發表於華語文能力測驗成果發表會，國立台灣師範大學，2011 年 6 月 25 日

「華語文寫作能力測驗」評分者一致性探討

林佩樺 彭淑惠 藍珮君

一、測驗簡介

華語文寫作能力測驗(前稱為 Test Of Proficiency-Huayu Writing，現更名為 Test of Chinese as a Foreign Language- Writing, 簡稱 TOCFL- Writing)，是專為母語非華語之人士所設計的一種外語/第二語言寫作能力測驗。本測驗於 2007 年著手研發，等級規劃主要參照 CEFR(The Common European Framework of Reference for Languages：Learning, Teaching, Assessment) 寫作能力指標描述（見表 1），以「溝通任務」為導向，依溝通任務的難度與複雜度劃分為基礎級、進階級、高階級、流利級四個等級，分別對應到 CEFR 的 A2、B1、B2、C1 等級。自 2007 年迄今，基礎級和進階級已舉辦過 11 場預試，將於今年 10 月正式施測，高階級則舉辦了 5 場預試(含寫作比賽)，與流利級皆尚在研發階段。本文將就基礎級與進階級說明寫作命題及評分理念與方式。

表1 與CEFR寫作能力對應及華語文寫作能力指標

TOCFL	CEFR	華語文寫作能力指標
基礎級	A2	能運用短語、連結句子，撰寫和自己生活經驗相關便條或簡單的私人信件。
		能運用短語、連結句子，撰寫和自己生活經驗相關的短文。
進階級	B1	能書寫詳細私人信件，傳達切身相關的訊息。
		能對具體或抽象事件、主題，描寫經驗、感受與反應。
高階級	B2	能藉由書信強調個人在事件或經驗中的重要性，並對通信者提供建議與評論。
		能統整不同論點，並針對問題來評析優劣，提出支持或反對的理由。
流利級	C1	研擬中

二、寫作命題

寫作測驗屬主觀性測驗（subjective test），影響其信度、效度的主要因素，包括試題選樣誤差與評分者間一致性的問題，Weigle (2002)指出，寫作測驗牽涉到兩種基本成分：設定 1 或 2 個任務、告訴考生要寫什麼，以及評量寫作樣本的

方法。盧雪梅（2005）認為，寫作題目的功能在引發考生的寫作表現，以從中觀察和評估考生的寫作能力。題目設計的良窳不僅涉及測驗目的是否達成，也影響到後續的評分工作。

在命題方面，本會依據華語文寫作能力指標，針對不同等級的考生設計與實際生活與工作相關的不同任務，力求試題能確實符合該等級能力指標的敘述內容，使考生能最大限度地展現其寫作能力，並避免非寫作能力之因素干擾。為求更確實地評量考生不同面向之寫作能力，所有等級均必須完成兩種文體寫作。寫作試題形式包含三個部分：情境敘述（如：陳述事件引導考生）、寫作任務（題目設定的寫作內容）、注意事項（字數、時間等相關規定）。各等級題型見表2。

表2 華語文寫作能力測驗各等級題型

等級	文體	字數規定	測驗時間
基礎級	應用文（便條、邀請函、感謝信、道歉信、卡片等）	70-120	20分鐘
	記敘文（看圖作文）	150-200	40分鐘
進階級	應用文（私人書信）	300-400	50分鐘
	記敘文（真實事件或虛構故事）	300-400	50分鐘
高階級	應用文（申訴信、建議信等）	500-600	60分鐘
	論說文	500-600	60分鐘
流利級	研擬中		

現今以電腦作為書面表達媒介的使用已日益普及，在文書處理上，如：剪下、複製、貼上等功能，符合真實寫作需經構思、重新組織、檢視文句適切與否等過程，而資料儲存與即時寄送等功能，亦帶給測驗單位極大的便利性與時效性。基於上述諸多優點，本測驗採取電腦寫作方式，寫作題目與作答欄左右並列(題目呈現方式，如圖 1 所示)。考生可運用考試系統上的工具列編輯文章、加註標點符號，也能知道已打字數和測驗的剩餘時間。考完後，將考生文本傳輸至主機。雖然本測驗以電腦為媒介，但考生僅需具備基本的中文輸入能力。而為幫助考生熟悉題型與作答方式，本會在網站上設置「線上寫作練習平台」供其練習。

