談教育現場的教學評量

羅 豪 章

Professor

Department of Digital Content and Technology
National Taichung University of Education

教育現場為什麼要教學評量?

- 被動的理由
- 主動的理由

想了解自己的教學

學生的學習經驗

- 學生學會了沒?

  認知、技能、情意等

- 學生喜歡這個課程嗎?

  滿意度、態度、動機等

教學者的教學經驗

- 是否達到課程設計的教學目標?
- 自己的教學想法是否得到驗證?

如何回答前述問題?

- 量化：透過測量工具取得數據，用統計讓數字說話
- 質性：觀察、晤談、問卷

測量工具

- 在自然科學領域中，我們對實體已發展出精確、完整的測量概念(concept)和儀器(例如:身高、體重、時間等)。
- 在社會科學中，我們也需要測量非實體的世界(例如:滿意、喜歡、態度、美感、動機等) ，然而這些都是屬於人類建構出來的抽象概念(構念，construct)，故無法使用自然科學領域測量實體的工具進行測量，結果也無法那麼地精確。

社會科學常用的測量工具

- 測驗(test)
- 問卷(questionnaire)
- 指數(index)
- 量表(scale)

測驗(test)

在心理與教育測驗中，測驗可分

- 客觀測驗(objective test)
- 主觀測驗(subjective test)

客觀測驗

- 客觀測驗包括是非題、選擇題、配合題、排列組合、填充題等，試題經過仔細的篩選，且有事先預定、固定而客觀的評分標準或正確答案，因而具有診斷學生學習困難的作用，並有助於增加測驗的效度，而且測驗項目簡短，受試者的反應有一定的型式，故可在有限的時間內答完。不論評閱者是誰，所得的結果都會一樣，故一分試卷應得的分數相當確定，如此可避免評分者的主觀判斷，使得測驗的結果不致有太大的差異性，而引起爭議，有助於提高測驗的信度。

主觀測驗

- 主觀測驗測驗是教師提出一些問題，由學生依其處理問題、使用和組織資料的方式自由回答。這種測量方式可以測量出學生的「組織」、「統整」及「表達觀念」的能力，這些能力通常無法以客觀測驗來測量。口試(oral test)、論文測驗(essay test)與投射測驗(projective test)等均之。
- 這種測量方式在評分時易受評分者的主觀意識所影響，不同評分者或是同一位評分者在不同情境下可能結果不盡相同。再者，這種測量和評分方式比較費時。

問卷(questionnaire)

- 問卷通常是根據測量者想要知道的內容(可以回答所要了解的問題)而設計，可以不需要理論的依據，只要符合探究的主題即可。
- 問卷在分析時，大部分是以每一題為分析單位。

問卷範例

教師想知道在翻轉教學時，學生課前自學教材的情形：

- 1.每次上課前你平均會花費多少時間閱讀教材?
- 2.你大部分會在上課多久前閱讀教材?
- 3.你如果閱讀教材遇到問題，你會如何尋求解答?
- 4.
- .
- .

指數(index)和量表(scale)

- 在測量構念時，往往不是一個問項就可以測得，通常我們會採用複合測量（composite measure），就使用數個問項組合，間接測得這個構念。
- 指數和量表都是社會科學研究常用的複合測量工具。
- 在建構指數和量表時，必須根據理論為依據去設計。一般而言，一個構念可能會涵蓋數個構面(dimensions)，每一個構面就會有一個分量表。
- 計分時則是以每一分量表(即包含分量表中所有題目)為分析單位。

- 指數是根據受測者對每一個問項的回答加以計分，再計算各問項得分所得到的分數。一般而言，我們會嘗試以這個分數去描述一個我們想了解的現象。
- 例如:藉由健康狀況問卷（Patient Health Questionnaire-9, PHQ-9），透過PHQ-9中9個問項所得積分，一個人便可進行自我檢測自己是否有憂鬱的傾向。
  https://www.cgh.org.tw/rwd1320/store/f4/CGHPHD_2.pdf

