發表 2008 全美外語教學協會(ACTFL)年會，美國佛羅里達州，2008 年 11 月 21-23 日

使用共同題不等組設計之華語文能力測驗水平等化研究

摘要

本研究旨在針對國家華語測驗推動工作委員會研發之華語文能力測驗(Test Of Proficiency-Huayu，簡稱 TOP)初等、中等以及高等測驗進行水平等化研究，以瞭解初、中、高三等測驗相同等級不同內容的試卷，其估計出的測驗難度是否相近。

本研究以參加華語文能力測驗預試的華語學習者為對象，研究工具為華語文能力測驗初、中、高三等測驗。使用的等化設計為共同題不等組設計法(Common-Item Nonequivalent Groups Design)，在編擬預試試卷時，初、中、高三個等級各組成二份試卷，且挑選相同的試題將其置於同一等級二份試卷中相同的位置。

資料分析方面，採用「試題反應理論」(Item Response Theory)的等化方法，回收考生作答反應後，除計算考生答對題數外，並使用 ConQuest 軟體，以同時估計法(concurrent calibration)估計每道試題的難度參數(item difficulty parameter)以及每位考生的能力參數(ability parameter)，再將所得資料以皮爾森相關、百分比一致性進行分析。分析結果主要針對以下三個部分進行探討：1. 相同等級不同試卷之間的測驗難度是否相近；2. 於相同等級不同試卷答對同樣題數的考生，所估計出來的能力參數是否有正相關；3. 相同考生在同一等級不同試卷中的表現是否一致。研究結果可作為支持華語文能力測驗難度穩定性及測驗公正性的實徵證據。

關鍵字：共同題、華語文能力測驗、測驗等化

緒論

近年來全球掀起一股華語學習熱潮，已有許多外國人士開始學習華語，美國、法國、日本以及韓國也有不少學校陸續開設中文課程(香港文匯報，2007 年)；美國聯邦眾議員亦提出「美中語言交流法案」，要求政府撥款補助中小學開辦中文教學課程 (奇摩新聞，民 96 年)；加上美國於 2007 年開辦 AP Chinese 考試(大紀元，2007 年)、2008 年奧運在北京盛大舉行，亦連帶引起各國對於中華文化的興趣，上述跡象皆顯示「華語熱」正方興未艾。學習華語一段時間後，學習者若想瞭解自己的華語能力，或欲證明自身的語言程度，就需仰賴客觀的測量工具，因此各界對於華語測驗的需求也越來越大。

由國家華語測驗推動工作委員會(簡稱華測會)研發之華語文能力測驗(Test Of Proficiency-Huayu，簡稱 TOP)，分為初、中、高三等測驗，自 2003 年 12 月舉辦第一次正式考試以來，至今已超過四千人到考，考生國籍遍布六十餘國家；2007 年 11 月在初等測驗之下，新增適合華語初學者參加的基礎測驗。承蒙教育部國際文教處駐外代表處文化組，以及海外各地華語教學機構的大力協助與支援，從 2006 年開始，華測會也陸續在海外各地舉辦正式考試與試測(中央社，民 97a；中央社，民 97b)。

由於華語文能力測驗為一能力測驗，各個等級皆訂有通過標準，如何維持每一份測驗難度的穩定性便顯得相當重要；若每次考試難度不一，將會嚴重影響測驗的公正性以及考生的權益。過去華測會研發人員為確保測驗難度的穩定，除了依照固定且嚴密的流程進行審題及組卷外，也將試卷送交華語教學專家進行內容審查。然而質化的審查方法仍有其限制，難以呈現量化的數據結果供外界參考。國外許多大型考試，如： GRE、SAT 、TOEFL，都是使用量尺分數，並藉由測驗等化的程序，使參加不同版本測驗的考生成績可以互相比較。華語文能力測驗在目前使用原始分數的情形下，更需要透過等化研究，以確認測驗版本間難度的一致性。

