Establishing test form comparability: a case study of an English-Chinese translation test

Rachel Yi-fen Wu

LTTC® 財團法人語言訓練測驗中心

THE LANGUAGE TRAINING & TESTING CENTER

Background of the study

- This presentation will report the results of an empirical study which investigated test form comparability in the context of English-Chinese translation exams developed by Taiwan's Ministry of Education.
- Comparability of translation tasks is traditionally monitored based on expert judgment through a holistic interpretation of source texts and translation texts.
- When translation tasks are developed for multiple versions of a test, it is important to incorporate systematic procedures to define what is being tested and to determine whether the tasks are parallel in terms of difficulty level.

Focus of the study

- The focus of this study is to examine the characteristics attributable to task difficulty and to identify ‘test points’ in order to help stablise the degree of difficulty of translation tasks across test versions.

Test points in this study

- Translation is defined as ‘the replacement of a source text by a semantically and pragmatically equivalent target text’ (House, 2014: 254).
- In the process of producing a translation, the translator encounters problems of various sorts and uses a set of strategies to solve them (Levý, 1967; Reiss, 2000).
- ‘Test points’ refer to problems translators encounters.

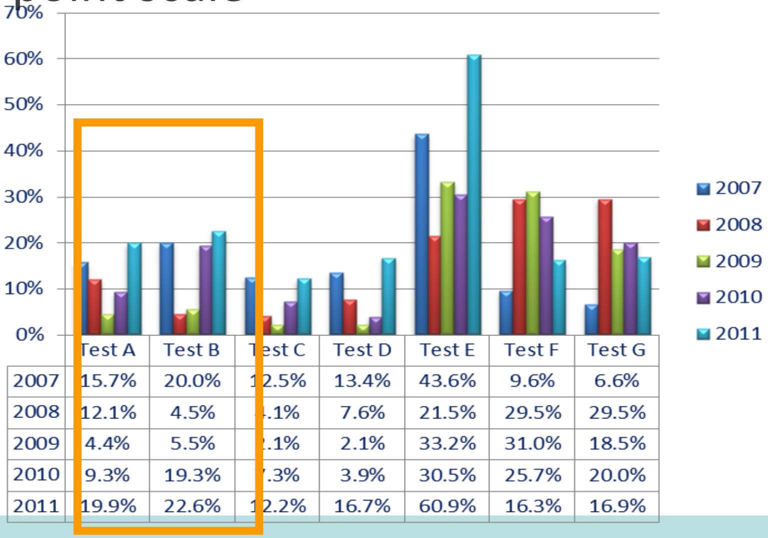
Translation problems

- **not limited to the word or sentence level**
  - neologisms, metaphors, passage titles, and dialects (Newmark 2005)
  - grammatical equivalence, textual equivalence, pragmatic equivalence, and codes of ethics (Baker 1992)
  - relationship between text types, language dimension, and translation methods (Reiss 1977/1989; Snell-Hornby 1988, 1995)
  - linguistic problems, textual problems, extralinguistic problems, problems of intentionality, and problems relating to the translation brief and/or the target-text reader (PACTE 2011)

Context of the study

- ▶ **MoE Translation and Interpretation exams:** consist of 7 tests
- ▶ **Target examinees**: level of English at CEFR C1 level or above
- ▶ **Structure of the EC translation tests**: one text of 250 words
- ▶ **Time allocation**: 60 minutes
- ▶ **Scoring criteria**:
  - Accuracy of information (message): 60%, using a 6-point scale
  - Expression (delivery): 40%, using a 4-point scale

✓**Certificates of English-to-Chinese Translation**: those who pass both Tests A and B within 3 years



	Test A	Test B	Test C	Test D	Test E	Test F	Test G
2007	15.7%	20.0%	12.5%	13.4%	43.6%	9.6%	6.6%
2008	12.1%	4.5%	4.1%	7.6%	21.5%	29.5%	29.5%
2009	4.4%	5.5%	2.1%	2.1%	33.2%	31.0%	18.5%
2010	9.3%	19.3%	7.3%	3.9%	30.5%	25.7%	20.0%
2011	19.9%	22.6%	12.2%	16.7%	60.9%	16.3%	16.9%

2007-2011 MoE English-Chinese translation test specifications

- **Topics to choose from**:

Test A	Commerce, Finance, Education, Cultural Affairs, etc.
Test B	Popular Science, Healthcare, Information Technology, etc.

