編譯論叢
第一卷 第一期（2008 年 9 月），71-92

四種翻譯評量工具的比較

賴慈芸

本研究比較下列四種翻譯評量工具的評分結果：根據國立編譯館「建立國家翻譯人才評鑑標準第二期研究」所提出的量表評分方法（「忠實」/「通順」各五分量表，獨立評分）、錯誤扣分法，以及兩種修正的量表評分方法。第一種修正方法是比例不變，但改爲合併評分（「正確」/「表達」各五分量表，合併評分）；第二種修正方法是加重訊息正確的比例（「訊息準確」六分，「表達風格」四分，合併評分）。研究者從前述研究中抽取 30 份答卷作爲評分樣本，共有 12 位翻譯教師 / 專業譯者參與評分。研究結果發現，在英譯中組的部分，修正後的兩種量表評分法都與錯誤扣分法達到高度相關，但第二期研究的量表評分法與錯誤扣分法只有中度相關，表示合併評分的三種方法較爲一致；其中「六 / 四評分法」的評分人間信度最高，與錯誤評分法的相關度也最高，可知爲穩定而有效的工具。在中譯英組部分，「六 / 四評分法」的評分人間信度也是最高，但四種評分法的結果都達到高度相關，差異不大。

關鍵詞：翻譯測驗、翻譯評量、評量工具、量表評分、錯誤扣分法

收件：2008 年 3 月 31 日；修改：2008 年 6 月 30 日；接受：2008 年 6 月 30 日

賴慈芸，國立臺灣師範大學翻譯研究所助理教授，E-mail : tysharon@ms29.hinet.net。

Compilation and Translation Review
Vol. 1, No. 1 (September 2008), 71-92

A Comparison of Four Assessment Tools for Translation Tests

Tzu-yun Lai

This study compares four assessments used in translation tests: a scale-based method proposed by Liu Minhua et al in "A Study on the Establishment of National Assessment Criteria of Translator and Interpreters, Phase II" (2005), the error-analysis-based method applied by most schools and institutions, and two modified assessments based on Liu's method. In the present study, twelve graders were invited to re-grade 30 papers in Liu's experiment by the other three methods. The result of the English-Chinese group showed that the two modified scale methods both reached a high correlation with the error-analysis method while Liu's scales only reached a medium correlation. The inter-rater correlation of the 6/4 scale (6 grades for "Accuracy" and 4 grades for "Expression") was the highest among all the methods used in the research. The correlation between the 6/4 scale and error-analysis method was also the highest. It showed the 6/4 scale method was a reliable and valid assessment tool. In the Chinese-English group, however, the results of the four methods were similar, although the inter-rater correlation of the 6/4 scale was still the highest among the four.

Keywords: translation test, translation assessment, assessment tools, assessment scales, error-analysis

Received: March 31, 2008; Revised: June 30, 2008; Accepted: June 30, 2008

Tzu-yun Lai, Assistant Professor, GITI, National Taiwan Normal University, E-mail: tysharon@ms29.hinet.net

壹、導言

語言測驗已有可觀的研究，但是在翻譯測驗領域，研究的數量遠遠不如語言測驗；目前可見的文獻或研究個別譯作的品質、探討業界翻譯品質管理（Translation Quality Assessment, TQA）、側重於翻譯教學領域之內的評量[1]，或以翻譯作爲評量語言能力的手段[2]，而較少針對翻譯測驗本身進行研究。雖然從 1970 年代開始，美國、加拿大、澳洲、英國都陸續開辦翻譯專業考試，但針對此種考試的研究卻仍然不多。Stansfield（1992）因此在 1987 年爲美國聯邦調查局設計翻譯考試時，抱怨幾無可用的翻譯測驗研究文獻。國內以翻譯作爲評量語言能力工具行之有年，翻譯系所內也有各項考試，但針對翻譯測驗的研究一樣不多見，尤其是實證研究。

本研究即爲翻譯測驗中評量工具的實證研究。本文主要比較量表（級分制）與錯誤扣分法；但量表又分爲 3 種，包括劉敏華、張武昌、林世華、陳碧珠、葉舒白、駱香潔（2005）提出的五 / 五量表（「忠實」五級分 /「通順」五級分，兩組獨立評分）、修正後的五 / 五量表（「正確」五級分 /「表達」五級分，兩組合併評分）、六 / 四量表（「訊息準確」六級分 /「表達風格」四級分），因此一共有 4 種評量結果。研究目的爲改進現有的評量工具，讓翻譯測驗更爲可靠和有效。