- 相較於指數，量表則更強調所欲測量構念的理論架構，它能針對各個問項之間的結構提供更有精準的順序排列功能，並考慮被列入複合測量的問項可能會有不同的強度。
- 量表有很多不同的類型，如:李克特量表、鮑氏社會距離量表、語意差異法量表、瑟式量表、古特曼量表等，其中最常見的就是李克特量表(Likert scale)。

- 李克特量表(Likert scale)由一組陳述組成，每一陳述有"非常同意"、"同意"、"不一定"、"不同意"、"非常不同意"五種回答，分別記為1，2，3，4，5，每個被調查者的態度總分就是他對各道題的回答所的分數的加總，這一總分可說明他的態度強弱或她在這一量表上的不同狀態。
- Likert scale將屬同一構念的問項用加總方式來計分，單獨或個別問項是無意義的。

量表範例

想了解大學生禮儀課程之學習動機與學習滿意度，設計「禮儀學習動機量表」與「禮儀學習滿意度量表」為測量工具。「禮儀學習動機量表」包含價值、期望和情感三個構面，「禮儀學習滿意度量表」則包含教師教學、 課程內容、人際關係、學習成果及學習環境五個構面。

http://www.feu.edu.tw/adms/aao/aao95/jfeu/25/2504/250301.pdf

量表範例

體育課滿意度量表(發表於輔仁大學體育學刊第五期)：
將學生對於體育課滿意度的評量問項，分成場地與器材、教師能力與素養、體育教學效果、 身體與能力發展、同儕關係和體育教學行政等六個構面，共 38 題。

http://www.phed.fju.edu.tw/article/publication-5/04%E9%99%B3%E6%98%A5%E5%AE%89-%E9%AB%94%E8%82%B2%E8%AA%B2%E6%BB%BF%E6%84%8F%E5%BA%A6%E9%87%8F%E8%A1%A8%E7%B7%A8%E8%A3%BD%E7%A0%94%E7%A9%B6%EF%BC%8D%E4%BB%A5%E5%8D%97%E5%8F%B0%E7%A7%91%E6%8A%80%E5%A4%A7%E5%AD%B8%E7%82%BA%E4%BE%8B.pdf

量表範例

學術研究動機量表：

研究者根據依Pintrich（1989）與 Wigfield 與 Eccles（2000）對動機之分類與觀點，將學術研究動機分成三個構面「價值成份」、「期望成份」和「價情成份」分別包含 6 題、9 題與 5 題，共20 題。

http://www3.nccu.edu.tw/~ycyeh/instrument-english/2007%20academic%20research%20motivation.pdf

