發表於第七屆國際電腦漢語教學研討會，2012 年 5 月 25-27 日，美國夏威夷州。

華語寫作測驗線上評分系統之運用

Shuhwei Peng　　Peihua Lin
（彭淑惠）　　（林佩樺）
Steering Committee for the Test of Proficiency-Huayu
speng0620@sc-top.org.tw　cecilin@sc-top.org.tw

摘要

寫作測驗爲主觀性測驗，難免發生評分結果不一致的情形。有鑒於此，本會研發一套「線上評分系統」，將寫作文本、考題、評分表單與樣卷皆上傳於系統中，評閱時可同時瀏覽考題及樣卷；並支援後端監控功能，培訓人員於評分者上線評閱時，即能檢測評分是否一致。本研究以某次評分培訓之統計分析，以及評分者對該系統使用滿意度問卷調查結果，作爲優化該系統之參考，以期提高評分一致性。

關鍵字：寫作測驗、線上評分系統

1. 研究動機與目的

華語文寫作能力測驗（Test of Chinese as a Foreign Language-Writing，簡稱 TOCFL- Writing），是專爲母語非華語之人士所設計的外語／第二語言寫作能力測驗，考試時採取鍵盤輸入方式，考生將文本送出後，直接上傳至考試系統。評分會議前，研發人員將文本匯入評分系統中，由評分者進行評分。本會在多次的評分實務中發現，分析式評分法（analytic rating）的評分結果能提供較多的訊息，供研發人員瞭解評分者的評分思維，但較爲費時。因此本會爲兼顧訊息量之取得與評閱效率，設計了評分表單與標註方式，將之運用於線上評分系統中。

本會基礎級寫作測驗的主要評量內容包括「情境任務的符合度與充實度」、「結構組織句法表現」、「詞語表現」三個向度；次要評量項目則是字數、錯別字和標點符號的使用。其中「情境任務的符合度與充實度」的評分較爲主觀，因此研發人員在每一次的寫作評閱工作中，皆需針對該次的寫作題目，訂出相應的任務評分細則。而「結構組織句法表現」與「詞語表現」的評分則相對客觀，不論題目爲何，其給分標準不變，因此可將其錯誤量化。本會根據過去多次的評分結果訂出各細項[1]的量化數值，作爲給分的標準。建置於線上評分系統中的評分表單即根據上述評量內容表格化而成。

在正式評閱時，評分者需全面關注考生的各項表現。然而要求尚在培訓階段的評分者同時兼顧各個面向的評量實屬不易，若依培訓目的針對某評量向度進行密集訓練，預期可達更佳的培訓效果。

[1]「結構組織句法表現」向度包含全文結構、前後文銜接和句內結構三細項；「詞語表現」向度包含詞語適切度、詞語簡潔度和詞語完整度三細項。

基於此，本研究請評分者僅針對可量化的「結構組織句法表現」與「詞語表現」兩個向度進行評分，研發人員透過評分系統後端監控介面與統計分析結果，掌握評分者的評分一致性與嚴格度，並以線上評分系統滿意度問卷調查評分者意見，作爲優化該系統之參考，以利凝聚評分者共識，提升一致性。

2. 文獻探討

寫作測驗爲主觀性測驗（subjective test），評分者的主觀看法對考生分數的影響甚鉅，若未經訓練，難免發生評分結果不一致之問題。許多學者對此議題得出不少研究成果，例如：Anderson 與 Follman（1967）提到優良的評分程序可讓評分者產生共識；French 認爲透過密集訓練與監控可大幅度提高評分信度，評分者間一致性也可因使用定義良好的評分規則和周延訓練而提升（鄒慧英，2003）。