貳、文獻回顧

誠如爲加拿大翻譯局設計評量標準的 Williams（1989）所說，翻譯評量工具必須是 一、可靠的，即評量者的判斷是一致的，判準是穩定的，不能有時過於嚴格，有時過於寬鬆；二、有效的，即是否足以判斷樣本的品質等級、優點和缺點。第一項要求，只要同一評分人的兩次評分結果相差不遠，或不同評分人間信度達到高度相關，即可證明；但第二項要求

卻較難證明。

Stansfield（1992）是較早的翻譯測驗實證研究之一。該研究係針對美國聯邦調查局委託研發的 SEVTE（Spanish into English Verbatim Translation Exam 西班牙語譯英語翻譯測驗）進行效度研究。這個測驗分兩部分來評估譯者的能力，也就是「正確」（Accuracy）和「表達」（Expression）。參與測試的有 58 人，全數都是美國聯邦調查局的雇員。測驗分成兩部分，一部分是選擇題，另一部分爲翻譯。翻譯的部分又分爲單字片語翻譯、單句翻譯和段落翻譯。單字片語和單句翻譯都只評估「正確」能力，段落翻譯加上「表達」能力的評估。研究者利用受試者的自評問卷、其他西班牙語（聽力、閱讀）和英語（口語）等語言考試資料，研究「正確」、「表達」兩項指標的效度。結論是，「正確」與所有其他變數都高度相關，但與「表達」的相關性較低，也比較不一致。也就是說，「正確」的效度得到證實，但似乎比較難以證明「表達」的效度。

Waddington（2001）也是針對翻譯測驗效度的實證研究，研究對象是 64 位西班牙大二學生，測驗西班牙文翻譯爲英文的能力，由 5 位評分人利用 4 種不同的方法評分。Waddington 用了 17 種外部指標，包括英語寫作成績、母語能力、其他翻譯課成績、教師排名、自評等等，想要證明錯誤扣分法比整體評分法（分爲 5 級）可靠；結合錯誤扣分法與整體評分法最能正確評量學生成績。結果發現這些外部指標與評分法間並沒有明確的關連性，倒是意外發現 4 種評分方法結果都很一致，達到高度相關，因此該研究無法證明哪一種評分法比較有效，只能說明這 4 種評分方法都是有效的。

劉敏華等人在 2005 年接受國立編譯館委託，爲教育部的翻譯人才能力考試（從 2007 年開辦）開發了一套翻譯評量工具，改良 Carroll（1966）設計的翻譯品質評量量表，以單句爲評分單位，由兩組評分人獨立評分，一組專評「通順」，一組專評「忠實」，各採用五分量表，1 分爲最低，5 分爲最高。劉敏華的研究團隊並進行了實證測試，結果在信度方面，各組

評分人的「忠實」得分大多達到高度相關，證實就「忠實」而言，單句的量表評分法是穩定而一致的；但「通順」則出現中度相關多於高度相關的情形，而且「『通順』之評分人信度普遍低於『忠實』，顯示判斷譯文是否通順，比較容易受到評分人主觀標準的影響」（劉敏華等，2005，頁114）。在效度方面，該研究以系別和年級分析，英譯中組的英語系四年級學生優於非英語系四年級，但非英語系四年級學生又優於翻譯系三年級；中譯英組的碩一學生優於大三學生，大三學生卻優於大四學生。由於這次參與學生來自不同學校，在沒有其他語言測試能力成績支持之下，系別和年級的鑑別力並不充分，因此該研究的效度並沒有令人滿意的分析。

參、研究問題

劉敏華等人所設計的評分工具雖然經過實證測試，但如果要在正式考試中採用，還必須回答下面兩個問題：第一，如何改善「通順」組的評分人間信度，使整個工具更爲穩定可靠；第二，如何證明這是有效的評量工具。

關於第一個問題，本研究改變評分方法，由同一位評分人評兩項分數，而不採兩組獨立評分。參與國立編譯館第二期研究的一些評分人反應[3]，單獨評「通順」很困難，即使一開始沒有對照原文，但改了幾份以後也能夠推知大致訊息爲何；若明知道內容有嚴重錯誤，仍根據評分原則，只要「句子的陳述清晰明白，在用詞、表達或句法上，沒有或極少不當之處」就給滿分，似乎並不符合翻譯專業要求，因此評分上會有猶疑及不一致的狀況。在世界各翻譯測試機構的評分方法中，即使有分開考量規準（如英國 IOL 筆譯證書），也從無由兩組評分人獨立評分的作法。因此我們改由一人評兩項分數，希望評分人能夠由整體表現判斷，一個未達滿分的譯句，究竟是「正確」還是「表達」上的失誤。