圖 1 考題與作答欄呈現方式

三、寫作評分

在評分方面，華測會為確保寫作能力測驗品質，從理論與實務兩方面出發，除了參考寫作評量相關文獻，汲取他人經驗，並擴大蒐集寫作樣本範圍，根據不同考生群的寫作樣本特徵，訂出該等級的評分原則，作為評分教師評分依據，同時，擬訂閱卷流程、研發線上評分系統，以提高評分一致性，並透過百分比一致性(percentage agreement)與斯皮爾曼等級相關分析(Spearman rank correlation)，檢測「評分者間信度」(inter-rater reliability)。此外，並採用多面向 Rasch 測量模式(many-facet Rasch measurement)分析軟體 FACETS，分析評分者嚴格度(rater severity)，以了解評分教師與會內嚴格度的差異，並從評分者面向之適配度(fit)探討評分者穩定性，即評分者內信度，藉此檢視本會寫作測驗評分品質。寫作評分相關內容說明如下：

（一）評量概念與評分方式

研究顯示，「級分制」較能確保評分者信度，因此本會以五個等級評定考生的寫作能力，由優至劣區分為五級至一級分。其中，三級分表示已達一般水準，為通過門檻。另外，針對完全不符合情境與任務、只抄題目、完全以對話形式呈現、條列式、字數極少等（幾乎）無法達成書面溝通任務之文本，給予 0 級分。

華語文寫作能力測驗評量向度分為「寫作任務完成度」和「文意表達」兩大部分。前者以檢視考生是否針對題目所設定的情境與各項任務敘寫，以其切合題意的程度與發展性給分；後者檢視「全文結構組織表現」與「詞語運用能力」兩項能力。結構組織表現主要觀察分段與整體結構的適切度、脈絡的條理清晰度與前後句的銜接策略，而語言運用則是觀察詞語的掌握度與豐富度。至於句內語法結構，考量到不同等級的評量重點不同與配分問題，將基礎級與進階級納入結構

組織表現，高階級納入語言運用能力。此三個向度的五個等級表現即構成評量考生文章優劣的標準，即「評分原則」(scoring rubric)，見附錄一。

大型寫作測驗之計分方式，一般採用整體式計分（holistic scoring）或分析式計分（analytic scoring)。前者是由評分教師根據文本整體印象，給予一個單一分數；後者是由評分教師根據文本特徵，給予數個向度或層面的分數，最後將各項分數依計分規定加總，得到一個總分。

本會在研發初期,考慮到未來正式施測時,可能一次需要評閱的文本數量大,但評閱時間有限，故決定採取整體式計分。然而在多次評分實務中發現，此種計分方法雖然執行起來簡單、快速，但是較主觀，而且不論是對測驗單位、教學單位或是考生，其評分結果的回饋訊息皆不夠明確，也缺少診斷功能。以成績相同之文本為例，因考生寫作風格殊異、文本特徵多樣，若只給一個分數，不僅不利於了解評分教師個別的評閱概念與給分標準，當考生質疑成績時，也不易說明給分理由。而分析式計分雖能避免上述缺點，但各面向得分的加總並不能真實反映整體寫作表現，同時，此種計分方式也相當耗費時間與財力，恐帶來評閱的時間壓力。為了解決上述問題，本會設計評分表單與標註方式，將之運用在評分教師培訓工作上。

（二）監控評分一致性機制

2.1 評分表單的設計與運用

為提高評分一致性與了解評分教師的評分思維,本會依照評分原則設計寫作文本與評分表格左右並列的評分表單，並訂出統一的標註方式。評分教師必須按照規定將考生的錯誤標記在左邊的文本上，或填寫在右邊的細項說明欄，並給予整體分數(見圖 2)。例如詞語不當以灰底標示於文本上；某些語法偏誤,如虛詞、實詞、補語及某些特殊句式等，將偏誤部分複製於右邊的評分表單中。最後計算出各細項得分以及「情境任務關聯度」、「全文結構組織表現」、「句子層面語言運用能力」三個向度的分數，再打上整體分數。

評分表單可運用於評分教師培訓階段，亦可運用於未來的正式評閱階段。在培訓階段，評閱方式以分析式評分法為主，即必須在文本與表單上明確標示出偏誤部分，並給予細項分數與向度分數，最後再依整體寫作表現給分。透過評分表單，可了解評分教師的評分概念與標準是否與本會的標準相同，而評分表現好的評分教師可成為本會的核心評分教師。在正式評閱階段，評分方式則只需給予細項、向度與整體分數，不需標註偏誤內容。因此不但能節省許多評分時間，還能得到足夠的回饋訊息。