隨著電腦和網路技術的快速發展，許多學者認爲使用線上評分的方式亦有助於提高一致性，尤能避免傳統紙本的人工評分方式無法隨時記錄所有的評分結果，造成評分持續偏離的情況。英國測驗研究專家 Shaw（2007:14）提到電腦輔助評分能够即時提供數據，這些數據對信度尤其重要。使用電腦亦使新的評分模式易於操作，且能嚴格控制品質與成本。他同時指出，在線上評分時不僅應提供數據，亦應展示、測試各面向以供研究。高丙成、秦旭芳（2007）進一步針對降低線上評分時的評分差距，提出四點建議：即時向評分者反饋詳細數據、嚴格管理工作要求、加强培訓以及減少誤差參數。鄭丹丹等（2011）建議採用網路雙評機制，同時要控制每位評分者的速度及每日的評閱數量，通過背對背雙評，可以很容易地發現評分誤差，同時通過設置一些分析模式還可以發現評分者自身評分行爲的不一致，從而控制評分誤差。黃燕（2007）針對評分者對線上評分的態度、感受以及對閱卷系統的看法進行調查，結果顯示：線上評分使注意力更爲集中，有利於評分品質和效率的提高，而其不足之處在於系統提供的數據會影響打分；就系統本身來說，在作文詞數統計功能和參照卷的設置方面還需改進；線上評分對身體有傷害，尤其是眼睛容易疲勞。

目前已有寫作測驗單位關注線上評分系統的研發。例如：美國教育測驗中心（ETS）開發之線上評分網路系統（Online Scoring Network）；劍橋大學考試委員會（UCLES）於 2001 年之前即調查線上評分的可行性及其對於評分過程的品質和時間的影響；國中基本學力測驗推動工作委員會（BCTEST）寫作測驗於 2006 年首次採取線上評分方式；中國大陸英語專業考試四級的評分於 2009 年 5 月 12 日首次啓用電腦輔助人工閱卷（陸遠，2010）等。

綜上所述，可發現寫作測驗使用線上評分漸成趨勢，其即時收集數據與立即回饋之功能可彌補傳統紙本評分方式之不足，能有效提高評分一致性。然評分者須在電腦螢幕前長時間工作，易影響視力，若使用設計良好的操作介面，則可減輕視覺負擔。

3. 研究方法

3.1 研究對象

探討線上評分系統運用成效的研究對象爲參加本會於 2012 年 2 月舉辦之基礎級寫作測驗評分培訓工作的 4 名評分者；探討線上評分系統使用狀況調查結果的研究對象爲曾參加本會評分培訓的 10 名評分者（包含前述之 4 名評分者），他們皆具 10 至 20 年華語教學經驗。

3.2 研究工具

研究工具包括「線上評分系統」與「線上評分系統使用調查問卷」，分述如下。

3.2.1 線上評分系統

本會的「線上評分系統」，能即時儲存評分記錄，還具備顯示閱卷數量、評分是否一致、已列印與否及轉成 excel 檔等功能；平時培訓評分者時，亦可經由後端的監控，瞭解其評分一致性及嚴格度。本會爲提高評分系統的可操作性，將評分細項表格化，並稱之爲「評分表單」，置於評分者端操作介面的右邊，評分者須將各細項分數、向度分數、整體分數輸入表單之中；左邊爲考生的寫作文本，文本上方顯示錯誤標註工具列，評分者依會內的規定標示出錯誤之處，例如：「語序錯誤」以底線標註、「詞語錯誤」以灰底標註……等，研發人員可藉此瞭解評分者對於文本錯誤的分類概念；下方有考題和樣卷的連結，可供評分者瀏覽。評分者端評分操作介面見圖 1。