關於第二個問題，我們從 Waddington（2001）的研究得到啓發：如

果外部指標難以證明翻譯評量工具的效度，是否可以用不同的評量方法來證明？因此我們徵募了兩組評分人，一組以全文爲評分單位，根據錯誤扣分法評分，作爲對照組；另一組則採用劉敏華開發的工具，以單句爲評分單位，根據量表評分，但把兩組獨立評分改爲合併評分。

肆、研究設計

我們從國立編譯館第二期研究計畫之正式筆譯測試的試卷中，抽樣做爲本研究的評分樣本，並於 2007 年 7 月舉行評分工作坊，包括評分訓練、當場評分及座談會。在評分工作坊之後，研究團隊根據評分人的建議和討論，重新修訂量表（六 / 四量表），再次試評，不過這次採用的是通訊評分。最後再將第二期評分結果（五 / 五量表，兩組獨立評分）、工作坊的兩組評分結果（錯誤扣分法和五 / 五量表合併評分）、第二次評分（六 / 四量表）等 4 種評分結果一起比較分析。

以下說明樣本來源、評分人背景、第一次試評的兩種評分工具、以及第二次試評的評分工具。「第二期研究」有近兩百名受試者，分別來自臺灣北部及南部 3 所大學共 10 個班級，包含翻譯系三年級、英語系四年級、非英語系四年級，以及碩士班一年級。本研究的評分樣本挑選原則，係根據「第二期研究」的得分，由高分至低分分爲 6 等份，每份抽取 5 份試卷，中譯英組及英譯中組各 30 份。不過，由於中譯英組在第二期研究中無人及格，爲了讓各種分數能平均呈現，便於評分討論，本研究最後決定只從第二期研究抽取 26 份試卷，另外請了 4 位英文程度較佳的研究生製作了 4 份試卷加入評分。爲求反映考試真實情境，試卷樣本由電子檔請人抄錄爲手寫稿。爲了避免因爲評分順序可能造成的誤差，每組試評評分人的答卷順序都稍做調整。30 份試卷分爲 3 等份，前後順序錯開，使每個評分人所改的答卷順序都有所不同，以求客觀公平。

第二期研究中，英譯中組及中譯英組各考兩題。限於工作坊的評分

時間有限，本研究各採用其中一題評分，英譯中組採用的是 “Dwindling readership: Are tabloids the answer?”，英文字數共 242 字，測驗時間約爲 60 分鐘[4]；中譯英組則是「讓自己出頭天」，中文字數 402 字，測驗時間也是 60 分鐘（見附錄）。

評分人是從研究團隊提出的合格人選當中，徵求自願協助評分的筆譯專業人士及大專相關系所教師。參與的評分人總計有 16 人次，英譯中及中譯英組各包括錯誤分析倒扣 3 人、五／五量表評分 3 人、六／四量表評分 2 人。

本研究所採用的 3 種評分方法簡介如下：

一、錯誤扣分法

這是最多翻譯機構採用的評分法，也是最多翻譯教師使用的評分方法。以全篇譯文爲評分單位，就重大錯誤及次要錯誤逐一扣分，扣分處直接標示在試卷上。研究團隊參考輔大／臺師大聯合專業考的評分原則，編寫簡要的評分準則如下：

錯誤分析扣分法評分原則：以 80 分爲及格分數

(一) 曲解原意，扭曲訊息：扣 2 ～ 8 分（視影響範圍扣分）

(二) 用詞不當、語域、搭配、風格不佳：扣 1 分

錯別字 (拼字錯誤)、贅字、標點：扣 1 分

(三) 翻譯策略或風格：加 1 ～ 2 分

二、修訂過的五 / 五量表評分法

國立編譯館第二期研究以單句爲評分單位，採用「忠實／通順」兩項規準，由兩組評分人各根據五分量表給分。本研究先修訂評分方法，將「忠實／通順」改爲「正確／表達」[5]，援用各五分量表，逐句評分，但改爲由同一人評分。本研究並設計了評分表供單句評分法使用，讓評分單位一目瞭然，節省自行斷句的時間，也避免出現評分單位不一致的情況。

評分表將原文分句列印在表格上，每句有兩個評分格，一爲「正確」得分（1 到 5 分），一爲「表達」得分（1 到 5 分）。本次所用試題英譯中分爲 12 句，中譯英分爲 9 句。最後「正確」、「表達」總分再除以句數，可以得到兩項規準的平均成績。最後這個成績再換算成百分制，80 分爲及格。本研究所採用的五 / 五量表如下：