文獻探討

1. 測驗等化的定義

等化(equating)是指使用統計方法，將一測驗的分數轉換至另一測驗分數量尺，以比較兩個測驗分數關係的過程，其目的是為了校準試題難度的差異。而且，等化不受試題施測內容和受試者能力分佈的影響。等化具有下列幾項特性(Kolen & Brennan，1995；引自曾玉琳、王暄博、郭伯臣、許天維，民 94)：

(1) 對稱性(symmetry)：等化之轉換必須是可逆的，即由 X 測驗等化至 Y 測驗的結果，與由 Y 測驗等化至 X 測驗之結果相同。
(2) 相同試題規格(same specifications property)：兩測驗欲進行等化，測驗內容需皆測量相同之能力特質。
(3) 相等性(equity)：受試者不論接受 X 測驗或 Y 測驗，所測得結果並無差異。

(4) 團體不變性(group invariance property)：等化過程中受試者不論來自何種團體樣本，所轉換出來的結果均相同。

2. 測驗等化的種類

Hambleton 與 Swaminathan(1985)指出測驗等化可分為水平等化(horizontal equating)與垂直等化(vertical equating)兩種，介紹如後(引自張鈺卿，民 96)：

(1) 水平等化

水平等化是指對兩個以上測量相同特質、相同能力且難度相近的測驗，將其原始分數轉換至同一量尺的過程。水平等化常被應用在許多大型測驗中，例如：托福、GRE，以及基本學力測驗等考試，就有多種複本測驗，可以在一年中實施多次的考試，然後藉由水平等化的過程，將不同複本測驗的成績轉換為同一量尺以進行比較。

(2) 垂直等化

垂直等化是指對兩個以上測量相同特質、相同能力但難度不一的測驗，將其原始分數轉換至同一量尺的過程。此類測驗的受試者的能力通常是屬於不同年齡或年級，如美國的加州成就測驗(California Achievement Tests，CAT)、愛奧華基本技能測驗(Iowa Test of Basic Skills)等，就是透過垂直等化的方式，將測驗與測驗之間的分數進行連結。

3. 測驗等化設計

測驗等化設計是指收集等化資料的方法。一般常見的等化設計包括單組設計(single group design)、平衡對抗隨機組設計(counterbalanced equivalent groups design)、等群組設計(equivalent group design)、平衡不完全區塊設計(balanced incomplete block design)、試題預先等化設計，以及共同題不等組設計(Common-item nonequivalent group design)等(王寶墉，民 84；Kolen & Brennan，1995)。

本研究採用的是共同題不等組設計，作法是在兩個不同能力分佈受試者母群體 P 和 Q 中，分別隨機抽取受試者樣本 P1 和 Q1。其中，P1 受試者接受 X 測驗，Q1 受試者接受 Y 測驗。P1、Q1 兩樣本受試者群另外需接受共同題(common items)A 測驗的施測。共同題不等組設計經常使用於一測驗只能被施測一次的測驗形式。通常共同題於兩個樣本群體中的施測順序是相同的，以避免施測順序的影響(order effects)，且共同題的內容和難度與 X、Y 測驗相似。共同題不等組設計如表 1 所示(Kolen & Brennan，1995；引自曾玉琳，民 94)。

表 1　共同題不等組等化設計

受試者群	X 測驗	Y 測驗	共同題 A 測驗
P1	O		O
Q1		O	O

由於共同題不等組設計無須讓考生在短時間內作答二份試卷，不會造成考生生理上的負擔與疲勞；可於進行預試時，事先規劃好共同題，安置於欲進行等化的測驗中，在實際實施上較為可行，故本研究採用共同題不等組等化設計。

4. 試題反應理論等化方法

測驗等化方法是指測驗等化資料收集完畢，進行測驗等化之時，連結測驗量尺之間的方法。常用的試題反應理論等化方法包括：(1)同時估計法(concurrent estimation)；(2)分離估計法(separate estimation)：包含平均數法(mean method)、平均數與標準差法(mean and sigma method)、特徵曲線法(characteristic curve method) (Kolen & Brennan，1995)，以下介紹本研究所使用的「同時估計法」。