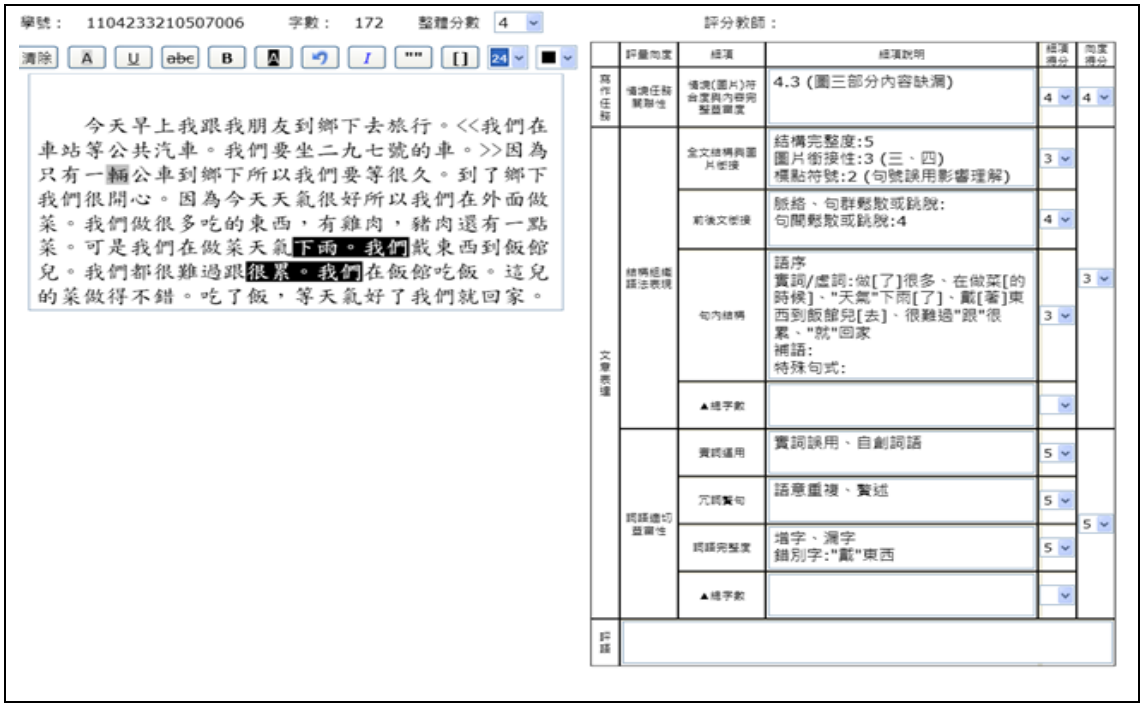

圖 2 評分表單

2.2 線上評分系統

為便於比對評分教師所送出之成績，本會研發出線上評分系統，支援後端監控功能，教師於線上評閱並送出成績後，本會可立即於後端之監控畫面得知所有細項得分和整體分數，藉以檢視評分教師的一致性（見圖 3），若某位教師偏離了評分標準，則可立刻將其已評閱之文本列印出來，並針對與本會標準不同之處，與該老師互相討論以達成共識。

