

國立臺中師範學院進修暨推廣部九十二年度
教育測驗統計研究所國民小學教師在職進修教學碩士學位班(暑期班)招生考試題本

測驗與評量

說明：一、本卷分為四大題；共四十小題，每小題二.五分，總計一百分。

二、請依題號順序將答案書寫於答案紙內頁。

閱讀選擇題一

下面是一篇關於應用測驗與評量方法進行研究的報導，此敘述內容不盡完整，請您閱讀完後，根據文章中的訊息做適當推論，並回答下列問題。

(註：此內容中有關人名等部分為化名)

一、研究目的

本研究旨在進行數學學習日誌的研究，研究者想要探究接受數學學習日誌教學的學生，其數學學習成就以及數學態度是否較佳，以及學生對於數學學習日誌的反應情形。

本研究之樣本為國小二年級的學生，在實施過程中，亦考量了智力因素可能會影響數學學習成就與數學態度。在整個研究設計中，研究者自編數學學習成就紙筆評量、數學學習日誌以及數學態度量表。具體言之，本研究目的如下列數項：

- (一) 探究數學學習日誌教學對於數學學習成就的效果。
- (二) 分析數學學習日誌教學對於數學學習態度的影響。
- (三) 分析學生在數學學習日誌中的內容，和學生對於數學學習日誌的反應。

二、研究架構

本研究設計架構如圖 1 所示：

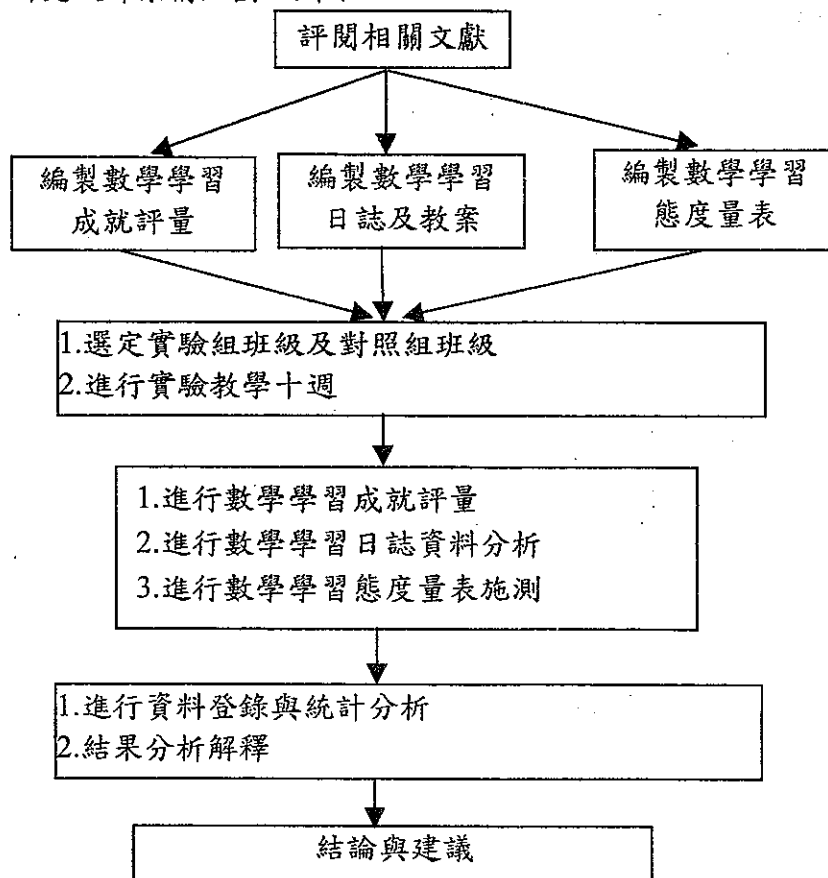


圖 1 研究架構

三、研究步驟

根據研究架構，本研究步驟如下：

- (一) 評閱有關數學學習日誌對於學生數學學習成就，以及數學態度影響的相關理論與文獻。
- (二) 自編 A、B 兩份數學學習成就紙筆評量，研究者設計雙向細目表(two-way specification table)，此兩份紙筆評量在試題內容與試題品質上務求相近。研究者以大慧調查法(Delphi technique) 商請專家根據雙向細目表審核題目適合度，此兩份自編測驗包含填充題、配合題以及應用題，其中填充題、配合題為二元計分，而應用題為多元計分。經預試後修改題目，並計算 A、B 兩份自編數學學習成就紙筆評量的效標關聯效度與信度。
- (三) 自編數學學習日誌檔案評量資料二十份，以及設計相關的教案，以大慧調查法商請專家審核檔案評量資料以及教案的意見。
- (四) 根據王小明(民 88)提出的數學態度三因論，研究者修改自李小華(民 91)編製的「國中生數學學習態度量表」。原李小華(民 91)的「國中生數學學習態度量表」填答設計形式，舉例如下：

