「批判思考測驗第二級」(CTT-II)簡介

編制者：葉玉珠(2005)
參考資料：葉玉珠(2005)。**網路學習融入師資培育課程對提升職前教師批判思考教學能力之探討**(1/2) (NSC93-2520-S-004-002-)。(執行期限：2004/07-2005/08)

批判思考測驗第二級(Critical Thinking Test, Level II, CTT-II)原名「成人批判思考技巧測驗」(葉玉珠、陳月梅、葉碧玲、謝佳蓁，2001)。由於有部分題目的語意與描述有所爭議，因此進行修訂。爲確保提數最少爲 25 題，因此每一大題增加一題，共 30 題。經以 493 位大學生及研究生爲樣本，進行預試及以五向度 IRT 模式進行資料分析，發現所修訂測驗具有不錯的配適度。

一、信度與效度分析

不同於傳統古典測驗理論下的基本假設，現代測驗理論 IRT 強調量尺分數具有等測量標準誤的優點，不同分數之間的比較更爲有意義及合理；同時，現代測驗理論 IRT 使用「邏輯斯」（logit）爲量尺單位，使得項目難度與潛在能力的單位一致，可以放在一起相互比較，當某試題的邏輯斯值低於某個人的能力邏輯斯值時，我們可以說，這個人有 50%以上的機率可能會答對該試題。總的來說，使用 IRT 的目的就是爲了獲得更爲精確的參數訊息，減少測量誤差所造成的測驗不精準現象。使用 IRT 的原理即是以一個複雜且適當的數理模式（在本研究中使用五向度單參數 IRT）來契合我們實際上所收集到的資料，當檢定契合程度的指標顯示出模式與資料的適配程度是可以接受時，我們即可以用此模式來估計出我們想要的且模式可以提供的參數訊息，例如試題難度、個人的潛在能力。

本研究以五向度 IRT 模式來適配「批判思考測驗第二級」中各題目的難度，發現各個題目大都符合兩種卡方形式的均方統計適配標準（見表 1）；這兩種卡方形式的均方統計適配標準分別是：（1）未加權均方誤適配統計量，即將所有作答者在該題的標準化殘差之平方和除以作答人數；（2）加權均方誤適配統計量，即將所有作答者在該題的標準化殘差之平方以其變異數加權後再加總，最後再除以作答人數。這兩種考驗的 MNSQ 值皆以 1 爲準，大於 1 表示作答組型的變異較理論模型預期的大，相對地，小於 1 則表示作答組型的變異較理論模型預期的小。對選擇題而言，其值在 0.7～1.3 之間則視作該題的作答組型與理論模式之間是適配的，在此範圍之外，則較不理想；另外，這兩種均方統計量經過公式的轉換(Wilson-Hilferty transformation）後，可以用傳統的 t 檢定來考驗，在此，我們以 0.01 雙尾考驗爲臨界標準，t 值在±2.57 內都表示該題的作答組型與 IRT 理論模式是適配的，超過這個範圍則較爲不佳（Bond, Fox, 2001, pp176-179）。

