《教育測驗與統計》

試題評析

今年考題仍依循往年的測驗學與統計學各兩題的比例。其中，測驗學跳脫古典測驗理論的主宰情況，而以電腦適性測驗與一般測驗的倫理規範爲命題方向；統計學考的則是敘述統計的統計量數與推論統計的變異數分析及t考驗。

第一題：考生必須將適性測驗與試題反映理論連結，並提及試題訊息函數、電腦題庫、不同試題組合等關鍵詞。本題對於本班考生應無太大問題。

第二題：考的是測驗倫理，雖然在測驗學的第一章已經提及可由四大構面來思考，考生仍可依試題文字的指引，發揮邏輯與文字能力，取得基本分數。

第三題：考的是敘述統計中變異量數裡的相對差，計算容易，應拿滿分。

第四題：考的是平日上課強調的兩種檢定方法之間的比較，不需計算，只要在電腦報表中找尋適當數據，進行解釋說明，本班考生也應該可以分數全拿；唯一可能的挑戰是，報表以英文呈現，若無準備，可能影響臨場表現！

總結：在其他科目表現差不多的情況下，今年要能上榜的分數應須至少75分以上。

一、隨著資訊科技的發展，目前已有很多測驗利用電腦施測，其中適性測驗（adaptive test）是一種較受重視的施測方式，試就適性測驗回答下列問題：

（一）適性測驗係依據何種理論？（10分）

（二）適性測驗的基本原理為何？（10分）

（三）適性測驗何以必須要用電腦施測？（5分）

答：

(一)當測驗的難度能夠適合考生的能力程度，該測驗所測量到的考生能力才會精確。然而，實際上，任何一次施測結果，都難以針對每位考生提供最精確的能力測量。最理想的施測狀況是能夠針對每位考生不同的能力程度，提供適合個別情況的測驗方式，這就是電腦化適性測驗(computerized adaptive testing, CAT)。適性測驗的理論基礎正是試題反應理論(Item Response Theory, IRT)。由於試題反應理論中的試題反應模式，可以獲得不受不同施測試題影響的能力估計值，也就是說不同的考生考不同的試題，只要試題性質相同，不同能力考生的能力估計值同樣可以被精確的估計出來，因而可以互相比較。

(二)應用試題反應理論，必須先滿足該測驗只測量單一主要因素的基本假設，而這個假設在適性測驗裡，通常都能夠獲得滿足。另外，試題訊息函數在適性測驗扮演著很重要的角色。一般而言，能讓考生大約有 50% 或 60%答對機率的試題，也就是難易適中的試題，通常都是屬於能夠提供最大訊息量的試題。這些能對測量精確性發揮最大貢獻的試題，會被優先挑選爲施測試題。Lord 認爲，如果施測的試題都能針對每位考生的能力提供最大的參考訊息，即使縮短測驗的長度，也應該不會降低對每位考生能力的精確測量。因此，理論上，每位考生所接受的施測試題，應該都是不同的試題組合。

(三)要實施適性測驗，因爲每位考生的試題組合與題數不同，必須準備大量題庫，施測時，要根據個別考生在先前試題作答的表現好壞，由電腦立即試算，由題庫中挑出對考生能力估計精確性最有貢獻的最大訊息量的試題，以接著呈現給考生，因此，適性測驗必須運用電腦才有可能達成縮短測驗長度，卻不犧牲測量精確性的目的；此外，電腦科技的發達，特別是超大容量可以貯存測驗訊息、編製、施測、和記錄測驗分數，有利於適性測驗的運用。

【參考書目】傅立葉「教育測驗與統計」講義第四回第34頁；考猜第11頁。

二、任何一種工具，在善於使用它的人手中，可以發揮其正面的、積極性的功能，產生有益於人的效果。若是落入不善於使用它的人手中，就常會發生誤用和濫用的情形，形成有害的後果；測驗工具亦復如是。因此測驗必須由具備適當條件的人，才能使用。目前政府並無對此項具有專門訓練人員，加以考銓認定，以建立其專業地位；為使不具備是項資格者，無從干擾測驗之推

展，關於測驗使用者的資格，應有測驗倫理的規範，試從下列三方面說明測驗使用者的資格應有那些規範？

(一)使用測驗的必要性。（10分）

(二)選用適當的測驗。（10分）

(三)測驗的發行。（5分）

答：

根據 APA(1981)訂定的心理學專業人員倫理信條(Ethical principles of Psychologists) 中，與心理測驗有關的部份，以及 AERA、APA 與 NCME(1985)訂定的教育及心理測驗準則(Standards for Educational and Psychological Testing)，可將測驗倫理的涵義，歸納爲專業訓練與專業素養兩部份。廣義來說，測驗倫理的專業訓練包括測驗的發展、出版與使用等專業評估技術，它是維持良好測驗品質必需具備的條件。狹義的專業訓練，是指對使用測驗者的資格所規範的權限。專業素養指的是測驗使用者的責任，包括測驗資料的保密性，測驗資料得正確使用，以維護受試者的權益。

(一)就使用測驗的必然性而言，使用測驗的人，有責任去了解，要使用該測驗的理由、必備的程序、以及施測的後果，才能使測驗的效果達到最大，並且使不當之處，即使無法避免，也能讓它減到最低。

1.測驗的目的是否得以解釋現象、解決問題？

根據APA的調查報告，幾乎任何測驗，如用於正確的情境，都能有利於測驗的目的；反過來說如果使用不當，那麼即使再好的測驗，也會對受試者造成傷害(Kaplan And Saccuzzo, 1989)。

