P1-4

第一章　緒論

壹、計劃緣起

國外許多先進國家的教育系統，對於學生基本能力表現都有相當深切的關懷及具體明確的認知，因此，這些國家持續地進行教育資料庫的建置。國內教育系統也漸漸重視教育資料庫之建置，特別是在九年一貫課程實施之後，陸續有全國性學生學習成就資料庫之建置計畫，例如：臺灣教育長期追蹤資料庫（Taiwan Education Panel Survey, TEPS）、臺灣高等教育資料庫之建置及相關議題之探討、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）等，足以見到國內教育對此重視之端倪，其中臺灣學生學習成就評量資料庫為結合國內大學院校、學術研究機構等學者專家之學術專長以及資深教師的經驗，建置臺灣地區國小四年級、國小六年級、國中二年級、高中二年級與高職二年級學生數學科學習成就評量資料庫。其主要目的如下：

一、　建立國民中小學、 高中及高職學生學習成就長期資料庫，以追蹤、分析學生在學習上變遷之趨勢，進而檢視目前課程與教學實施成效。

二、　提供完整、標準化 的學習成就資料，作為分析學生學習成就上差異表現變項資料，以評估學生未來在學術方面能力之發展與社會期許。

三、　瞭解國內學校教學 及學生學習成效之現況，作為課程與教學政策改進之參考，並為縣市政府教育局及學校推動補救教學之重要參據。

四、　提供各縣市學生學 習表現資料，建立與縣市合作機制，以擴大資料庫應用效益。

五、　以資料庫的量化資 料，提供國內外相關研究人員，深入探討學生學習成就方面的相關政策議題。

六、　建立本國學生學習 成就評量資料庫，同時考慮與國際接軌，利於加入國際比較行列，藉以瞭解臺灣教育之獨特面與優缺點。

然而，國內教育資料庫目前仍是以仿造國外資料庫之建置方式，例如：試題研發、等化設計、施測流程、背景變項調查等，卻沒有制定一套符合國內教育資料庫之標準化測驗建置流程。除了使得資料庫僅侷限於國內學生間相互比較，無法達到與國外教育資料庫進行連結對照之功效外，亦無法將其價值與貢獻發揮至

極致。

近幾年國內積極參加一些國際評比之大型測驗（large-scale assessments），而國家教育進展評量（National Assessment of Educational Progress, NAEP）由於實施的時間較早，成為許多大型測驗之參考依據，分別說明如下：

一、國家教育進展評量

美國全國教育統計中心（National Center for Education Statistics, NCES）的最高行政長官負責執行 NAEP 政策，並由全國評量管理委員會（National Assessment Governing Board, NAGB）所制定之政策指導下執行其功能。NAEP 是美國評量學生成就之代表，自 1969 年便開始定期地對 4 年級、8 年級及 12 年級學生進行閱讀、數學、科學等科目之評量（The Nation's Report Card, 2005）。2005 TASA 英語科與數學科即是參考 NAEP 之評量架構，以檢測學生之能力。

二、國際學生評量（**The Programme for International Student Assessment, PISA**）

PISA 測驗是由「經濟合作與發展組織（Organization for Economic Co-operation and Development, OECD）」主辦，至從 1997 年起分別進行三次跨國的學生評量測驗（PISA 2000、PISA 2003 及 PISA 2006），國內參與第三次跨國學生評量測驗（PISA 2006），並積極參與 PISA 2009 前置作業。

三、國際數學與科學教育成就趨勢調查（**Trends in International Mathematics and Science Study, TIMSS**）

行政院國科會於 1999 年起參與「國際成就調查委員會（The International Association for the Evaluation of Education Achievement, IEA）」主辦之「第三次國際數學與科學教育成就研究後續調查（Third International Mathematics and Science Study Repeat, TIMSS-R）」，之後並改名為 TIMSS。國內學童除參與 TIMSS-R、TIMSS 2003 及 TIMSS 2007 之測驗外，國內許多研究人員亦參與此大型調查研究之相關工作。

由上述可知，國內教育系統已邁向國際化，進行與國際評比接軌之動作。目前國內參與之國際評量，仍是以抽樣設計、題本設計、測驗編製與內容等前置作業為主，國內大型標準化測驗之建置，對於較為核心之技術仍較