表 1　五 / 五量表

正確量表	
5 分	譯句所傳達的訊息，與原文完全相同，沒有誤譯、缺漏或添譯。
4 分	譯句所傳達的訊息，與原文非常接近，可能有一兩處次要的錯誤。
3 分	譯句所傳達的訊息，與原文相當不同，有一處重大的或多處次要的誤譯、漏譯或添譯。
2 分	譯句所傳達的訊息，與原文極爲不同，有多處重大的誤譯、缺漏或添譯。
1 分	譯句所傳達的訊息，與原文根本不同。
表達量表	
5 分	句子的陳述清晰明白，在用詞、表達或句法上，沒有或極少不當之處。
4 分	句子的陳述大致清晰明白，不過在用詞、表達或句法上，有少數不當之處。
3 分	句子的陳述還算清楚，但在用詞、表達或句法上，有許多不當之處。
2 分	句子的陳述大體上看不懂，在用詞、表達或句法上，有許多不當之處。
1 分	句子的陳述完全無法理解，全句的用詞、表達或句法嚴重不當。

資料來源：研究者自行整理。

五分量表中，設定 4 分以上爲及格成績，3 分以下爲不及格，評分前亦告知評分人。

三、六 / 四量表評分法

第一次評分測試後，經過評分人討論，建議將五 / 五量表修正爲六 / 四量表，也就是加重「正確」的比例。因爲在評分過程中，我們發現一些訊息完全錯誤，但表達沒有太大問題的譯句。若根據五 / 五量表，這樣的譯句應該拿到 6 分（正確 1 分 / 表達 5 分）；但大多數的評分人都認爲這樣的句子給 6 分過高，總分有可能會因而及格，但實際上卻是完全不能接受的譯文。例如英譯中第六個評分單位，原文爲 “But struggling publishers often seek the quickest method to cut costs and increase circulation without harming ad revenues.” 考生若譯成「但發行者發現，最快的方法來減少開支，便是在不影響年收入的範圍增加發行量。」由於訊息完全錯誤，「正確」只能拿到 1 分；但這句的中文陳述清晰明白，在用詞、表達或句法上，並沒有什麼不當之處，因此有可能拿到 4 分或 5 分。研究團隊與評分人都認爲，一般文件翻譯（不包含文學作品）工作中訊息準確的重要性高於文辭表達，這樣的句子不應該拿到總分 5 分或 6 分，甚至有些評分人建議「正確」不及格者，「表達」應不予計分。而且「表達」的評分人間信度較低，也表示這是兩項規準中較主觀的一項。因此我們修訂了量表，讓「正確」的區分增加爲 6 級，更爲敏銳，一方面也降低「表達」的區分，只分 4 級，讓評分人更易於判斷，以期提高評分人間信度。

另外，「正確 / 表達」兩項規準也依照評分人建議，「正確」似乎容易與「語言正確」相混淆，因此改稱「訊息準確 / 表達風格」，以期更準確呈現評分內涵。表 2 爲第二輪評分所用的六 / 四量表：

表 2 六 / 四量表

訊息準確量表	
6 分	譯文所傳達的訊息，與原文完全相同，沒有錯誤。
5 分	譯文所傳達的訊息，與原文大致相同，但有一處次要的錯誤。
4 分	譯文所傳達的訊息，與原文有所不同，有兩處以上的次要錯誤。
3 分	譯文所傳達的訊息，與原文相當不同，有一處重大錯誤，或三處以上次要錯誤。
2 分	譯文所傳達的訊息與原文極爲不同，有兩處以上重大錯誤，或只是堆砌字詞的字面解釋。
1 分	譯文所傳達的訊息，與原文根本不同，或完全缺譯。
表達風格量表	
4 分	陳述清晰明白，且用詞、語域、搭配、標點方面皆無不當。
3 分	陳述大致清晰明白，但用詞與表達方面有一兩處不當，或有錯別字、贅字等。
2 分	陳述勉強可以理解，但有句法上的錯誤，用詞與表達也有多處不當。
1 分	陳述不符合句法，難以理解，或完全缺譯。

資料來源：研究者自行整理。

六 / 四量表中，設定訊息準確 5 分以上爲及格成績，4 分以下爲不及格；表達風格 3 分以上爲及格成績，2 分以下爲不及格。評分前皆告知評分人。

伍、研究結果與分析

我們的研究假設是：

第一、修正後的評分方法，即由同一評分人同時評「訊息準確」與「表達風格」，而且比例改爲六 / 四，可以提高評分人間信度。

第二、修正後的評分方法與錯誤扣分法的結果相近，因此是有效的評分工具。

以下分爲英譯中組及中譯英組分別討論。

一、英譯中組

本研究英譯中組的 3 種評分方式，各項得分如表 3 所示，並與國立編譯館第二期研究的評分結果相對照。

表 3　英譯中組各項分數與及格人數

	最高分	最低分	全距	平均	標準差	及格人數
錯誤扣分	86.7	47.0	39.7	67.7	10.33	2/30
五 / 五量表	84.2	54.2	30.0	69.5	6.57	1/30
六 / 四量表	85.0	57.9	27.1	74.0	6.56	3/30
第二期研究	92.9	57.1	35.8	74.2	8.94	5/30