2.測驗是否被濫用？

當測驗被用來彰顯某些專業人員的影響力，或是不一定需要施測，甚至被張冠李戴的不當使用；更有甚者，當測驗施測被用來牟取不當利益時，這些情境都隱含著測驗施測的不必然性。

3.測驗的品質前提假設

唯有經過專業訓練，評鑑已出版測驗的品質無虞，才可施測。以免未達施測目的，反而對受試者造成傷害。

(二)如何能夠挑選適合特定目的，同時也適合受試者的測驗？選用測驗者應做到以下幾點：

1.了解與該測驗有關的研究文獻；

2.評鑑該測驗就專業技術層面的優點，例如信度、效度、常模等；

3.根據測驗目的或所欲測量的目標特質，進行選擇；

4.審視該測驗的市場評價與受試者意見；

(三)廣義來說，測驗倫理的專業訓練包括測驗的發展、出版與使用等專業評估技術，它是維持良好測驗品質必需具備的條件。從測驗發行的職業規範來看，一份測驗在未經完成嚴謹的編製程序之前，不應先行出版使用；如果沒有充分的客觀證據，不應宣稱該測驗具備了某些特質；同時隨著測驗所附的使用手冊，應含有如何施測、計分與結果解釋有關的充足且完全的訊息。測驗在發行前，應先經過預試、信效度檢視、修題、再測、題本確定等程序步驟。

【參考書目】傅立葉「教育測驗與統計」講義第三回第7頁。

三、有一教育資料庫蒐集了某地區從小一到大四學生的體重資料，其中小一、國一及大一的體重的平均數及標準差如下表：

年級別	平均數	標準差
小一	20.29	2.37
國一	50.78	5.52
大一	61.54	6.25

(一)依上表資料，試就年級別的體重的個別差異由大至小排序。（10分）

(二)試舉證支持你排序的理由。（15分）

答：

(一)不同年級別的體重差異，應以計算變異係數(或稱相對差)加以判斷：

小一 $CV = \frac{2.37}{20.29} = 0.1168 = 11.68\%$

國一 $CV = \frac{5.52}{50.78} = 0.1087 = 10.87\%$

大一 $CV = \frac{6.25}{61.54} = 0.1016 = 10.16\%$

因此，依大小排序爲：小一、國一、大一，其代表的意義爲，隨著年齡的成長，研究之對象學生的體重變異愈來愈小。

(二)以上排序，應爲小孩成長過程中，因接觸不同飲食與發育時期的交互作用，造成體重變異較大；大學生已經發育完全，體重變化較爲一致。

【參考書目】傅立葉「教育測驗與統計」講義第一回第四章。

四、有一企業請專家為其員工講授兩天16小時的「資訊安全」課程，且規定要課後評量，評量成績將做為年度考績的參考。業主主觀上認為上課坐前排者較認真，成績應比較好；坐後排者則反之。課後評量後，業主刻意把前兩排及後兩排的成績分開處理，然後請其助理群比較前、後排的得分是否有顯著差異。但其助理群對統計分析的方法有不同的看法。A助理認為以單因子變異數分析處理會較有深度，而B助理則認為用t檢定比較簡單，也能達到分析的效果。兩人根據同樣的資料運用電腦軟體分析出來的結果如表A及表B。

Group Statistics

座位	N	Mean	Std. Deviation	Std. Error Mean
前兩排	16	70.25	.931	.233
後兩排	17	69.53	1.663	.403

表A ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.280	1	4.280	2.318	.138
Within Groups	57.235	31	1.846		
Total	61.515	32			

表B Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95%Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper

分數	Equal variances assumed	9.040	.005	1.523	31	.138	.721	.473	-.245	1.686
	Equal variances not assumed			1.548	25.425	.134	.721	.466	-.238	1.679

試根據前述及表A、表B回答下列問題：

(一)檢定此問題的假設如何陳述？（5分）

(二)表A及表B雖是不同統計方法的結果，但兩者有何種關係存在？（10分）

(三)就統計分析的觀點看，你認為那一位助理的方法較妥適，並說明你的理由。(10分)

答：

(一)要檢定前後排員工的評量分數是否存在顯著差異，其假設可以陳述如下：

1.虛無假設：前後排員工的平均分數無顯著差異

2.對立假設：前後排員工的平均分數存在顯著差異

(二)以這兩種統計檢定方法而言，t檢定原本就是針對兩組獨立樣本的平均數差異的顯著性檢定；而變異數分析也具有相同的檢定功能；在資料固定下，兩種方法具有關係如下：

1.t檢定統計量的平方，等於F統計量

$$1.523^2 = 2.319 \cong 2.318$$

2.t臨界值的平方也會等於F臨界值(因爲統計報表未提供，而不用舉證說明)

3.t檢定與F檢定的結論應一致

兩個報表的顯著性(Sig. 2-tailed)，即p-值皆爲0.138，因大於0.05的顯著水準，都可結論爲前後排員工的平均分數存在顯著差異。

(三)從統計分析的觀點，B助理的t檢定方法應該是此一研究問題的較妥適方法。支持此論點的理由如下：

1.變異數同質性的檢定結果(Levene's test)，因p-值爲0.005，表示兩組分數的變異數同質性顯著不等，違反變異數分析之等變異性的假設要件。

2.兩組的員工人數分別爲16人與17人的小樣本，滿足t檢定方法的假設要件。

【參考書目】傅立葉「教育測驗與統計」講義第二回第十章與第十三章。