我認為數學是 很有用的科目	非常 不同意	←—————→								非常 同意
		請圈選	1	2	3	4	5	6	7	

由於原量表之對象為國中生，而本研究對象為國小二年級學生。所以在措辭及回答方式上需斟酌修改，並歷經預試及以其他非實驗組和控制組的同年級學生為正式施測樣本，以確定自編量表的信度和效度。本研究自編量表的填答設計形式，舉例如下：

我認為數學是 很有用的科目	非常 不同意	不 同意	同 意	非常 同意

- (五) 選定實驗組班級以及對照組班級，在對照組部份，徵得對照組導師的同意，在實驗組部份，研究者即為實驗組班級的導師並進行數學學習日誌教學。
- (六) 進行對照組以及實驗組的標準化智力測驗，此標準化智力測驗為正式出版的測驗工具，研究者徵得出版社授權同意使用。
- (七) 實驗組與對照組在實驗開始前，兩組皆進行 A 卷數學成就紙筆評量，以及數學學習態度量表。實驗組進行為期十週的實驗教學，實驗組教學中，每位學生每週填答數

學學習日誌二份，共計二十份。

- (八) 蒐集整理學生數學學習日誌檔案資料，就此資料商請同學年三名教師協助審視學生在這十週中數學學習日誌中的資料，據此將實驗組學生分為「喜歡數學學習日誌」、「不喜歡數學學習日誌」、「沒有表達對數學學習日誌的喜好」三類。
- (九) 實驗教學進行十週後，兩組皆進行 B 卷數學學習成就紙筆評量，以及數學學習態度量表。
- (十) 進行資料分析

四、測驗工具品質

- (一) A、B 兩份數學成就紙筆評量的效度分別為 .76 和 .78，皆達統計上顯著水準，信度分別為 .89 和 .92。
- (二) 自編數學學習日誌檔案評量資料二十份，經商請專家審核，經過三次的問卷回收後，各份數學學習日誌評量資料的同意程度，其平均分數皆在 4.1 以上（滿分為 5 分）。
- (三) 在數學學習態度量表方面，以線性結構方程模式（linear structural equation model）考驗研究資料的模式適合度，分析結果顯示資料適配（fit）良好。

五、重要結果與討論

- (一) 實驗組與控制組學生在教學前的 A 卷數學學習成就紙筆評量並無顯著差異，經過實驗教學後，實驗組與控制組學生在 B 卷數學學習成就紙筆評量已達顯著差異，而且實驗組高於控制組。
- (二) 若扣除智力的影響因素後，實驗組與控制組在教學前的 A 卷數學學習成就紙筆評量，以及實驗教學後的 B 卷數學學習成就紙筆評量都未達顯著差異。
- (三) 實驗組與控制組學生在教學前的數學學習態度量表並無顯著差異，經過實驗教學後，實驗組與控制組學生在數學學習態度量表已達顯著差異，而且實驗組高於控制組。
- (四) 就數學學習日誌的整體內容，三名教師將實驗組學生歸納為「喜歡數學學習日誌」、「不喜歡數學學習日誌」、「沒有表達對數學學習日誌的喜好」三類，其歸類結果經統計檢定，顯示具有一致性。而且實驗組學生在「喜歡數學日誌」、「不喜歡數學日誌」、「沒有表達對數學日誌的喜好」的人數比率上亦達顯著差異。

請根據以上文章，選擇最適當的選項，回答下面問題。

1. 就研究性質而言，文章所述內容，應為下列何種研究？
 - (1) 詮釋性研究
 - (2) 縱貫研究
 - (3) 內容分析研究
 - (4) 行動研究

2. 研究者在自編兩份數學學習成就紙筆評量中，其設計雙向細目表的目的為何？
- (1) 做為統計分析的依據
 - (2) 比較兩份數學學習成就紙筆評量的異同
 - (3) 確保該評量具有內容效度
 - (4) 做為效標關聯效度計算的依據
3. 根據本文的內容，假設學生因為知道本身是控制組，便力求認真努力，與實驗組競爭。這種情形應該稱為何者？
- (1) 強亨利效應 (John Henry Effect)
 - (2) 霍桑效應 (Hawthorne Effect)
 - (3) 月暈效應 (Halo Effect)
 - (4) 雙盲實驗 (Double Blind Experiment)
4. 在研究步驟 (二) 中，研究者應用大慧調查法，下列何者為該方法的特色？
- (1) 專家可面對面溝通討論以獲得共識
 - (2) 反覆詢問專家，可能造成專家的不便或不願作答
 - (3) 每次詢問專家務必求得一致共識
 - (4) 共識獲得快速且必定具有共識
5. 根據 A、B 兩份自編測驗的題型，要求得該測驗的信度係數，以下列何種信度較適宜？
- (1) 重測信度 (test-retest reliability)
 - (2) 庫李 20 信度 (KR20 reliability)
 - (3) α 信度 (Cronbach α coefficient)
 - (4) 折半信度 (split-half reliability)
6. 在研究步驟的第四 (四) 中，李小華 (民 91) 的「國中生數學學習態度量表」稱為何種量尺？
- (1) 李克特試量尺 (Likert scale)
 - (2) 數字評定量尺 (numerical rating scale)
 - (3) 敘述的圖示量尺 (descriptive graphic rating scale)
 - (4) 語意區別量尺 (semantic differential scale)
7. 在測驗工具品質的第三 (三) 中，研究者以線性結構方程模式分析數學學習態度量表資料，其主要目的可能為何？
- (1) 確認該量表具有內容效度 (content validity)
 - (2) 確認該量表具有建構效度 (construct validity)
 - (3) 確認樣本的反應資料符合常態分配
 - (4) 確認題目之間具有高度線性關係
8. 在重要結果與討論的第二 (二) 中，扣除智力的影響因素後，實驗組與控制組在 A、B 兩份數學學習成就紙筆評量都未達顯著差異。可據以推論研究者在這方面所使用的統計方法可能為何？
- (1) 變異數分析 (analysis of variance)
 - (2) 共變數分析 (analysis of covariance)
 - (3) 多元迴歸 (multiple regression)
 - (4) 因素分析 (factor analysis)