表 1 「批判思考測驗」第二級之各題目難度、標準誤與模式符合度指標

分測驗	題號	難度（logit）	標準誤	未加權均方誤適配考驗	加權均方誤適配考驗

				MNSQ	T	MNSQ	T
第一分測驗（辨認假設）	1	0.075	0.077	1.06	0.9	1.05	0.9
	2	0.166	0.077	0.95	-0.7	0.96	-0.7
	3	-2.275	0.099	1.00	0.1	1.01	0.1
	4	0.342	0.078	1.01	0.1	0.98	-0.3
	5	-0.144	0.077	1.07	1.1	1.07	1.3
	6	1.836*					
第二分測驗（推論）	7	0.120	0.077	1.03	0.6	1.02	0.3
	8	1.790	0.075	1.09	1.4	1.09	1.7
	9	-1.250	0.092	0.93	-1.1	0.91	-1.2
	10	-0.431	0.082	1.07	1.1	1.06	1.0
	11	0.054	0.078	1.04	0.6	1.03	0.5
	12	-0.283*					
第三分測驗（演繹）	13	-0.093	0.073	1.02	0.4	1.02	0.3
	14	-0.111	0.073	1.11	1.7	1.10	1.8
	15	-1.007	0.077	1.03	0.4	1.03	0.5
	16	-0.698	0.075	1.16	2.4	1.15	2.6
	17	0.479	0.073	1.08	1.3	1.05	0.8
	18	1.431*					
第四分測驗（解釋）	19	-0.470	0.067	1.31	4.5	1.30	5.6
	20	-0.375	0.067	0.89	-1.8	0.88	-2.5
	21	0.081	0.068	1.11	1.6	1.10	1.9
	22	-0.074	0.068	0.99	-0.1	0.99	-0.1
	23	0.581	0.071	1.18	2.8	1.18	3.1
	24	0.257*					
第五分測驗（評鑑）	25	0.265	0.071	1.04	0.6	1.04	0.8
	26	1.322	0.080	1.10	1.5	1.08	1.3
	27	-1.722	0.078	1.00	-0.0	1.00	-0.0
	28	-1.195	0.073	1.13	2.0	1.13	2.3
	29	0.407	0.072	0.91	-1.4	0.91	-1.7
	30	0.924*					

註：* 表示該題受限在分測驗中所有難度平均爲零的限制式下，這是爲了模式辨識的目的所常用的作法。

在表 1 中，所有 30 題的 MNSQ 值可以說都在 0.7～1.3 的標準內，不管是根據未加權均方誤適配統計量或是加權均方誤適配統計量來判斷；轉換爲 t 值的考驗後，在 0.01 雙尾考驗

的標準下，則只有第四分測驗的 19、23 兩題有嚴重的不適配情形，表示這兩題所測量到的能力（在此指批判思考中的「解釋」能力）摻雜其它未知的能力，而與其它四題在構念上有較大的不同。從理論上來說，當作答反應組型與模型不適配時，其估計出來的參數值之可信度較低，在此，19、23 兩題的難度估計值應謹慎對待。雖然 19、23 兩題也測量到預期以外的能力，但是綜合而論，這兩題至少在 MNSQ 統計量上符合標準，因此本測驗將之保留。

此外，從表 1 中的各題難度值可知，在第一分測驗中，以第 6 題最難（1.836logit），第 3 題最簡單（-2.275logit）；第二分測驗中，以第 8 題最難（1.790logit），第 9 題最簡單（-1.250logit）；第三分測驗中，以第 18 題最難（1.431logit），第 15 題最簡單（-1.007logit）；第四分測驗中，以 23 題最難（0.581logit），第 19 題最簡單（-0.470logit）；第五分測驗中，以第 26 題最難（1.322logit），第 27 題最簡單（-1.722logit）。

表 2 是批判思考測驗第二級中的五個向度之相關/共變數矩陣，以及各向度的變異數及平均數估計值(主要用來顯示使用較爲精確的 IRT 模式後所估計出來的各向度之間的真實相關或共變情形)。不同於古典測驗理論的是，這些估計值是排除測量誤差後所獲得的較準確估計；此爲 1950 年代提出 IRT 時的共識。表 2 中，除「評鑑」能力與「辨認假設」能力出現負值外，其它相關係數皆呈現正值。各向度的平均數及變異數也是排除測量誤差後的結果，因此較爲可信。

表 2 各向度之相關/共變數矩陣、變異數、平均數

	辨認假設	推論	演繹	解釋	評鑑
辨認假設	----	0.042	0.032	0.004	-0.014
推論	0.133	----	0.351	0.198	0.061
演繹	0.095	0.635	----	0.247	0.027
解釋	0.016	0.532	0.609	----	0.019
評鑑	-0.078	0.212	0.087	0.090	----
變異數	0.191	0.507	0.602	0.273	0.161
平均數 （標準誤）	-0.190 （0.020）	1.140 （0.032）	0.167 （0.035）	-0.404 （0.024）	-0.258 （0.018）

