TASA與其它國際評量之比較

曾建銘／國家教育研究院籌備處測驗與評量組組長

摘要

茲由國家教育研究院籌備處所建置之臺灣學生學習成就評量資料庫（TASA）與現行國外大型評比資料庫如TIMSS（Trends in International Mathematics and Science Study）、PISA（Programme for International Student Assessment）、NAEP（National Assessment of Educational Progress）之異同比較，希冀從國外建置之評量計畫中擷取經驗、技術作為TASA或國內大型資料庫建置之改進參考。

關鍵詞：臺灣學生學習成就評量資料庫、TASA、TIMSS、PISA、NAEP

前言

由國家教育研究院籌備處所建置之臺灣學生學習成就評量資料庫（TASA）與現行國外大型評比資料庫如TIMSS、PISA、NAEP之異同比較，希望從國外建置之評量計畫中擷取經驗、技術作為TASA或國內大型資料庫建置之參考標的。

以下分為下列幾點逐項討論：一、目的與性質；二、測驗題型、測量模式與分析軟體；三、測驗內容；四、施測年級與樣本；五、背景問卷；六、施測期程；七、主辦單位；八、分析技術；九、結論與建議。

一、目的與性質

TASA和TIMSS（Trends in International Mathematics and Science Study）、PISA（Programme for International Student Assessment）、NAEP（National Assessment of Educational Progress）等國外大型測驗皆在判定學生於學科之不同精熟程度，屬「標準參照測驗」。

TASA目的在於建立國民中小學及高中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，進而檢視目前國家教育體制與政策實施之成效（臺灣學生學習成就評量資料庫網站，2008）。

TIMSS屬國際評比，主要目的在於長期追蹤學生在數學與科學成就表現，並瞭解各參與國的教育改革、課程改革、及社會變遷的學習環境及改進措施的成效，以做為各國教育改革的方針（IAEP, 1991）。

PISA屬於國際評比，目的在評量學生是否有能力足以應付未來的世界，如在數學、科學或閱讀等面向，是以評量學生是否懂得應用或能擴充數學、科學等知識來解決真實世界中的問題（OECD, 2005）。

NAEP成立目的在瞭解美國學生學習進展情況，藉以促進教育改革與課程教學革新，並提供了解影響教育表現之因素（NAEP, 2009）。

二、測驗題型、測量模式與分析軟體

TASA和TIMSS（Trends in international Mathematics and Science Study）、PISA（Programme for International Student Assessment）、NAEP（National Assessment of Educational Progress）等國外大型測驗皆採嚴僅標準化流程進行試題編製。

TASA題型包含選擇題、題組題、應用題、

作文等建構反應試題。測量模式採3PL，分析軟體為BILOG-MG與SCORIGHT。

TIMSS題型包含選擇題、建構反應試題。測量模式採3PL、2PL與GPCM（Generalized partial credit model），分析軟體為BILOG-MG與PARSCALE。

PISA題型包含選擇題、建構反應試題（含closed 和 opened）。測量模式採MRCML （Multidimensional Random Coefficients Multinomial Logit model），分析軟體為ConQuest。

NAEP題型隨科目不同而有差異，主要為選擇題、建構反應試題與動手做工作單（hands-on task），探討學生利用工具觀察、進行調查、評估實驗結果及解決，而目的是以能朝向能真實在學校內、外碰到的情境為主。測量模式採3PL、2PL與GPCM（Generalized partial credit model），分析軟體為BILOG-MG與PARSCALE。

三、測驗內容

TASA包含國文、英文、數學、社會與自然。

TIMSS包含數學、科學。PISA包含數學、科學、閱讀、問題解決。

NAEP涵蓋十一項學科（包括閱讀、寫作、數學、科學、社會、公民、美國歷史、地理、文學、音樂、電腦教育）。

四、施測年級與樣本

TASA含括國小四年級、六年級、國中二年級、高中二年級與高職二年級。

樣本抽樣上採兩階段分層群集抽樣方式，每個學科作答樣本高達七、八千人，幾乎可等同推論至整個母群表現狀況。

TIMSS屬國際評比，於2007年參與國達六十五個國家，針對國小四年級、國中八年級實施。樣本抽樣上採兩階段分層群集抽樣方式。

PISA國際評比，於2009年參與國達六十七個國家，主要針對十五歲學生實施。樣本抽樣上採兩階段分層群集抽樣方式。

NAEP針對美國境內四、六、八、及十二年級學生實施。樣本抽樣上採兩階段分層群集抽樣方式，所得之美國樣本約達十六萬人附近。

五、背景問卷

TASA採學生、學校問卷。

TIMSS採學生、教師以及家長問卷，此外還加了課本分析（textbook analysis）以及一些課堂錄影分析（video study）以便了解「教什麼」以及「如何教」對學習成果的影響。