- **Texts to avoid**:

  - texts with strong religious, political, or moral ideology
  - technical leaflets or research papers
  - texts that include many oral features

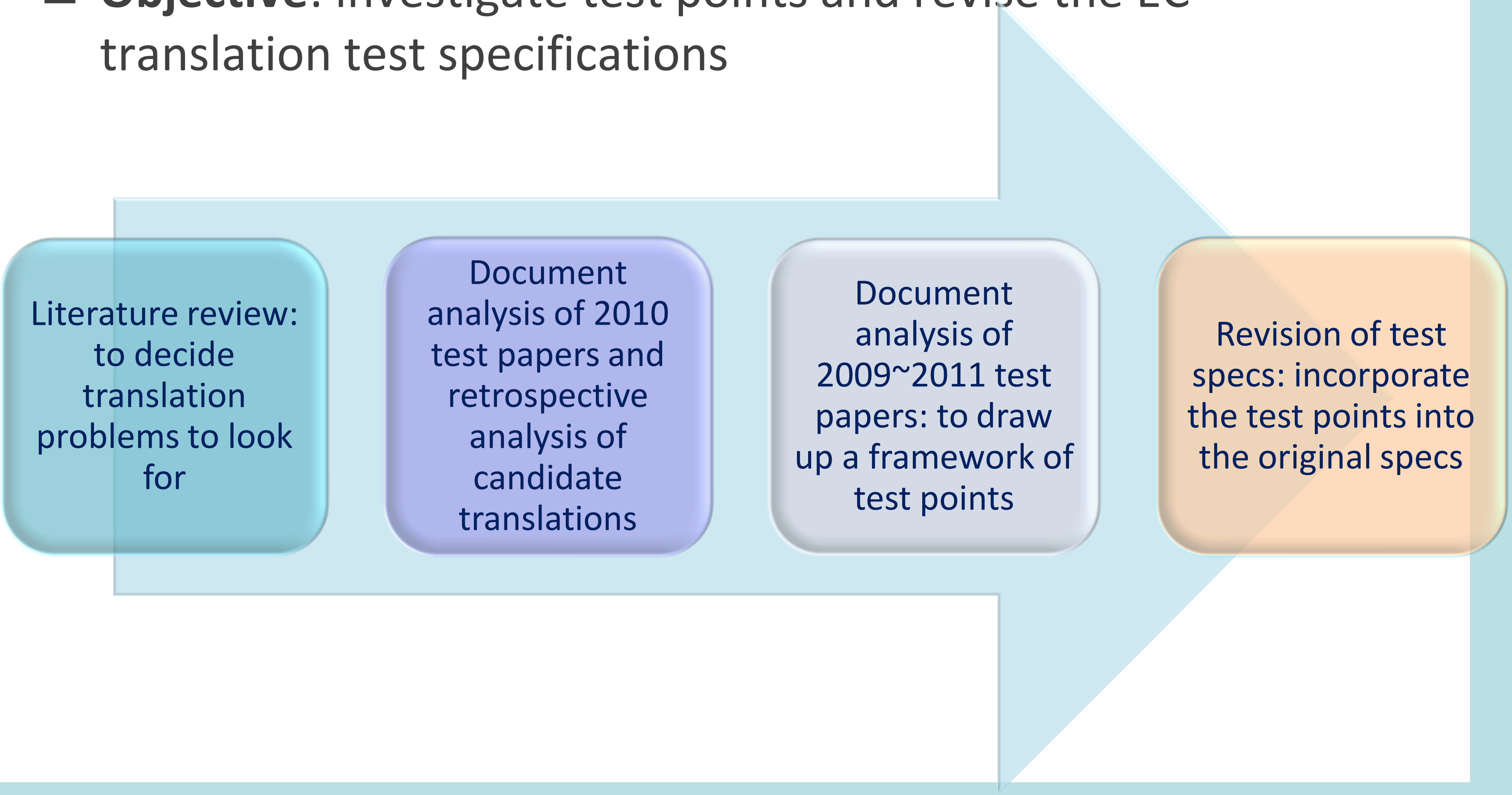
- **Specialized background knowledge**: not required
- **Organization**: logical, coherent, with introduction and conclusion
- **Reading difficulty**: can be understood by the educated general reader with language proficiency at CEFR C1 or above
- **Source material**:
  - published within the last 5 years
  - not translated from another language

Research design

- This study involved both the qualitative analysis of the test prompts of earlier versions of the examinations and candidates’ translations, and the quantitative examination of test performance, including classical descriptive analysis and MFRM (many-faceted Rasch model) analysis.