註：對角線以下爲相關係數，對角線以上爲共變數

logit	辨認假設	推論	演繹	解釋	評鑑	各分測驗題目
3						
		X				
		X				
		X				
		X				
		XX				
		XX				
2		XXX				
		XXX	X			6
		XXXX				8
		XXXXX	X			
		XXXX	X			
		XXXXX	XX			18
		XXXXXX	X			26
		XXXXX	XX			
1		XXXXXX	XX			
		XXXX	XXX			30
	X	XXXXX	XXX			
	X	XXXX	XXX	X		
	XX	XXXX	XXXXX	X	X	23
	XXX	XXX	XXXX	XX	XX	17
	XXXX	XXX	XXXX	XXXX	XXX	4 29
	XXXXX	XX	XXXXXX	XXX	XXXX	24 25
	XXXXXXX	XX	XXXXX	XXXXX	XXXXX	1 2 7 21
0	XXXXXXX	XX	XXXXX	XXXXXX	XXXXXXX	11
	XXXXXXXXX	X	XXXXX	XXXXXXXXXXXXXXX		5 13 14 22
	XXXXXXX	X	XXXX	XXXXXXX	XXXXXXXXX	
	XXXXXXXXX	X	XXXX	XXXXXXX	XXXXXXXXX	12 20
	XXXXXXX		XXX	XXXXXXXX	XXXXXXXXX	10 19
	XXXXXX		XXX	XXXXXX	XXXXXXXX	
	XXXXX		XXX	XXXXX	XXXXX	16
	XXX		XX	XXXXX	XXX	
	XX		XX	XXXXX	XX	
-1	XX		X	XXX	X	15
			X	XXX	X	28
			X	XX	X	9
			X	X		
				X		
						27
-2						
						3

註：每一個'X'代表六位受試者

圖　受試者在各向度上的分佈情形與各題目難度值的對應關係

根據 IRT 模式來估計題目難度及受試者能力的優點之一，就是兩者的原點及單位是相同的，都是以 logit 爲量尺，因此可以同時放在圖中一起比較(此爲基本 IRT 的假設)。在由圖 1 可知，每一位受試者在每個向度上各有一個能力分數，因此每一位受試者都有五個能力分數，分別是辨認假設、推論、演繹、解釋、評鑑的能力分數，最右邊則是各題目的難度估計值。其意義是，在某個向度下，當某位受試者的能力在圖中的位置與該向度中的某一題相同時，表示該爲受試者有 50%的機率可以答對該題。要注意的是，哪些題目歸屬於哪一個向度下，在比較時要注意不要錯置。亦即在比較時，應分別以每一個向度中受試者的能力來與圖中最右側的所屬題目之難度比較，例如，對第一向度的「辨認假設」而言，其所屬的 1～6 題中，以第 6 題最難，幾乎沒有人可以答對（但不一定是指沒有任何人答對）；而以第 3 題最簡單，幾乎人人皆可答對（但不一定是指所有人都答對）。另外，在解釋此圖時要特別注意一點，千萬不可作如下的結論：受試者在「批判思考測驗」第二級中的表現，其「推論」能力（平均數爲 1.140 logit）優於「解釋」能力（平均數爲-0.404logit）。之所以不能作如此結論的理由是，這五個向度並沒有共同的單位及原點，因此不能做出上述結論，我們只能根據每一個向度的受試者能力與該向度所屬的題目之難度作比較，才是合理的。