同時估計法是藉由測驗等化設計與 IRT 電腦軟體所提供之功能作連結，將所有測驗之測驗資料同時進行試題校準，經由校準後，就能將所有測驗之受試者能力值與試題參數放置在相同量尺上。其主要的原理為：將測驗等化設計測驗題本中之試題參數估計值同時對應於相同能力量尺上(許天維，民 95)。簡言之，同時估計法是利用各測驗具有的共同試題，將所有測驗試題串在一起，並同時估計試題參數。

由於同時估計法所得到的試題參數量尺只有一種，減少了測驗間連結時產生的誤差。且國內外許多文獻亦證實，採用同時估計法能獲得較佳的精準度(Hanson & Beguin, 2002；Kim & Cohen, 1998；陳煥文，民 93；引自許天維，民 95)，因此本研究採用同時估計法。

研究目的與假設

綜上所述，本研究旨在探討使用共同題不等組設計進行水平等化研究之華語文能力測驗同一等級不同試卷，其估計出的測驗難度是否相近。研究假設如下：

1. 初、中、高三等測驗中，同一等級不同試卷之間的測驗難度(β 值)相近。
2. 初、中、高三等測驗中，於同一等級不同試卷答對相同題數的考生，所估計出來的能力參數(θ 值)有高度正相關。
3. 初、中、高三等測驗中，相同考生在同一等級不同試卷中所估計出來的能力參數(θ 值)有高度正相關。

研究方法

1. 試卷設計

本研究為瞭解同一等級二份試卷之間的整體測驗難度是否相近，採用共同題等化設計(common items design)，於 2007 年 1 月份進行預試組卷時，在初、中、高三等測驗皆編擬二套試卷，並在每一等級二份試卷中相同位置(即相同題號)，放入同樣的試題以進行水平等化。華語文能力測驗總題數為一百二十題，學者 Livingston(2004)建議超過一百題的試卷，共同題題數至少為 20 題；此外，共同題宜為原測驗之迷你版本(mini form)，各題型試題比例宜與原測驗相仿。根據上

述二原則，本次初、中、高三等測驗預試共同題題數分別為 25 題、25 題以及 24 題。詳細分測驗與題型題數分佈，及共同題題數分佈如表 2 所示。

表 2　TOP 初中高等測驗共同題分佈情形[1]

等級		聽力理解測驗			詞彙語法測驗		閱讀理解測驗		
		單句	對話	段落	詞彙	語法	單句	材料	短文
初	原始題數	20	20	10	20	20	10	20	—
	共同題數	4	4	2	5	4	2	4	—
中	原始題數	15	20	15	10	20	10	10	20
	共同題數	3	5	3	2	4	2	2	4
高	原始題數	15	20	15	20	10	10	—	30
	共同題數	3	4	3	4	2	2	—	6

2. 研究參與者

本研究參與者為 2007 年 1 月份參加 TOP 測驗預試之考生，考生可根據華測會建議的學習時數或詞彙量[2]，選擇適合的等級報考，各等級各卷別到考人數如表 3 所示。由於本次預試每一等級均提供二份試卷，且不限制考生報考卷別，因此有部分考生同一等級二份試卷都報考，初等有 57 人，中等與高等分別有 57 人及 52 人。

表 3　TOP 初中高等測驗到考人數一覽表

測驗等級	卷一	卷二	合計	跨考同等測驗[3]
初等	183	115	298	57
中等	196	147	343	57
高等	101	105	206	52

3. 資料分析

本研究使用 ConQuest 軟體進行「同時估計法」分析，採用試題作答理論的單參數模式(one parameter model)，分別估計同一等級所有試題的難度參數以及考生的能力參數。再比較相同等級不同試卷間的測驗難度(β 值平均數)；以及用皮爾森積差相關分析相同等級不同試卷答對同樣題數的考生，所估計出來的能力參數(θ 值)是否達到正相關。最後，針對跨考同一等級不同試卷的考生，其測驗成績進行比較，檢視其於二份測驗的表現是否一致。

[1] 初等測驗沒有短文題型；高等測驗沒有材料題型。
[2] 初、中、高三等測驗適用對象，請參見華測會官網 http://www.sc-top.org.tw/。
[3] 跨考同等測驗指的是同一名考生跨考同一等級不同試卷的人數。