PISA採學生、學校問卷。

NAEP學生、教師以及家長問卷，後來又偶爾加收學生在校學籍記錄之資料（school transcripts information），以了解學生在校課程和學習活動與學習表現之關係。

六、施測期程

2005年只施測國小六年級國、英、數三科目，2006年、2007年針對五個年段、五個科目（小四社會除外）全面施測，2009年起每3年一個循環，分別對國小、國中與高中職進行施測。

TIMSS每4年施測一次，自1995年起，施測年度分別為1995、1999、2003和2007。PISA每3年施測一次，自2003年起，施測年度分別為2003、2006及接下來的2009年。

美國的NAEP依不同科目性質，分二年、三年、五年甚至八年或十年才重覆一次，而在閱讀能力、數、理能力評量，則是每二至三年重複一次。

七、主辦單位

TASA由教育部委託國家教育研究院籌備處辦理。

TIMSS由 IEA（International Association for the Evaluation of Educational Achievement）主辦。

PISA由OECD（Organization for Economic Co-operation and Development）主辦。NAEP由美國教育部的教育統計中心（U.S. Department of Education National Centerfor Education Statistics）主辦。

八、分析技術

TASA、TIMSS、PISA和NAEP等評量皆：

（一）採嚴僅現代測驗理論進行能力估計。

（二）施測採試題等化技術（BIB），可含括較廣的內容領域。

（三）經等化設計，所以可將受試者測驗的分數轉換到同一量尺上，以進行跨年級、跨學科、甚至跨年度的比較。

TASA、TIMSS、PISA、PIRLS和NAEP等評量，資料經整理釋出後，可供學術界作深入教育議題之探討，可應用之延伸價值高。

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，此類計分規則較為複雜之試題。而且一份測驗或是一題組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」才能從題目反應中萃取出所要瞭解的高層次數學能力。從PISA、NAEP、TIMSS所公布的試題範例（以數學科試題為例），清楚呈現其測量之能力不單純的只有單一能力，即測驗可能是多向度。然而，當試題是測量多向度能力，但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TASA、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, et al., 2007; Mullis, Martin, Ruddock, O`Sullivan, Arora, Erberber, 2005），對各學科所屬之次級量尺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 & 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。

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。

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（Mislevey,1991; Mislevey, et al., 1992）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。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）。

此外，量尺建立與標準設定之程序亦是大型標準化測驗之重要議題。在NAEP、TIMSS及PISA中，即是透過量尺化程序，將各重要測驗之分數轉換到一個共同的標竿上（例如：進行數理或閱讀成就表現之比較連結），以期作為教育者比較各國、各州、各地區、各學校、甚至個人表現的評比依據（National Research Council, 1999；Kolen, 2000）。然而，若能透過標準設定，使用決斷分數來定義成就或精熟水準，則不但可以減輕學習者對於受試學生能力理解或解釋上的負擔外，亦可以經由受試學生的成就水準或表現水準，來了解學生成就的趨勢。TASA除數學科依照NAEP將學生表現水準的描述分為三級，也就是基本的、精熟的、和進階的（洪碧霞、林素微、林娟如，2006）外，其餘各科並不一致。但依據國外測驗之量尺分數進行分級，是否適用於國內大型標準化測驗使用，各級之間的級距是否合理，決斷分數（cut scores）或稱表現標準（performance standards）是否適當，TASA尚未進行完整的檢視。

九、結論與建議

由以上國外大型測驗（NAEP、TIMSS、PISA）之比較分析，可以發現大型標準化測驗實施時之重要程序，有測量模式、量尺化程序、及標準設定，初步歸納出適合TASA模式，分別說明如下：