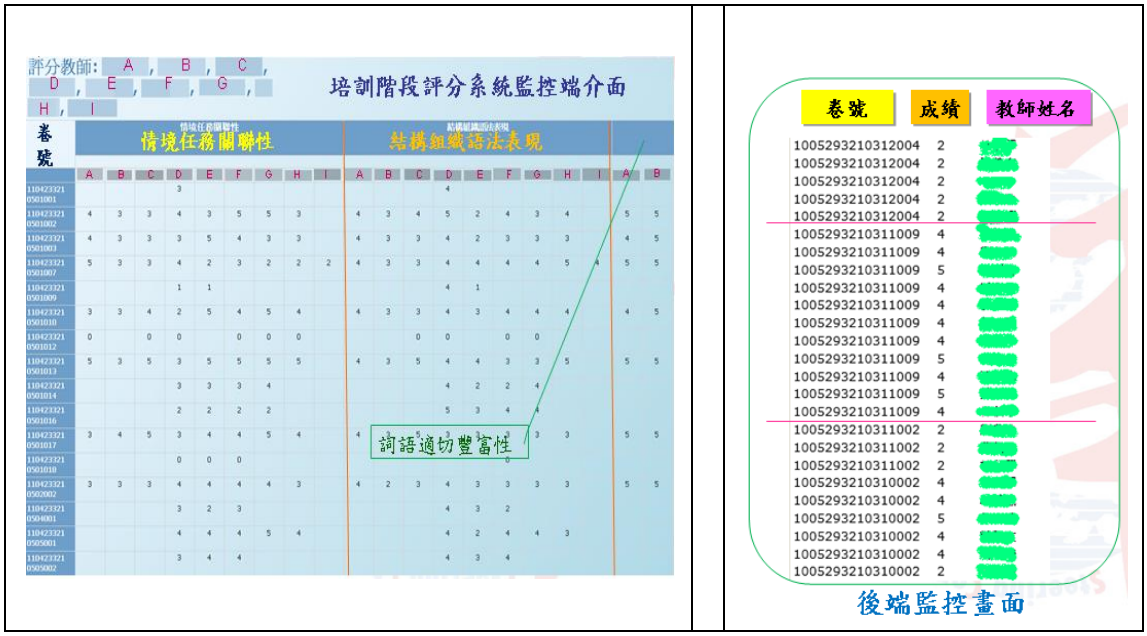

圖 3 線上評分系統後端監視畫面

此外，在正式評閱時，透過評分系統自動將寫作答卷隨機分配給事先分組的教師評閱，並隨機插入標準卷[1]。如此一來，不僅可確保所有評分教師均與內部評分標準一致，在評分不一致時，透過觀察各自評閱標準卷的表現，也可瞭解是否其中一位教師偏離了評分標準。

（三）評分教師培訓計畫

關於如何提高寫作測驗信度，過去也有不少研究成果。例如：Follman 與 Anderson(1967)認為，測驗單位應擬訂評分程序，因為評分者間信度低的主要原因是評分者來自不同的教育背景和經驗，評分時也可能會用不同的態度。因此，設定良好的評分程序，則可讓評分者在評分時有一個共同的方向與認知。French (1996)指出，透過密集訓練和監控，一位評分者的信度可達到 0.7，而未受過訓練的評分者，若依自己的判斷去評分，其信度僅達 0.31。國民中學基本學力測驗寫作測驗題庫發展組亦提出，為加強評分者間閱卷標準的一致性，「評分規準（scoring rubrics）的建立」與「評分人員的訓練」是現階段寫作測驗的研發重點。綜上所述，評分者間一致性會因為使用定義良好的評分規則和周延訓練的評分者而提升（鄒慧英，2003）。

雖然任何一種評量方式都無法達到絕對的客觀，但研究顯示，嚴謹的評分教師培訓計畫有助於提高評分一致性。以下說明評分研習相關內容與成果。

3.1 舉辦評分研習

舉辦評分研習可使參加者了解寫作評分原則內容，透過實際評閱也易於掌握評分要領，測驗單位亦能從中挑選有熱忱且穩定性高的評分老師，做為核心教師。

3.2 寫作評分研習流程

在進入評閱之前，測驗單位必須提供評分原則，作為可共同遵循的評分標準，並透過實際評分活動，觀察評分原則內容是否周延，同時參考評分教師的建議，修訂出更為完備的評分原則（見附錄一），以利於後續的評分培訓工作。

評分原則內容為「任務完成度」與「文意表達」兩大評量向度之五個等級的寫作能力表現特徵。其中，文意表達能力向度所要評量的是考生的結構組織表現以及對於詞語與語法的掌握度，儘管每次的考題不同，但對這兩項能力的要求是一致的。而「任務完成度」則不然，必須依照每一個題目設定的各項任務涵蓋面與主次關係，訂出相應的詳細等級評定標準，本會稱之為「評分細則」。

為求達到最大共識，本會在舉辦評分研習之前，會邀請數位經驗較為豐富的

[1]標準卷是在樣卷會議中，選出所有核心教師評分結果相同的試卷。

評分教師根據本會草擬的「評分細則」試評，並共同討論與修訂，而後，測驗研發人員再從試評卷中挑選出較無爭議的答卷作為寫作樣卷與練習卷。希冀透過多次的評閱與溝通，能夠調整評分教師的評分寬嚴度，使其趨近於華測會的標準，而使得測驗能更具信效度。寫作評分研習的籌備工作內容，詳列如表 3。