Phase 1 (from March 2011-July 2012): document analysis

- **Objective**: investigate test points and revise the EC translation test specifications



Literature review: to decide translation problems to look for

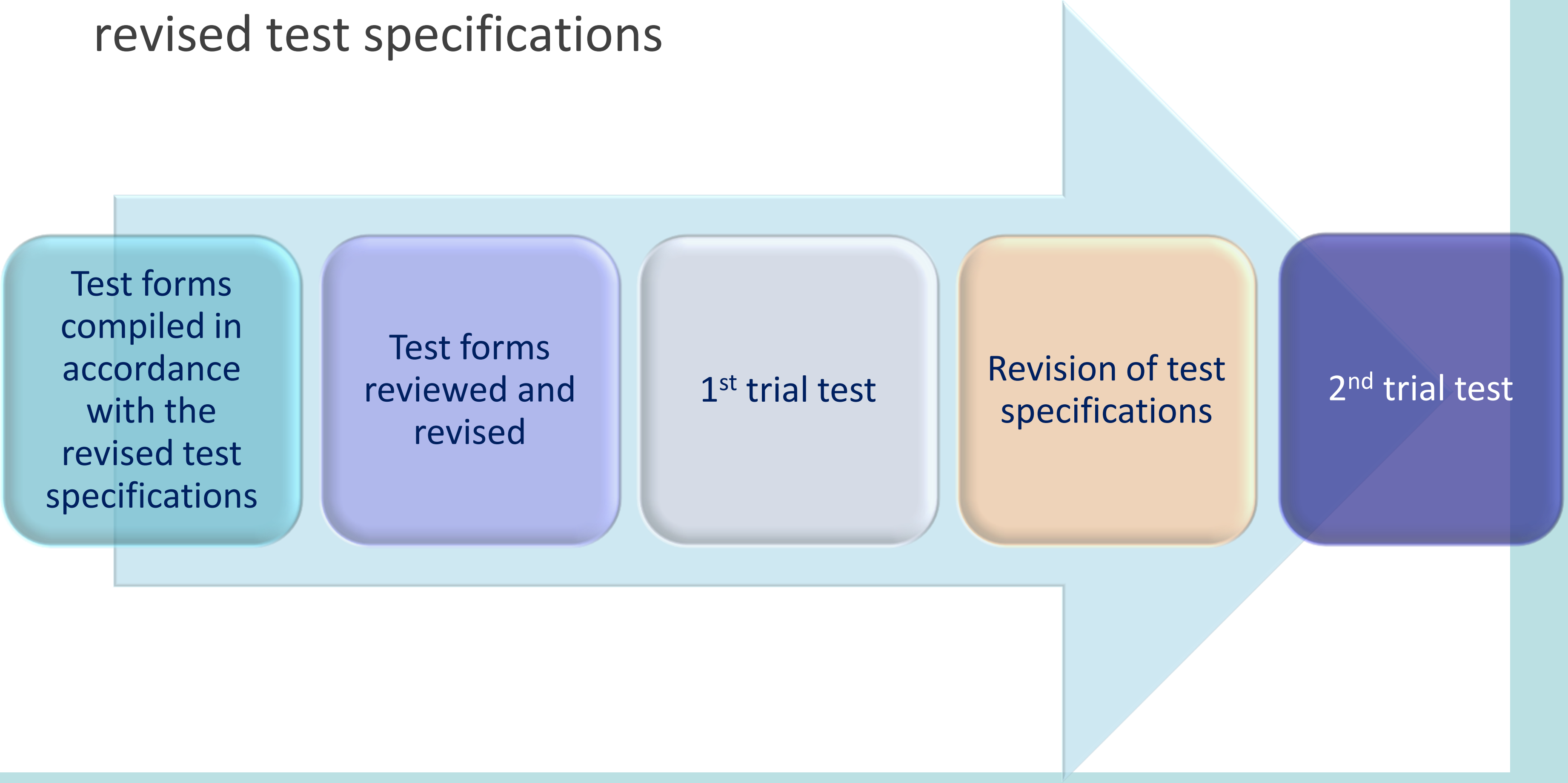
Document analysis of 2010 test papers and retrospective analysis of candidate translations

Document analysis of 2009~2011 test papers: to draw up a framework of test points

Revision of test specs: incorporate the test points into the original specs

Phase 2 (September 2012~March 2013): empirical validation

- **Objective**: validate the appropriateness and usefulness of the revised test specifications



Test forms compiled in accordance with the revised test specifications

Test forms reviewed and revised

1st trial test

Revision of test specifications

2nd trial test

Data Collection—Phase 1

- **Objective**: investigation of test points
- **Participants**: 2 researchers
- **Instruments**:
  - 2010 test papers
  - 9 examinee translations
    - 3 passing (score=8)
    - 3 borderline (score=7)
    - 3 failing (score=6)

Procedures to identify test points

- **Test point selection**:
  - Linguistic units that 2 thirds of or more candidates in the passing and borderline groups had problem translating were considered test points.