資料來源：研究者自行整理。

全篇譯文的平均得分，以錯誤扣分法爲最低（67.7），第二期研究最高（74.2），及格人數也以第二期研究爲最多。我們推測第二期研究的評分方法，由於「忠實」與「通順」由兩個評分人獨立評分，而且比重相等，可能是比較寬鬆的。也就是說，有些考生誤解甚多，若用錯誤扣分法一定是低分卷；但可能因爲文字通順流暢，「通順」得到相當高分，平均起來就拉高了成績。由於樣本數太少，閱卷人數也少，因此研究團隊雖懷疑第二期研究的評分方法及格率容易偏高[6]，但必須等更大規模的資料才能檢驗。

錯誤扣分法的全距最大，標準差也最大。本期研究無論是五 / 五量表或是六 / 四量表，全距和標準差都較小。我們認爲，錯誤扣分法由於評分人的扣分權限較大，比較容易出現高低分極端，符合一般學者教師對錯誤扣分法易流於主觀的批評（McAlester, 2000；劉敏華等，2005），但或許也可解釋爲最爲敏感。

表 4 英譯中組各評分人間相關

	錯誤扣分法	五五量表	六四量表
評分人 1 及評分人 2	.735**	.616**	.786**
評分人 1 及評分人 3	.602**	.617**	-
評分人 2 及評分人 3	.733**	.722**	-

** p < .01

資料來源：研究者自行整理。

以 Pearson 積差相關係數計算各組評分人之間的相關（見表 4），六／四量表組 2 位評分人達到高度相關（.786），錯誤倒扣及五／五量表組各 3 位評分人則爲中度或高度相關（.602 ～ .735）。對照第二期研究，「通順」組的評分人相關係數，從 .256 到 .811 都有（2005, p. 99），起伏甚大。因此可以證明六／四量表是較穩定的評分工具[7]。

表 5 英譯中組不同評分方法之相關

	錯誤扣分法	五五量表	六四量表	第二期研究
錯誤扣分法	1	-	-	-
五五量表	.846**	1	-	-
六四量表	.870**	.921**	1	-
第二期研究	.615**	.608**	.595**	1

** p < .01

資料來源：研究者自行整理。

以 Pearson 積差相關係數計算各評分法之相關（見表 5），發現本期研究所使用的 3 種評分方式，結果都達到高度相關（r>.8），可以證實這 3 種評分方法都有效。但這 3 種評分結果與第二期的評分結果都僅達到中度相關，表示由兩位評分人獨立評「訊息準確」與「表達風格」，結果與其他的評分方法差異較大。因此本研究建議，如果採用量表評分法，將兩組分數由同一位評分人評分，會比較符合多數評分人的預期。

表 6　英譯中組各項之相關

	錯誤扣分	五五量表	六四量表	第二期研究
總分及「訊息準確」	-	.958**	.984**	.886**
總分及「表達風格」	-	.891**	.856**	.836**
「訊息準確」及「表達風格」	-	.724**	.751**	.486**

** $p < .01$

資料來源：研究者自行整理。

在 3 種量表評分法當中，總分與「訊息準確」和「表達風格」都達到高度相關，且總分與「訊息準確」的相關性都高於「表達風格」，與第二期研究結果一致。表示在英譯中組，「訊息準確」與翻譯能力較爲相關，支持本研究將「訊息準確」分數比例提高的作法。本研究的兩組量表評分，「訊息準確」及「表達風格」達到高度相關（r>.7），第二期研究則爲中度相關（r<.5）。我們認爲「訊息準確」與「表達風格」的相關性是合理的，相關程度也接近 Stansfield（1992）的實驗：在該研究中，「正確」與「表達」的相關度是介於 .74 到 .75 之間。「訊息準確」與「表達風格」約略可代表「外語閱讀能力」和「母語寫作能力」；在筆者個人的教學經驗上（賴慈芸，2003），也曾發現英文閱讀理解能力不佳的學生，中文表達能力往往較弱。[8]

二、中譯英組

本研究的中譯英組同樣採用了 3 種評分方式，各項得分如表 7 所示。由於國編館第二期研究的中譯英組測試結果無人及格，本研究另外徵求了 4 名英文翻譯寫作能力較佳的參與者製作答卷，總計評分試卷爲 30 份，其中第二期研究僅占 26 份，因此第二期研究的得分情形只能做爲參考，無法與本研究 3 項評分法直接比較。

