出，傳統的小學科學教學只將科學視爲是一門教材內容與一系列的步驟，學習的目標只是要獲得概念與精熟程序，然而這卻無法產生高層次的科學素養。許多研究指出，科學教育應使學生對於核心科學概念有更深入的瞭解，並且增進科學推理的能力。

（二）實驗方式

該研究採集群隨機化試驗的設計，整個計畫根據洛杉磯某學區內之八個社區的劃分，針對 191 所被提名學校中，從各社區隨機選出 10 所學校，其中 5 所學校隨機分派到實驗組，另外 5 所學校到對照組，總共選出 80 所學校進行研究（即 40 所實驗組的學校及 40 所對照組學校），而分析的樣本則爲來自此 80 所學校之四年級學生。實驗組的學校實行教育介入，提供小學四、五年級科學課的領導教師（lead teachers）專業發展機會，教導他們如何運用該課程的沉浸式單元（immersion unit），及後由他們再向校內其他科學課的教師分享，另外學期間還會持續提供輔導。至於被分派到對照組的學校，只爲它們的教師提供沉浸式的單元，但不提供相關的專業發展訓練機會。該研究並非要剝奪對照組教師受專業發展訓練機會，而是基於資源等緣故，先讓實驗組教師接受訓練。

該課程之所以被稱爲沉浸式，是因爲該專業發展研習可以訓練教師，在科學教室內實施內容廣泛並以探索爲導向（content-rich, inquiry-based）的科學課程，讓老師與學生沉浸在科學探索的完整過程中。Borman 等（2008）試圖研究地方政府努力推動的教師專業發展訓練，會

爲四年級學生的科學成就帶來什麼樣的影響，並藉由階層線性模式分析，檢視在學校層級之教師專業發展介入的效果。至於那些參與研究的學校，它們有各式各樣不同種族的學生，而且有許多是來自低收入戶的家庭。就讀這些學校的學生有四分之三是西班牙裔，將近 9%是非裔美國人，大約 9%是白種人，接近 3%是亞洲人，3%是菲律賓人，而美洲印第安裔和太平洋群島裔則少於 1%。此外，大概有 81%的學生具免費或減價午餐的資格，約 12%接受特殊教育，超過 42%是母語非英語者。爲了瞭解在正式實驗之前，實驗組學校與對照組學校的學生背景是否相似，不會干擾到以後研究結果的詮釋，Borman 等人蒐集包括種族的組成，具有免費或減價午餐資格學生的百分比，以及科學、閱讀與數學測驗成績等方面的資料，並對兩組學生背景進行統計考驗，結果發現兩組並無顯著差異。

（三）實驗結果

該研究有不少結果，比較特別的是研究發現教師接受科學沉浸式教學的專業發展，對於學生科學成就的影響並不如預期，此專業發展方式對於教學經驗低於 3 年的生手教師所教之學生有正面的影響，但是對於經驗豐富的教師所教之學生，反而有負向的影響。

爲了要解釋此負向影響，Borman 等（2008）推測了幾個原因，其中包括可能是因爲實驗組與控制組的退出率不同所致，或者是實驗組學校所用之範圍廣泛的測驗方式有關，因此造成較低的成績。後來根據分析的資料顯示，他們認爲這些解釋可能都說不通。Borman 等人認爲最有可能的解釋爲，在實行教學改革的階段，狀況會先變壞再變好。因爲革新需要教師發展出新的技巧與知識，而此過程通常約需 2-3 年的時間（Fullan & Miles, 1992）。故即使該研究的初步結果並不理想，這有可能只是暫時的情況，待將來教師熟練此教學方法時，便能克服此問題。其他研究結果請參原文。

陸、抗拒進行隨機化試驗的原因和與之抗衡的論點

本節將扼要交待爲何教育研究較少進行隨機化試驗的可能原因，以及推動隨機化試驗之學者所提出抗衡的理由，由於本文目的並非要做全面性的回顧，因此只選擇學者 Cook 的文章。之所以挑選 Cook 的原因在於他具代表性且有重要的學術地位，更特別的是，他雖然推動以隨機化試驗進行研究，卻能瞭解到教育環境的限制比較多，較難進行隨機化試驗，但他在一一分析教育界抗拒隨機化試驗的原因之後，仍然提得出相抗衡的論點。此外，Cook 因應不同的邀約及需求，撰寫了幾篇針對此議題的論文，性質皆相近，其中以 2002 年的論文最常被引用，本節遂偏重其內容以利做簡明的介紹。