- **Translation problems**:
  - **Comprehension**: e.g. polysemy and complex sentence structures
  - **Reformulation:** e.g. polysemy, words that require change in word class when translated, and words that require background knowledge about the source-text culture

Test points identified

Vocabulary	1.1 Non-technical low frequency words
	1.2 Idioms and fixed expressions
	1.3 Polysemy
	1.4 Other words that require translation strategies
Sentence structure	2.1 Adverbial phrases
	2.2 Adjectival phrases
	2.3 Prepositional phrases
	2.4 Active/passive voice
	2.5 False subject
	2.6 Double negatives
	2.7 Other complex structures
Organization	3.1 Cohesion
	3.2 Coherence
Background knowledge	4 Culture and current events
Connotations	5.1 Rhetoric
	5.2 Author's stance and attitude

Vocabulary

- **1.3 Polysemy**
  - Meaning in context may be different from its definition in dictionaries
  - Example: No wonder some of the industry's biggest recent success **stories**, including Nintendo's Wii, have emphasized playability over mind-blowing graphics.→**產品(products)**

- **1.4 Other words that require translation strategies**
  - Example: Individually, the disclosures are trivial: some would be barely newsworthy if published legally. But collectively, they are corrosive. **America appears humiliatingly** unable to keep its own or other people's secrets.→**讓美國顏面無光 (It humiliated America)**

Test points for MoE English-to-Chinese translation Exams

Test point	Test purpose	Number
Vocabulary	Examine test-takers' vocabulary depth and breadth and their ability to reformulate ST lexis appropriately.	7-10
Sentence structure	Examine test-takers' understanding of and translation strategy for rendering complex sentence structures.	5-7
Organization	Examine test-takers' ability to use connectors appropriately to ensure cohesion and coherence of TT.	0-3
Background knowledge	Examine test-takers' understanding of and translation strategy for rendering culture-specific terms and terms related to current events.	1-3
Connotations	Examine test-takers' understanding of author's stance and attitude and test-takers' ability to render rhetorical devices appropriately.	
Total		18-20

Inter-coder reliability: Kappa = 0.91 ($p<.001$)

Phase 2:
1st Trial — Data Collection

- Participants
  - 6 candidates
    - Expected to pass: 3 [translation-major students who received certificates in English-to-Chinese translation]
    - Expected to fail: 3 [English-major graduates who did not receive substantial training in translation]
  - 3 raters
    - 3 university instructors who took part in the MOE translation exam rating in the previous years
    - Rater training before marking
    - Final marks determined by the average of 2 scores closest to each other

Phase 2:
1st Trial — Instruments

- Test forms
  - 1 set of test forms, including 1 Test A and 1 Test B
- Questionnaire for candidates
  - 6 questions on their background
  - 9 questions on the face validity, difficulty and administration of the translation exam

Phase 2:
1st Trial — Results

Sample candidates	Test A	Test B
Expected to pass (N=3)	2* passing	2* passing
Expected to fail (N=3)	3 failing	3 failing

Inter-rater reliability: 0.97 ($p<.01$) for both Tests A and B

*The higher-proficiency candidate that failed was the borderline test-taker whose performance might not be consistent across occasions.

Sample translations

Source text:

A second revelation: when we expect to be able to find information again later on, we don't remember it as well as when we think it might become unavailable.

Candidates' responses:

- **High-pass:** 另一項研究發現是：如果我們預期還能夠在之後再次取得資訊，則我們記憶的表現，將不如我們預想無法取得資訊的情況。

Another research finding is that when we expect to be able to find information again later on, our memory performs more poorly than when we think the information might become unavailable.

- **Pass:** 第二，當我們指望資訊能在稍後再被找到，我們就不會花心思努力記住。

Secondly, when we hope that the information can be found later on, we won't make efforts to remember it.

- **Fail:** 特點二：就算覺得之後一定找的到資料，後來可能就忘記方法，誤以為其實找不到資料。

Even if we expect to be able to find the information later on, we may forget how to locate it, and we may wrongly believe that we couldn't find the information.