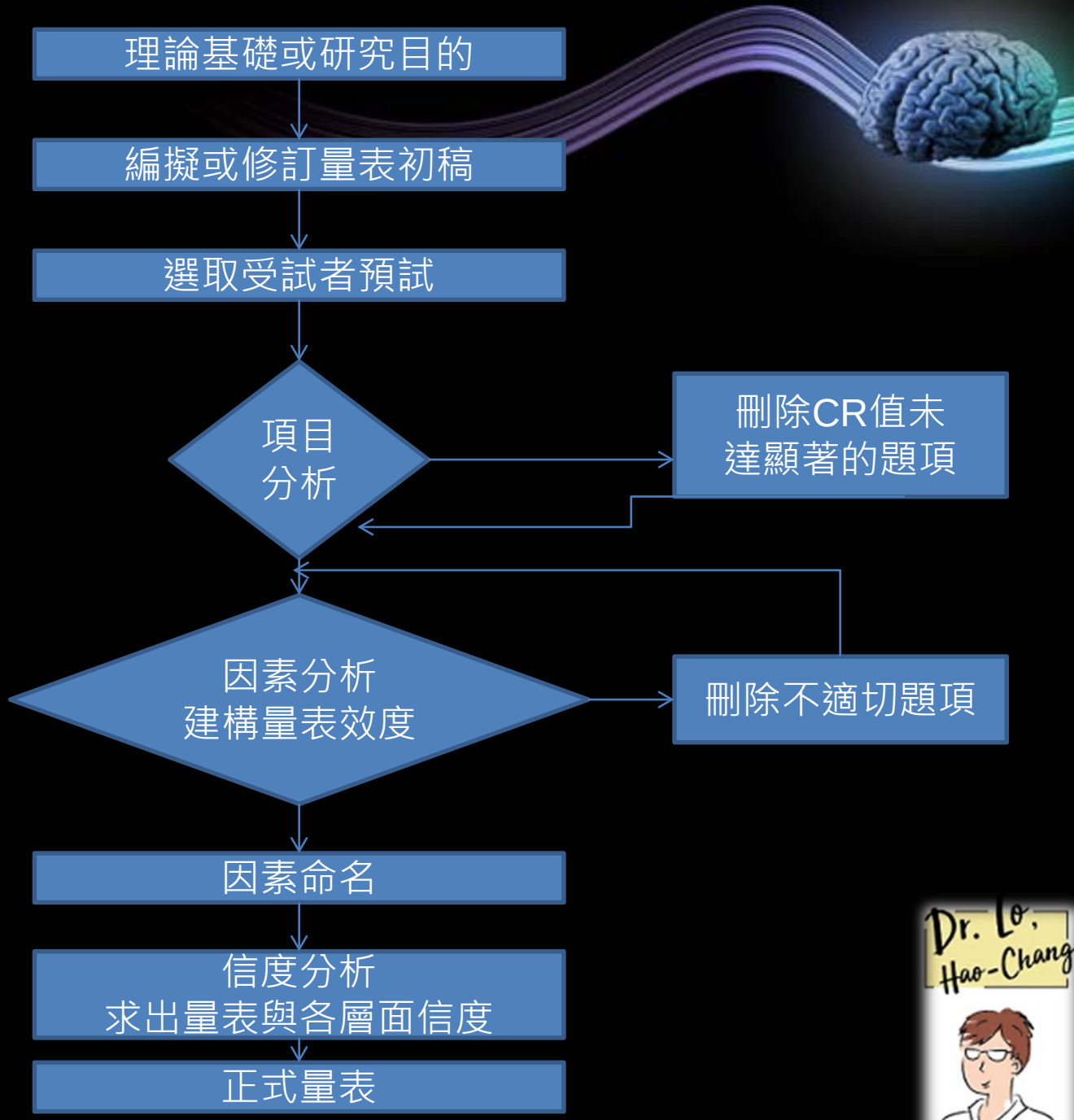
量表+指數範例

魏氏兒童智力量表第五版(WISC-V)：

透過該量表，受測者可獲得一項全量表智商(FSIQ)和語文理解指數(VCI) 、視覺空間指數(VSI) 、流體推理指數(FRI) 、工作記憶指數(WMI) 和處理速度指數(PSI)五項主要指數分數，以及五項選擇性指數分數（數量推理、聽覺工作記憶、非語文、一般能力、認知效能）。共包含十六個分測驗。

http://www.mytest.com.tw/All_page.aspx?title=I_WISCV

量表編製建構流程圖


理論基礎或研究目的
編擬或修訂量表初稿
選取受試者預試
項目
分析
刪除CR值未
達顯著的題項
因素分析
建構量表效度
刪除不適切題項
因素命名
信度分析
求出量表與各層面信度
正式量表

Dr. Lo,
Hao-Chang

量化測量工具的信效度

信度與效度的基本概念

- 信度(Reliability)是觀察分數與真實分數的相關程度，當觀察分數與真實分數的相關性愈高，表示研究者使用的問卷信度越好
- 效度(Validity)是研究者採用的測量工具是否能真正測量到研究者想要測量的變項，亦即研究者得到的觀察分數能夠測量到所想測變項的特質的程度。

測量工具的信度

古典測驗理論中，基本上有以下幾種不同的信度係數種類:

- 穩定係數（跨時間的一致性）：受施測時間間隔影響，當測量時間間距愈短，通常信度愈高
- 複本信度：分析兩份相似的測驗之間的一致度
- 內部一致性係數（跨問項的一致性）：一般能力測驗常使用的信度分析方法
- 評分者信度