根據 Cook（2001）的分析，教育學術界對隨機化試驗敬而遠之的原因之一，是在於很多教育學院的評估課程中，常傳達一些看似合理的反對隨機化試驗的見解。例如，他們認爲隨機化試驗是科學實證主義觀點所標榜的研究方式，其發展的起源來自農業及公共衛生，而與

教育研究無關。他們認爲教育的問題原本就十分複雜，每 1 所學校、每 1 間教室都有可能擁有其獨特的文化，同一種教育介入（intervention）的效果，亦會因學校的組織或人事的不同而有所差異，而且教育的改變應是很有創意地融合各種不同來源的知識，故此並不一定需要用實驗的方式來建立知識（參 Morrison, 2009）。可是，如果研究的目的是要瞭解某教育介入是否確實有效，隨機化試驗仍然是最有力的方法（Cook, 2001, 2002; Mosteller, 1995; Mosteller & Boruch, 2002）。

Cook（2001）認爲另外一個原因，是由於隨機分派被視爲是源自一種已經不適當的世界觀，它會遮蔽每所學校的獨特性，將學校內複雜的因果關係過分地化約。現時大部分教育評鑑學者擁護從後實證主義、民主、技能的模式來建立知識，他們認爲這種方式會優於來自科學的模式，後者對於改善學習並不能建立有效用的知識。

Cook（2001, 2002）進一步指出，反對的另一原因是很多事物的因果關係十分複雜。隨機化試驗只能處理少數幾個的原因及其交互作用，故此認爲隨機化試驗只宜應用於單一且明確的因果關係問題之上。而現實世界的教育問題卻是受許多因素所影響，因此很難從可觀察的改變中隔離出一、兩個作爲主要原因。而且在現實世界中，也不大可能存在一種教育介入，其效果可以無視不同的環境而放諸四海皆一致。換句話說，欲透過隨機化試驗建立一些教育方面的通則，對於我們身處這個複雜且多元的世界，並不見得是一件可行的事。對此，Cook 承認這是一個合理的批評，但他基本上認爲，這個批評並不僅限於隨機化試驗，許多採用其他方法的教育研究著作中，常會提及一些語氣肯定的因果關係命題，例如，有要求回家作業就可以改善成績，這些方法亦會遭到同樣的質疑。Cook 認爲要評估效果，研究就應該朝建立一套具因果解釋能力的設計，避免以爲使用某些特殊的介入即可以得到既直接且即時的效果。

另外一個原因是有些研究者會認爲過去就曾流行使用量化研究，然而並未能建立強而有力的通則。不過，Cook 回應指出，這主要是針對六、七〇年代量化研究的批評，其時一些重要的研究結果並無法被複製，亦沒有持久的效果，原因在於該類型的量化研究並沒有運用到隨機分派這要素。因此 Cook 覺得隨機化試驗並非如教育評估學者所說，是已經嘗試過且知道是行不通的方法。他認爲被嚴重批評的是六、七〇年代運用準實驗研究法的量化研究，該等研究用今天的眼光來說，其實並不算是質優的準實驗研究。

Cook 還討論了一個原因，他猜測教育研究者很少採用隨機化試驗，是因爲他們誤認爲隨機分派對教育研究而言是不切實際的，無論是在管理面或者是倫理面皆不可行。例如有研究者會擔心，如果運用隨機分派，隨機讓部分學生獲得教育介入的機會，另外一部分學生卻因爲實驗的緣故，而無法獲得教育介入的機會，這樣的處理有不公平的意味，尤其是當這些介入表面上看起來是很有效的時候，這將會引起家長的抗議，學校職員亦因此不敢做這類的安排。一般學校會希望能夠自主決定採用什麼方案或教育介入，而且好的方案應對所有學生一視同仁實施。要應付這一點，Cook 基本上認爲需要從政策面來切入，如果資助的單位能夠要

求明確的因果結論，這將可推動從事隨機化試驗的研究，正如公共衛生領域即是如此；美國幼教方面的研究亦因國會的要求而使用到隨機分派。相對來說，教育評估的研究者並不在乎隨機化試驗，當他們推薦一些教學最佳的進行方式時，並不在意其背後研究設計的品質，而比較在意其是否來自教育工作者的共識。