9.在重要結果與討論的第(四)中，三名教師的歸類經統計檢定，顯示具有一致性。試問其計算的一致性係數可能為何？

- (1) Spearman rank correlation
- (2) Kendall coefficient of Concordance
- (3) Kappa coefficient
- (4) Pearson product-moment correlation

10.在重要結果與討論的第(四)中，指出實驗組學生在「喜歡數學日誌」、「不喜歡數學日誌」、「沒有表達對數學日誌的喜好」的人數比率上達顯著差異。試問在這方面所使用的統計方法可能為何？

- (1) t-test
- (2) Chi-square test
- (3) ANOVA
- (4) Linear Regression

閱讀選擇題二

以下為虛構的文章，請在詳細閱讀後，依文章內的相關資料以及正確的測驗評量概念，回答文末的選擇題。

某標準化測驗的研究發展中心宣稱其新上市的某一測驗工具符合以下特點：
本測驗為一多因素智力測驗工具(Multi-dimensional intelligent testing battery，簡稱 MITB)，內容為非例行性的問題，範圍與目前國內所使用的課程無關，MITB 包括：語文理解、數學運算、科學知識、空間覺知、社會規範常識等五個分測驗(subscale)，MITB 之目的在同時測量受測者的多方面之心智能力發展狀況，MITB 之施測結果可以提供教育單位或心理評鑑機構在職業輔導或諮商診斷時的重要資訊。本測驗共編制國中及高中兩種版本，國中版本之 MITB 適用對象為國中二、三年級學生，高中版本之 MITB 適用對象為高國中一、二、三年級學生；每種版本各有兩個複本，MITB 甲式與 MITB 乙式，可供選用。

在排除有明顯的生理、心理或心智障礙的樣本之後，MITB 測驗之標準化樣本包括了國中二年級學生 950 人、國中三年級學生 1000 人、高中一年級學生 900 人、高中二年級學生 920 人以及高中三年級學生 960 人。MITB 測驗工具可提供各分測驗之原始分數(以 X 表示)、百分等級(以 PR 表示)、 T 分數(以 T 表示)、MITB 分項量尺分數(MITB subscale scores，簡稱 MITB-SS 分數)以及 MITB 量尺總分(MITB scaled total scores，簡稱 MITB-STS 分數)。各個分數間有以下的關係：

$$Z = \frac{X - \mu}{\sigma}$$

其中 X 表示原始分數、 μ 為某一組群的原始分數之平均分數、 σ 為其標準差。 Z 是標準分數，其平均數為 0、標準差為 1。另外，

$$T = 10Z + 50。$$

而 MITB-STS 分數的平均分數以 μ_{STS} 表示，其標準差為 σ_{STS} 。

MITB 在信度(reliability)方面的資料彙整如下。重測信度為 0.32 ~ 0.92 間，庫李信度(KR-20 reliability)為 0.45 ~ 0.88 間。

MITB 在效度(validity)方面的資料彙整如下表；表中所示為分測驗的分數間的 Pearson 相關係數表。

	語文理解	數學運算	科學知識	空間覺知	社會規範常識
語文理解	1.00				
數學運算	0.54	1.00			
科學知識	0.61	0.73	1.00		
空間覺知	0.30	0.35	0.25	1.00	
社會規範常識	0.21	0.10	0.13	0.09	1.00

由上表得知：語文理解、數學運算、科學知識三者之間的相關皆在 0.5 以上(0.54, 0.61, 0.73)符合原測驗的設計目的，而空間覺知、社會規範常識與其他分測驗的相關皆在 0.5 以下(0.09~0.35)亦符合原測驗的設計目的。本測驗的效度因此獲得確認。