Phase 2:
1st Trial — Survey results

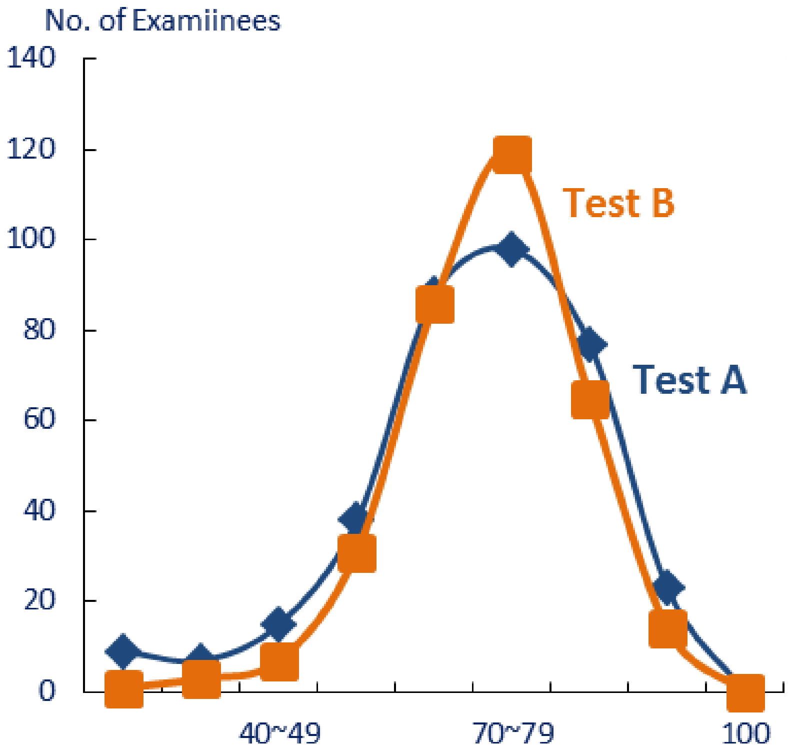
- **Can assess their competence in translation**: 100%
- **Perceived difficulty of the test**: 87% right level of difficulty to easy, 17% difficult (no clear relationship was observed between test takers' judgments of difficulty and their actual performance, which corresponds to the results of previous studies)
- **Comparable to school exams in terms of difficulty**: 100%
- **Comparable to projects done outside of school**: 75% easier (shorter in length, fewer technical terms included)

Phase 2:
2st Trial — Results

	Test A	Test B
Mean	69.37	72.30
SD	14.80	14.95
Max	86.67	91.11
Min	40.63	48.33
Candidates expected to pass (N=9)	67%	78%
Candidates expected to fail (N=9)	0%	11%

Inter-rater reliability: 0.92 for Test A and 0.91 for Test B

Results of 2013 MoE EC exams



	Test A	Test B
No of examinees	355	326
Mean	**70.30**	**72.90**
SD	14.84	11.46
Min	20	28
Max	95	96