再者，Cook（2001）指出研究者對學校的觀念，也會對研究方法的選取有影響。以前的研究者視學校爲一個實地場所，它包含許多教室，教師在其內用他認爲可增加學生表現的教學方法，努力傳遞有效的課程。但七〇年代左右，研究者開始視學校爲一個複雜的社會機構，因此美國很多教育學院覺得有需要聘請政治學家與研究機構的社會學家加入，協助瞭解學校的管理與運作的方式，這些學者比較喜好質性的研究方法，也經歷過他們領域內的量化與質性研究之爭。此一觀念轉變帶來的其中一個後果，就是認爲隨機分派的先備條件尚欠成熟，因爲它必須依賴學校的管理與有高品質方案的實施。除此之外，則是認爲學校的文化及管理，必須透過深度的個案研究來進行，但量化研究對此的貢獻將極其有限。不過，Cook（2002）認爲隨機分派並不需要有非常明確的方案、理論，良好的學校管理或者是實驗處理，完全忠於方案的理論。實驗的主要目的是要避免做因果估計時有所偏誤（bias），其次才是避免估計值的不精確性。在研究設計上，學校的複雜度與不一致性會需要採用較大的樣本（即較多學校），並且要分析出差異的來源，以減少它們對於詮釋方案效果時所帶來的干擾。另一方面，在執行方案所指定的教育介入時，其品質可能因學校或教師的不同已有所差異，但可以視此爲依變項，以瞭解哪一類學校或教師執行得比較好。Cook 還提醒在教育的環境之下，介入必定無法完全標準化，即使某些介入成爲政策之後，還是會無法標準化地執行。因此，如果研究的目的是要瞭解在不太能夠標準化地執行某些介入的場所中，該等介入會有什麼效果，則研究仍可以進行。

Cook（2001, 2002）認爲，實驗的目的並不是要解釋依變項所有變異的來源，而是在於探查學校改革的措施是否會產生實質差異，儘管學校、教師、學生或其他因素本來就會有所差異。在瞭解到更多關於學校管理與實施之前，教育研究即可採用隨機分派的研究方式，並不需要迴避。

柒、討論

雖然近年來，隨機化試驗有備受學術界注意的趨勢，但從某一個角度來說，隨機化試驗帶有一點爭議性，這並不是說該方法有嚴重的問題，而是因爲社會的環境引發不同背景的研究者想要表達他們對該方法的不同意見。由於美國的教育部以至於國會都曾明顯出力推動隨機化試驗，雖則在一些相關的文獻中，亦有提及教育研究需要嚴謹的量化與質性的研究方法（例如 National Research Council, 2002），但仍然免不了讓很多研究者認爲，隨機化試驗有被特別高舉的感覺。有些質性研究者甚至覺得這一波推動隨機化試驗的努力，有把科學窄化爲

實證主義觀點下的科學，而有排斥其他觀點的趨勢。

因此除了前述 Cook 所整理出來的原因之外，有些研究者會針對隨機化試驗執行時可能被忽略的地方，進而質疑隨機化試驗所可能做出的結論，是否真能如倡議者所言的明朗確鑿。例如他們會指出，假設某隨機化試驗發現某教學法有效，然而在現實世界中，即使安排與實驗相同的教師用相同的教學法教導下一學期同年級的學生，也很可能會得到與預期效果有異的結果。與此相似的另一個質疑則認爲，由於隨機化試驗需要小心控制條件的緣故，所以隨機化試驗是在一個頗爲封閉的系統中進行。可是事實卻相反，我們所存活的真實世界卻是一個複雜的系統，內中元素與元素之間擁有很多網絡和連結，而且與周遭環境還會產生交互作用，因此隨機化試驗所得結果可應用之處會很有限。再加上參與者是會有感受的，在參與的過程中，他們的參與動機不同、態度不同，而且還可能會有所期望，因而影響到研究的結果，然而，這些因素通常會被一些採用隨機化試驗的研究者所忽略（參 Morrison, 2009）。