表 3 寫作評分研習流程

評閱前	1. 制定研習流程。 2. 研發人員依據題目制定任務完成度評分細則。 3. 邀請經驗較豐富的閱卷教師試評與討論爭議卷。 4. 研發人員確定各級樣卷、練習卷。 5. 由資訊人員將各級樣卷、練習卷及待評卷置入系統。
評閱中	1. 說明評閱流程。 2. 說明評閱重點、評分規準與細則。 3. 監控閱卷教師評分一致性情況。
評閱後	1. 分析評分一致性。 2. 依需要修訂評分細則。 3. 提交評分結果。

3.3 培訓階段評分結果分析

表 4 為某次第一階段培訓八位評分教師評分結果的等級相關結果，主要在了解經過評分研習後，評分教師與會內評分標準的關聯性，並進一步透過三個給分向度，瞭解評分教師可能在某一向度需要做進一步的溝通，釐清概念。以下表為例，評分教師 E 與 F 與本會研發人員給分之間的相關值較高，為所有評分教師中表現較為理想的；針對評分教師 C 則會與其進行溝通，瞭解評分時是否遇到什麼樣的困難；評分教師 H 在「情境任務關連性」與「結構組織語法表現」二個向度與本會給分關連性佳，但在「詞語適切豐富性」的相關值未達顯著($r=0.347$, $p>0.05$)，需要就此一向度的給分多做瞭解。

表4　評分教師與華測會研發人員各向度給分之斯皮爾曼等級相關

評分教師	A	B	C	D	E	F	G	H
篇數	15	15	15	36	35	29	22	20
情境任務關聯性	.579*	.778**	.480	.671**	.911**	.847**	.785**	.827**
結構組織語法表現	.576*	.545*	-.242	.325	.809**	.631**	-.078	.720**
詞語適切豐富性	.723**	-.173	.704**	.328	.719**	.669**	.832**	.347
整體分數	.759**	.763**	.514*	.630**	.909**	.853**	.620**	.684**

*p<0.05；**p<0.01

除了等級相關外，亦會針對評分教師與會內研發人員整體分數進行百分比一致性分析，表 5 與表 4 為同一次培訓後的分析結果，由下表可知，此次八位評分教師，在整體分數給分上，與華測會研發人員相差一級分以內，且都評為通過或不通過的百分比(P_{0+1})普遍表現不錯，除了評分教師 D、G 外，其餘教師與會內給分的一致性均達到 85.0%以上。

表 5　基礎級預試看圖寫作評分結果百分比一致性結果（培訓階段）

評分組合	篇數	P_0	P_1	P_{0+1}
A & 華測會	15	9 (60.0%)	5 (33.3%)	14 (93.3%)
B & 華測會	15	9 (60.0%)	5 (33.3%)	14 (93.3%)
C & 華測會	15	5 (33.3%)	8 (53.3%)	13 (86.7%)
D & 華測會	36	20 (55.6%)	5 (13.9%)	25 (69.4%)
E & 華測會	35	28 (80.0%)	3 (8.6%)	31 (88.6%)
F & 華測會	29	22 (75.9%)	4 (13.8%)	26 (89.7%)
G & 華測會	22	13 (59.1%)	2 (9.1%)	15 (68.2%)
H & 華測會	20	12 (60.0%)	5 (25.0%)	17 (85.0%)

註：P_0 在此指評分結果完全相同的比例；P_1 指的是評分結果相差一級分，且都評為通過或不通過的比例；P_{0+1} 表示 P_0 與 P_1 的和。

透過多面向 Rasch 測量模式，則可以進一步瞭解評分教師與會內研究人員給分的嚴格度差異，以及評分教師自身給分的一致性。從表 6 可知，在評分者嚴格度方面，與華測會人員相較，較為接近的是評分教師 B 和 F；評分教師 G 給分較為寬鬆，其餘評分教師與華測會人員差異並不大。評分者內的穩定性方面，幾乎所有評分教師均符合 INFIT 值大於 0.7 或小於 1.3 的標準(McNamara, 1996；Bond & Fox, 2001；引自 Eckes, 2005)，顯示經過評分培訓，評分教師在給分時能維持自身的標準。

表6　向度分數的評分者嚴格度（培訓階段）

評分教師	觀察的平均值	調整過平均值	嚴格度	與華測會差異	標準誤 (S.E.)	INFIT MNSQ	OUTFIT MNSQ
H	3.5	3.42	0.38	0.25	0.18	1.30	1.78
E	3.5	3.44	0.37	0.24	0.13	0.99	0.93
B	3.7	3.55	0.21	0.08	0.20	1.01	0.90
華	**3.5**	**3.61**	**0.13**	**—**	**0.11**	**0.73**	**0.76**
F	3.8	3.69	0.00	-0.13	0.15	0.83	0.79
C	3.9	3.74	-0.07	-0.20	0.20	0.85	0.81
D	3.8	3.81	-0.19	-0.32	0.14	1.31	1.45
A	3.9	3.82	-0.20	-0.33	0.21	0.75	0.62
G	4.1	4.08	-0.63	-0.76	0.19	1.13	1.00

RMSE 0.17　Adj S.D. 0.27　Separation 1.59　Reliability 0.72
Fixed (all same) chi-sqaure: 28.7　d.f.: 8　sig: 0.00

經過第一階段培訓後，表現較為穩定的評分教師，則可進入第二階段培訓，此一階段的評分，採二人一組。進行評分後，使用百分比一致性分析瞭解二位評分教師的一致性。由表 7 可知，二位評分教師給分的百分比一致性不錯，在 19 篇文本中，有 14 篇給分完全相同，2 篇相差一級分，但都評為通過或不通過。

表 7　進階級記敘文第四組二位評分教師評分結果百分比一致性結果

篇數	P_0	P_1	P_{0+1}
19	14 (73.7%)	2 (10.5%)	16 (84.2%)

註：P_0 在此指評分結果完全相同的比例；P_1 指的是評分結果相差一級分，且都評為通過或不通過的比例；P_{0+1} 表示 P_0 與 P_1 的和。

第二階段的培訓工作，會內研發人員會適時抽取一至數位評分教師評閱的文本，進行評分，以持續追蹤評分教師給分的標準，表 8 與表 9 即為某位評分教師與會內人員的評分分析結果。從下表可知，進入到第二階段的評分教師，在等級相關以及百分比一致性的分析結果，均顯示其與會內研發人員的給分標準相當一致。

表8　進階級記敘文第五組二位評分教師各向度給分之斯皮爾曼等級相關

篇數	情境任務關聯性	結構組織語法表現	詞語適切豐富性	整體分數
18	0.854**	0.804**	0.721**	0.834**

**$p<0.01$

表 9　進階級記敘文第五組二位評分教師評分結果百分比一致性結果

篇數	P_0	P_1	P_{0+1}
18	14 (77.8%)	2 (11.1%)	16 (88.9%)

註：P_0 在此指評分結果完全相同的比例；P_1 指的是評分結果相差一級分，且都評為通過或不通過的比例；P_{0+1} 表示 P_0 與 P_1 的和。

四、未來研發方向

華語文寫作能力測驗歷經四年研發，在試題方面已確立基礎級、進階級和高階級三個等級的題型，並設置電腦考試系統；評分方面，舉辦了近百場評分會議，累積了不少評分經驗，藉此修訂之評分標準與評分方式，線上評分系統的設置，皆有助於凝聚評分教師共識，提升評分一致性。未來將在此基礎上，密集舉辦評分研習，使評分老師更熟悉評分要領，並透過線上評分系統的控管，以期提升評分一致性。