總結而言，MITB 的施測結果，符合標準化測驗的要求，是一份優良的測驗工具，歡迎學術研究或諮商輔導等相關人員，依測驗使用規則申請使用。

請依文內的相關資料以及正確的測驗評量概念，並針對以下問題的內容，每題選出一個相對之下最合理或最佳的選項。

11. 在一般的測驗評量理論中，信度(reliability)主要在顯示測驗的何種性質？
 - (1) 真實程度(relevance)
 - (2) 效率程度(effectiveness)
 - (3) 實用程度(practicality)
 - (4) 一致程度(consistency)
12. 在一般的測驗評量理論中，效度(validity)主要在顯示測驗的何種性質？
 - (1) 真實程度(relevance)
 - (2) 效率程度(effectiveness)
 - (3) 實用程度(practicality)
 - (4) 一致程度(consistency)
13. 若依「MITB 的重測信度是介於 0.32~0.92 間」之施測結果為主要判斷依據，以下何者較有可能是正確的敘述？
 - (1) MITB 的分測驗(subscale)有良好的信度
 - (2) MITB 的某些分測驗可能穩定度不夠
 - (3) MITB 的分測驗有良好的效度
 - (4) MITB 的施測約有 8% 的誤差
14. 若依 MITB 對效度證據的說明為判斷依據，MITB 的效度說明是以下列何種效度理論為依據？
 - (1) 專家效度
 - (2) 內容效度
 - (3) 實徵效度
 - (4) 表面效度
15. 若依 MITB 的效度證據的說明中：「語文理解、數學運算、科學知識三者之間的相關皆在 0.5 以上」，這是何種效度的表現？
 - (1) 幅合(convergent)效度
 - (2) 辨別(discriminant)效度
 - (3) 顯著(significant)效度
 - (4) 一致(consistency)效度

34. 假設某一題的參數 $b_i = 0.7, c_i = 0.2$ ，試問圖 1 中何者與此題的試題特徵曲線最為接近？

- (1) item 1
- (2) item 2
- (3) item 3
- (4) item 4

35. 圖 1 中何者最適合鑑別 $\theta = 1$ 附近之受試者？

- (1) item 1
- (2) item 3
- (3) item 4
- (4) item 5

36. 圖 1 中 item 2 最適合以下列哪一種模式來描述？

- (1) Ramsay 模式
- (2) 二參數 logistic 模式
- (3) 三參數 logistic 模式
- (4) 四參數 logistic 模式

37. 圖 1 中 item 5 最適合以下列哪一種模式來描述？

- (1) Ramsay 模式
- (2) 二參數 logistic 模式
- (3) 三參數 logistic 模式
- (4) 四參數 logistic 模式

38. 在樣本足夠時，圖 1 中 item 3 最不適合以下列哪一種模式來描述？

- (1) Ramsay 模式
- (2) 二參數 logistic 模式
- (3) 三參數 logistic 模式
- (4) 四參數 logistic 模式

39. 下列何者最不可能為造成圖 1 中 item 5 曲線的原因？

- (1) 樣本不足
- (2) 測驗不具單維性
- (3) 測驗時間不足
- (4) 局部獨立不成立

40. 有關 Ramsay 模式中 h 參數的敘述以下列何者最正確？

- (1) h 大時則曲線較平滑
- (2) h 大時則曲線較不平滑
- (3) h 大時則曲線較陡峭
- (4) h 大時則曲線較不陡峭

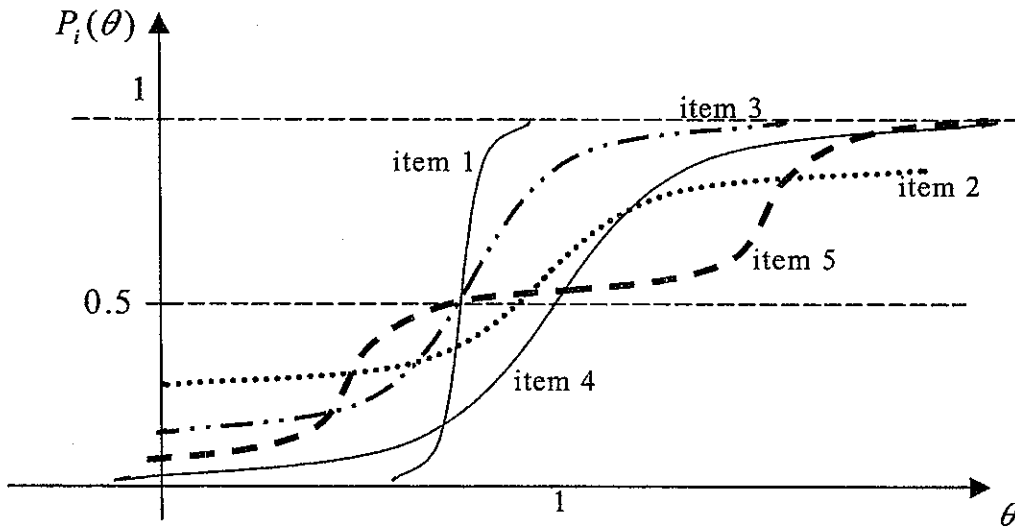


圖 1、 試題特徵曲線

31. 相較於 CTT，下列何者非 IRT 之優點？

- (1) 參數估計所需樣本數較少
- (2) 對試題的特性描述較為精確
- (3) 能力估計較為精準
- (4) 較適合用於電腦適性測驗

32. 下列何者非使用上述 logistic IRT 模式應控制之條件？

- (1) 能力值介於-3 與+3 間
- (2) 局部獨立
- (3) 非速度測驗
- (4) 單維性

33. 三參數 logistic 模式假設某一題的參數 $a_i=1, b_i=0, c_i=0.2$ ，則 $\theta=1$ 時，此題之期望答對率 $P_i(\theta)$ 為下列何者？