表 7 中譯英組各項分數與及格人數

	最大值	最小值	全距	平均數	標準差	及格人數
錯誤扣分法	94.0	50.0	44.0	68.4	10.48	5/30
五 / 五量表	97.8	46.3	51.5	65.7	13.75	5/30
六 / 四量表	96.1	41.1	55.0	65.1	14.46	4/30
第二期研究	81.1	45.0	36.1	62.1	9.91	0/26

資料來源：研究者自行整理。

全篇譯文的平均得分，以錯誤扣分法爲最高（68.4），兩項量表評分法相近（各爲 65.7 及 65.1）。第二期研究因爲少了 4 份較高分的試卷，因此平均分數略低（62.1）。

表 8 中譯英組各評分人間相關

	錯誤扣分法	五五量表	六四量表
評分人 1 及評分人 2	.710**	.831**	.832**
評分人 1 及評分人 3	.625**	.781**	-
評分人 2 及評分人 3	.832**	.819**	-

** $p < .01$

資料來源：研究者自行整理。

以 Pearson 積差相關係數計算各組評分人之間的相關（見表 8），六 / 四量表組 2 位評分人達到高度相關（$r = .832$），與英譯中組一致。錯誤扣分法的 3 位評分人相關差異較大[9]。但比起第二期研究，「通順」組的評分人間信度最低爲 .587，「忠實」組也有一組評分人信度只有 .563，本研究兩種量表的評分人間相關都比較穩定，證明都是穩定的評分工具。

表 9 中譯英組不同評分方法之相關

	錯誤扣分法	五五量表	六四量表	第二期研究
錯誤扣分法	1	-	-	-
五五量表	.959**	1	-	-
六四量表	.891**	.920**	1	-
第二期研究	.827**	.868**	.746**	1

** p < .01

資料來源：研究者自行整理。

以 Pearson 積差相關係數計算各評分法之相關（見表 9），發現本期研究中的 3 種評分方式，都達到高度相關（r>.8），可以證實這 3 種評分方法都有效。但與英譯中組不同的是，這 3 種評分結果與第二期的評分結果也達到高度相關，表示「訊息準確」和「表達風格」無論是分開評分或是一起評分，結果似乎都差不多。我們認爲，這可能與低分卷偏多有關。第二期研究參與試考的是大學生，全部以中文爲母語，因此中譯英都是譯入外語。Waddington（2001）的研究對象也是大學生，以西班牙文爲母語，考試內容是譯入英語，一樣是譯入外語；Waddington 的實驗中 4 種評分方法得到差不多的結果，我們的 4 種評分結果也差不多。

表 10 中譯英組各項之相關

	錯誤倒扣評分	五五量表評分	六四量表評分	第二期評分
總分及「訊息準確」	-	.970**	.990**	.920**
總分及「表達風格」	-	.983**	.980**	.936**
「訊息準確」及「表達風格」	-	.909**	.941**	.724**

** p < .01

資料來源：研究者自行整理。

在 3 種量表評分法中，總分與「訊息準確」和「表達風格」都達到高度相關（r>.9）。總分與「表達風格」的相關又都高於「訊息準確」。這點與英譯中組相反，應與譯入語非母語有關。「訊息準確」及「表達風格」兩項規準之間，達到高度相關（r>.9），遠高於第二期研究（r=.724），似乎反映出由 1 位評分人給兩項分數，容易交互影響。我們認爲，第二期研究中參與試考的都是中文母語的學生，中譯英表現不佳，分數偏低，沒有 1 個人及格。在這次評分時，我們雖然邀請了 4 位程度較佳的學生（包括雙語人士）加入試考，但整體還是低分卷偏多。根據評分人表示，許多句子完全不知所云，除了表達風格低分之外，連訊息是否準確也無從判斷，因此易造成兩項分數極爲相近的情況。所以如果正式考試中考生來源增加，包括英語母語人士、雙語人士、或已在職場工作的現職譯者，在整體表現較好的情況下，兩項分數應不致於難以分辨。

陸、結論

在評分人間信度方面，英譯中組和中譯英組一樣，相關度最高的都是六 / 四量表的評分人。證實由一位評分人給兩項分數的方法，可以提高信度。也證實六 / 四量表是穩定可靠的評分方法。

