TIMSS 2003 國小四年級數學新試題的開發及建構反應試題診斷性編碼系統的製定

林碧珍、蔡文煥
國立新竹師範學院　數學教育學系

壹、背景

國際教育學習成就調查委員會（IEA）繼 TIMSS 1995 與 TIMSS 1999 之後，繼續辦理國際數學與科學教育成就趨勢調查 2003（Trends in International Mathematics and Science Study, 2003）。我國在國科會和教育部共同資助下正式參與 TIMSS 2003 調查，藉此瞭解我國學生數學與科學學習的成就表現，並瞭解國際評量的新趨勢及新的評量方法。「國際數學與科學教育成就趨勢調查 2003」是一個四年期的研究計畫，由國立台灣師範大學理學院院長張秋男教授主持，組織一個研究團隊，這個研究團隊由研究組及實務組組成，研究組包括中學數學組及科學組，小學數學組及科學組、和資料分析組。實務組包括試務組、抽樣及資料處理組、調查工具組、及印刷組，研究成員合計 19 人。本文的兩位作者是這個研究群小學數學組的成員，主要負責與國小四年級數學相關的研究發展工作；諸如：新試題的開發、試題翻譯工作、試題的閱卷等。

參與此國際性的大型研究工作，從參與的過程中，學習了一個國際性的大型組織如何做好其嚴謹的研究流程，兩位作者希望藉由本文將我們獲得的珍貴經驗，與國內從事數學教育的研究者共同分享。由於 TIMSS 2003 的正式施測是在 2003 年的 5 月舉行，美國國際測驗中心（ISC）為了各國能如期順利的舉行，及各國能蒐集到可靠的資料，因此進行了三年的準備及發展相關的研究工作。在這期間，國際測驗中心在世界各地每年輪流舉辦兩次的國際研究協調代表（NRC）會議。從參與 NRC 國際會議過程中，我們學到了國際測驗中心如何發展試題、編製試題時如何決定診斷性編碼、試題閱卷過程的信度考驗，以及如何提高跨國之間、國家之內、和追蹤試題之評分者信度。因此，本文描寫的重點放在如何開發新試題及建構-反映試題評分編碼系統的製定。另外，對填充題或需要學生說明作法的簡答題如何進行評分也是本文描述的重點。

貳、新試題的開發

一、開發新試題的格式

國際測驗中心鼓勵這次參與 TIMSS 2003 的世界各國開發新的評量試題，以充實國際測驗中心的評量試題資料庫。試題的開發需依據 TIMSS 2003 評量架構及細

目（TIMSS Assessment Framework and Specifications 2003）（Mullis, et al.,2001），這個評量架構在數學科分成兩個向度，一個向度是數學內容領域；另一個向度是數學認知領域；數學內容領域是用來說明評量試題應該包含的數學內容，諸如：數、代數、測量、幾何、資料。每一個主題下各包含許多的數學子概念，例如：正整數與 0、分數和小數、比值、比例和百分比等，詳細的內容請參閱其他文章（林碧珍、蔡文煥，2003a;2003b）。數學認知領域是用來說明當學生探究數學內容時，我們所期望的學生認知行為，這些認知行為包括：知道數學的事實和過程、使用概念、解例行性問題、和推理。這個評量試題架構之各項數學內容及各項數學認知的分配比例不同，數、代數、測量、幾何、資料分別為 40%、15%、20%、15%、10%；知道事實和過程，使用概念、解例行性問題、推理的百分比分別為 20%、20%、40%、20%。

在開發試題時，必須依循試題架構項目分析表，命題者必須考慮要對數學內容哪一個子概念命題，同時必需考慮試題屬於哪一個認知領域。國際測驗中心在訓練開發試題時，要求各國開發試題的命題者依試題內容、形式、適用年級、試題特性、試題難度、標準答案、和各選項的背後想法，詳細填入規定的格式，格式的各項內容分述如下：