（一）適用於TASA之測量模式

國外大型測驗所使用之測量模式並不一致，例如：NAEP與TIMSS使用2PL、3PL、及GPCM之IRT模式；PISA使用MRCML與PCM之IRT模式。然而，TASA測驗二元計分試題包含單選題試題與題組試題，且學科測驗可能是測量多向度能力。因此，TASA擬進一步探究適合之測量模式，以降低能力參數與試題參數估計之偏差。

（二）量尺化程序

量尺化程序是大型標準化測驗相當重要的部分，主要是使用多重插補法或可能值的方法論來估計能力分布的特徵，以下針對TASA量尺化程序的幾個重要步驟進行說明。

1.試題量尺化

TASA量尺需由試題類型與適合之測量模式結合，且不同學科領域有其不同的量尺。因此，不同年級與不同學科領域之試題參數是分開估計。

2.產生量尺的可能值

量尺建立後，可能值是由預測量尺分數分布抽取而得。然而，對於不同學科量尺而言，多變量可能值向量是由能力值的聯合分布抽取而得，這些多變量可能值對於不同年級而言是分開計算的。最後，所有量尺的可能值將藉由適當的線性轉換機制，轉換至最終的量尺上。

3.建立量尺轉換機制

將估計之量尺經由線性轉換至TASA量尺，以建立屬於TASA之量尺分數。

4.定義合成的多變量量尺

此步驟是將個人於不同學科量尺的可能值進行合成，以產生學生整體能力（overall proficiency）之測量。然而，多變量量尺的合成必須藉由個別量尺平均可能值之權重，此權重反應出量尺相對之重要性與提供學科的發展架構。

5.分數報告

TASA個別能力估計使用MLE（maximum

likelihood estimation）估計受試者能力表現；群體能力估計將使用可能值的方法。

（三）標準設定

教育測驗情境中，不只是受試者甚至是學校、學區、縣市都需要被評估分類，如決定受試者是否通過或失敗，因此，標準設定之相關議題受到重視與討論。然而，藉由NAEP、TIMSS、PISA之探討，可發現各評量設定之通過標準分數（成就水準）並不相同。因此，TASA擬對標準設定之流程與標準設定之方法進行探討，以定義適合TASA之成就水準。

參考文獻

洪碧霞、林素微、林娟如（2006）。認知複雜度分析架構對TASA-MAT六年級線上測驗試題難度的解釋力。教育研究與發展期刊，2（4），69-86。

臺灣學生學習成就評量資料庫網站（2009）。臺灣學生學習成就評量資料庫。線上檢索日期：2009年6月10日。網址：http://tasa.naer.edu.tw/brief.htm.

IAEP (1989). A World of Differences. Princeton, NJ: Educational Testing Service.

IAEP (1991). The 1991 IAEP Assessment-Objectives for Mathematics, Science, and Geography. Princeton, NJ: Educational Testing Service.

NAEP (2009). The Nation's Report Card. 線上檢索日期：2009年11月10日。網址：http://nces.ed.gov/nationsreportcard/

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. Applied Psychological Measurement. 13, 113-127.

Lee, J., Grigg, W., & Dion, G.. (2007). The Nation's Report Card: Mathematics 2007. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Mullis, I. V.S., Martin, M. O., Ruddock, G. J., O`Sullivan, C. Y., Arora, A., & Erberber, E. (2005). TIMSS 2007 Assessment Frameworks. Chestnut Hill, MA: Boston College.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a facet random-effects facet model. Applied Psychological Measurement, 29, 296-318.

Mislevey, R. J. (1991). Randomization-based Inferences about Latent Variables from Complex Samples. Psychometrika, 56, 177-196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133-161.

OECD (2005). PISA 2003 Technical Report. OCED. Paris.

Nattional Research Council. (1999). Uncommon Measures: Equivalency and Linkage of Educational Tests. Washington, DC: Author.

Kolen, M. J. (2000). Issues in Combing State NAEP and Main NAEP. In J. W. Pellegrino, L. R. Jones, & K. J. Mitchell, (Eds.), Grading the Nation's Report card: Research from the Evaluation of NAEP. Committee on the Evaluation of National and State Assessments of Educational Progress.