在效度方面，英譯中組和中譯英組結果不同。在英譯中組，我們的假設得到證實：六 / 四量表評分法，與錯誤扣分法達到高度相關（r=.870），與修正的五五量表也達到高度相關（r=.921）；可以證實是有效的評分法；反觀第二期研究使用的五 / 五量表法，與其他評分法都只有中度相關（r=.595 ～ .615），在效度上比較有疑慮。但在中譯英組的部分，我們卻發現所有評分法都達到高度相關，也就是說無論是否由同一評分人給分，結果都很接近。雖然也證實了六 / 四量表的效度，但的確與原先的假設是不同的。

但雖然我們證實，改變評分方法，可以提高評分人間信度，但也有

兩項規準混淆的疑慮。若使用六 / 四量表評分，無論是哪一組，「訊息準確」與「表達風格」都達到高度相關。究竟這是翻譯能力的本質（本來兩項能力就很密切），還是評分方法的設計問題（因合併評分而在某一項目有高估或低估的情況），還有待後續研究澄清。

另一個問題是，雖然兩組評分人都贊成提高「訊息準確」的比重，但就中譯英組的結果看來，與錯誤扣分法最爲接近的其實是五 / 五量表而非六 / 四量表。是否因爲此次考生絕大多數都是譯入外語，外語寫作能力才是決定分數高低的關鍵，因此「表達風格」只分 4 級則不易辨別高下？這問題也有待進一步實證研究。但翻譯專業考試的目的是篩選有能力進入翻譯市場的譯者或準譯者，而不是爲了教學需求。第二期研究無人及格，六 / 四量表的及格者則爲 4 人，應可判定就是研究團隊後來加入的 4 份高分卷；就篩選效果來說是很有效的方法。六 / 四量表無法細分不及格卷的高低，在教學上或許是個缺點，對大規模的翻譯能力考試來說卻未必是個缺點。因此本研究認爲，六 / 四量表是穩定可靠而有效的評分方法，可以在大型考試中採用。

最後一個問題是，如果錯誤扣分法和量表評分法一樣有效，爲何不直接採用錯誤扣分法就好？筆者認爲，使用錯誤扣分法，評分人對翻譯的主觀看法（哪一種錯誤算嚴重錯誤，哪一種錯誤較爲輕微）較易影響結果，適用於需要排名次，如翻譯比賽或入學考，或一個評分人即可完成的評量，如課堂成績或專業考。量表評分法雖無法完全排除主觀的成分，但不同評分人間要判斷一個句子是「訊息準確 5 分句或 4 分句」，比起「譯錯這個單字在全文中該扣幾分」容易達成共識；閱卷數量相當大時，也較不會因爲疲倦或閱卷時間相隔太久等因素而影響扣分輕重。錯誤扣分法還有長度的問題。如果不計長度，越長的文本當然越容易被扣分；因此像美國譯者協會 ATA 採用錯誤扣分法的機構都有長度基準（如 250 字原文，超過此長度或不足的要按比例計分）；單句量表則完全沒有長度的問題，不需因爲長度而調整扣分比例，未來若考試長度有所更動，

也較能保障歷次考試的公平性。因此量表似乎比錯誤扣分法更適用於大型的翻譯能力考試。

註釋

1. 關於翻譯教學領域內的評量研究，Sonia Colina（2003）的第五章有詳盡的文獻回顧可供參考。
2. 例如全民英檢寫作能力測驗中的「翻譯」項目就是此類。中國大陸的文獻多半屬於此類。
3. 根據參與國立編譯館第二期研究人員告知。筆者也是該研究的評分人之一。
4. 劉敏華等（2005）是用兩個小時時間考兩題，兩題長度差不多，因此一題的作答時間約為 60 分鐘。中譯英組亦同。
5. 改變用語的理由是，「忠實」和「通順」用語較為主觀，有較多解釋的空間。
6. 筆者曾以兩種方法試評翻譯所資格考的試卷，總共 14 份試卷，錯誤扣分法有 5 人不及格，但改用五 / 五量表評分法則全數及格，顯示五 / 五量表似乎比錯誤扣分法寬鬆。
7. 由於六 / 四量表組的兩位評分人皆是第二次評這 30 份試卷，或許可視為充分教育訓練的結果。
8. 母語能力是否會影響外語學習，是外語教學上很複雜的議題，也有許多研究。例如 Ganschow（1991）就發現，外語學習成績不佳的大學生，母語閱讀的成績偏低。不過在翻譯教學上的探討還不多。
9. 本研究第一輪評分的 12 位評分人中，英譯中組 6 人到齊，但中譯英組有兩位因故無法到場，後來都以錯誤扣分法通訊評分。沒有參與評分訓練或許是造成評分人間信度較低的原因之一。