（1）試題難度估計

試題的難易度，需要在編製試題時一併列入考慮，這些難易度可以提供給國際測驗中心選擇題目的依據，因為一份評量試卷所包含的題目除了項目分析之外，試題的容易度均勻分配是一項不可或缺的考量。命題者依自己的經驗，預估該題的難易程度，其容易程度由 90%，80%，¼至 10%。對學生而言，百分比越高代表該題難度越低，越容易。

（2）試題描述

當題目編製好之後，要針對該題要評量什麼需做簡單的描述。例如：

下面哪一個分數比 $\frac{1}{2}$ 大？

① $\frac{3}{5}$　② $\frac{3}{6}$　③ $\frac{3}{8}$　④ $\frac{3}{10}$

對該試題的描述，其內容為：「這個問題主要是評量學生對分數 $\frac{1}{2}$ 的量感，而非進行異分母分數的比較。」

（3）選項的考量

選擇題的各個選項在編製內容時，需要有特殊的考量零理由，因為我們企圖從各選項的百分比能瞭解學生在各選項的答題狀況，雖然學生沒有選對正確的答案，但是也可以從其他選項看出學生的錯誤或迷思概念是什麼？例如：從上面分數的例子，有 47.3%的四年級學生選④，從答對率可以看出學生的錯誤是因為他們認為分母越大，則分數越大。

（4）備註欄

此欄位是用來說明該試題是否需要做文化國情上的考量，例如：若是美國的

山嶽，長又難讀，若需要翻譯成台灣的施測試題，則可以考慮更改山名。以下將台灣所開發的新試題中，舉一個選擇題和簡答題，來說明封面頁及內容頁的撰寫方式及內容。

（一）新試題開發的封面頁-選擇題

右欄各填選一項	年級	形式	數學內容領域	數學認知領域
	●四年級 ○八年級	●選擇題 ○填充題 ○簡答題	○數 ○代數 ●測量 ○幾何 ○資料 **主要子概念：** 工具、方法、公式	○知道事實及過程 ○使用概念 ●解例行性問題 ○推理

試題難度估計（學生可以正確回答的百分比）

90% 80% 70% (60%) 50% 40% 30% 20% 10%

容易 ◄——— 中等 ———► 困難

試題的描述：

（如下）

標準答案：(1)　　（僅供選擇題用）

選項的理由：

（2）忘了算"差 6 公分"。

（3）沒有釐清差 6 公分及剩下 6 公分。

（4）沒有理由。

備註欄：

（略）

試題來源辨識：

國家：台灣　　日期：6,28,2001

試題編號：3

（二）新試題開發的內容頁-選擇題

國恩用長 30 公分的尺量黑板，還差 6 公分就剛好量 9 次，請問黑板的長度是

（1）264 公分

（2）270 公分

（3）276 公分

（4）279 公分

試題來源辨識：

國家：台灣　　日期：6,28,2001

試題編號：3

（三）新試題開發的封面頁-簡答題

右欄各填選一項	年級	形式	數學內容領域	數學認知領域
	●四年級 ○八年級	○選擇題 ○填充題 ●簡答題	○數 ○代數 ●測量 ○資料 **主要子概念：** 工具、方法、公式	○知道事實及過程 ○使用概念 ○解例行性問題 ●推理

試題難度估計（學生可以正確回答的百分比）

90%　80%　70%　(60%)　50%　40%　30%　20%　10%

容易 ◄——— 中等 ———► 困難

標準答案：　（僅供選擇題用）

選項的理由：

（略）

備註欄：

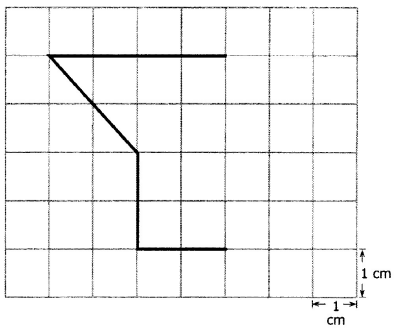
（略）

試題來源辨識：

國家：台灣　　日期：6,28,2001

試題編號：5

（四）新試題開發的內容頁-簡答題



每一小格的面積為 1 平方公分，請依照現有的粗線條上面繼續完成使它圍出來的圖形面積為 13 平方公分。

試題來源辨識：

國家：台灣　　日期：6,28,2001

試題編號：5

（五）簡答題的評分編碼系統

編碼	學生反應
正確回答	
10	畫出的圖形面積為 13 平方公分。
不正確的回答	
70	誤將半格計算為一格
71	直接畫出一條直線使這個圖形形成一個封閉的圖形
72	圖形為一個對稱的圖形
79	其他不正確的類型（包括擦掉、刪掉、或留下與本題無關的筆跡）
沒有回答	
99	空白