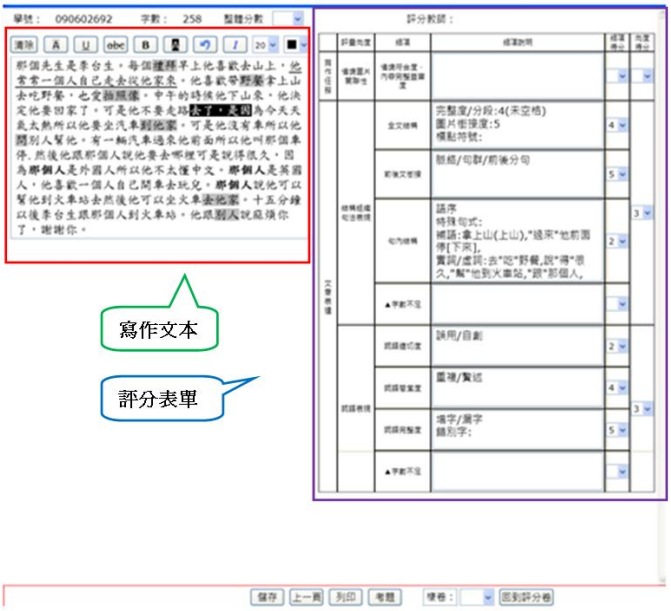

圖 1：評分者端評分操作介面

本系統可用於正式評閱及培訓階段。在正式評閱時，研發人員於評分者登入系統後，即透過後端的監控介面掌握其評分情況。當評分者批閱完一份卷子時，

其成績即呈現於後端畫面，供研發人員比對、觀察，如發生偏離的情況，則將偏離者之文本標註和表單頁面列印出來，查出偏離的原因，而後與之討論、溝通，俟其概念釐清後，再繼續評閱；在培訓階段，本系統亦可依培訓目的彈性選擇需要的功能加以運用。圖 2 爲本次評分培訓分項成績的監控介面，上方的欄位由左而右分別爲「情境圖片關聯性」[2]、「結構組織句法表現」、「詞語表現」和「整體」（總成績）；左邊的欄位顯示考生文本編號。

圖 2：後端監控介面

3.2.2 線上評分系統使用調查問卷

本研究亦調查使用者對評分系統的看法，調查對象總計 10 名。問卷內容爲對於整體操作介面、錯誤標註工具列、評分表單、考題連結和樣卷連結的滿意度。

3.3 研究步驟

由研發人員先向 4 位評分者說明本會的評分方式與評分標準，並介紹線上評分系統操作方式，接著請評分者依據評分規定，僅針對可量化的「結構組織句法表現」與「詞語表現」兩個向度評閱 20 篇文本，不須評閱「情境圖片關聯性」，亦不須送出整體成績。本次培訓之目的在於檢視在本會說明相關評分規定之後，評分者對於說明內容的掌握，雖然研發人員在評分過程中可從系統後端瞭解其不一致情況，但此時不中斷其評閱。研習結束後將評閱結果匯出，並以電子郵件將線上評分系統使用調查問卷寄給評分者，請他們填妥寄回，最後由統計同仁進行分析。

4. 評分結果分析

本研究透過皮爾森積差相關分析（pearson correlation）與百分比一致性（percentage agreement），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），以瞭解評分者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

[2] 此次培訓使用基礎級看圖寫作之考生文本，故評分系統將「情境任務的符合度與充實度」設定爲「情境圖片關聯性」。

作測驗評分品質，分析結果說明如下。

由評分者與研發人員各向度給分之皮爾森積差相關分析結果，可得知評分者與會內評分標準的關聯性，而進一步透過「結構向度」與「詞語向度」兩個向度的給分結果，可瞭解評分者可能在哪一個向度需要進一步溝通，以釐清概念。如下表 1 所示，評分者 B 在「詞語向度」上與研發人員給分之間的相關值較高，爲所有評分者中表現較爲理想的；而其餘達顯著相關之值，皆約爲 0.550；但評分者 A、C 在「結構向度」與評分者 C、D 在「詞語向度」上的相關值皆未達到顯著水準，以下進一步以百分比一致性觀察之。

表1：評分者與研發人員各向度給分之pearson相關

評分者	A	B	C	D
篇數	20	20	20	20
結構向度	.411	.580**	.359	.554*
詞語向度	.566**	.867**	.343	.426