對支持者而言，這些反對意見是可以理解的，但他們想要強調的是，如果研究者十分在乎因果關係的建立，比方說他們很想知道小班級是否能讓學生獲得更好的成績，還是得透過隨機分派學生至不同大小的班級，以進行研究。至於隨機化試驗在執行上可能會有瑕疵，該方法的支持者亦有留意到這些問題，而且會研究改善的策略，尤其是關於集群隨機化試驗，近年來引起很多關注者的興趣。要執行集群隨機化試驗不但費用不菲，其中設計相關的考量要點甚多，要進行的觀察項目及評量的次數也非常多，因此這方面的研究者會探討如何有效訓練參與評量的觀察者。此外，爲了要能做出因果性的結論，在實驗的過程中，必須確保實驗組與對照組的教育介入，兩者在執行方面有足夠的忠精度（implementation fidelity），而不會中途走樣，因此該如何觀察與評估教育介入在執行時的忠精度，是近年來備受注意的研究主題。另一方面，有些參與者雖經隨機分派被列在名單之中，卻因各種可能緣故並沒有真正接受到教育介入，有不少研究探討如何可鼓勵參與，以及如何在此限制下，發展統計方法對相關介入的效果進行估計。再者，這方面的研究者十分留意實驗組與對照組參與者的流失率是否有出現不同（differential attrition），即使兩組一開始時有採用隨機分派，但因爲兩組的流失率有差異，兩組也可能變得不相等了，干擾到研究結果的詮釋。這該如何處理，亦爲相關學者所關心的研究主題，例如前述 Borman 等（2008）的研究，即有遇到流失率不同的問題。

在現實世界中，很多隨機化試驗會因爲受到現實環境的各種限制，因而必須針對原有的研究設計做出不同程度的改變，才能執行，可是這樣的處理方式是可接受的？抑或是不可取的？該如何看待？如果從正規實驗設計的角度來說，這樣的處理方式有可能會讓研究損失一些內在效度，因此可以說是減分的。但如果從現實的角度來說，爲了讓隨機化試驗能在並非單純的環境與限制之下仍可進行，從而部分放寬實驗控制的嚴謹性，這是迫不得已的作法。此時，研究者有責任將實驗環境與所做的任何改變詳實報導，並交待做這些改變的必要性。除此之外，如果能夠輔以質性資料的報導，將有助於對實驗結果的詮釋，尤其是對於要判斷

資料問題之所在這一層面。再者，如果能夠將研究問題及假設愈詳細說明，對於研究的設計愈有幫助。必須留意的是，在現實的環境下，很多實驗的研究設計並不能如一般研究方法課本或課程所陳述理想的方式來實行，而是需要進行修改，在可接受的範圍內做適度的妥協。

由於有些時候因受不同條件的限制，並不容易採用隨機分派的方式來進行研究，例如家長可能會被實驗的教育介入所吸引，從而要求將其被分派到對照組的子女轉介至實驗組，因此有些學者提出可以改而考量使用觀察性研究（observational studies）的方式來探討因果關係，然而在這種研究方式之下，爲了要排除組別間的樣本選擇偏誤（selection bias），近年來學術界提出了多種方法作爲因應，其中頗爲熱門的是採用傾向分數配對法（propensity score matching）的方式做配對，其基本想法是根據實驗組成員的某些屬性視之爲共變數，然後透過統計方法挑選出屬性與其最相似的個案作爲對照組的成員，藉以減少在研究起步時兩組成員在相關屬性上的差異（Rosenbaum & Rubin, 1983, 1984, 1985），以利後續探討實驗處理與結果之間的因果關係時，在估計實驗處理效果方面能夠減少偏誤。然而，此方法也有其限制，例如，它需要應用於大型的調查資料之上，如果資料的筆數不夠多，用傾向分數配對的效果可能會欠佳。此外，此方法只能夠針對有觀察到的屬性（或共變數）做調整，對於沒有被觀察到或考慮到的屬性就不能夠做調整，換句話說，用傾向分數配對法所選出來對照組的成員，可能在一些沒有被觀察到的屬性上，與實驗組成員並不相配。因此，傾向分數配對法對於該如何選取共變數以進行配對的問題十分重視，可是學者們對這方面的看法尚無一致的共識。再者，學術界針對傾向分數配對法的效果是否可與隨機分派相比較，也漸漸累積了一些實徵研究的成果，有興趣的讀者可參考 Glazerman、Levy 與 Myers（2003），Shadish、Clark 與 Steiner（2008）及 Sutherland（2010）的研究。另外還有一種名爲迴歸不連續設計（regression discontinuity design）的方法，得到部分學者（例如 Thomas Cook）的支持，但其統計考驗力比隨機化試驗略低，有興趣的讀者請參考 Shadish 等（2002）或其他相關的著作。