簡答題在命題時，除了題目內容之外，其各題的評分編碼系統亦需同時編製。不管學生所畫的圖形爲何，只要滿足面積爲 13 平方公分，都給予編碼 10。除了這種類型之外，其他的圖形若其面積不爲 13 平方公分,則再依不同的錯誤類型做分類。第一種錯誤類型是誤將半格計數爲一格，其編碼爲 70。第二種錯誤類型是直接畫出一條直線使這個圖形形成一個封閉的圖形，其編碼爲 71。第三種錯誤類型是所畫出的圖形是一對稱圖形，其編碼爲 72。不屬於以上所述的其他錯誤類型，其

編碼爲 79。如果學生沒有作答，留下空白的答案，其編碼爲 99。從各個編碼的次數，可以統計出學生解該題的各種方法之百分比。

二、編製新試題的原則

（一）選擇題的命題原則

TIMSS 2003 的試題分爲選擇題與建構-反應（constructed-responses）試題兩種形式，建構-反應試題主要是瞭解學生如何解題，瞭解學生的想法，故需要將做法寫出來。這兩種不同形式的試題，在命題的原則上有所不同。選擇題在命題時，必須很清楚的描述試題的題幹及四個選項的內容，並且要提供標準答案。以下是國際測驗中心提供給各國在編製選擇題時的命題原則。

（1）題幹的說明要足夠清楚，要讓作答者在還沒有閱讀選項之前就能清楚地知道要回答什麼問題。

不適當試題	適當試題
求解方程式 25-X＝19。	25-（ ）＝19，（ ）中的數字要多少才能使等式成立。

（2）題幹不要包含過多的訊息，除非該題的目的是要選出相關的資訊。

不適當試題	適當試題
養雞場的主人收回了 180 個雞蛋，主人將這些蛋送到 3 公里外的市場去賣。在送到市場去賣之前，主人先將這些蛋，每 12 個蛋裝成一盒，請問主人共需要幾個蛋盒才夠裝？ （1）13 （2）14 （3）15 （4）18	有 180 個蛋，裝成 12 個為一盒的蛋盒，問需要多少個蛋盒才夠裝？ （1）13 （2）14 （3）15 （4）16

（3）題幹要以提問問題方式呈現，盡量少用命令語句。

不適當試題	適當試題
求算長為 2 公分，寬為 6 公分的長方形面積。 （1）8 平方公分 （2）12 平方公分 （3）16 平方公分 （4）20 平方公分	長為 2 公分，寬為 6 公分的長方形面積為多少？ （1）8 平方公分 （2）12 平方公分 （3）16 平方公分 （4）20 平方公分

（4）若題幹中的答案不只一個，在題幹上最好加上"下列哪一個"的敘述。

不適當試題	適當試題
哪一個數比 4 大？ （1）5 （2）2 （3）1 （4）3 （意味著除了 5 之外，還有其他比 4 大的數）	下列哪一個數比 4 大？ （1）5 （2）2 （3）1 （4）3

（5）命題時，選項要確定僅有唯一的正確答案，避免有多種的可能答案。

不適當試題	適當試題
假如有一個數加 6 後比 7 大，這個數是多少？ （1）這個數大於 –1 （2）這個數大於 0 （3）這個數大於 1 （4）這個數大於 7 【(1)(2)(3) 皆為正確答案，(1) 與 (2) 的數可能包含於 (3)】	假如有一個數加 6 後比 7 大，這個數是多少？ （1）這個數小於 –1 （2）這個數等於 –1 （3）這個數等於 1 （4）這個數大於 1