*p<.05；**p<.01

由下表 2 兩個向度的百分比一致性可得知，在「結構向度」上，評分者與研發人員級分完全相同者（P_0）僅在 55%左右；相差一級分且皆評爲通過或未通過的百分比（P_1）爲 42%左右；P_{0+1} 皆在 95%以上。在「詞語向度」上，評分者與研發人員級分完全相同者皆在 75%以上，包含相差一級分以內，且皆評爲通過或未通過的百分比（P_{0+1}）皆在 95%以上。

整體而言，「結構向度」與「詞語向度」評分的一致性（P_{0+1}）都在 95%以上，但在「結構向度」上，評分者與研發人員級分完全相同者，明顯低於「詞語向度」。其中造成與研發人員相關低或是無顯著相關的原因在於 P_1 部分的百分比較高。

表 2：評分者與研發人員各向度評分者一致性

分析項目	評分組合	篇數	P_0	P_1	P_{0+1}
結構向度	A & 會內	20	11(55.0%)	8(40.0%)	19(95.0%)
	B & 會內	20	10(50.0%)	10(50.0%)	20(100.0%)
	C & 會內	20	12(60.0%)	7(35.0%)	19(95.0%)
	D & 會內	20	10(50.0%)	9(45.0%)	19(95.0%)
詞語向度	A & 會內	20	15(75.0%)	4(20.0%)	19(95.0%)
	B & 會內	20	18(90.0%)	2(10.0%)	20(100.0%)
	C & 會內	20	15(75.0%)	4(20.0%)	19(95.0%)
	D & 會內	20	16(80.0%)	3(15.0%)	19(95.0%)

透過多面向 Rasch 測量模式，可進一步瞭解評分者與研發人員給分的嚴格度差異，以及評分者自身給分的一致性。從表 3 可知，在結構向度分數的評分者嚴格度方面，與研發人員相較，較爲接近的是評分者 C 和 D；評分者 B 給分較爲寬鬆，評分者 A 給分較爲嚴格。評分者內的穩定性方面，評分者 A、B、D 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），顯示評分者 C 較不能維持自身給分的一致性。至於詞語向度分數的評分者嚴格度方面，並未達顯著水準（sig=0.57），顯示評分者與研發人員給分的嚴格度並沒有差異。

表 3：結構向度分數的評分者嚴格度

評分者	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
A	3.5	3.45	0.99	0.98	0.44	0.93	0.92
C	3.7	3.67	0.21	0.20	0.44	1.61	1.63
華	**3.7**	**3.72**	**0.01**	**—**	**0.45**	**0.47**	**0.44**
D	3.7	3.72	0.01	0	0.45	0.95	0.94
B	4.0	4.02	-1.22	-1.23	0.46	1.11	1.06

RMSE 0.45 Adj S.D. 0.66 Separation 1.46 Reliability 0.68

Fixed (all same) chi-sqaure: 12.1 d.f.: 4 sig: 0.02

由上述分析得知，在「詞語向度」上，評分者與研發人員較爲一致，故以下僅就「結構向度」加以討論。

表 4 是評分者與研發人員在「全文結構」、「前後文銜接」、「句內結構」三個細項的百分比一致性。整體來看，可以發現 4 位評分者在「全文結構」與「前後文銜接」的一致性（P_{0+1}）較高，皆在 80%以上；而在「句內結構」上的一致性（P_{0+1}）僅有 65%～80%，顯示「句內結構」的一致性較爲不佳。

若以評分者來看，評分者 A、B、C 在 P_1 的比例都差不多，但評分者 D 的 P_1 比例在「全文結構」、「前後文銜接」、「句內結構」三個細項皆比 A、B、C 高，顯示評分者 D 的標準有改進之空間。