附錄一

基礎級 看圖寫故事 評分原則

＊ 5 級分並非要求完美無誤的文本，而是其寫作表現在該等級已屬高水準。

＊ **3 級分為通過門檻**。三項評量向度**均至少須達 3 級分**。

＊ 若是高於 3 級分，則視此卷**整體表現**偏向何種等級的評分原則敘述給分。

向度 / 級分	任務完成度	文意表達	
	內容取材適切性與豐富度	結構組織完整性與語法正確度	詞語適切度
5 級分	完成所有寫作任務。	5.1 結構可；圖片內容銜接良好；脈絡清楚。 【標點符號使用大致正確】 5.2 句間銜接良好（轉折詞適切運用）。 5.3 句內結構幾乎無誤。	5.1. 適切運用基礎級實詞。 （容許極少數超越等級的實詞使用不當） （容許極少數錯別字或增、漏字） 5.2 冗詞贅句極少。
4 級分	大致完成所有寫作任務。	4.1 段落開頭未空兩格；圖片銜接大致良好，內容偶有不合邏輯之處。 【標點符號使用尚可（偶爾未斷句）】 4.2 句間銜接大致良好（轉折詞運用大致適切）。 4.3 少數句內結構錯誤。	4.1 基礎級實詞運用大致正確。 （容許少數基礎級實詞使用不當） （容許少數錯別字或增、漏字） 4.2 偶有冗詞贅句（偶有語意重複或贅述）。
3 級分	完成大部分寫作任務。	3.1 分段較多或該分段未分段；對話形式較多；部分圖片銜接不甚理想；內容少部分不合邏輯。 【標點符號使用不甚理想（多數未斷句）】 3.2 句間銜接較差（較鬆散或少部分跳脫）。 3.3 句內結構錯誤較多。 ▲ **3.4**：字數不足(**121 – 149** 字)，具 4、5 級分表現。	3.1 基礎級實詞掌握不甚理想，但不影響理解。 3.2 冗詞贅句略多（語意重複或贅述略多）。 3.3 詞語偶不完整（錯別字或增、漏字較多），基本上不影響閱讀。 ▲ **3.4**：字數不足(**121 – 149** 字)，具 4、5 級分表現。
2 級分	完成少部分寫作任務。	2.1 非篇章形式或分段過多；部分圖片銜接不佳；內容多處不合邏輯。 【標點符號使用錯誤多，影響理解】 2.2 句間銜接差（鬆散或部分跳脫）。 2.3 句內結構掌握差。 ▲ **2.4**：字數不足(**121 – 149** 字)，具 3 級分表現。 ▲ **2.5**：字數少(**100 – 120** 字)，具 4、5 級分表現。	2.1 基礎級實詞掌握差，影響理解。 2.2 冗詞贅句多（語意重複或贅述較嚴重）。 2.3 詞語多處不完整（錯別字或增、漏字很多），影響理解。 ▲ **2.4**：字數不足(**121 – 149** 字)，具 3 級分表現。 ▲ **2.5**：字數少(**100 – 120** 字)，具 4、5 級分表現。
1 級分	完成極少的寫作任務。	1.1 圖片銜接極差；脈絡凌亂。 【標點符號使用錯誤極多，嚴重影響理解】 1.2 句間銜接極差。 1.3 句內結構掌握極差。 ▲ **1.4**：字數不足(**100 - 149** 字)，**未達 3 級分**。 ▲ **1.5**：字數過少(**50 -99** 字)。	1.1 基礎級實詞掌握極差，嚴重妨礙理解。 1.2 冗詞贅句嚴重（語意重複或贅述嚴重）。 1.3 詞語極不完整（錯別字或增、漏字極多），嚴重影響理解。 ▲ **1.4**：字數不足(**100 - 149** 字)，**未達 3 級分**。 ▲ **1.5**：字數過少(**50 -99** 字)。
0 級分	**完全空白；僅抄題目；文不對題；使用真實姓名；全文以對話形式呈現；全文條列式（如清單）、字數極少（含標點符號，50 個字以下）**		

參考文獻

鄒慧英（2003）。測驗與評量─在教學上的應用。台北：洪葉文化事業有限公司。

盧雪梅 (民 2005)。淺析寫作測驗的重要課題。國中基本學力測驗專刊-飛揚，第 35 期。http://www.bctest.ntnu.edu.tw

國中基本學力測驗推動工作委員會題庫發展組（2006）。寫作測驗結果的使用。國中基本學力測驗專刊-飛揚，第 37 期。http://www.bctest.ntnu.edu.tw

Cooper, C. R., and Odell, L.(eds.). *Evaluating Writing*. Urbana, Ill.: National Council of Teachers of English, 1977.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

Follman, John. C., and Anderson, James A. "An Investigation of the Reliability of Five Procedures for Grading English Themes." *Research in the Teaching of English.* 1967.

Robert, W, *Assessment and Testing: A Survey of Research.* Cambridge University Press. 1993.

Sara Cushing Weigle. (2002). Assessing Writing. Cambridge: Cambridge University Press.

Veal, L. R. & Hudson, S. A. Direct and indirect measurements for large- scale evaluation of writing. Research in the Teaching of English. 1983.