（6）命題時避免能從選項中的數字倒推答案，若是屬於這種問題，可以改爲填充題或簡答題方式。

不適當試題	適當試題
下列哪一個數可以滿足 3 ×（ ）＝17 （1）4 （2）5 （3）6 （4）7	滿足 3 ×（ ）＝17 的（ ）為多少？

（7）在設計選項時，應依作答者最有可能發生的錯誤或迷思概念來考量，以避免學生用消去法猜得答案。

不適當試題	適當試題
下列哪一個合乎三角形的性質？ （1）有四個角 （2）有四個邊	下列哪一個合乎三角形的性質？ （1）三邊長為 2、4、6 公分，可以形成一個銳角三角形

（3）內角和為 180° （4）有一雙對邊平行且相等。	（2）有一個大於 90°角為銳角三角形 （3）內角和為 180° （4）等腰三角形也是正三角形的一種。

（8）命題時，避免用錯誤的方法也能巧合算出正確答案。

不適當試題	適當試題
半徑為 2 公分的圓面積為多少平方公分？ （1）4 （2）2π （3）8 （4）4π （誤將周長當成面積也能得出 2πr ＝ 2π×2＝4π）	半徑為 3 公分的圓面積為多少平方公分？ （1）6 （2）9 （3）6π （4）9π

（9）選項的排列應考慮邏輯順序（數字由小到大或由大到小）。

不適當試題	適當試題
下列哪一個數可以同時除盡 21 和 35？ （1）7 （2）3 （3）9 （4）5（將正確答案排在第一個）	下列哪一個數可以同時除盡 21 和 35？ （1）3 （2）5 （3）7 （4）9

（10）每一個選項的內容要平行處理，敘述的長度要相當，避免對正確的選項多做描述。

不適當試題	適當試題
下列何者正確？ （1）三角形至少有一個角大於 90°。 （2）正三角形也是等腰三角形。因為正三角形至少有兩個邊等長，所以也是等腰三角形。 （3）三角形不一定有高。 （4）銳角三角形一定有兩個角相等。	下列哪一個正確？ （1）三角形至少有一個角大於 90°。 （2）正三角形也是等腰三角形。 （3）三角形不一定有高。 （4）銳角三角形一定有兩個角相等。

（11）在題目中盡量避免重複出現的語詞。

不適當試題	適當試題
691+208 的和最接近？ （1）600+200 的和	691+208 的和最接近下列哪兩個數的和？

（2）700+200 的和 （3）700+300 的和 （4）900+200 的和	（1）600+200 （2）700+200 （3）700+300 （4）900+200

（12）題幹避免使用否定詞，如：不對、至少、最差、除...之外。假若題幹以否定詞描述，則選項盡量避免再使用否定詞。

不適當試題	適當試題
下列哪一個敘述不真？ （1）長和寬不相等的長方形不是正方形。 （2）等腰三角形不是三角形。 （3）正方形不是菱形的一種。 （4）平行四邊形不是菱形的一種。	下列哪一個敘述是對的？ (1)長和寬不相等的長方形是正方形。 （2）等腰三角形是三角形。 （3）正方形是菱形的一種。 （4）平行四邊形是菱形的一種。

（13）一個選擇題最好一次只評量一個概念，避免涉及多個概念。

不適當試題	適當試題
下列哪一個是 4+0.04 和 0.6x6 之間的中間數？ （1）0.44 （2）3.6 （3）3.82 （4）4.04	下列哪一個是在 0.2 和 1.6 之間的中間數？ （1）0.6 （2）0.7 （3）0.9 （4）1.4

（14）題目儘量避免使用您或您們。

不適當試題	適當試題
您最近四次數學考試成績為 77，85，79，83，請問您的平均分數是多少？ （1）77 （2）81 （3）82 （4）85	偉明最近的四次數學考試成績為 77，85，79，83，請問偉明的平均分數是多少？ （1）77 （2）81 （3）82 （4）85