內部一致性係數

內部一致性係數的大小主要反映一份試題兩種性質：

- 1. 內容取樣（content sampling）的誤差：因為題目選擇的隨機因素所造成的分數變異
- 2. 題目之間的同質性程度：依據受試者對整份測驗所有題目的反應，分析題目間的一致性，以確定測驗中的題目是否測量相同的特質。

內部一致性係數較常見的有：

- 1.折半信度(split-half method)
- 2. Cronbach's α 係數
- 3. KR20 公式（Kuder-Richardson 20）

折半信度

- 研究者可依照隨機的方式將一分試卷所含的試題分成兩半(例如將奇數和偶數題折半或是前後折半)，再求取這兩半的分數之間的皮爾森積差相關(Pearson product moment correlation)。由於此法所求得之信度其實只是半個測驗的信度，因此信度係數乃有低估的現象，這種低估的現象常以斯布公式(Spearman-Brown formula)加以校正。折半信度愈高即代表兩個半測驗的內容愈一致。當試題數愈多時，其折半信度值也會相對的愈高。

Cronbach’s α 信度

- 在社會科學的研究領域編製測驗或量表時，Cronbach’s α 常作為測量信度之依據。
- α係數法是由Cronbach(1951)年創用，他以α係數來代表量表的內部一致性信度，也是試題間相互關連程度的函數，α係數愈高，代表量表的內部一致性（關連性）愈佳。一般而言，如果內在信度α係數在.80以上，就表示該量表具有很好的信度(Bryman & Cramer，1997)。

- Cronbach α係數 < 0.3（不可信）
- 0.3≦Cronbach α係數 < 0.4（勉強可信 ）
- 0.4≦Cronbach α係數 < 0.5（可信）
- 0.5≦Cronbach α係數 < 0.7（很可信/最常見）
- 0.7≦Cronbach α係數 < 0.9（很可信/次常見）
- 0.9≦Cronbach α係數（十分可信 ）

庫李(Kuder-Richardson)信度

- 庫德(G.F. Kuder)和李查遜(M.W. Richardson)所提出之「庫李信度」是根據受試者對所有試題的反應來分析題目間的內部均質性，如果題目內容的同質性愈高，則其內部均質性也就愈高。只要測驗的題目是同質性的，採**非對即錯**的計分方式，亦即不給部分分數，且該測驗非「速度測驗」(speed test)，就可利用庫李20號（簡稱KR20）或庫李21號（簡稱KR21）的公式來估計測驗的信度。。

測量工具的效度

- 效度是指測量分數的正確性，也就是測量工具能夠測量到所欲測量特質的程度。
- 古典測驗理論中，效度一般有內容效度、效標關聯效度，以及建構效度三種。
- 其中，建構效度是 一個範圍比較廣的概念，涵蓋內容效度和效標關聯效度，是指測驗能夠測量到理論上的構念或特質的程度（Anastasi, 1982）。

內容效度（Content Validity）

- 內容效度係指測驗能測出所欲測量行為領域的程度。涉及測驗之題目內容是否周延、具代表性、適切性、並確實包含所欲測量主題的內涵。
- 通常透過文獻探討形成測驗初稿，再商請專家進行檢核，研究者再根據專家的意見進行測驗的修訂。

效標效度（Criterion validity）

- 效標指的是衡量測驗有效性的外在標準
- 效標效度是利用某些標準或準則來正確地指明某個構念，通常可以藉著一個新的測量工具和一個已經建立且有效的測量方式進行比較，當兩者測量結果相近時，則稱其具效標效度。

根據搜集效標的時間，可以將效標效度分為預測效度和同時效度：

- 同時效度（**Concurrent Validity**）：指的是所使用的指標是當前的資料，研究者的發展一個新的智力測驗，用魏氏智力測量再作一次測驗，如果研究者自己設計的測驗結果和魏氏智力測量結果一致，就可以說研究者有較高的效標效度。
- 預測效度（Criterion validity）： 指一個指標預測未來的事件，如美國高中生的SAT智力與性向測驗，是測學生將來在大學是否能念得好。

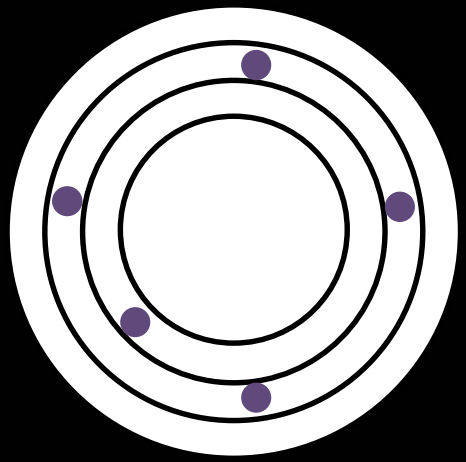
建構效度（Construct Validity）

- 建構效度係指測驗能測量理論的概念或特質之程度而言。此種效度旨在以心理學的理論概念來說明並分析研究所測量數值的意義，並探討研究測量工具是否能夠真正測量到所要測量的研究構念。
- 所謂「建構」或「構念」，就是心理學理論所涉及之抽象而屬假設性的概念、特質或變項，如智力、焦慮、性向、成就動機等。

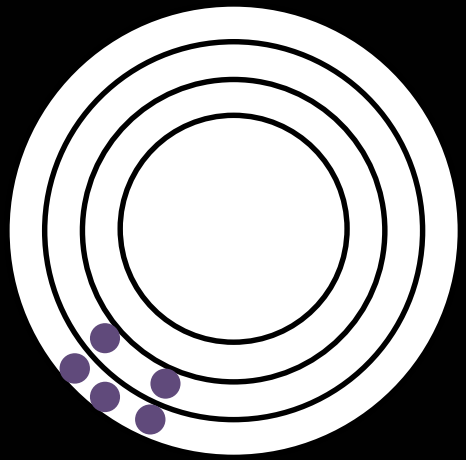
- 建構效度用於具有多重指標（題目）的測量方法上，探討利用多重指標測量某一構念時，這些多重指標測量結果的相似或相異情形。
- 當某些多重指標同時都是在測量同一個構念時，他們測量的結果應該是相似的，稱為輻合效度（Convergent Validity）；當某些多重指標和所測量構念不同時，他們測量的結果應該是相異的，稱為區別效度（Discriminant Validity）。

信度與效度之間的關係

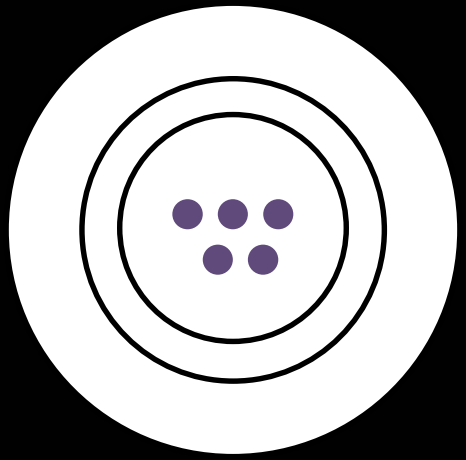
靶心=完美的量數



低信度與低效度



高信度與低效度



高信度與高效度

資料來源:引自Babbie (1995:128)

測驗試題的分析

出題者想知道自己所出的測驗試題是否適切，可以透過以下幾種試題分析得知：

- 難度
- 鑑別率
- 雙向細目表

試題難度分析

難度分析的主要目的在於確定每一個題目的難易程度，題目的難易程度可以兩種方式計算：

- 1. 以全體受試者答對或通過該題的百分比(percentage passing)表示。這個百分比即稱為難度指數P = R/N，P代表題目的難度指數，R為答對該試題的人數，N為全體受試者的人數。
- 例如：如在50名受試的學生中，答對某一題目者有25人，則其難度為：20/50=.40（或40%）

- 2.先將受試者依照測驗總分的高低次序排列，然後把得分最高與得分最低的，各取全體總人數的27%，定為高分組和低分組，再分別求出此兩組在某一題目上通過人數的百分比，以兩組百分比的平均數作為該試題的難度。其計算公式：

$$P = (PH+PL) / 2$$

上式中P代表題目的難度指數，PH為高分組通過該題人數的百分比，PL為低分組通過該題人數的百分比。

- 例如：如某題高分組有74%答對，低分組有22%答對，則該題的難度指數為：(.74+.22)/2 = .48（或48%）。

- 不論使用哪一種計算方式，P值越大，難度越低；P值越小，難度越高。
- 一份試題中，若目的是測驗學生是否已學會課程所教的基本能力，全部試題的平均難度可安排於0.8 左右；若是要區分學生在某學科方面的個別差異， 全部試題的平均難度可安排於0.5左右。

測驗試題鑑別度分析

- 鑑別度分析目的在於試題能否反應不同能力學生答題的差異，鑑別度越高，表示該試題能區別受試者能力高低程度的越高。
- 鑑別度指標(D)

  = PH(高分組答對百分比)– PL (低分組答對百分比)

- 通常鑑別度指標大於0.4(40%)為優良的試題，小於0.2者為劣的試題，介於其中者則建議修改
- 例如：某試題高分組答對百分比為45%，低分組答對百分比15%，則D= 45%- 15%= 30%

測驗試題的雙向細目表

- 試題的雙向細目表在於檢視一份測驗試題的內容效度，雙向細目表可以幫助命題者釐清教學目標和學習內容的關係，以確保測驗能反映教材的內容，並能夠真正評量到預期之學習結果。
- 雙向細目表示測驗的架構藍圖，是以教學目標(橫軸)和學習內容(縱軸)為兩個軸，分別說明各項評量目標。
- 教學目標：常以 Bloom 所提的認知領域知識、理解、應用、分析、綜合、評鑑六個教學目標為依據 (或是以Anderson 與 Krathwohl等人修訂Bloom所得之記憶、了解、應用、分析、評鑑、創造)

認知層級	範例說明
知識	能默寫出牛頓第二運動定律F= m * a
理解	了解m為物體的質量，外力(F)與加速度成正比(a)
應用	以F= m * a 解決「施力10牛頓的力水平外力，推動5kg的物體，可產生的加速度為何?」
分析	面對「施F牛頓的力於甲，產生8m/s^{2}的加速度，若施此力於乙物體，產生16m/s^{2}的加速度，求甲、乙兩物體的質量比」的問題時，可以從F= m * a了解到相同的力作用於兩物，物體的質量與加速度成反比，進而完成解題
綜合	結合運動相關其他概念，解決「2 公斤的物體在光滑平面上，受8牛頓的作用力後，由靜止開始運動，則物體經5秒後的速度為_____米 / 秒」
評鑑	經由相關資料，認定其實牛頓本人所提出第二運動定律並非F= m * a而是F=dP/dt，會變成F= m * a是因為…..

教材內容 \ 教學目標 / 試題型式		記憶	理解	應用	分析	綜合	評鑑	合計
第1章 第1節 （名稱）	選擇題	9（3）		3（1）	3（1）			15（5）
	填充題	8（2）				8（2）		16（4）
	計算題		6（1）					6（1）
第1章 第2節 （名稱）	選擇題	9（3）		3（1）		3（1）		15（5）
	填充題		8（2）		4（1）			12（3）
	計算題			5（1）				5（1）
第2章 第1節 （名稱）	選擇題	6（2）	6（2）		3（1）			15（5）
	填充題			4（1）				4（1）
	計算題		6（1）	6（1）				12（2）
合計	選擇題	24（8）	6（2）	6（2）	6（2）	3（1）		45（15）
	填充題	8（2）	8（2）	4（1）	4（1）	8（2）		32（8）
	計算題		12（2）	11（2）				23（4）
	小計	32（10）	26（6）	21（5）	10（3）	11（3）		100

~ 感謝您的聆聽 ~