(1) $0.2 + \frac{0.8e^{1.7}}{1+e^{-1.7}}$

(2) $0.2 + \frac{0.8e^{1.7}}{1+e^{1.7}}$

(3) $0.2 + \frac{0.8}{1+e^{1.7}}$

(4) $0.2 + \frac{0.8e^{-1.7}}{1+e^{1.7}}$

閱讀選擇題四

在古典測驗理論(conventional test theory, 簡稱 CTT)的試題分析方法中以該試題答對人數除以總受試人數, 表示試題的難度係數, 表示此一群受試者在該題的通過此一試題的比率, 而第 i 題的難度係數可寫成下面的數式:

$$P_i = \frac{\text{第 } i \text{ 題答對人數}}{\text{全體受試人數}(N)}$$

試題反應理論(item response theory, 簡稱 IRT)改進古典測驗理論, 使得試題分析技術進步至可獲得個別受試者 (能力值 θ) 在第 i 題的期望答對率 $P_i(\theta)$, 下列為常見的試題特徵曲線模式。

單參數 logistic 模式:

$$P_i(\theta) = \frac{1}{1 + \exp(-1.7(\theta - b_i))}$$

二參數 logistic 模式:

$$P_i(\theta) = \frac{1}{1 + \exp(-1.7a_i(\theta - b_i))}$$

三參數 logistic 模式:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\theta - b_i))}$$

四參數 logistic 模式:

$$P_i(\theta) = c_i + \frac{1 - c_i - d_i}{1 + \exp(-1.7a_i(\theta - b_i))}$$

Ramsay 模式:

$$P_i(\theta) = \frac{\sum_{s=1}^N \exp\left[-\frac{(q_s - \theta)^2}{2h^2}\right] y_{is}}{\sum_{s=1}^N \exp\left[-\frac{(q_s - \theta)^2}{2h^2}\right]}$$

其中

a_i, b_i, c_i, d_i 分別表第 i 題的鑑別度參數、難度參數、猜測度參數、天井參數

y_{is} 為受試者 s 在第 i 題的作答情形 (0 表錯誤, 1 表正確)

q_s 為受試者 s 在第 i 題的能力估計值, h 為帶寬參數

圖 1 為試題特徵曲線之舉例。

27.下列哪一個選項是科學態度的評量方法之一？

- (1) 情境測驗
- (2) 實際操作
- (3) 行為觀察
- (4) 口頭報告

28.依本文所述，下列哪一個選項並非為「評量」的工作？

- (1) 根據教學設計，評斷學生的學習改變內容
- (2) 依照教學目標，研究學生的學習表現情形
- (3) 運用科學方法，蒐集學生的學習行為訊息
- (4) 使用科學技術，彙整學生的學習成就資料

29.依本文所述，下列哪一個選項是目前中小學教育的評量偏失現象？

- (1) 忽略紙筆測驗的功能。
- (2) 忽略了解學習的過程。
- (3) 忽略學科知識的評量。
- (4) 忽略評量的比較用途。

30.從評量取向觀點來看，下列哪一個選項是美國 The Eight-year Study 時代之前的取向？

- (1) 描述取向
- (2) 協商取向
- (3) 測量取向
- (4) 判斷取向

試根據上面文章，選擇最適當的選項，以回答問題。

21. 依本文所述，下列哪一個選項是教學評量三大內涵之一？

- (1) 行政的辦學效率之評量
- (2) 課程的目標本位之評量
- (3) 教師的學科能力之評量
- (4) 學生的學習成就之評量

22. 下列哪一個選項不是教育部明確公布的「國民中學學生成績考查辦法」之一？

- (1) 機械記錄
- (2) 校外學習
- (3) 口頭問答
- (4) 紙筆測驗

23. 依本文所述，下列哪一個選項是「因材施教自我比較」的觀點之一？

- (1) 觀察學生在學科上的成績之相對意義，可以看出學生在學科上的「努力」程度。
- (2) 比較學生在學科上之歷次成績的前後關係，可以看出學生在學科上的「進退步」。
- (3) 觀察全部學生在學科上的成績側面剖析圖，可以看出學生在學科上的「長短處」。
- (4) 比較全部學生在學科上的成績相對地位，可以看出學生在學科上的「優劣」潛能。