研究結果與討論

1. 描述統計結果

初中高等測驗預試試題難度參數(β 值)描述統計結果如表 4 所示。難度參數為平均數為 0，標準差為 1 的數值，若難度參數為正值表示題目較難；若為負值表示題目較為簡單。因此，由表 4 可以得知，初等和高等測驗都是卷一的平均難度比卷二稍微困難一些；中等測驗則是卷一比卷二簡單一點。雖然初中高三等測驗的卷一、卷二平均難度不完全相等，但二者之間難度差異極小，僅分別相差 0.12、0.104 以及 0.063。顯示經由研發人員仔細審題、修改，並交由華語教學專家進行審查後的 TOP 測驗，同一等級不同試卷的難度相當接近，差異並不大。

表 4　TOP 初中高等測驗試題難度參數描述統計

測驗等級	卷別	平均數	標準差
初等	卷一	0.083	1.063
	卷二	-0.037	1.560
中等	卷一[4]	-0.064	0.919
	卷二	0.040	1.024
高等	卷一	0.048	1.107
	卷二	-0.015	1.113

2. 相關分析結果

為瞭解 TOP 三等測驗同一等級卷一、卷二二份試卷所測得考生的華語文能力是否相近，研究者進一步分別檢視在初等、中等以及高等測驗卷一、卷二答對相同題數之考生，其能力參數[5]是否接近。採用方法為先使用 ConQuest 分析軟體估計所有預試考生能力參數(θ 值)，接著從卷一與卷二篩選出答對相同題數考生之能力參數，再以統計套裝軟體 SPSS 15.0 進行皮爾森積差相關分析。若能力參數之間達顯著相關，則表示卷一與卷二測得的華語文能力具有一致性。

結果顯示，初、中、高三個等級分別有 49 人、49 人以及 46 人在同等級不同試卷答對相同題數，考生之間的能力參數相關係數均高達 1.0[6] ($p<0.001$；$p<0.001$；$p<0.001$)。此一高度正相關的結果顯示於卷一、卷二答對相同題數之考生，其估計出的能力參數相當接近(見下圖 1)；亦即，在同一等級測驗卷一、卷二試卷答對相同題數的考生，所測得華語文能力是極為接近的。

---

[4] 中等測驗卷一第 116 題送分，不列入難度參數估計。

[5] 試題反應理論(item response theory)中，能力參數(θ 值)代表測驗測得考生的能力程度(ability level)，在此可視為過去傳統測驗中的總分。

[6] 初等測驗相關值為 0.99993；中等測驗相關值為 0.9999331；高等測驗相關值達到 0.999992。經四捨五入至小數點第三位，均為 1.000。