r = .66 (N=297); inter-rater reliability = .88

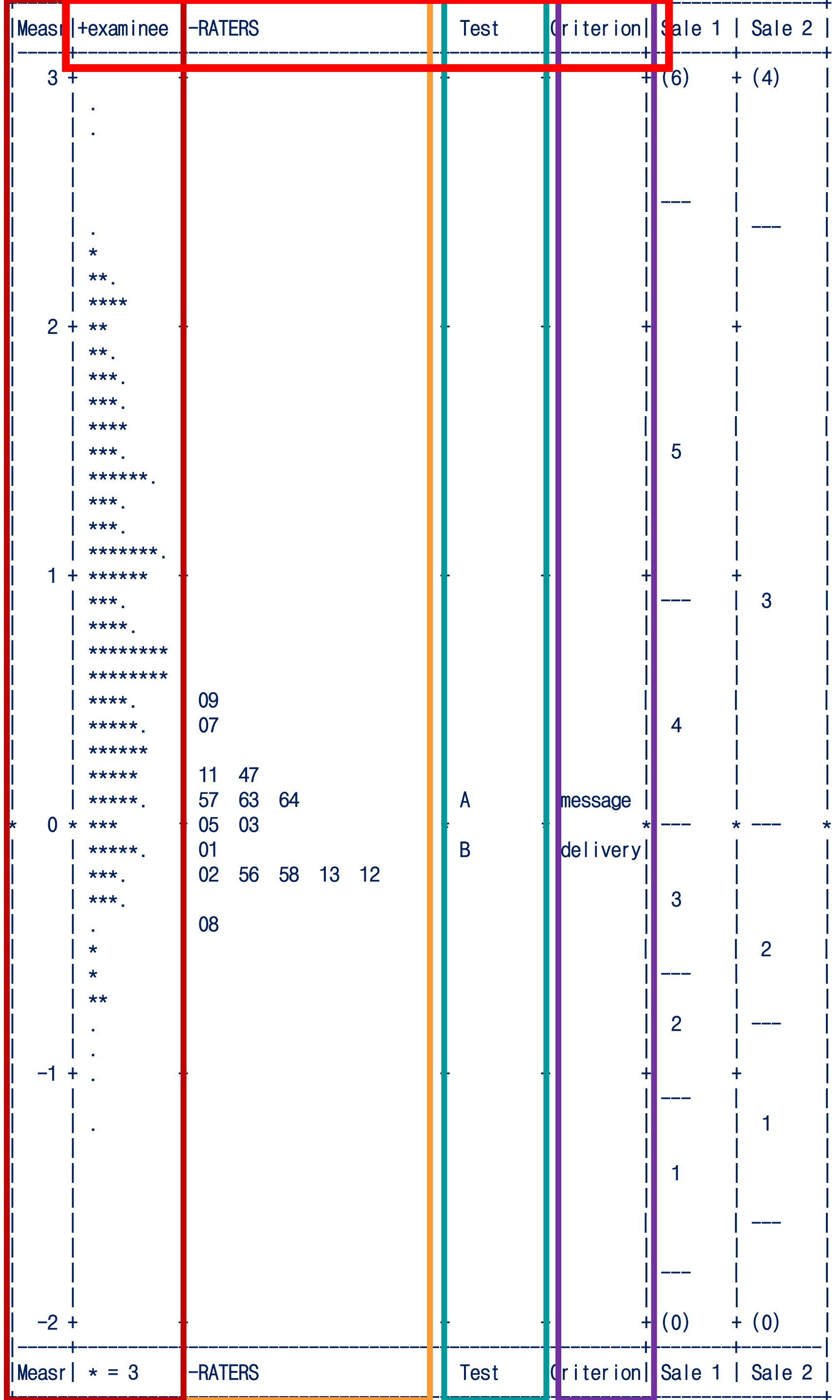
MFRM Results

Considerable variation in candidate ability, ranging from the highest logit of about 3 to the lowest of -1

Significant variation in harshness among the raters

Test A and Test B appeared to be similar in terms of difficulty, while Test A was a little more difficult than Test B, same as the CTT results.

The two criteria appeared to be close in terms of difficulty.



```
|Measr|+examinee -RATERS               Test    Criterion|Sale 1 | Sale 2 |
|-----+-------------------------------------------------+-------+--------|
|   3 +                                                 +(6)    + (4)    |
|     | .                                               |       |        |
|     | .                                               |       |        |
|     |                                                 |       |        |
|     |                                                 |       |        |
|     |                                                 | ---   |        |
|     | .                                               |       | ---    |
|     | *                                               |       |        |
|     | **.                                             |       |        |
|     | ****                                            |       |        |
|   2 + **                                              +       +        |
|     | **.                                             |       |        |
|     | ***.                                            |       |        |
|     | ***.                                            |       |        |
|     | ****                                            |       |        |
|     | ***.                                            | 5     |        |
|     | ******.                                         |       |        |
|     | ***.                                            |       |        |
|     | ***.                                            |       |        |
|     | *******.                                        |       |        |
|   1 + ******                                          +       +        |
|     | ***.                                            | ---   | 3      |
|     | ****.                                           |       |        |
|     | ********                                        |       |        |
|     | ********                                        |       |        |
|     | ****.    09                                     |       |        |
|     | *****.   07                                     | 4     |        |
|     | ******                                          |       |        |
|     | *****    11  47                                 |       |        |
|     | *****.   57  63  64            A       message  |       |        |
*   0 * ***      05  03                                 * ---   * ---    *
|     | *****.   01                    B       delivery |       |        |
|     | ***.     02  56  58  13  12                     |       |        |
|     | ***.                                            | 3     |        |
|     | .        08                                     |       |        |
|     | *                                               |       | 2      |
|     | *                                               | ---   |        |
|     | **                                              |       |        |
|     | .                                               | 2     | ---    |
|     | .                                               |       |        |
|  -1 + .                                               +       +        |
|     |                                                 | ---   |        |
|     | .                                               |       | 1      |
|     |                                                 |       |        |
|     |                                                 | 1     |        |
|     |                                                 |       |        |
|     |                                                 |       | ---    |
|     |                                                 |       |        |
|     |                                                 | ---   |        |
|     |                                                 |       |        |
|  -2 +                                                 +(0)    + (0)    |
|-----+-------------------------------------------------+-------+--------|
|Measr| * = 3   -RATERS               Test    Criterion|Sale 1 | Sale 2 |
```

Rater facet analysis

Rater	Measure logit	Infit mean square
08	-.35	1.09
13	-.23	.82
12	-.22	1.00
02	-.22	1.19
56	-.20	1.16
58	-.16	1.05
01	-.12	1.20
05	.00	1.38
03	.01	.87
64	.09	.91
63	.10	1.40
57	.13	.76
11	.17	1.27
47	.18	.98
07	.38	.84
09	.46	.82
Mean	.00	1.05
SD	.23	.20

.81

Separation =7.39; Separation reliability=.98

Fixed (all same) chi-square=1090.6, d.f.=15, p=.00

【1.05 + .20 x 2 = 1.45】

All the infit mean squares were below 1.45. => Raters were self-consistent in their own scoring.

The likelihood that the raters consistently differ from one another in overall severity.

Raters were not equally severe or lenient.

Test facet analysis

Test	Measure logit	Infit mean square
B	0.12 -.06	1.01
A	.06	1.05
Mean	.00	1.03 【**1.03 +**
SD	.08	.02 **.02 x 2 =**

Separation =5.58; Separation reliability=.97 **1.07**】
Fixed (all same) chi-square=64.3, d.f.=1, p=.00

The infit mean squares were both below 1.07. => Neither of the tests was misfitting.
The two test forms were not equally difficult or easy.