為缺乏，例如：抽樣設計與抽樣權重（sampling weights）、測量模式、試題特性與背景變項資料分析、量尺化程序（scaling procedures）等，這一方面國外大型測驗提供良好的範例，因此，本計劃欲藉由分析這些大型測驗(TIMSS、PISA、NAEP）的測驗發展程序與量尺建立之方式，以幫助本計劃如何使用標準化程序建立嚴謹之大型標準化測驗。

近年來，隨著資訊科技快速進步、測驗形式的改變及測量的概念日趨複雜，大型測驗之評量亦開始採用較複雜之測驗題型，例如：填充題、簡答題之類的建構反應試題（constructed response item），或是題組試題等，此類試題計分規則較為複雜且一份測驗或是一題試題可能測量許多不同的能力或特質，因此，必須配合適當的「測驗理論」才能從學生答題反應中萃取出所要瞭解的認知能力。

由 PISA、NAEP、TIMSS 技術報告公佈的評量架構，清楚呈現其測量之能力不單純的只有單一能力，這樣的測驗可能是多向度（OECD, 2006; The Nation's Report Card, 2009; Mullis, Martin, Ruddock, O`Sullivan, Arora, & Erberber, 2007）；然而，當試題是測量多向度能力，卻仍以單向度試題反應理論（unidimensional item response theory, UIRT）進行參數估計，將會產生偏差的試題參數估計和能力參數（Ackermean, 1991）。目前 NAEP、TIMSS 仍以 UIRT 為主要使用之測量模式，僅能對各個學科能力以單一能力值進行描述（Lee, Grigg, & Dion, 2007; Mullis, et al., 2007），對各學科所屬之次級量尺(subscales)表現較無法做精確描述；PISA 使用多向度試題反應理論（multidimensional item response theory, MIRT）中之多向度隨機係數多項 logit 模式（multidimensional random coefficients multinomial logit model, MRCML）進行測驗分析並對各學科之次級量尺進行估計；然而，PISA 使用多點計分模式對題組試題進行分析（OECD, 2005），未考慮題組試題對於參數估計之影響。Wang 和 Wilson（2005）研究結果顯示：如果測驗為題組試題之測驗題型，但卻忽略試題之間彼此可能相依之情形，則會高估能力參數且造成試題參數估計之偏差。

在 NAEP、TIMSS 及 PISA 中，除了對個別（individual）受試者之能力表現進行估計外，母群或母群中某些群體之能力表現亦為大家所關注之議題，國內常見的方式是直接使用個別受試者的成績（能力值）對母群或個別群體的表現進行

估計，常以個別受試者的成績（能力值）平均值或變異數代表該群體之某一能力表現及其分散程度，更進一步進行各種假設檢定，例如： TASA 數學科即採用此方式（洪碧霞、林素微、林娟如，2006）。依據 Mislevey 等人（Mislevey,1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; OECD, 2005; Lee, et al., 2007）之研究結果顯示，此種推論母群表現之方式容易造成偏誤。根據 Mislevey 等人之研究，可能值（plausible values）包含隨機誤差成分，不適合描述個體分數，但可能值具有良好群體估計一致性，適合描述群體之特性（Mislevey,1991; Mislevey, et al., 1992）。因此，目前國際上大型測驗皆以此種技術進行群體統計特性描述（OECD, 2005; Lee, et al., 2007），本研究擬深入探討此一技術之使用方式，運用於臺灣學生學習成就評量資料庫（TASA），進而介紹國內相關研究領域使用。

綜合上述可知，目前國際上較知名的大型標準化測驗在評量架構、試題與測量模式之配合上仍有不一致與不足之處。因此，本計劃擬探討 NAEP、TIMSS 及 PISA 等測驗之資料分析步驟與方法，進而提出適用於國內 TASA 之標準化資料分析步驟與方法。

貳、研究目的

針對上述背景及動機，特提出本研究計畫，研究目的如下：

1. 藉由國外大型標準化測驗（NAEP、TIMSS 及 PISA）之文獻蒐集，以建立一套適合 TASA 之標準化流程，包含抽樣設計與抽樣權重、測量模式、試題特性與背景變項資料分析、量尺化程序等。
2. 針對某一些議題進行實徵資料實驗，以評估所研發標準化測驗之成效。