24. 依本文所述，下列哪一個選項是學習評量的正確演進發展階段？

- (1) 第一階段評量 Assessment，第二階段評量 Evaluation，第三階段評量 Measurement
- (2) 第一階段評量 Measurement，第二階段評量 Assessment，第三階段評量 Evaluation
- (3) 第一階段評量 Evaluation，第二階段評量 Measurement，第三階段評量 Assessment
- (4) 第一階段評量 Measurement，第二階段評量 Evaluation，第三階段評量 Assessment

25. 下列哪一個選項不是成績考查的範圍之一？

- (1) associate learning
- (2) assistant learning
- (3) concomitant learning
- (4) primary learning

26. 下列哪一個選項是「MSPAP」的評量實例？

- (1) Portfolio Assessment
- (2) Alternative Assessment
- (3) Performance Assessment
- (4) Authentic Assessment

- 11.自我評量：學生就自己學習情形、成果及行為表現，做自我評量與比較。
- 12.同儕互評：學生之間就行為或作品予以相互評量。
- 13.校外學習：就學生之校外參觀、訪問等學習活動予以考查。
- 14.實踐：就學生之日常行為表現予以考查。
- 15.其他。

例如：自然科學課程強調科學概念、科學方法和科學態度之教學，故考查學生的學習結果時，必須兼顧這三方面的評量，不可有所偏廢。科學概念的評量，可採用口頭考問、紙筆測驗或情境測驗等方式進行之；科學方法注重過程技能（process skills）之訓練，此方面的評量宜以實際操作方式實施之；至於科學態度方面的評量技術，較難達到客觀衡鑑的要求，必須同時並用行為觀察法、評定量表、自陳量表、項目檢核表及個別晤談等方式，設法蒐集多方面的資料，以便比較分析與綜合研判。

陸、因材評量與自我比較

教育最基本的理念與方法，就是「有教無類」和「因材施教」。學齡兒童及青少年來自不同的家庭社經背景，在身心發展及行為表現各方面，都有個別差異的現象。在當前以班級教學為主要的教育體制下，應先充分了解學生個別差異，以便普遍實施適性教育，促使學生個別的才華，都能因多樣化的學習情境，而獲得有尊嚴且快樂的成長。既然配合學生的個別差異，實施適性教育，把每個學生帶上來，理應把握「因材施教」及「因材評量」的原則，進行多元化評量措施，採自我比較的方式，解釋分析各方面整合評量的結果。

學校教師應該提供各種教育活動的機會，採用各種適性評量的方式，讓每個學生都能充分發揮才華，把努力學習的成果從各方面表現出來；再就各科成績作適當的比較分析，以顯示出相對的意義。從教育和評量的觀點來說，最好採用自我比較的方式，也就是以學生本人的學習潛能及實際的學習表現相互參照作合理的解釋分析。通常有三種方式，第一基於學生個人的潛能來解釋他在學科成績上的相對意義，可以看出他「努力」的程度；第二把學生個人過去和現在歷次考試評量所得的成績前後加以比較，可以看出「進步」或「退步」的情形；第三把學生個人在許多不同學科所得到的成績畫成側面剖析圖，從中可以清楚地看出各科分數的高低，可以了解他在各學科領域學習成果的相對優劣，也可以凸顯出整個學習結果在各方面的「長處」和「短處」，作為進一步學習輔導的依據。總而言之，我們應該注重因材評量，採用自我比較的方式。

柒、結語

學生學習成就的評量，必須顧及個別差異，符合因材施教、因材評量的原則。成績評量涉及教育各方面，因而必須採用多種方式。多元化評量是教學評鑑與心理計量學研究發展的最新趨勢，有其理論基礎與實務應用的意義。傳統的紙筆式測驗，因評分標準客觀，分數較少誤差，且施測方便，乃廣被採用，但有時過度僵化，並非最佳的評量方式。事實上，教學評量除了一般的紙筆式測驗之外，尚有許多其他變通的方式，可視學科性質、教學情境及評量重點等因素，酌採其中幾種不同的評量方法，作彈性的搭配運用，以發揮最大的評量功能。

《取材自簡茂發博士：多元化評量之理念與方法》

生方面，能達到「有意義的學習」(meaningful learning)。所謂「有意義的學習」，係指兒童及青少年所學的內容與方式力求與其日常生活情境切近且密切結合，同時有所謂的“Authentic Assessment”(真切性評量)。此種評量特別注重在真實的或模擬的情境中進行考查，強調評量的內容與方式應配合兒童當前的生活經驗，使其覺得自然親切，而不致格格不入，且能充分反映其真實的行為改變或成長，獲得全面的深切了解。

此外，尚有“Portfolio Assessment”(卷例評量)，係指在學生學習過程中彙集相關的各项資料，針對評量的需要，找出一些切要的項目作成適當的組合，進行成套的評量，而非零碎個別項目的考核，具有持續累積的評量效用。由於評量不只限於靜態的作品，還要評量作品產生的過程以及作品的優劣，是一種整體系統性的評量，因而又連帶出現一種“Work Sampling System”(工作取樣系統)，亦即從工作項目中找出重要的樣本，讓學生操作反應，再進行觀察和記錄。