總而言之，隨機化試驗並不是沒有缺點的研究方法，因此無論是閱讀相關的報告，或者是要進行隨機化試驗的研究，都必須非常小心。一般而言，隨機化試驗著意於建立因果關係，它並不能反映出因果機制（causal mechanism），有些研究則可能需要深入探究因變項是如何影響果變項的，而這將屬於另一個研究主題。此外，隨機化試驗並不是停滯沒有發展空間的研究方法，相反地，投身於這方面研究的學者爲數並不少，近年來甚至有一些中興的跡象。教育領域有很多不同的研究方法，如果有興趣的讀者想要進行的研究其目的是要建立因果關係，建議不妨先小心考慮是否可以進行隨機化試驗，並多注意這方面的理論層面以及統計技術方面的發展。

這類型的介紹文章，由於主題內容甚豐，常會有掛一漏萬之困，因此以下向有興趣的讀者，推薦 3 本對隨機化試驗有較爲深入介紹的著作，以及 2 篇近期採用隨機化試驗的論文，可作爲進一步瞭解的起點。

一、「Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.」：該書中文版由楊孟麗研究員翻譯，由心理出版社出版。

二、Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

三、Bloom, H. S. (Ed.). (2005). *Learning more from social experiments: Evolving analytical approaches*. New York: Russell Sage Foundation.

四、「Trenholm, C., Devaney, B., Fortson, K., Clark, M., Quay, L., & Wheeler, J. (2008). Impacts of abstinence education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. *Journal of Policy Analysis and Management, 27*(2), 255-276.」：由學者 Trenholm 領導的研究團隊在評估美國國會資助的貞潔教育方案，獲得美國評估學會（American Evaluation Association, AEA）所頒發的 2009 年傑出評估獎。AEA 肯定此進行了 9 年，採用隨機化試驗的評估計畫之嚴謹性，並表彰其處理的方式是評估敏感議題的楷模。

五、「Baum, H., Kirsch, I., & Yamamoto, K. (in press). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record.*」：該文採用隨機化試驗探討 NAEP 閱讀方面的題目是否可評量出學生精熟度的研究，將於 2011 年刊出。

參考文獻

一、中文文獻

方金雅、蘇姿云（2005）。童謠教學對幼兒聲韻覺識影響之研究。**高雄師大學報，19**（2），1-19。

【Fang, C.-Y., & Su, T.-Y. (2005). A study of the influence of nursery rhymes on children's phonological awareness in Taiwan. *Kaohsiung Normal University Journal, 19*(2), 1-19.】

杜祖貽、呂俊甫（2007）。**教育學詞彙**。香港：香港中文大學。

【Du, Z.-Y., & Lu, C.-F. (2007). *Chinese terms in educational studies*. Hong Kong: The Chinese University of Hong Kong.】

洪儷瑜、黃冠穎（2007）。兩種取向的部件識字教學法對國小低年級語文低成就學生之成效比較。**特殊教育研究學刊，31**，43-71。

【Hung, L.-Y., & Huang, K.-Y. (2007). Two different approaches to radical-based remedial Chinese reading for low-achieving beginning readers in primary school. *Bulletin of Special Education, 31*, 43-71.】

章勝傑、李冠蓉（2003）。一個綜合性中輟預防方案的實驗研究。**台東師院學報，14**（1），1-28。

【Chang, S.-J., & Li, G.-R. (2003). A study on the effectiveness of a comprehensive dropout prevention program. *Journal of National Taitung Teachers College, 14*(1), 1-28.】

二、外文文獻

Anderson, T. W., & Finn, J. D. (1996). *The new statistical analysis of data*. New York: Springer-Verlag.

Bloom, H. S., Bos, J. M., & Lee, S. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review, 23*(4), 445-469.

Borman, G. D. (2009). The use of randomized trials to inform education policy. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 129-138). New York: Routledge.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year effects. *Journal of Research on Educational Effectiveness, 1*(4), 237-264.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage.

