均分數及國際排名爲何？

爲進一步瞭解我國國二學生在各個科學主題的學習成就，擬回答問題如下：

(3) 分別以 TIMSS 2007 科學成就排名前十名國家/地區學生的平均表現以及新加坡學生的表現作爲參照，我國國二學生在哪些科學主題上相對表現較弱？

表 8：TIMSS 與 PISA 提供的試題資訊及量尺分數之類型

學習目標	描述層次	TIMSS		PISA	
		試題資料	量尺分數	試題資料	量尺分數
科學內容知識	學科領域	○	○	乂	乂
	科學主題	○	乂	乂	乂
關於科學的知識	面向	NA	NA	乂	乂
	次類別	NA	NA	乂	乂
認知能力/科學能力	領域/面向	○	○	○	○
	指標能力/次類別	乂	乂	乂†	乂

註：○表示有該層次的試題資訊或量尺分數。乂表示無該層次的試題資訊或量尺分數。NA 表示該學習目標根本不存在於評量架構中。

†：除了已公布試題之外，在未公布試題的資料中僅公布了各試題所屬的科學能力面向，沒有說明各試題所對應的科學能力面向次類別。

表 9：TIMSS 2007 與 PISA 2006 我國學生於各項學習成果之國際排名平均分數

	TIMSS 2007			PISA 2006		
	次項	排名	平均(標準誤)	次項	排名	平均(標準誤)
整體		2	561 (3.7)		4	532 (3.7)
科學內容知識	生物	3	549 (1.9)			
	化學	1	573 (4.2)			
	物理	4	554 (3.7)			
	地球科學	1	545 (2.9)			
認知能力	認識	2	560 (3.4)	辨識科學議題	17	509 (3.7)
	應用	1	565 (3.5)	科學解釋現象	3	545 (3.7)
	推理	5	541 (3.5)	運用科學證據	8	532 (3.7)

貳、研究方法

研究問題一及二所分析的是 PISA 的資料庫，研究問題三分析的是 TIMSS 的資料庫。二者有所不同，分述如下。

在研究問題一及二中，首先要利用「學生年級」這個變項將 PISA 資料庫中 7-9 年級

學生的部分挑選出來。其次在計算各國學生科學素養及分項科學能力平均分數時要考慮樣本權值，在計算標準誤時還要考慮取樣誤差及測量誤差。最後在比較我國與他國學生之間有無顯著差異時，以獨立樣本 t 檢定爲之。

根據 PISA 資料分析手冊 (OECD, 2009b)，標準誤中取樣誤差的部分透過複製法（replication）來估計。所謂複製法意思就是重複估計統計量。在每一次在估計平均值的時候，部分觀察體會被系統地去除，其觀察值視作缺漏值。如此重複計算平均值之後，求取平均值的標準差。實務上，PISA 採用所謂平衡式重複複製的 Fay 方法（Fay' s variant of the Balanced Repeated Replication）來系統地去除觀察體和處理缺漏值，重複計算的次數是 80 次。公式如下：

$$\sigma_{\hat{\theta}}^2 = \frac{1}{20}\sum_{j=1}^{80}\left(\hat{\theta}_j - \hat{\theta}\right)^2$$

式中 $\hat{\theta}$ 是沒有刪除任何觀察體時的平均值估計值，$\hat{\theta}_j$ 是刪除部分觀察體的平均值估計值。標準誤中測量誤差的部分則利用 PISA 提供的學生量尺分數的五組似真值（Plausible Value）來處理。亦即，利用每一組似真值都會求得一個平均值估計值，五個平均值估計值的標準差，就是標準誤中測量誤差的部分，公式如下：

$$\sigma_{test}^2 = \frac{1}{4}\sum_{j=1}^{5}\left(\hat{\mu}_j - \hat{\mu}\right)^2$$

式中 $\hat{\mu}_j$ 是第 j 組似真值求得的平均值，$\hat{\mu}$ 是五個 $\hat{\mu}_j$ 的平均，也就是最後的平均值估計值。最後考慮了取樣誤差和測量誤差的標準誤如下(OECD, 2009b, p.118)：

$$\sigma_{error}^2 = \frac{1}{5}\left(\sum_{j=1}^{5}\sigma_{\hat{\theta},j}^2\right) + 1.2 \cdot \sigma_{test}^2$$

式中 $\sigma_{\hat{\theta},j}^2$ 是以第 j 組似真值計算出來的取樣誤差。PISA 提供了 SPSS 的巨集檔可供利用，可方便地計算出平均值及標準誤。

在研究問題三中，學生在各個科學主題上的表現是以答對率作爲指標。亦即先計算學生在該主題上各個試題的答對率，然後求取該主題所有試題的平均答對率，以此作爲學生在該主題上的平均表現。跨國比較時，答對率的變異呈現在該主題各個試題的答對率上，由於不同國家都做了相同的試題，因此以配對 t 檢定來考驗差異的顯著性。