（15）涉及日常生活的題目，盡量問些與生活相關的問題。

（16）試題的選項避免使用“以上皆非”及“以上皆對”。

（17）命題時要確認選項是用來回答題幹上的問題，避免作爲解其他題目的線索。

（二）”建構一反應” 試題的命題原則及診斷性編碼系統的製定

有些概念無法用選擇題評量出來，這時可能需要以”建構一反應”試題的形式呈現，建構反應試題必須由學生自行解出答案。建構一反應試題分爲填充題（short-responses）和簡答題（free-responses），填充題需要學生寫出數字的答案，或完成一個表格或畫圖；簡答題需要學生對作法或答案進一步說明及解釋。

1.”建構一反應” 試題的命題原則

在編製 ”建構一反應” 試題，國際測驗中心提供下列的命題原則：

（1）命題時，使用的語言要考慮施測對象的年齡。

（2）編製試題時要考慮學生的作答時間，一般填充題作答時間估計每題約爲 1 至 2 分鐘，而簡答題的作答時間每題約爲 3 至 5 分鐘。

（3）試題盡量與日常生活情境相結合。

（4）題目的答案範圍不要太廣。

（5）要填寫的答案能明確的在題目上說明清楚。例如：“請給三個例子”，避免用“請給一些例子”。

（6）試題上不要有任何暗示的答案或可作爲解其他題目的線索。

（7）在編製建構反應試題時，要同時提供每一個試題的評分指南。

2、TIMSS 2003 “建構一反應” 試題的評分編碼系統

（1）簡答題的評分編碼系統

命題者在出題時，就要同時思索著學生解該題時所使用的解題策略與方法。完全正確的解法皆以 20、21 等碼號，十位數字的 2 代表完全正確的解法，得分爲 2 分，若該題的編碼有 20、21、22，代表著學生用了三種正確的方法解該題目，20、21、22 之個位數字有 0、1、2 代表診斷性編號。部分正確的解法，給予 10、11、12、13 等四個碼或更多碼，十位數字的 1 代表學生的解法有部分是正確的；10、11、12、13 的個位數字 0、1、2、3 有四個診斷編碼，代表學生用了四種的解法，每一種解法只有一部份是正確的。同樣的， 70、71、72 或更多的編碼，代表完全不正確的作法。編碼系統的不正確解法是依據題目的條件與數據，學生可能最常犯的錯誤或迷思概念來決定的。若非以上所列的錯誤解法或出現與題目無關的符號或數學符號，則給予編碼 79。若學生的答案卷在該題爲空白，則給予編碼 99。

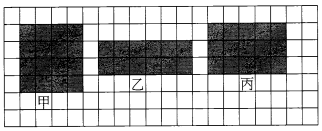
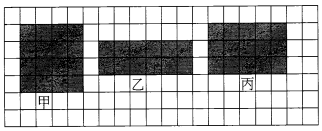
換言之，TIMSS 的診斷性編碼系統之所有的編碼皆爲兩位數字，十位數字 2、1、7 分別代表學生的回答爲全對、部分對、全錯。第二個編碼是診斷性編碼，數碼爲 0 到 5（例如：20 到 25，10 到 15，70 到 75）。第二個編碼 9 代表其他類型的作法（如：29、19、79）；編碼 99 代表學生的作答爲空白。由於各國之間有文化上的差異，各國之間也許對某一個題目有特定的解法，而沒有列在編碼系統表，此時各國可以針對實際需要在第二個數碼增加 7 和 8。

（2）填充題的評分編碼系統

填充題的評分編碼系統與簡答題的評分編碼系統一樣，以兩位數字來編碼，數碼 10 到 19 代表學生的回答全對，得一分；70 到 79 代表學生的回答不正確，得 0 分；99 代表空白答案，得 0 分。