表 4：評分者與研發人員在結構向度內之評分者一致性

分析項目	評分組合	篇數	P_0	P_1	P_{0+1}
全文結構	A & 會內	20	12(60.0%)	7(35.0%)	19(95.0%)
	B & 會內	20	11(55.0%)	6(30.0%)	17(85.0%)
	C & 會內	20	9(45.0%)	7(35.0%)	16(80.0%)
	D & 會內	20	12(60.0%)	8(40.0%)	20(100.0%)
前後文銜接	A & 會內	20	15(75.0%)	3(15.0%)	18(90.0%)
	B & 會內	20	14(70.0%)	4(20.0%)	18(90.0%)
	C & 會內	20	14(70.0%)	4(20.0%)	18(90.0%)
	D & 會內	20	6(30.0%)	10(50.0%)	16(80.0%)
句內結構	A & 會內	20	7(35.0%)	6(30.0%)	13(65.0%)
	B & 會內	20	10(50.0%)	5(25.0%)	15(75.0%)
	C & 會內	20	9(45.0%)	4(20.0%)	13(65.0%)
	D & 會內	20	5(25.0%)	11(55.0%)	16(80.0%)

從表 5 可知，在結構向度細項分數的評分者嚴格度方面，與研發人員相較，較爲接近的是評分者 A、C 和 D，而評分者 B 給分較爲寬鬆一點。評分者內的穩定性方面，幾乎所有評分者均符合 INFIT 值大於 0.7 或小於 1.3 的標準，顯示評分者能維持自身的標準。

表 5：結構向度細項分數的評分者嚴格度

評分者	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
C	3.8	3.77	0.34	0.31	0.18	0.85	0.81
A	3.8	3.87	0.17	0.14	0.19	0.93	0.87
D	3.9	3.89	0.14	0.11	0.19	1.34	1.25
華	**3.9**	**3.95**	**0.03**	**—**	**0.19**	**0.95**	**0.90**
B	4.2	4.32	-0.67	-0.7	0.21	0.95	0.80

RMSE 0.19 Adj S.D. 0.34 Separation 1.80 Reliability 0.76

Fixed (all same) chi-sqaure: 15.2 d.f.: 4 sig: 0.00

5. 問卷分析

本問卷分爲整體介面、錯誤標註工具列、評分表單、考題連結、樣卷連結五個調查面向，除錯誤標註工具列含 4 個子題之外，其餘皆爲 2 個子題。本問卷採用 Likert 五點量表，選項分爲等級 1-5，從「非常不滿意」到「非常滿意」，分數之敘述依題目略有差異，但不影響計分之方向與强度。

依各調查面向而言，在整體介面與評分表單面向上，同樣有 80%的評分者

表示滿意，20%認爲普通；在錯誤標註工具列面向上，則所有評分者皆滿意；另外，有 90%的評分者對考題連結面向感到滿意，10%認爲普通；而在樣卷連結上，有 70%的評分者滿意，30%覺得普通。詳細人數如下表 6 所示。

表 6：五大調查面向之滿意度

評分面向 / 整體得分	整體介面	錯誤標註工具列	評分表單	考題連結	樣卷連結
3.00	1(10.0%)	0(0.0%)	0(0.0%)	0(0.0%)	1(10.0%)
3.50	1(10.0%)	0(0.0%)	1(10.0%)	1(10.0%)	2(20.0%)
3.75	—	—	1(10.0%)	—	—
4.00	1(10.0%)	3(30.0%)	3(30.0%)	1(10.0%)	2(20.0%)
4.25	—	—	2(20.0%)	—	—
4.50	5(50.0%)	0(0.0%)	0(0.0%)	3(30.0%)	0(0.0%)
4.75	—	—	1(10.0%)	—	—
5.00	2(20.0%)	7(70.0%)	2(20.0%)	5(50.0%)	5(50.0%)