綜上所述，在五向度單參數 IRT 模式下的參數估計不包括鑑別度或猜測參數，或許有人會質疑這樣豈不是試題分析不完整嗎？爲什麼不使用五向度多參數的模式來適配呢？回答這個問題涉及理論上的爭論：首先，多參數 IRT 並未發展出適合兩個向度以上的模式，因此現階段無法應用包含鑑別度及猜測情形的多向度 IRT；再者，多參數 IRT 模式在學術界尚有爭議，雖然可以使用，但部分學者認爲從數理推導過程中，發現多參數 IRT 不再具有量尺單位一致的優點，所以使用上必需謹慎。最後，我們曾試著使用單向度三參數的 IRT 模式來適配資料，雖然該模式能夠提供鑑別度及猜測參數的估計，但是適配指標告訴我們，這個模式與資料的契合程度不佳，也就是我們蒐集到的受試者作答資料無法被這個模式解釋到其應有的訊息，故我們捨棄這個模式，改用目前的五向度單參數 IRT，結果是令人信服的。

必須說明的是，在五向度單參數 IRT 下，所有試題的鑑別度都是假定一樣的。或許有人仍然對此感到不解，怎麼可能鑑別度都一樣呢？要解決這個疑惑必須先問，是甚麼理論讓你認爲各試題鑑別度應該不一樣？鑑別度一不一樣是建立在不同的模型理論假設下來說才有意義，在傳統的古典真分數理論下，我們建立簡單的 X＝T＋E 模式來說明觀察分數（X）與潛在能力或真分數（T）之間的關係，且認定各試題的鑑別度可能是不一樣的，然後我們定義出一套簡單的公式來計算它；但是，在現代測驗理論下，我們使用更複雜的數理模型來解釋資料，雖然每個模式都有其限制或不足之處，但這正是每個模式對試題的基本假定不同所致，我們關心的是這個模式是否能夠完美地解釋資料，若是可以，則我們必須接受這個模式對試題的基本假定是無誤的。總而言之，在五向度的單參數 IRT 模式下，我們獲得不錯的適配結果，因此，我們應該接受該模式的基本假設，認爲各試題的鑑別度是可以被視爲一樣的。

此外，以 CTT-II 總分進行 T 考驗檢驗高低分組（M 上下 27%）在每一分測驗及總分上的差異（ps < .001），也發現 CTT-II 具有良好的鑑別度。

Independent Samples Test

	Group	N	Mean	SD	t	df	Sig.
ASSU_Z	1.00	128	2.30	.99	-6.916	244	.000
	2.00	118	3.25	1.18			
INFER_Z	1.00	128	3.30	1.10	-15.260	229.014	.000
	2.00	118	5.14	.78			
DEDU_Z	1.00	128	2.05	1.23	-15.964	244	.000
	2.00	118	4.40	1.06			
EXPLA_Z	1.00	128	1.64	.98	-14.425	234.937	.000
	2.00	118	3.56	1.10			
EVALU_Z	1.00	128	2.08	.98	-7.305	244	.000
	2.00	118	2.99	.98			
CTABI_Z	1.00	128	11.38	1.67	-39.615	244	.000
	2.00	118	19.35	1.49			

CTT-II 總分與分測驗分數之間有低度到中度相關，其相關係數爲.352 ~ .665，*p*s < .001（見表）。相隔三個月的重測信度爲.458，*p* < .01（N = 100）

表 3 ：CTT-II 各分測驗與總分之間的相關（N = 493）

量表	辨認假設	歸納	演繹	解釋	評鑑
總量表	.352***	.621***	.665***	.597***	.360***

*** *p* < .001。

表 3 ：CTT-II 相隔三個月的重測信度（N = 100）

量表	辨認假設	歸納	演繹	解釋	評鑑	總量表
辨認假設 1	.327***					
歸納 1		.304**				
演繹 1			.329***			
解釋 1				.355***		
評鑑 1					.008	
總量表 1						.458**