參、從開發到使用之間在試題內容的改變

我國小學四年級這一次 TIMSS 2003 所開發的十四個試題，送到國際測驗中心的試題資料庫，其中有四個題目被放入 TIMSS 2003 的試測試題，這四個題目有些被國際測驗中心修改過，試題的開發到試題被接受使用之間的過程是：首先我們將所開發的中文試題翻譯爲英文試題寄至國際測驗中心，並由國際測驗中心修改英文語詞與文法，成爲正式的英文版試題；然後，放在題本的試測英文版試題由各個國家依自己使用的語言，如台灣將英文試題翻譯成中文試題，再交由國際測驗中心尋找國際語言中心，對我們所翻譯的中文試題再做最後的確認。今將我們開發的四個中文試題到試測題本上的中文試題，分別一一呈現，以方便做對照與比較兩者之間的不同。

台灣開發的中文試題	題本上的中文試題
 上面三個圖，哪一個圖形周長最大？ （1）甲 （2）乙 （3）丙 （4）一樣大	 下面哪一個對矩形甲、乙、丙的敘述是正確的？ 台灣　國際 ①甲的周長最大　63.5%　51.5% ②乙的周長最大　13.7%　9.8% ③丙的周長最大　4.6%　8.6% ❹甲、乙、丙的周長相等　16.9%　24.0%

依據上面兩個題目的對照，國際測驗中心將我們開發的試題保留原來的圖形及題幹中的周長概念，只是將選項改變了敘述的方式。從 TIMSS 2003 試測的學生答題百分比顯示，我國四年級學生誤將面積視爲是周長；其他國家也至少有五成的學生有相同的迷思概念。

台灣開發的中文試題	題本上的中文試題
下面為「臥虎藏龍」電影的播放時間表： 場次 | 電影開始時間 第一場 | 9：00 第二場 | 11：40	以下是平日電影播放的時間表： 場次 | 電影播放時間 第一場 | 下午 2：00 第二場 | 下午 3：30

第三場	2：20
第四場	？

假如任意兩個場次撥放的時間間格一樣，請問第四場電影開始的時間是：

（1）3：20

（2）4：40

（3）5：00

（4）5：20

第三場	下午 5：00
第四場	？

如果依照這個規則繼續下去，那麼第四場電影開始播放的時間是？

	台灣	國際
①下午 5：30	13.6%	11.6%
②下午 6：00	11.4%	8.2%
❸下午 6：30	61.4%	66.0%
④下午 7：00	12.3%	10.9%

依據上面兩個題目的對照分析，國際測驗中心將我們開發的試題內容之電影播放時間表格做了稍微的修飾，並且把電影播放的時間由上午的時段全部改爲下午的時段，而且每場電影的間隔時段由 2 小時 40 分縮短爲 1 小時 30 分。我們提供的試題明確的指出片名爲「臥虎藏龍」，而該片實際播放的時間爲 2 小時 40 分；而國際測驗中心將片長改爲一般兒童電影的片長，每一場電影的播放時間縮短爲 1 小時 30 分。

從以上的題目，台灣學生在我們國家開發的試題之試測表現，平均通過率皆低於國際平均通過率。依常理而言，這些試題的開發來源都是第一現場的老師所設計的，台灣學生在這些題目的表現應當較爲理想，但事實卻不盡然，學生的表現反而不如其他國家學生的表現；相反的，台灣學生在其他國家所開發的試題之表現卻比其他國家的學生平均通過率高，其原因值得進一步深究。

誌謝

本論文之所以能完成，首先要感謝張秋男院長主持的「國際數學與科學教育成就趨勢調查 2003」專題研究計畫的研究群伙伴同意兩位作者使用試測蒐集的小學四年級數學的資料。

肆、參考文獻

1.林碧珍、蔡文煥（2003a）：四年級學生在國際教育成就調查試測的數學成就表現。科學教育月刊，第 258 期,2-20。

2.林碧珍、蔡文煥（2003b）：我國國小四年級學生在國際教育成就 2003 試測的數學成就表現。論文收錄於九十二學年度師範學院教育學術論文發表會論文集（編號 92115）。論文發表於 10 月 24～25 日。國立台南師範學院編印。

3.Mullis, I. S, Martin, M.O., Smith, T.A, Garden, R.A, Gregory, K.D, Gonzalez, E.J,Chrostowski, S.J & O'conner, K.M. (2001). *TIMSS Assessment Framework and specification 2003*, International Study Center. Lynch School of Education, Boston College.