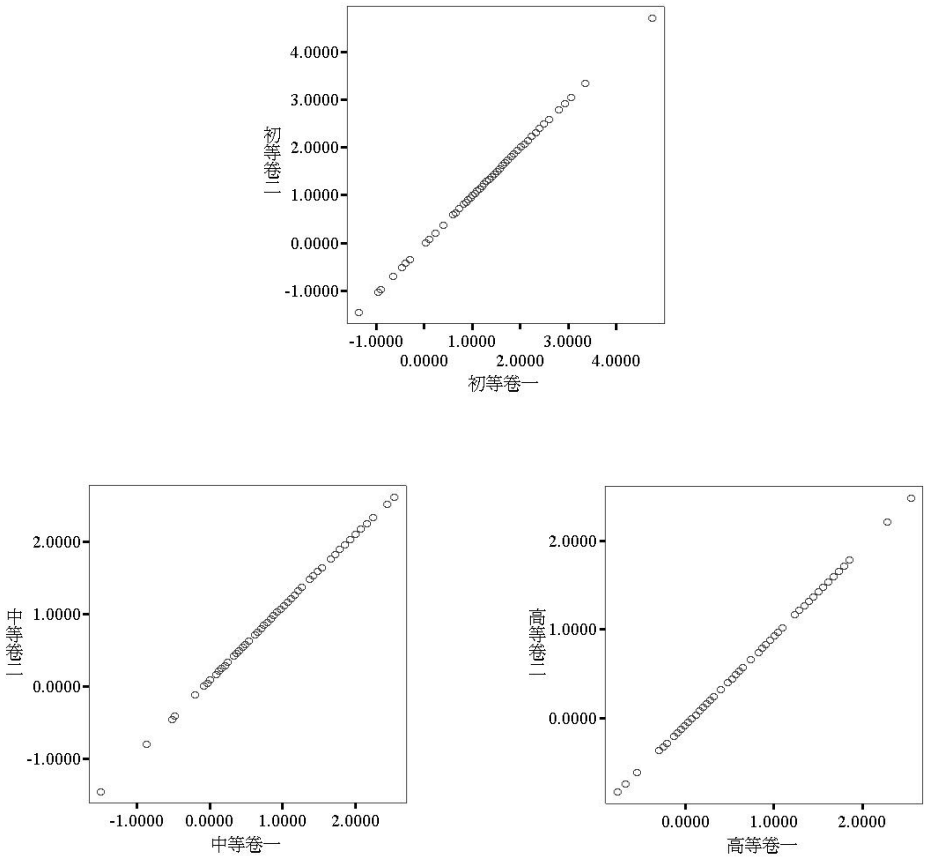

圖 1　初、中、高等測驗卷一卷二答對相同題數考生能力參數(*θ*)相關

3. 相同考生跨考同等測驗之成績表現

(1) 能力參數及答對總題數之相關

採用 IRT 所估計出的能力參數(*θ* 值)，代表此名考生的華語文能力，因此若跨考同等級不同試卷的考生，其能力參數有高度相關，便表示該生在二份同等級試卷所測得的能力非常接近，考生的華語文能力不因其報考不同卷別而有差異。初、中、高三等測驗分別有 57 人、57 人以及 52 人跨考同等級不同卷別之預試，這些考生能力參數之皮爾森積差相關值分別為 0.947 ($p < 0.001$)、0.822 ($p < 0.001$)以及 0.911 ($p < 0.001$)，顯示跨考考生的能力參數之間有高度正相關。

由於此次預試成績，仍提供考生原始分數，因此也針對考生測驗總分(即答對總題數)進行相關分析，結果顯示，初、中、高三等測驗跨考考生之測驗總分相關值為 0.947 ($p < 0.001$)、0.848 ($p < 0.001$)以及 0.903 ($p < 0.001$)，同樣達到高度正相關。

(2) 測驗通過情形

TOP 考試為一能力測驗，於初、中、高每一等級均設有通過門檻，考生若通過門檻可獲得證書，惟預試成績僅供參考用，不發放證書。因此除了考生測驗總分的相關情形外，研究者也關心跨考考生在同一等測驗是否有一致的測驗通過情形，故針對三等測驗分別進行百分比一致性分析。

如表 5 所示，初等測驗部分，57 人之中，有 43 人在卷一和卷二均為相同等級，比例為 75.44%；若單看獲得證書與否，則有 87.72%的考生在二份試卷上的表現是一致的。

表 5　初等測驗卷一、卷二通過等級交叉分析表

		卷二			合計
		未通過	初等 1 級	初等 2 級	
卷一	未通過	26	5	0	31
	初等 1 級	2	9	5	16
	初等 2 級	0	2	8	10
	合計	28	16	13	57

表 6 為中等測驗跨考考生通過情形交叉分析表，57 人之中，有 47 人在二份試卷獲得相同等級，百分比為 82.46%；若單看獲得證書與否，則有 84.21%的考生在二份試卷上的表現是一致的。

表 6　中等測驗卷一、卷二通過等級交叉分析表

		卷二			合計
		未通過	中等 3 級	中等 4 級	
卷一	未通過	31	5	1	37
	中等 3 級	3	13	1	17
	中等 4 級	0	0	3	3
	合計	34	18	5	57

高等測驗的部分，如表 7 所示。52 人中有 47 人在二份試卷獲得相同等級，百分比為 84.62%；若單看獲得證書與否，則有高達 92.31%的考生在二份試卷上的表現是一致的。