上述各種評量方法都是相關的，主要強調的都是要讓學生在切近生活經驗的活動中實際操作學習，再由教師進行有系統的觀察、記錄和評分。另外，針對學科的學習，現在有所謂的“Curriculum-based Competency testing”(課程本位能力測驗)。每一個科目或課程，都有其教材大綱及教學重點，同時也要考慮教學目標，而教學目標又分為情意、認知、技能三方面。因此，在評量時，課程本位能力測驗應考慮針對該科學習的活動，包括教材內容、行為目標、學習情境、以及學生個人的背景能力等各方面；亦即課程設計、課程實施、學習情境、個人等因素都應納入考慮。

近年來，國立臺灣師範大學科學教育中心與美國馬利蘭州教育廳合作進行學習評量的改進計畫—MSPAP (Maryland School Performance Assessment Program)。MSPAP 就是一種超脫選擇題紙筆式測驗的評量方式，採用先舉行一個活動，讓學生參與、觀察，然後詢問他一些相關的問題，讓學生有所表現，再根據其表現來評分。這是上述「實作評量」典型的實例，在心理計量學與教學評量領域研究發展上，頗受稱許和肯定。

教育部於民國八十七年八月二十六日修正公布的「國民中學學生成績考查辦法」規定：學校對國中學生成績之考查，應視學生身心發展與個別差異，以獎勵及輔導為原則，並依各學科及活動性質，得就下列十五種評量方式選擇辦理。

1. 紙筆測驗：就學生經由教師依教學目標、教材內容所編訂之測驗予以考查。
2. 口試：就學生之口頭問答結果予以考查。
3. 表演：就學生之表演活動予以考查。
4. 實作：就學生之實際操作及解決問題等行為表現予以考查。
5. 作業：就學生各種習作予以考查。
6. 設計製作：就學生之創造過程及實際表現予以考查。
7. 報告：就學生閱讀、觀察、實驗、調查等所得結果之書面或口頭報告予以考查。
8. 資料蒐集整理：就學生對資料之蒐集、整理、分析及應用等活動予以考查。
9. 鑑賞：就學生由資料或活動中之鑑賞領悟情形予以考查。
10. 晤談：就學生與教師晤談過程，了解學生反應情形予以考查。

參、教學評量的內涵與類型

教學評量包括三大部分，即教師的教學效率之評量 (evaluation of teacher's teaching effectiveness)、學生的學習成就之評量 (evaluation of students' learning achievement)、課程的設計與實施之評量 (evaluation of curriculum program)。從評量的時機和性能而言，教學評量可分為「形成性評量」(formative evaluation)和「總結性評量」(summative evaluation)；從評量資料的解釋方式而言，教學評量又可分為「常模參照評量」(norm-referenced evaluation)和「標準參照評量」(criterion-referenced evaluation)。美國教育學者 R.K. Watkins 曾將學校中通常所採用的成績評量措施，歸納為下列九種方法：1.教師的評判 (teacher's judgement)；2.口頭述誦 (oral recitation)；3.論文考試 (essay examination)；4.標準化客觀測驗 (standardized objective test)；5.教師自編客觀測驗 (informal teacher-made objective test)；6.學生作品的評定 (the rating of sampling of products of pupils' work)；7.操作的評定 (performance rating)；8.非正式記述的評量 (informal descriptive evaluation)；9.機械記錄 (mechanical recording)。

肆、多元化評量的理念

由於教學和輔導的範圍非常廣泛，除知識技能之外，尚包括學生的品行、人際關係、學習態度、興趣、方法和習慣等方面，所以必須採用不同的考查方法，以配合學科性質和教材內容進行多種方式的評量。

一、成績評量是多方面的：從前一般人以為學校的成績評量只限於知識技能的考查，事實上現代國民中小學注重德、智、體、群、美五育均衡發展，應該把評量的範圍擴大至教育的全面。當我們從事某一學科的教學時，雖然以該學科本身為重點，但也不能忽略與該學科有關的其他知識，而且透過各學科之學習，學生在其態度及理想上可能有所改變或進步，因而副學習 (associate learning) 與附學習 (concomitant learning) 之結果，也應納入成績考查的範圍，不要僅以主學習 (primary learning) 作為成績評量之唯一對象。最近美國心理學者認為學校教學包括：1.認知方面—以學生心智能力之發展為核心；2.情感方面—以學生之態度、興趣、對人對事之方式以及各種鑑賞能力為重點；3.心理動作方面—以各種技能為主。

二、成績評量必須採用多種方式：因為成績評量之範圍非常廣泛，所以應該採取不同的考查方法，以配合教材之性質與內容，進行多種方式的評量。評量時所採用的方法愈多，蒐集資料愈齊全，其結果愈客觀正確，愈能符合成績考查的要求。