感謝詞

本文爲國立編譯館與臺灣師範大學翻譯研究所合作的《建立國家中英文翻譯人才能力檢定考試「一般文件筆譯」評分機制之研究》（2007）的一部分。特此感謝國立編譯館、研究團隊及參與評分的老師和譯者。

參考文獻

劉敏華、張武昌、林世華、陳碧珠、葉舒白、駱香潔（2005）。**「建立國家翻譯人才評鑑基準之研究」期末報告**。臺北：國立編譯館。

賴慈芸（2003）。他們走了多遠？——大學部學生、翻譯所學生與專業譯者的翻譯表現比較。**第八屆口筆譯教學研討會**。臺北：國立臺灣師範大學。

Colina, S.(2003). *Translation teaching, from research to the classroom: A handbook for teachers.* Boston: McGraw-Hill.

Ganschow, L. (1991). Identifying native language difficulties among foreign language learners in college: A "foreign" language learning disability? *Journal of Learning Disabilities, 24*(9), 530-541.

McAlester, G. (2000). Translation into a foreign language. In C. Schäffner and B. Adab (Eds.), *Developing translation competence* (pp. 229-241). Amsterdam: John Benjamins.

Stansfield, C. W., Scott, M. L., & Kenyon, D. M. (1992). The measurement of translation ability. *The Modern Language Journal, 76*(iv), 455-467.

Waddington, C. (2001). Different methods of evaluating student translations: The question of validity. *Meta, XLVI,* 311-325.

Williams, M. (1989). Creating credibility out of chaos: The assessment of translation quality. *TTR, 2-2,* 13-33.

附錄 1：英譯中考題及六 / 四評分表

資料來源：原文摘自 Jacques R. Bughin & Henrik Poppe (2005). Dwindling readership: Are tabloids the answer? *Mckinsey Quarterly, January.*

評分單位	原　　文	訊息準確 1-2-3-4-5-6	表達風格 1-2-3-4
01	As consumers have increasingly turned to television and the Internet for news, the circulation of paid newspapers has declined—by 2 to 4% annually for more than a decade in most developed markets.		
02	The trend is set to continue, particularly as growing broad-band penetration encourages the wider use of online media.		
03	In addition, free commuter tabloids, available in many big European and US cities, have lured away some paying custom-ers.		
04	As a result, the revenues and profits of traditional newspapers are under intense pressure.		
05	Newspapers have fought back with free subscription trials and other promotions, with advertising platforms such as new or expanded feature sections, and with better home and newsstand distribution.		
06	But struggling publishers often seek the quickest method to cut costs and increase circulation without harming ad revenues.		
07	Many see their salvation in changing formats: they believe that switching to a more compact one, such as the tabloid format, may lift circulation by attracting disaffected newspaper readers, particularly teens and women.		
08	Higher circulation, in turn, stimulates demand for advertising, so newspapers can raise their ad rates.		
09	But changing the format of a newspaper carries big risks.		
10	Despite the potential for a quick uptick in circulation, churn among one profitable category of readers—subscribers—may rise because of their reluctance to accept the change.		
11	And most newspapers see an initial drop-off in advertising revenues when they make the change.		
12	These problems, as well as resistance by employees, can derail the process.		
Total			

附錄 2：中譯英考題及六 / 四評分表

資料來源：原文摘自施君蘭（2005）。讓自己出頭天。**天下雜誌，316 期**。

評分單位	原　　文	訊息準確 1-2-3-4-5-6	表達風格 1-2-3-4
01	國際化與全球化的潮流席捲而來，帶動產業快速變遷，也爲個人帶來新挑戰。		
02	管理大師 Peter Drucker 指出，知識將成爲生存與競爭的主要資源。		
03	無論產業或個人，不但要具備知識，還必須隨時更新，才能在競爭激烈的環境中克敵致勝。		
04	致勝的關鍵在「學習」，不斷學習、提升自我，已成爲知識經濟時代個人求生存、儲備競爭力的不二法門。		
05	提升專業能力是自我加值的首要任務，根據一家人力資源公司統計，語言能力、金融證照、MBA 學歷是 2004 年下半年前三大熱門的進修領域。		
06	學英語成了全民運動，因爲從外商到本土，從民營企業到公家機關，英語能力都是招募人才或升遷的必要條件。		
07	除了專業能力之外，面對工作或人生，積極的態度都是決勝負的標準。		
08	正面的處世態度是面對現實、堅持理想的關鍵，甚至比背景、教育、環境重要，也比外表、天分或技能重要。		
09	資深經理人就指出，近十年來企業對專業能力的要求一直更新，但是對態度的要求跟十年前一樣，要熱情敬業、全力以赴，不管年紀老或輕，都要肯學習。		
Total			