二.人口變項在 CTT-II 得分的差異比較

由ANOVA分析得知：性別及就讀層級對CTT-II總分皆無顯著效果，*F*s(1, 501)依次爲 0.000 及 1.764，*p*s = .990 與.185。由MANOVA分析得知：性別對CTT-II五項指標的整體效果達顯著(Λ = .977，p = .043，η^2 = .023)，進一步分析發現性別僅對評鑑一指標有顯著效果(*F*(1, 491) = 4.065，p = .044，η^2 = .008)；然而，就讀層級對CTT-II五項指標的整體效果並未達顯著(Λ = .990，p = .454，η^2 = .010)。

三、批判思考測驗第二級常模的建立

男性與女性大學生、研究生及全體參與者的詳細得分情形見表 。

表 5：不同層級及全體參與者在 CTT-II 的得分

層級	性別	量表	人數	最小值	最大值	平均數	標準差
大學							
	男生	辨認假設	49	.00	6.00	2.59	1.32
		歸納	49	2.00	6.00	4.16	1.14
		演繹	49	.00	6.00	3.31	1.37
		解釋	49	.00	6.00	2.67	1.53
		評鑑	49	.00	5.00	2.43	1.02
		總量表	49	10.00	25.00	15.16	3.21
	女生	辨認假設	293	.00	6.00	2.70	1.04
		歸納	293	1.00	6.00	4.32	1.11
		演繹	293	.00	6.00	3.15	1.39
		解釋	293	.00	6.00	2.42	1.21
		評鑑	293	.00	6.00	2.66	1.05
		總量表	293	4.00	24.00	15.25	3.17
	全體	辨認假設	342	.00	6.00	2.68	1.09
		歸納	342	1.00	6.00	4.30	1.12
		演繹	342	.00	6.00	3.17	1.39
		解釋	342	.00	6.00	2.46	1.26
		評鑑	342	.00	6.00	2.63	1.05
		總量表	342	4.00	25.00	15.24	3.17
研究所							
	男生	辨認假設	47	1.00	6.00	2.74	1.26
		歸納	47	1.00	6.00	4.19	1.14
		演繹	47	.00	6.00	3.57	1.46
		解釋	47	.00	5.00	2.55	1.25
		評鑑	47	1.00	5.00	2.51	1.00
		總量表	47	9.00	23.00	15.57	3.04
	女生	辨認假設	104	1.00	5.00	2.86	.96
		歸納	104	1.00	6.00	4.36	1.23
		演繹	104	.00	6.00	3.22	1.31
		解釋	104	.00	5.00	2.40	1.26
		評鑑	104	1.00	5.00	2.84	1.00
		總量表	104	8.00	23.00	15.67	3.02

	全體	辨認假設	151	1.00	6.00	2.82	1.06
		歸納	151	1.00	6.00	4.30	1.20
		演繹	151	.00	6.00	3.33	1.36
		解釋	151	.00	5.00	2.45	1.25
		評鑑	151	1.00	5.00	2.74	1.00
		總量表	151	8.00	23.00	15.64	3.02
全部參與者							
	男生	辨認假設	96	.00	6.00	2.67	1.29
		歸納	96	1.00	6.00	4.18	1.13
		演繹	96	.00	6.00	3.44	1.41
		解釋	96	.00	6.00	2.61	1.39
		評鑑	96	.00	5.00	2.47	1.00
		總量表	96	9.00	25.00	15.36	3.12
	女生	辨認假設	397	.00	6.00	2.74	1.02
		歸納	397	1.00	6.00	4.33	1.14
		演繹	397	.00	6.00	3.17	1.37
		解釋	397	.00	6.00	2.42	1.22
		評鑑	397	.00	6.00	2.71	1.04
		總量表	397	4.00	24.00	15.36	3.13
	全體	辨認假設	493	.00	6.00	2.73	1.08
		歸納	493	1.00	6.00	4.30	1.14
		演繹	493	.00	6.00	3.22	1.38
		解釋	493	.00	6.00	2.45	1.26
		評鑑	493	.00	6.00	2.66	1.03
		總量表	493	4.00	25.00	15.36	3.18