伍、多元化評量的方法

以往傳統的學習評量，大都採用標準化紙筆式測驗 (paper-and-pencil test) 或教師自編的課堂測驗 (teacher-made classroom test)，為時已久，習以為常。由於紙筆式測驗較為通俗，使用普遍，易於被非專業的人所誤用或濫用，而造成許多不良的結果。因此，現今教學評量的研究發展趨勢更注重彈性的、變通的、多元化的評量，並且強調動態的過程，乃出現 "Alternative Assessment" (另類評量、變通性評量)、"Dynamic Assessment" (動態評量) 和 "Performance Assessment" (實作評量) 等新名詞。

教學包含教師的「教」和學生的「學」。過去只要求「教學正常化」，但現在更強調「教學卓越化」。換言之，希望在教師方面，能達到「有效率的教學」(effective teaching)；在學

閱讀選擇題三

教學是師生共同參予而產生交互影響的動態過程；而評量則是運用科學方法和技術，蒐集有關學生學習行為及其成就的正確資料，再根據教學目標，就學生學習表現的情形，予以分析、研究和評斷的一系列工作。在整個教學歷程中，評量是承接轉合的關鍵部份，而不是教學歷程的終點站，亦非表示教學活動的結束。教學評量的主要目的，在於分析教學得失及診斷學習困難，作為實施補救教學和個別輔導的依據。

由於各級學校長期使用紙筆式測驗作為學習成就評量的主要工具，簡便易行，固然發揮了相當的功能，但沿襲已久，殊少變通，流弊所及，也成為當前教育改革中熱烈討論的重要課題之一。教育部在民國八十七年九月三十日公布的「國民教育階段九年一貫課程總綱綱要」中曾提及評鑑方法應採多元化方式實施，兼重形成性和總結性評鑑。因此，多元化教學評量理念與方法之評介，在課程及教學的研究發展上，實為當務之急，有其必要。

壹、學習評量概念的演進

檢視過去和最近的相關文獻，學習評量的發展演進可由其所用的名詞與涵義，分為三個階段。美國在「八年研究」(The Eight-year Study) 時代之前，強調的是“Measurement”以量化的方法取得正確可靠的數據；到後來則認為應該從教育的目標、人格的發展各方面來進行評量，亦即除了客觀的數字之外，尚須有一些價值標準來加以衡鑑，而將“Measurement”提昇至“Evaluation”；晚近，學者們又將“Evaluation”提昇至“Assessment”，強調評量時應考量各種相關的整體情境，從各種可行的途徑，蒐集全面性、多元化的資料，再從各個角度和不同觀點加以比較分析與綜合研判，進行整合性的詮釋，獲致充分的了解。

貳、教學評量問題的檢討

目前中小學教育由於升學競爭而導致教學未能正常化的結果，尤其是在教學評量方面產生許多的流弊，這些有所偏失的現象，值得大家關心、省思和切實的檢討。舉其要者，包括下列各項：

- 一、偏重智育或學科知識的評量。
- 二、評量偏重學習結果，而忽略學習過程的了解。
- 三、考試次數太多。在個人方面，課業負擔重，無信心，壓力大，造成考試焦慮症候群；在人際關係方面，因惡性競爭而對立，彼此疏遠，不能相互尊重與合作。
- 四、常用紙筆式測驗，以坊間普遍印行的測驗卷為評量工具，千篇一律，習以為常，覺得厭惡，又無法拒絕，真有強其所難之感。
- 五、無法充分了解考試分數的意義及其所隱藏的訊息，評量的診斷功能尚未發揮。人人競逐高分，考試淪為競賽的工具，學生成為考試的機器，結果失敗挫折者居多，考試的負面效應層出不窮，如作弊、逃學、……。
- 六、考試領導教學，教學未能正常化。
- 七、不適當的比較，滋生許多流弊。
- 八、升學取向，惡性補習，戕害兒童及青少年身心健康。

16. 若某一學生甲之 MITB 的標準分數(Z)為零分，則其相對之 T 分數為何？
- (1) 45
 - (2) 10
 - (3) 50
 - (4) 100
17. 設 MITB-STS 分數的平均分數 μ_{STS} 為 60，其標準差 σ_{STS} 為 12，若某一學生 A 之 MITB-STS 為 50，而學生 B 之 MITB 之 T 分數亦為 50，請問以下何者為真？
- (1) A 的原始分數較高
 - (2) B 的原始分數較高
 - (3) A 與 B 的原始分數相等
 - (4) 無法判定 A 與 B 之關係
18. 依 MITB 的測驗說明，MITB 的研發程序是依據以下何種測驗的定義而進行的？
- (1) 成就參照測驗
 - (2) 常模參照測驗
 - (3) 標準參照測驗
 - (4) 建構參照測驗
19. 依 MITB 的測驗的研發程序說明，MITB 較適用於以下何種情境？
- (1) 診斷人格障礙
 - (2) 篩選行為異常
 - (3) 了解學生潛能
 - (4) 評量教學效能
20. 該測驗中心最需要應再提供以下何種訊息，使該測驗的使用者不管是 MITB 甲式與 MITB 乙式，都能安心使用？
- (1) 複本信度、等化公式
 - (2) 建構效度、構念效度
 - (3) 試題難度、專家效度
 - (4) 折半信度、標準九常模