*數字表示人數，括號內表示百分比，”—“表示無該分數出現可能

針對 12 題子題做次數分配後，發現評分者對於本系統的各面向皆給予極大的肯定，僅在 4 個子題中有評分者表示不滿意，其數據與意見分別如表 7、表 8 所示。

表 7：評分者對系統不滿意之試題報表

題號 / 得分	第 2 題	第 7 題	第 10 題	第 12 題
1	1(10.0%)	0(0.0%)	0(0.0%)	0(0.0%)
2	0(0.0%)	2(20.0%)	1(10.0%)	1(10.0%)
3	1(10.0%)	1(10.0%)	0(0.0%)	2(20.0%)
4	6(60.0%)	4(40.0%)	3(30.0%)	2(20.0%)
5	2(20.0%)	3(30.0%)	6(60.0%)	5(50.0%)

*數字表示人數，括號內表示百分比

表 8：評分者對系統之意見

題號	題目	評分者意見
2	整體介面的背景色調與亮度，是否讓您感覺舒適？	背景亮度偏高，影響視力。
7	在評量「結構組織句法表現」時，由大結構至小結構的評閱順序是否符合您的習慣?	習慣從「句內結構」開始評閱。
10	考題介面（如：排版、字體大小……）您覺得如何？	希望背景色再柔和些、字體再大一些。
12	樣卷介面（如：排版、字體大小……）您覺得如何？	希望背景色再柔和些、字體再大一些。

6. 結論與建議

根據上述分析結果可得知，在評分者一致性方面，「詞語向度」在 P_0 的比例明顯優於「結構向度」。在評分者嚴格度方面，評分者 B 在「結構向度」的嚴格度較研發人員略微寬鬆；在「詞語向度」方面，所有評分者的嚴格度與研發人員並無顯著差異，故未來培訓評分教師時，可針對評分結果分歧較大的「結構組織句法表現」向度進一步與評分者溝通。而在評分者內的穩定性方面，評分者 C 穩定性不佳，其 INFIT 值略大於標準值，顯示其評分者內的一致性略低，在給分時無法維持自身的標準，可提醒該評分者注意評分寬嚴的穩定度。

至於寫作線上評分系統的使用滿意度問卷調查結果，各評分者大致皆滿意，顯示此系統雖已趨於完備，但仍有改善的空間，因此未來將在評分者端的介面上，調整螢幕色調和考題、樣卷介面的字體，以提高評分者之視覺舒適度；錯誤標註工具列的順序也將略作調整，以加強評分者由大結構評閱至小結構的概念。此外，在後方監控端上，由於此系統原本是針對正式評分所設置，僅能檢視三個向度的分數和總成績，但爲因應評分培訓工作的需求，未來將加上各細項的評分欄，如此研發人員將更能有效掌握評分者的評分思維。

參考文獻

高丙成、秦旭芳.(2007).成人高考網上閱卷的評分員差異研究.烏魯木齊職業大學學報 2007 年第四期:96-99.

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

黃燕.(2007).大學英語四、六級考試網上閱卷情況調查.外語界 2007 年第 2 期:82-96.

陸遠.(2010).網閱環境下的英語專業四級考試作文評分員偏頗研究.上海外國語大學博士論文.

鄭丹丹、陳睿、張開、趙靜宇. (2011). 兩種評分量表的評分效應比較研究. 教育研究與實驗.2011 年第 4 期.

鄒慧英.(2003).測驗與評量-在教學上的應用.臺北洪葉文化事業有限公司.

ETS 線上評分網路系統. 參見 http://www.toeic.com.tw/sw/about_sw.jsp

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. (1967). An Investigation of the Reliability

of Five Procedures for Grading English Themes. *Research in the teaching of English.*

Shaw, S. D. (2001). Issues in the assessment of second language writing. Cambridge ESOL: *Research Notes, Issue 6:2-5*

Shaw, S. D. (2007). Modelling facets of the assessment of Writing within an ESM environment. Cambridge ESOL: *Research Notes, Issue 27:14-19*