Criterion facet analysis

Criterion	Measure logit	Infit mean square
Delivery	-.14	.89
Message	.14	1.10
Mean	.00	.99
SD	.19	.14

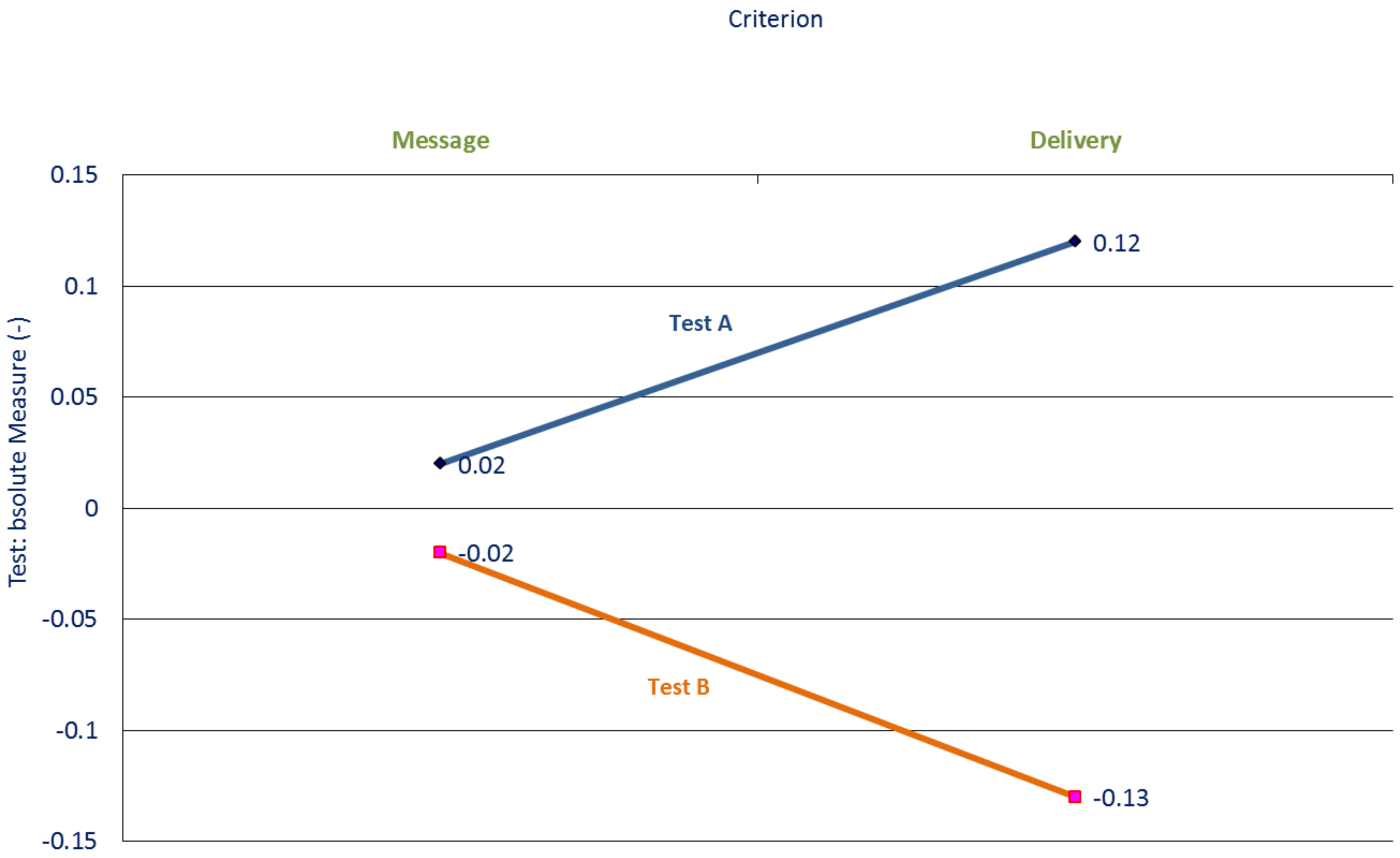
0.28

【.99 + .14 x 2 = 1.27】

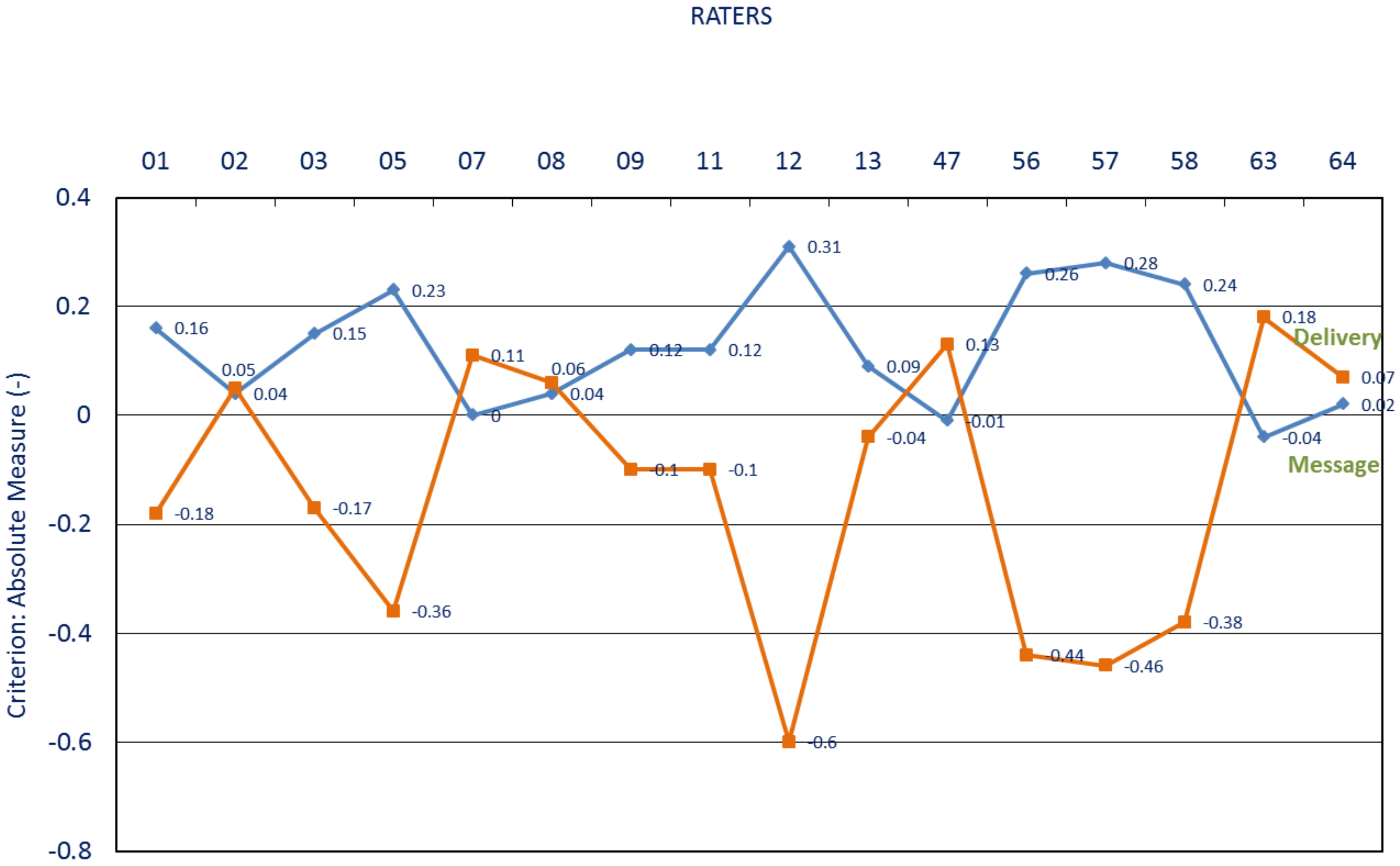
Separation =12.92; Separation reliability=.99
Fixed (all same) chi-square=336.1, d.f.=1, p=.00

All the infit mean squares were below 1.27.
The two criteria differed in terms of difficulty.

Test by criterion bias/interaction



Rater by Criterion bias/interaction



Discussion

- **The average scores of 2013 Tests A and B were very close.**
  - ➩Specifying test points can help stablise the MoE English to Chinese Exams in terms of difficulty to considerable extent.

- **Nevertheless, based on MFRM analysis, difference between Tests A and B in terms of difficulty reached significance level.**
  - ➩ Topics have a considerable impact on examinee performance.

- **The rater by criterion interaction analysis showed only a very slight variation in message.**
  - ➩ Specifying test points enables raters to agree more on message than on delivery.

Limitations

- **The data on which the test points was based are limited.**
  - The qualitative analysis was based on earlier versions of the MoE EC translation test papers only.
  - Construct validation is beyond the scope of this study.
  - The generalisability of the results obtained in this study in terms of test points is limited.
- **Impact of different types of test points is not investigated.**
  - Further qualitative analysis, such as coding examinees' responses and collecting introspective data from raters, to explore to what extent each type of test points affect candidates' performance is needed.

Thank you for your listening.