Brownlee, K. A. (1955). Statistics of the 1954 polio vaccine trials. *Journal of the American Statistical Association, 50*(272), 1005-1013.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*.

Chicago, IL: Rand McNally.

Cook, T. D. (2001). Sciencephobia: Why educational researchers reject randomized experiments. *Education Next, 1*(3), 63-68.

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis, 24*(3), 175-199.

Cook, T. D., & Payne, M. R. (2002). Objecting to the objections to using random assignment in educational research. In R. Boruch & F. Mosteller (Eds.), *Eridence matters: Randomized trials in education research* (pp. 150-178). Washington, DC: Brookings Institution Press.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Education Sciences Reform Act, 108 Cong. 2nd Sess. 48 (2002).

Finn, J. D., & Achilles, C. A. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*(3), 557-577.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1957). Dangers of cigarette-smoking. *British Medical Journal, 2,* 297-298.

Francis, T., Napier, J. A., Voight, R. B., Hemphill, R. M., Wenner, H. A., Korns, R. F. et al. (1957). *Evaluation of 1954 field trials of poliomyelitis vaccine: Final report*. Ann Arbor, MI: Poliomyelitis Vaccine Evaluation Center, University of Michigan.

Fullan, M. G., & Miles, M. B. (1992). Getting reform right: What works and what doesn't. *Phi Delta Kappan, 73*(10), 744-752.

Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science, 589*(1), 63-93.

Greenberg, D. H., & Robins, P. K. (1986). The changing role of social experiments in policy analysis. *Journal of Policy Analysis and Management, 5*(2), 340-362.

Knapp, T. (1998). *Quantitative nursing research*. Thousand Oaks, CA: Sage.

Morrison, K. (2009). *Causation in educational research*. London: Routledge.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children: Critical Issues for Children and Youths, 5*(2), 113-127.

Mosteller, F., & Boruch, R. (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings Institution Press.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill

grouping and class size. *Harvard Educational Review, 66*(4), 797-842.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Pearls, J. (2000). *Causality: Models, reasoning, inference*. New York: Cambridge University Press.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster-randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.

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

Slavin, R. (2003). A reader's guide to scientifically based research. *Educational Leadership, 60*(5), 12-16.

Spybrook, J., & Raudenbush, S. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis, 31*(3), 298-318.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martínez, A. (2009). *Optimal Design for*

longitudinal and multilevel research: Documentation for the Optimal Design software version 2.0. Retrieved April 16, 2010, from http://www.wtgrantfoundation.org

Sutherland, S. (2010). Using propensity scores to reduce selection bias in mathematics education Research. *Journal for Research in Mathematics Foundation, 41*(2), 147-168.

U.S. Department of Education (2002). *Strategic plan 2002-2007*. Retrieved December 7, 2002, from http://www.ed.gov/pubs/stratplan2002-07/index.html

Williamson, J. D., Karp, D. A., Dalphin, J. R., & Gray, P. S. (1982). *The research craft* (2nd ed.). Boston, MA: Little, Brown.

Journal of Research in Education Sciences
2010, 55(4), 69-95

Randomized Trials: Usage in Educational Research

Hak-Ping Tam
Graduate Institute of Science Education,
National Taiwan Normal University
Associate Professor

Abstract

Randomized trials have been around for many years as a research design. However, many factors have contributed to an emphasis on their use recently in the field of educational research and this has stirred much discussion from different aspects. The purpose of this paper is to offer an extended introduction to randomized trial design for educational researchers. It begins with an introduction to the social background that contributed to the current stress on the design. Next, the components that constitute a randomized trial will be explained with particular attention to the relationship between random assignment and causation. Two formats of randomized trials, namely, randomized controlled trials and cluster randomized trials will then be described. Afterwards, examples are given to illustrate the intricacies and the kinds of considerations needed for real world applications. Next, some concerns about randomized trials are briefly summarized followed by a presentation of counterarguments against them. The text closes with a discussion of relevant issues pertaining to randomized trial and observational studies. With awareness of the features of and the issues in such designs, interested researchers will be more comprehensive in planning a randomized trial.

Keywords: causation, cluster randomized trials, random assignment, randomized controlled trials

Corresponding Author: Hak-Ping Tam, E-mail: t45003@ntnu.edu.tw