表 7　高等測驗卷一、卷二通過等級交叉分析表

		卷二				合計
		未通過	高等 5 級	高等 6 級	高等 7 級	
卷一	未通過	28	1	1	0	30
	高等 5 級	2	5	2	0	9
	高等 6 級	0	1	7	1	9
	高等 7 級	0	0	0	4	4
	合計	30	7	10	5	52

結論與建議

由上述分析結果可知，TOP 考試初、中、高三等測驗，同一等級的二份試卷試題平均難度差異極小。於二份試卷答對相同題數的考生，能力參數之間的相關值均逼近 1.0；顯示考生無論參加哪一份試卷，所獲得的成績幾乎可以說是相等的，是能夠互相比較的。從跨考考生的測驗成績，也顯現 TOP 考試的測驗難度是一致的，考生在二週內參加同一等級不同內容的測驗，其測驗表現落差不大，測驗總分均達到高度正相關，通過測驗獲得證書的比例也相當一致。

整體來看，本研究三項假設都獲得了支持，分析結果均肯定 TOP 測驗初、中、高三等考試的測驗難度相當穩定。加上本研究使用的試題內容為預試試卷，正式考試的內容，還會從預試的試題中，篩選掉品質不佳的試題後再進行組卷，更能確保考生的權益以及測驗的公平性。

未來可進一步採用共同題不等組設計，進行 TOP 初、中、高三等測驗的垂直等化研究，將三等測驗試題難度參數置於同一量尺上進行比較，以瞭解三個等級的試題難度分佈情形，及各等測驗的難度是否區分良好。還可藉由不同等級獲得相同能力參數考生的答對總題數，獲得 TOP 三等測驗之間的分數對應情形，並建立初、中、高三等測驗的分數對照表，提供給華語教師及華語學習者參考。

參考文獻

1. 中文部分

王寶墉(民 84)：**現代測驗理論**。台北市：心理出版社。

北京奧運將引發新「漢語潮」(2007 年 9 月 14 日)。**香港文匯報**。民 96 年 9 月 15 日，取自：http://news.wenweipo.com/2007/09/14/IN0709140045.htm

美國首創中文 AP 考試 影響深遠（2007 年 2 月 23 日）。**大紀元**。民 97 年 5 月 8 日，取自：http://news.epochtimes.com/b5/7/2/23/n1628467.htm

美國會議員提案加強中小學中文教學（民 96 年 8 月 15 日）。**奇摩新聞**。民 96 年 8 月 15 日，取自：http://tw.news.yahoo.com/article/url/d/a/070815/5/iro9.html

張鈺卿(民 96)。**BIB 與 NEAT 設計在不同年度測驗連結效果之比較**。國立台灣教育大學教育測驗統計研究所碩士論文，未出版。

推動繁體中文　台灣在泰舉辦第三屆華文測驗（民 97b 年 4 月 27 日）。**中央社**。民國 97 年 5 月 8 日，取自：http://tw.news.yahoo.com/article/url/d/a/080427/5/y38j.html

許天維(民 95)。**大型教育測驗等化設計及效果之研究**。國科會專題研究計畫。

曾玉琳(民 94)。**不同配置設計下測驗等化效果之模擬研究**。國立台中師範學院數學教育學系碩士論文，未出版。

曾玉琳、王暄博、郭伯臣、許天維(民 94)。**不同 BIB 設計對測驗等化的影響**。測驗統計年刊，第十三輯下期，209-229 頁。

蔡良庭、施懿珊(民 94)。**具差異試題功能之定錨試題對測驗等化之影響**。測驗統計年刊，第十三輯上期，95-109 頁。

駐紐約文化組將再舉辦三場華語文能力測驗（民 97a 年 4 月 11 日）。**中央社**。民 97 年 5 月 8 日，取自：http://tw.news.yahoo.com/article/url/d/a/080411/5/x3bj.html

2. 英文部分

Kolen, M.J. & Brennan, R.J. (1995). *Test Equating: Methods and Practices*. New York: Springer-Verlag.

Livingston, S. A. (2004). *Equating test scores*. Retrieved March 14, 2007, from http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf