

# **Building the validity foundation for interpreter certification performance testing**

**Chao Han**

Bachelor of Arts

Master of Arts

A thesis submitted in fulfillment of the requirements for the degree of  
Doctor of Philosophy

Department of Linguistics

Faculty of Human Sciences

Macquarie University

Sydney, Australia

March 2015

©Copyright by Chao Han, 2015

## Table of Contents

Table of Contents.....	ii
List of Abbreviations .....	x
List of Tables .....	xi
List of Figures .....	xiii
List of Appendices.....	xiv
Abstract.....	xv
Statement of candidate.....	xvii
Acknowledgements .....	xviii
Dedication .....	xx
Chapter 1 Introduction.....	1
1.1 Introduction .....	1
1.2 Research background .....	1
1.2.1 Interpreter certification .....	1
1.2.2 Interpreter certification performance testing (ICPT).....	2
1.2.3 Validation research in ICPT .....	3
1.3 Theoretical underpinnings for validity investigation .....	5
1.4 Research scope, research purpose and research questions.....	8
1.5 Research design .....	11
1.5.1 RQ 1: Profile conference interpreting practice in China .....	11
1.5.2 RQ 2: Exploring the interplay between task characteristics, interpreting ability and interpreting performance quality .....	12
1.5.3 RQ 3: Methodological exploration of modern measurement theory to examine rater/score reliability .....	13
1.6 Thesis structure.....	14
1.7 Potential contribution of the research .....	16
1.8 References .....	17
An introductory note to Chapter 2.....	21
Chapter 2 Building a validity argument for interpreter certification performance testing .....	22
2.1 Introduction .....	22
2.2 The evolving nature of the validity concept and validation methods .....	25
2.2.1 Criterion validity: Correlation-based approach.....	26

2.2.2 Content validity: Judgment-based evaluation.....	26
2.2.3 Construct validity: An alternative approach .....	27
2.2.4 Trinitarian doctrine: A toolkit approach .....	28
2.2.5 Construct validity as a unitary concept: An integrated approach .....	28
2.3 An argument-based approach to validity.....	29
2.3.1 Toulmin's argument structure: Foundation building .....	30
2.3.2 An argument-based approach to validation in educational and language testing .....	31
2.3.2.1 Kane's interpretive argument and validity argument .....	31
2.3.2.2 Bachman's assessment use argument .....	33
2.3.2.3 Chapelle et al.'s validity argument for the TOEFL® .....	34
2.4 An argument-based approach to validation of the ICPTs.....	34
2.4.1 An overview of current practice in ICPT .....	35
2.4.2 Constructing a validity argument for ICPTs' test score interpretations .....	36
2.4.2.1 A chain of inferences and data .....	36
2.4.2.2 Warrants and underlying assumptions .....	38
2.4.2.3 Backing (validity) evidence .....	39
2.4.3 Constructing an assessment utilization argument for ICPT score use .....	45
2.4.4 Challenging ICPT score interpretations and uses .....	47
2.5 Implications .....	48
2.6 Conclusion.....	50
2.7 References .....	50
An introductory note to Chapter 3 .....	59
Chapter 3 An interactionalist approach to construct definition for English/Chinese interpreter certification performance testing' .....	60
3.1 Introduction .....	60
3.2 Literature review.....	63
3.2.1 Construct definition.....	63
3.2.2 A behaviorist approach .....	63
3.2.3 A trait approach .....	65
3.3 Proposing and articulating an interactionalist construct model for ICPTs.....	66
3.3.1 Rationales .....	66

3.3.2 An interactionist approach to construct definition.....	67
3.3.3 Components of interpreting ability.....	68
3.3.3.1 Knowledge of languages .....	68
3.3.3.2 Interpreting strategies.....	69
3.3.3.3 Topical knowledge .....	69
3.3.3.4 (Meta-)cognitive processes .....	70
3.3.4 A framework of SI task characteristics .....	70
3.3.4.1 Situational context .....	71
3.3.4.2 Physical (booth) condition .....	71
3.3.4.3 Characteristics of SI tasks .....	71
3.3.4.3.1 Task-level characteristics.....	72
3.3.4.3.2 Linguistic characteristics of source speeches .....	73
3.3.4.3.3 Paralinguistic characteristics of source speeches.....	74
3.3.4.3.4 Kinesic characteristics of speakers.....	75
3.3.4.3.5 Expected response from interpreters .....	75
3.4 Implications .....	75
3.5 Conclusion.....	78
3.6 References .....	78
An introductory note to Chapter 4.....	86
Chapter 4 Profiling conference interpreting practice in China: A sequential-exploratory mixed-methods design study.....	87
4.1 Introduction .....	87
4.2 Descriptions of the interpreting practice in China and in other countries.....	89
4.2.1 Surveys on the interpreting profession in China .....	89
4.2.2 Detailed descriptions of conference interpreting practice .....	91
4.3 An exploratory qualitative diary study.....	92
4.4 Method .....	93
4.4.1 Survey design.....	93
4.4.2 Sampling .....	94
4.4.3 Procedure .....	94
4.4.4 Data analysis .....	95
4.5 Results .....	95

4.5.1 Demographic information.....	95
4.5.2 Part I: Results.....	98
4.5.3 Part II: Results.....	101
4.5.3.1 Employment status .....	103
4.5.3.2 SI experience .....	104
4.5.3.3 Working location .....	104
4.5.4 Part III: Results .....	105
4.6 Discussion .....	109
4.6.1 Demographic information.....	109
4.6.2 Discussion of the results from Part I, II and III .....	111
4.7 Conclusion.....	113
4.8 References .....	113
An introductory note to Chapter 5 .....	117
Chapter 5 The effects of interpreting task characteristics on the quality of simultaneous interpreting: A convergent parallel mixed-methods approach.....	118
5.1 Introduction .....	118
5.2 Literature review.....	120
5.2.1 Review of empirical studies: Speed factor .....	120
5.2.2 Review of empirical studies: Accent factor .....	121
5.2.3 Discussion of the empirical studies.....	121
5.3 Research purpose and questions .....	123
5.4 Methods.....	124
5.4.1 Mixed sampling design: Participants .....	124
5.4.2 Experimental design.....	125
5.4.3 SI materials .....	125
5.4.3.1 Development of SI tasks .....	125
5.4.3.2 Manipulating the IV: Accent .....	126
5.4.3.3 Manipulating the IV: Speech rate.....	126
5.4.4 Other instruments.....	127
5.4.4.1 Post-task interview .....	127
5.4.4.2 Post-hoc questionnaire.....	127
5.4.5 Data collection procedure.....	128

5.4.6 Performance assessment .....	129
5.4.7 Data analysis .....	129
5.5 Results and discussions .....	131
5.5.1 Results I and discussion I.....	131
5.5.1.1 Quantitative results: Performance data .....	131
5.5.1.2 Discussion I: Performance data .....	134
5.5.2 Results II and discussion II .....	135
5.5.2.1 Quantitative results: Perception data .....	135
5.5.2.2 Discussion II: Perception data .....	136
5.5.3 Results III and discussion III .....	137
5.5.3.1 Qualitative results: Task development and IVs manipulation.....	137
5.5.3.2 Discussion III: Prominent difficulty factors.....	139
5.5.4 Results IV and discussion IV .....	139
5.5.4.1 Qualitative results: Interpreters' reflections on SI performance.....	139
5.5.4.2 Discussion IV: Perceived effects on SI .....	142
5.5.5 Summary discussion: Triangulation and meta-inference.....	143
5.6 Limitations of the study.....	146
5.7 Conclusion .....	146
5.8 References .....	146
An introductory note to Chapter 6.....	150
Chapter 6 Exploring the relationship between task characteristics, strategy use and performance in English-to-Chinese simultaneous interpreting.....	151
6.1 Introduction .....	151
6.2 Literature review .....	152
6.2.1 Defining, identifying and categorizing interpreting strategies .....	152
6.2.2 The effect of SI task characteristics on strategy use.....	153
6.2.3 Relationship between strategy use and SI performance .....	154
6.3 Research questions .....	155
6.4 Method .....	155
6.4.1 Participant recruitment .....	155
6.4.2 Experimental design .....	155
6.4.3 SI tasks.....	156

6.4.4 Experiment procedure .....	156
6.4.5 Strategy coding .....	156
6.4.6 English glosses for the target language interpretations .....	157
6.4.7 Data analysis .....	158
6.5 Results.....	159
6.5.1 Strategy use: Illustrated examples.....	159
6.5.2 Characteristics of strategy use .....	164
6.5.2.1 Patterns of strategy use .....	164
6.5.2.2 Strategy clusters .....	165
6.5.3 Effect of speech rate and accent on strategy use.....	167
6.5.4 Relationship between strategy use and SI performance .....	168
6.6 Discussion .....	170
6.6.1 Characteristics of strategy use .....	170
6.6.2 The effect of speech rate and accent on strategy use .....	171
6.6.3 Relationship between strategy use and SI performance quality .....	172
6.7 Limitations .....	173
6.8 Conclusion.....	173
6.9 References .....	174
An introductory note to Chapter 7 .....	176
Chapter 7 Investigating rater severity/leniency in interpreter performance testing: Using multifaceted Rasch measurement.....	177
7.1 Introduction .....	177
7.2 A review of rater training and rater variability in ICPT .....	179
7.3 A brief introduction to multifaceted Rasch measurement .....	181
7.4 Context of the present study.....	181
7.5 Foci of MFRM analysis in the study .....	182
7.6 Method .....	183
7.6.1 Raters.....	183
7.6.2 Materials .....	183
7.6.2.1 Data source: Interpretation recordings.....	183
7.6.2.2 Rating scale .....	184
7.6.3 Procedure.....	184

7.6.4 Measurement design.....	185
7.6.5 Rasch models .....	185
7.6.5.1 Rasch model variants .....	185
7.6.5.2 Unidimensionality assumption .....	185
7.6.5.3 Choice of Rasch model .....	186
7.6.6 Data analysis .....	187
7.7 Results .....	187
7.7.1 Global model fit .....	187
7.7.2 Effectiveness of the rating scale.....	187
7.7.3 Rater calibration reports: Rater severity and internal self-consistency .....	188
7.7.4 Bias analysis.....	190
7.7.4.1 Summary statistics on two-way interaction .....	190
7.7.4.2 Rater × Interpreter interaction .....	191
7.7.4.3 Rater × Criterion interaction .....	195
7.8 Discussion and implications .....	196
7.8.1 Sample size issue in the present study.....	196
7.8.2 Rater severity/leniency in the present assessment context .....	197
7.8.3 MFRM: Implications for ICPTs.....	199
7.8.3.1 Practical measurement designs.....	199
7.8.3.2 Rater training and monitoring .....	201
7.8.3.3 Utility of Rasch measurement .....	202
7.9 Conclusion.....	202
7.10 References .....	202
An introductory note to Chapter 8.....	209
Chapter 8 Investigating score reliability in English/Chinese interpreter certification performance testing: A generalizability theory approach .....	210
8.1 Introduction .....	210
8.2 Literature review .....	212
8.2.1 An overview of the ICPT practice .....	212
8.2.2 G theory in second language testing and its relevance to ICPT.....	213
8.3 Research questions.....	215
8.4 Method .....	216

8.4.1 Participants: Interpreters and raters.....	216
8.4.2 Materials: SI tasks and rating scales .....	216
8.4.3 Rater training .....	216
8.4.4 Procedures .....	217
8.4.5 Data analysis .....	217
8.5 Results.....	219
8.5.1 Univariate G-theory analysis .....	219
8.5.1.1 Univariate G study.....	219
8.5.1.2 Univariate D studies .....	220
8.5.1.3 Alternative measurement designs .....	221
8.5.2 Multivariate G-theory analysis .....	223
8.6 Discussion and implications.....	226
8.7 Limitations and conclusion .....	229
8.8 References .....	230
Chapter 9 Summary and Conclusions .....	234
9.1 Introduction .....	234
9.2 Summary of the thesis.....	234
9.3 Linking the empirical and methodological findings to the theoretical ICPT validity argument .....	237
9.4 Strengths, contributions and limitations of the research.....	238
9.4.1 Strengths and contributions of the research.....	239
9.4.2 Limitations of the research .....	240
9.5 Implications and recommendations .....	242
9.6 Future research .....	243
9.7 References .....	245
Appendices.....	247

## **List of Abbreviations**

AUA: Assessment use argument

ABA: Argument-based approach

CC: Covariance component

CTT: Classical test theory

EV: Extraneous variables

FluDel: Fluency of delivery

FSR: Fast speech rate

ICPT: Interpreter certification performance testing

ICPTs: Interpreter certification performance tests

InfoCom: Information completeness

MFRM: Multifaceted Rasch measurement

MMR: Mixed-methods design

NNS: Non-native speaker

NS: Native speaker

RMPA: Rater-mediated performance assessment

RQ: Research question

SCV: Statistical conclusion validity

SEM: Standard error of measurement

SI: Simultaneous interpreting

SSR: Slow speech rate

StrA: Strong accent

TCs: Treatment conditions

TLQual: Target language quality

VC: Variance component

## List of Tables

Table 2.1 Conceptualizations of "validity" and associated validation methods .....	27
Table 2.2 Warrants and underlying assumptions in the validity argument for ICPT score interpretations .....	40
Table 2.3 Expected validity evidence, available validity evidence and future studies .....	42
Table 2.3 ( <i>continued</i> ).....	43
Table 2.4 Warrant, assumptions and backing evidence in the assessment utilization argument .....	46
Table 3.1 A framework of SI task characteristics.....	72
Table 4.1 General information on the three surveys .....	90
Table 4.2 Basic demographic information.....	96
Table 4.3 Descriptive statistics of the three demographic variables .....	97
Table 4.4 The number of different materials received by different types of the interpreters .	100
Table 4.5 Mean frequency scores and standard deviations for the 18 varieties of SI task .....	102
Table 4.6 Relationship between employment status and SI task varieties .....	103
Table 4.7 Relationship between interpreting experience and SI task varieties.....	104
Table 4.8 Relationship between location and SI task varieties .....	105
Table 4.9 Who interpreted for over 31 minutes in one interpreting turn? .....	107
Table 4.10 The frequency (and percentage) of a certain difficulty factor identified by different types of the interpreters .....	109
Table 5.1 Demographic information of the participants in the main study .....	125
Table 5.2 The 2×2 factorial design.....	125
Table 5.3 An overview of the experiment procedure .....	128
Table 5.4 SEM & $\rho^2$ for the scores of the three rating dimensions .....	129
Table 5.5 Data source & data type matrix .....	130
Table 5.6 Descriptive statistics for the performance data.....	131
Table 5.7 Univariate ANOVA effects .....	132
Table 5.8 ANOVA results for the effects of the TCs on the quality criteria .....	134
Table 5.9 Descriptive statistics for the perception data.....	135
Table 5.10 Difficulty factors mentioned by the interpreters.....	138
Table 5.11 Number of codings & percentage for each task-by-criterion condition .....	141
Table 5.12 Comparison and triangulation of the results from the quantitative and the qualitative	

components .....	145
Table 6.1 Glosses for the Chinese interpretations .....	157
Table 6.2 Frequency count of strategy use in the interpreting tasks .....	165
Table 6.3 A comparison of strategy use between the fast and the slow conditions .....	167
Table 6.4 A comparison of strategy use between the native and the accented speech conditions .....	168
Table 6.5 A comparison of strategy use between the three interpreters .....	169
Table 7.1 A 2×2 factorial design used in the experiment .....	182
Table 7.2 Infit statistics of the rating criteria to support unidimensionality .....	186
Table 7.3 Separation values and variance statistics in RSM and PCM .....	186
Table 7.4 Overall data-model fit statistics .....	187
Table 7.5 Statistics relating to FluDel rating subscale .....	188
Table 7.6 Logit estimates for overall rater severity .....	189
Table 7.7 Tight and loose fit ranges to determine misfit and overfit of a rater-facet element	190
Table 7.8 Summary statistics for three two-way interactions .....	191
Table 7.9 Statistical information on all significant Rater × Interpreter interactions .....	192
Table 7.10 Statistical information on significant R08 × Interpreter interactions .....	193
Table 7.11 Statistical information on all significant Rater × Criterion interactions .....	196
Table 7.12 (a) Fully crossed/complete design – Connected .....	199
Table 7.12 (b) Incomplete design – Connected .....	200
Table 7.12 (c) Incomplete design – Connected .....	200
Table 7.12 (d) Incomplete design – Disconnected .....	201
Table 8.1 Decomposition of total variance into variance components for the InfoCom ratings .....	218
Table 8.2 Univariate G study for one task and one rater for each rating dimension .....	220
Table 8.3 D study results based on four SI tasks and two raters .....	221
Table 8.4 Estimated variance-covariance components (VCC) from the multivariate G study for one task and one rater .....	224
Table 8.5 Dependability of composite scores based on the three weighting schemes .....	225
Table 8.6 Composite score analysis results: Effective weights for the design of four tasks and two raters .....	226

## List of Figures

Figure 1.1 Linkages between research questions and studies .....	12
Figure 2.1 Toulmin's model of argument, based on Toulmin (2003) .....	30
Figure 2.2 A chain of inferences linking data to score interpretations and uses, based on Kane (1994) .....	32
Figure 2.3 Relationship between an interpretive and a validity argument, based on Kane (2006) .....	32
Figure 2.4 The structure of assessment use argument (AUA), based on Bachman & Palmer (2010) .....	33
Figure 2.5 A chain of inferences and related data in the validity argument .....	37
Figure 2.6 The AUA for the ICPT score interpretations and use .....	49
Figure 3.1 Two types of score-based inferences .....	61
Figure 3.2 A behaviorist approach to interpreting performance consistency .....	64
Figure 3.3 A trait approach to interpreting performance consistency .....	66
Figure 3.4 An interactionalist approach to interpreting performance consistency .....	68
Figure 3.5 Structure and components of language knowledge (Bachman & Palmer, 1996)....	69
Figure 3.6 The interactionalist construct model and score interpretations .....	77
Figure 4.1 Conference-related materials or information obtained in advance .....	99
Figure 4.2 When conference-related materials were received.....	101
Figure 4.3 Directionality of SI .....	106
Figure 4.4 Duration of an interpreting turn .....	106
Figure 4.5 Factors contributing to SI task difficulty .....	108
Figure 5.1(a) Performance scores (Speed).....	133
Figure 5.1(b) Performance scores (Accent).....	133
Figure 5.2 Averaged perceived difficulty ratings .....	136
Figure 7.1 Bias assessment map for R08.....	194
Figure 7.2 Bias assessment maps for all the raters.....	195
Figure 8.1 $SEM_{Abs}$ as a function of No. of tasks and raters.....	222
Figure 8.2 Index of dependability as a function of No. of tasks and raters .....	223
Figure 9.1 A visual display of linking the empirical and methodological findings to the theoretical validity argument .....	238

## List of Appendices

Appendix A: Macquarie University guidelines for thesis by publication.....	247
Appendix B: Final ethics approval .....	249
Appendix C: Part of a completed diary form.....	253
Appendix D: SI task categorization and descriptions.....	255
Appendix E English speech scripts for the four SI tasks .....	258
Appendix F: Indices for characteristics of the four source texts.....	259
Appendix G: Interview questions in the experiment .....	260
Appendix H: The <i>post-hoc</i> questionnaire .....	261
Appendix I: Background reading material in the experiment.....	263
Appendix J: Background information sheet .....	264
Appendix K: Interpreting performance assessment sheet.....	266
Appendix L: Strategies and definitions .....	267
Appendix M: Certification tests reviewed and associated literature .....	269

## **Abstract**

Interpreter certification performance testing (ICPT) has developed rapidly over the past decade. Yet there has been very limited discussion and systematic research conducted to enhance reliability and validity of high-stakes interpreter certification performance tests (ICPTs). This interdisciplinary mixed-methods research was therefore initiated to build theoretical and methodological foundations for rater-mediated ICPTs, with a special focus on test validation, construct definition, and rater/score reliability for English/Chinese ICPTs in China.

Presented in a thesis-by-publication format, the research follows a multi-phase mixed-methods research (MMR) design, in which research results from a previous study inform and build to a subsequent study.

To begin with, given the lack of guidance on rigorous validation of ICPTs, this thesis draws upon an argument-based approach to build a validity argument for ICPTs. The validity argument could serve as a roadmap to help testers collect validity evidence. Based on Interpreting Studies literature, two particular types of evidence are generally lacking: evidence supporting substantive score interpretations based on a strong construct theory, and evidence supporting test score generalizability, especially across raters.

To help generate evidence that justifies the substantive score interpretations intended by certification authorities in China, an interactionalist approach to construct definition is therefore proposed and articulated for English/Chinese ICPTs. Essentially, the interactionalist construct model contends that performance consistency (i.e., interpreting performance) is as a function of context (i.e., characteristics of test tasks), trait (i.e., interpreting ability), and interactions between the two. The theoretical construct model gives rise to two research questions (RQ). RQ 1: What are the characteristics of interpreting tasks in the real-life practice domain in China? RQ 2: What is the possible interplay between characteristics of interpreting tasks, interpreting ability, and interpreting performance quality?

To address RQ 1, an exploratory qualitative diary (n = 11) and a follow-up quantitative survey (n = 140) were conducted to generate empirical data that describe the characteristics of the interpreting practice in China. Main results include that the interpreters performed a greater variety of simultaneous interpreting (SI) tasks than previously thought, and encountered a number of prominent factors contributing to SI difficulty, such as fast speech rate (FSR) and strong accent (StrA).

To investigate RQ 2, a factorial repeated-measures experiment was conducted. Specifically, informed by the diary and the survey findings, the experiment sought to address the interactions between SI tasks (characterized by FSR and StrA), strategy use (regarded as a crucial component of interpreting ability), and SI performance quality (measured by information completeness, fluency of delivery and target language quality). In the experiment, 32 interpreters were asked to perform English-to-Chinese SI in four manipulated tasks. A crossed measurement design was then implemented in which nine trained raters assessed each performance by each interpreter on each rating dimension. Results show that 1) the speed factor had a pattern of mixed impacts on information completeness, fluency of delivery and target language quality of SI performance, while the accent factor had a consistent pattern of detrimental impacts across the three dimensions; 2) the strategies of syntactic transformation and substitution were used most frequently. It also would appear that while the speed factor greatly influenced the use of the two strategies, the accent factor did not; and 3) there seemed to be a general trend that the more strategies were used, the better SI performance was.

Finally, to help produce evidence supporting rater reliability and score generalizability (i.e., RQ 3), a methodological exploration was conducted to evaluate the utility of multifaceted Rasch measurement and generalizability theory in analyzing rater behavior, rater variability and its effects on score dependability. Data for the analyses were the rater-generated scores from the experiment. Results indicate that although the rating design produced reliable results, one of the raters was problematic, as s/he was not self-consistent, and provided significantly biased scores to a large proportion of the interpreters. The findings also show that increasing the number of raters and/or tasks would generally improve score reliability for each rating dimension, but the relative efficiency was different across the dimensions.

Ultimately, the empirical and methodological findings would contribute evidence to the ICPT validity argument, and their implications on ICPT design and validation were also discussed.

## **Statement of candidate**

I certify that the work in this thesis entitled “Building the validity foundation for interpreter certification performance testing” has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged.

In addition, I certify that all information sources and literature used are indicated in the thesis.

The research presented in this thesis was approved by Macquarie University Ethics Review Committee (Human Research) (see Appendix B) and conducted in accordance with the guideline stipulated.

Chao Han

Student ID: 42499542

March 2015

## Acknowledgements

First and foremost, I would like to express my deepest gratitude to Dr. Jing Chen, who supervised my postgraduate study at Xiamen University, inspired and introduced me to the field of language testing, especially interpreter certification testing.

I am deeply indebted to my supervisor, Dr. Mehdi Riazi, for his mentorship, unfailing support and generosity. Mehdi has exerted great influences on my research, in terms of research design and methodology, and language testing theories. I benefited tremendously from his Research Methodology unit, academic seminars, publications, numerous private and group meetings, and constructive feedback on my writing. I am also thankful for the opportunity to give lectures in the *Language Testing and Evaluation* course at Macquarie University convened by Dr. Riazi.

I am also extremely grateful to my associate supervisor, Dr. Helen Slatyer, for her guidance, valuable insight, constant encouragement and unfailing support. Helen has been closely involved in my research since the very beginning, imparted research know-how to me through countless meetings, and helped me go over difficult times during the candidature.

I would like to give special thanks to Dr. Wendy Noble. Her academic writing courses and feedback to my writing assignments helped me improve my writing skills.

I would like to acknowledge three close friends. My sincerest thanks go to my statistics mentor, Dr. Haining Wang, an inspiring young economist, for sharing with me his intellectual insight into general linear models. Thanks also go to Dr. Yang Shen, a role model for many early-career researchers and a rising star in Actuarial Science, for inspiring and encouraging me when I was down in the dumps. I also owe my gratitude to Sijia Chen, a Xiamen University alumna who is also conducting her PhD research at Macquarie University, for her generosity and time to help me with data analysis and provide constructive feedback on thesis chapters.

I am also grateful to technicians, speakers, volunteers, students, and conference interpreters involved in my empirical studies. Particularly, I would like to give heartfelt thanks to my old friends, Rui Bao and Qinxian Su, for providing accommodation to me when I was conducting the experiment in Beijing, and to Dr. Xiaoyan Xiao, Dr. Liuyan Yang, Yuping Chen, Wenhong Fang, Ruiling Gao, and Rongbo Fu, for their endless support while I was in Xiamen.

Special thanks go to Dr. Rita Green, Dr. Tim McNamara, Dr. Ute Knoch, Dr. John Michael Linacre, Dr. Micheline Chalhoub-Deville, Dr. Lyle Bachman, Dr. George Marcoulides, and Dr. Dina Tsagari who provided illuminating answers to my questions about generalizability theory

and Rasch measurement in seminars and workshops, and through emails. Thanks also go to Dr. Juliet Lum for organizing fantastic HDR learning workshops and training sessions at Macquarie University.

I would especially like to thank Collette Ryan, Sofia Robredo Moran, and consular officer Wenwu Liu for their professionalism to help me complete many administrative forms and to provide helpful answers to my various questions.

For the financial help, I would like to acknowledge the China Scholarship Council – International Macquarie University Research Excellence Scholarship that supported my PhD study; the Macquarie University Linguistics Department Research Enhancement Fund, for supporting the data collection in Beijing 2013, and my trip to an international conference at New York University 2014; the Macquarie University Postgraduate Research Fund, for sponsoring my trip to Lancaster University, UK for a two-week seminar in 2014; and the Student Travel Award provided by the Association for Language Testing and Assessment of Australia and New Zealand, for supporting my trip to a conference at University of Queensland 2014.

I am very thankful for the valuable, constructive and critical feedback from journal editors and anonymous reviewers of my published and submitted papers herein. I also appreciate the comments provided by scholars, researchers and peer students on my PhD research in national and international conferences I attended.

I would like to thank all my colleagues and friends for their kind help and mutual encouragement throughout my candidature. Particularly, I would like to acknowledge Dr. Xin Xiong, Dr. Fengyi Du, Dr. Yaxin Lu, Dr. Zhibi (Nick) Wei, Dr. Zhou Jiang, Dr. Jihong Wang, Xi Li, Sandra, Noon, Dariush, Nan Fan, Changpeng Huan, Jia Li, Wei Zhang, Wendy, Kun Fan, Sharon Yahalom, Mathew Book, and Yanna Cui.

Finally, my boundless gratitude goes to my family, my sister and my grandma for their love, support and care. I would particularly like to thank my parents for their deep and unconditional love. I have been tremendously influenced and shaped by my parents' strong work ethics, optimism, conscientiousness, and inspiring life stories. To my wife, Yuling Zhou, I love you.

## **Dedication**

This thesis is dedicated to my grandfather, late Yunnan Yuan, because of whose blessings this thesis is possible.

## **Chapter 1 Introduction**

... developing a valid and reliable test for translation and interpreting is of paramount importance. Both academe and the industry would benefit enormously from making accurate and sound decisions on translation ability and quality based on meaningful testing. (Angelelli, 2009, p. 14)

### **1.1 Introduction**

This opening chapter first introduces the research background and describes theoretical underpinnings for this PhD research. Against the backdrop, the research scope, research purpose and research questions are described. Next, the design of the research is presented, in which logical connections between research questions and individual studies are explained. Finally, the chapter outlines the thesis structure by briefly discussing each chapter.

### **1.2 Research background**

The central concern of this PhD research is validity of interpreter certification performance testing. This section provides the background to and contextualizes the research, by reviewing the current state of interpreter certification performance testing, and problematizing the research area.

#### **1.2.1 Interpreter certification**

Over the past 50 years or so, the professionalization of interpreting has been accelerated by the surging trend of economic globalization, international trade and investment, and increasing opportunities of technological, social and cultural exchanges between different parts of the world (Mackintosh, 2006). The process of professionalization is partly characterized by certification of interpreters to ensure the quality of professional services provided (Hlavac, 2013).

Interpreter certification practices around the world employ different procedures and different pathways to interpreter certification (for a detailed review, see Hlavac, 2013). In some countries (e.g., Australia), would-be interpreters who complete approved courses are eligible to

apply for certification without further requirements (National Accreditation Authority for Translators and Interpreters [NAATI], 2014). Certification may also be achieved by provision of relevant qualifications, evidence of professional association membership or professional experience in translation and interpreting (NAATI, 2015). However, one of the most important and widely used pathways to interpreter certification is testing (Roat, 2006), particularly interpreter performance testing, in which candidates perform a certain mode(s) of interpreting (e.g., dialogue interpreting, sight interpreting, simultaneous interpreting, whispered interpreting).<sup>1</sup> Test takers' performance is then assessed by human raters, using a predetermined scoring schedule. This type of assessment is called rater-mediated *interpreter certification performance testing* (ICPT).

### 1.2.2 Interpreter certification performance testing (ICPT)

This section provides an overview of the development of the ICPT around the world, with a special focus on China.

Over the past decades, ICPT has developed rapidly across the world. Accordingly, *interpreter certification performance tests* or ICPTs have been designed and administered to certify different types of interpreters working in different settings or domains. Hale, Garcia, Hlavac, Kim, Lai, Turner and Slatyer (2012), Hlavac (2013) and Roat (2006) present an informative summary of the use of ICPTs in different countries. Some ICPTs are used to certify interpreters working in legal settings (e.g., the US Federal Court Interpreter Certification Examination), some in medical settings (e.g., the Washington State Medical Interpreter Certification Examination), and others in general public services settings (e.g., the UK Diploma in Public Service Interpreting). ICPTs are also designed to accommodate different modalities of interpreting (i.e., signed & spoken language interpreting). For instance, the US National Association of the Deaf (NAD) and the Registry of Interpreters for the Deaf (RID) administer the National Interpreter Certification for American Sign Language (ASL) interpreters. Furthermore, some ICPTs are used specifically to certify interpreters working in only one language pair (e.g., the US National Association of Judiciary Interpreters and Translators' certification program currently only offers a Spanish/English test), and others in multiple language combinations (e.g., Australia's National Accreditation Authority for Translators and

---

<sup>1</sup> In the thesis, simultaneous interpreting (SI) is restricted to SI in conference contexts.

Interpreters or NAATI can potentially provide dozens of language combinations between English and a Language Other Than English).<sup>2</sup>

In China, the past ten years has witnessed a mushrooming of both national- and local-level ICPTs. Nationally, two tests figure prominently: the China Accreditation Test for Translators and Interpreters (CATTI) and the National Accreditation Examinations for Translators and Interpreters (NAETI). Several local ICPTs are also widely recognized such as the Shanghai Business Interpretation Accreditation Test (BIAT), the Shanghai Interpretation Accreditation test (SIA), and the English Interpreting Certificate (EIC) developed by Xiamen University.

Despite the diversity of the ICPTs, test scores are typically used as critical evidence to help certifying authorities make certification decisions. Only those who surpass a cut-off score become certified interpreters who are then allowed to practice the occupation in certain settings. As a result, the ICPT tends to play a gate-keeping role for the interpreting profession. In other words, the ICPT serves as a quality-control mechanism to protect recipients of interpreting services.

Given the pervasiveness and the high-stakes of ICPT, it is necessary to examine the validity of inferences and uses based on ICPT scores, and to evaluate whether ICPTs have functioned as intended (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). In other words, rigorous and thorough test validation research needs to be conducted for the ICPTs.

### 1.2.3 Validation research in ICPT

Despite the rapid expansion and the high-stakes of the ICPT, validation research driving and underpinning credible ICPTs is still lacking in the field of Interpreting Studies, with a few notable exceptions (Angelelli, 2009; Campbell & Hale, 2003; Chen, 2002, 2009; Clifford, 2005; Hale et al., 2012; Liu, 2013; Sawyer, 2004; Wu, 2010). The lack of proper research underpinning ICPTs has been echoed by prominent interpreting researchers who have a keen interest in interpreter performance testing (e.g., Angelelli, 2009; Hale et al., 2012; Sawyer, 2004). For example, although Sawyer (2004) foresees a vast potential for developing interpreter tests, the field of interpreter testing is still in its infancy. Hale et al. (2012) also observe that around the world interpreter and translator examinations have traditionally not been subjected

---

<sup>2</sup> [http://www.naati.com.au/PDF/Booklets/Accreditation\\_by\\_Testing\\_booklet.pdf](http://www.naati.com.au/PDF/Booklets/Accreditation_by_Testing_booklet.pdf)

to the same rigor as language proficiency tests. Specifically, there has been a paucity of in-depth discussions of and systematic investigations into reliability and validity of ICPTs. Describing the current state of interpreter test design, Sawyer (2000) cites a number of problems undermining test credibility such as arbitrary selection of test content, discrepant test administration practices, and inconsistent application of scoring criteria. Sawyer (2004) also states that a comprehensive analysis of validity in interpreter performance assessment has not been conducted. Similarly, Campbell and Hale (2003, p. 211) attribute the lack of literature on interpretation assessment to “the intuitive nature of test design and assessment criteria”.

There are three possible reasons for why validation research is lacking for interpretation testing and assessment. One reason is that developers of ICPTs are primarily interpreters and/or interpreter educators who do not necessarily have a full range of language testing expertise to launch rigorous test validation. Another possible reason is that the ICPT is a recent phenomenon, which has not attracted much attention from interpreting researchers. The last reason could be of the technical nature of the validity concept. Since its inception in the early 1900s, the notion of validity has developed into a sophisticated concept consisting of different “validities” and various approaches to validation, which may lead to confusion among interpreting researchers.

Against this background, a number of interpreting researchers and testers call for an overdue systematic and rigorous investigation into reliability and validity for interpreter performance testing, by learning from research and practice in mature disciplines such as educational measurement and language testing (Campbell & Hale, 2003; Clifford, 2005; Sawyer, 2004; Wu, 2010). Campbell and Hale (2003) suggest that the wider field of measurement and evaluation represents a solid source of knowledge that interpreting testers can use to understand and improve interpreter assessment practice.

Responding to the calls for rigorous validity investigation, this PhD research draws upon recent language testing and assessment research to provide preliminary solutions to some of the prominent problems that have undermined the progress of ICPT. Particularly, the research focuses on the fundamental issue of validity in the interpreter certification performance testing by advancing an argument-based approach to validity investigation and generating some initial empirical validity evidence.

### 1.3 Theoretical underpinnings for validity investigation

Broadly speaking, validity investigation or test validation is a process in which test developers and users collect and generate validity evidence to support and justify the inferences and actions based on test scores (Bachman, 1990; Messick, 1989). The validation process starts with identifying score-based inferences and actions,<sup>3</sup> and then embarks on a research program to generate evidence to justify the adequacy and appropriateness of intended inferences and actions.

For the ICPT, test scores are usually interpreted in both trait and performance-referenced (or task-based) approaches. On the one hand, test scores are used as an indicator of whether candidates have desirable traits such as the knowledge, skills, abilities and strategies (KSASs) required of an interpreter. On the other hand, test scores are also used to describe candidates' interpreting performance in a given context. Take the CATTI simultaneous interpreting (SI) tests for example,<sup>4</sup> test scores are used to indicate whether test candidates are able to “use Chinese and English languages with dexterity, have expansive background knowledge of politics, economics, culture, etc., apply SI skills adroitly, and demonstrate sound psychological qualities and coping tactics”. Test scores are also indicators of whether test takers have “rendered source-language content accurately and completely, pronounced correctly and clearly, delivered fluently and in a natural tone” in “various formal (conference) occasions”. In addition, test scores are used to make certification decisions so that only those candidates who outscore a cut-off point will become certified interpreters. Consequently, to validate the ICPTs, test developers and users need to provide credible validity evidence to link candidates' test performance to score-based inferences (i.e., both trait- and performance-based inferences), and finally to score-based actions (i.e., certification decisions).

In the field of Interpreting Studies, much of the theoretical discussion on validity issues (Angelelli, 2009; Campbell & Hale, 2003; Hale et al., 2012; Sawyer, 2004) is influenced by early work of validity theorists (e.g., Cronbach, 1971; Guion, 1980) and language testers (e.g., Bachman, 1990; Bachman & Palmer, 1996). Based on this traditional approach, the recipe for validity investigation is to collect whichever piece of evidence that is deemed as appropriate

---

<sup>3</sup> In language testing literature, score-based inferences refer to interpretations or explanations of test scores. Score-based actions refer to use of test scores to make certain decisions such as selection, replacement and certification.

<sup>4</sup> The syllabus for the CATTI SI test: [http://bbs.catti.china.com.cn/download/syllabus\\_EN\\_SI2.pdf](http://bbs.catti.china.com.cn/download/syllabus_EN_SI2.pdf) ; The quoted texts were originally in Chinese and translated into English by the author.

and practical from a list of separate “validities” or types of validity such as concurrent validity, predictive validity, content validity, construct validity, etc. However, this traditional approach to validity investigation has recently come under attack in the general field of educational assessment due to a number of weaknesses such as opportunistic choice of validity evidence, lack of systematicity, and lack of practical guidance (see Chapter 2 for the historical evolution of validity theory).

In response to the disconnected types of validity, Messick (1989) argues for construct validity as a unitary concept and as a unifying force integrating all aspects of validity. In his theoretical validity framework, Messick emphasizes on six distinguishable aspects of construct validity, including content, substantive, structural, generalizability, external, and consequential aspects of construct validity. These six aspects of unified construct validity could be gathered as validity evidence to support test score interpretations and uses.

To operationalize Messick’s theoretical validity framework, validation research in both educational measurement and language testing is increasingly guided and informed by an argument-based approach to validity (e.g., Bachman, 2005; Bachman and Palmer, 2010; Chapelle, 2008, 2012; Kane, 2006, 2012, 2013), which also forms the theoretical underpinning of this PhD research. The argument-based approach builds on Toulmin’s (1958, 2003) model of practical argument, and formulates a logical structure to link available evidence to score-based inferences and actions. By doing so, the plausibility and appropriateness of the inferences and uses can be evaluated and examined, that is, “validated”. Generally, to link examinees’ test performance to the final score interpretations, a chain of intermediate inferences is needed. Each intermediate inference has its associated warrants and underlying assumptions which need to be supported by validity evidence. When concrete and robust evidence is provided to back up the underlying assumptions, the intermediate inference is warranted and test validators proceed to examine the succeeding inference. This process goes on until all intermediate inferences are supported with validity evidence. That is, at the end of the process, the score interpretations and the certification decisions are scaffolded by a chain of intermediate inferences that are warranted and reinforced by validity evidence.

The argument-based approach to validity is drawn upon to advance a theoretical validity argument for the ICPTs, which could serve as a roadmap to guide interpreting testers in collecting and marshalling validity evidence for the ICPTs. Specifically, as will be seen in Chapter 2, six intermediate inferences are proposed for the ICPT validity argument that link a

given interpreting practice domain, to candidates' test performance, to observed test scores, to generalized test scores, to localized score interpretations, to conclusive score interpretations, and finally to certification decisions. These inferences include 1) domain analysis and modeling, 2) evaluation, 3) generalization, 4) explanation, 5) extrapolation, and 6) utilization. According to Bachman and Palmer (2010), Chapelle (2012) and Kane (2006, 2012), although in principle robust validity evidence should be collected for each inference, the most vulnerable or weakest inference(s) deserves special attention and should be prioritized in a validation research program, as it is most likely to undermine the scaffolded structure of the validity argument. In other words, the validity argument is only as strong as its weakest link. The two weakest inferences or links for the ICPT validity argument have been identified from the available literature in Interpreting Studies: 1) the explanation inference, which has to do with using a strong theory, or a model of test construct, to account for test score explanations; 2) the generalization inference, which pertains to rater variability and its effects on score generalizability.

To strengthen the explanation inference, a construct model tailored for ICPTs needs to be articulated, and evidence for the score explanations be generated. Based on a literature review (see Chapter 3), it is found that while there have been three general approaches to construct definition (i.e., a trait, a behaviorist, and an interactionalist approach) in the field of language testing and assessment (e.g., Bachman, 2007; Chapelle, 1998; Messick, 1981), there seems to be no articulated construct models specifically tailored for ICPTs. Given that the three approaches to construct definition postulate different relationships between entity (i.e., the phenomenon under observation) and context (i.e., where the researcher makes the observation), and thus lead to different score-based inferences (Bachman, 2006), and also given that the trait and behaviourist-based score interpretations are preferred by interpreter certification bodies, I have used the "interactionalist" or the "socio-cognitive" approach (Bachman, 2007; Chalhoub-Deville, 2003; Chapelle, 1998; Douglas, 2000; Read & Chapelle, 2001; Young, 2011) to design an interactionalist construct model for ICPTs, which accommodates both trait- and performance-referenced interpretations of test scores.

Essentially, the interactionalist approach postulates that part of test performance could be referenced to traits, another part to context, and still another part to interactions between the two (e.g., interactions between interpreting ability and characteristics of interpreting tasks), in a various and arguable proportion. In other words, according to Chapelle (1998, p. 43),

“performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts”. This approach to construct definition therefore offers a way to infer from performance something about both a practice-specific behavior and a practice-independent, person-specific trait (Young, 2000, 2011).

I have described the interactionist construct model for ICPTs in Chapter 3. In general, the construct model consists of two major components: 1) unobservable trait of “interpreting ability”, an umbrella term for knowledge, strategies, and (meta-)cognitive processes involved in the interpreting process, and 2) characteristics of interpreting tasks that constitute the context in which interpreting practice is undertaken. In addition, these two components and their interactions have consequences for observable interpreting performance. To a large extent, it is the interplay between characteristics of interpreting tasks, interpreting ability, and interpreting performance quality that this PhD research seeks to investigate and understand, which ultimately and hopefully generates empirical evidence to contribute to the ICPT validity argument.

#### **1.4 Research scope, research purpose and research questions**

Although this research is dedicated to ICPT, its scope is limited in four aspects. First, the language pair involved in the research is limited to English/Chinese. This is because I have the background of English/Chinese interpreting, thus being able to study this specific language pair. Second, despite different modalities of interpreting, the research focuses on spoken-language interpreting, because I was only trained to be and practiced as a spoken-language interpreter. Third, despite various modes of interpreting, only simultaneous interpreting (SI) is of the concern. This is because SI constitutes a major part of the current interpreting practice in China where interpreters primarily work in various conference settings (e.g., Dawrant & Jiang, 2001). Such interpreting practice differs from that of immigration countries such as Australia, the UK and the US where dialogue interpreting as part of community interpreting practice is prevailing. Accordingly, for analysis of ICPT, the emphasis is placed on certification of English/Chinese SI practitioners. Fourth, although the ICPTs in different parts of the world (e.g., Australia, Canada, China, the UK, the US) are reviewed and compared in the research, a special focus is on the ICPTs developed in China, because this is

where certification of English/Chinese SI practitioners primarily takes place. Despite the limited scope of the research, it is expected that the research findings could be relevant to the ICPTs administered in other modes of interpreting, in other language pairs, and in other settings.

In addition, it is worth noting that I have not selected a specific English/Chinese ICPT (e.g., CATTI, NAETI) for analysis for two reasons: firstly, the topics examined (i.e., test validation and construct definition) are foundational in nature and general in scope. Thus, there is no need to restrict such topics to a single English/Chinese ICPT, although in Chapter 8 some of the research results were discussed in reference to the CATTI SI test; secondly, there is an agreed paucity of published research that investigates English/Chinese ICPTs in China (Feng, 2005; Huang, 2005), and not all the authentic ICPTs administered in the past have been made available to the general public. Given the lack of information, it is very difficult, if not impossible, to launch an informed discussion on specific ICPTs.

Against the theoretical and practical background provided in sections 1.2.2 and 1.2.3, the purpose of the research is to generate empirical evidence for the proposed interactionist construct model, and ultimately contribute preliminary evidence to the ICPT validity argument. Specially, in the interactionist construct model, although the components of SI task characteristics were theorized, there have been no empirical studies conducted to profile real-life interpreting practice in China (see details in Chapter 4). Information is lacking as to what the most frequently performed types of SI tasks are, and what their associated characteristics in the real-life practice are. Therefore, the first major research question (RQ) is:

**RQ 1** What are the characteristics of the real-life English/Chinese conference interpreting practice in China, upon which design of test tasks could be drawn, and from which validity evidence could be derived?

In addition, as the interactionist construct model postulates that interpreting performance quality is as a function of SI task characteristics, interpreting ability and their interactions, the second major RQ is:

**RQ 2** What is the possible interplay between SI task characteristics, interpreting ability, and SI performance quality?

Subsumed under RQ 2 are three sub-RQs, they are as follows:

RQ 2.1 What are the effects of SI task characteristics on SI performance quality?

RQ 2.2 What are the effects of SI task characteristics on strategy use (as a crucial part of interpreting ability)?

RQ 2.3 What is the relationship between strategy use and SI performance quality?

In addition, a substantial amount of empirical data in the research was rater-generated quantitative scores that sum up interpreters' SI performance in an experiment designed to address RQ 2. Given that rater and score reliability underlies experiment results and subsequent conclusions, this PhD research also investigates rater variability and score generalizability. Specifically, the research attempts to evaluate the utility of two modern psychometric models (i.e., multifaceted Rasch measurement and generalizability theory) as methodological alternatives to classical test theory (CTT) approach to rater and score reliability. Therefore, the third major RQ is:

**RQ 3** How multifaceted Rasch measurement and generalizability theory can be incorporated into interpretation testing and assessment to investigate rater severity/leniency and score generalizability?

Particularly relevant to the PhD research are the following sub-RQs, given the use of multiple raters to assess SI performance and the large amount of rater-generated data in the experiment:

RQ3.1 Did the raters recruited in the study differ in overall severity/leniency when assessing the recorded interpreting performance?

RQ3.2 Did the recruited raters consistently use rating scales overall in operational rating?

RQ3.3 Did the recruited raters maintain a uniform level of severity/leniency across the interpreters, the SI tasks, and rating criteria in operational rating?

RQ3.4 What would be the impact of increasing the number of SI tasks and/or raters on the dependability of information completeness, fluency of delivery, and target language quality ratings?

RQ3.5 What would be the impact of different weighting schemes proposed *a priori* on the composite score dependability?

RQ3.6 What would be the empirical contributions of differentially weighted rating scales to the composite score variance?

The rationales and details are described in the following section which explains how the RQs are connected with one another.

## **1.5 Research design**

Overall, a multi-phase mixed-methods research (MMR) design was used (Creswell, 2013), in which results from a previous study inform RQs and design of a subsequent study.

The following sections of 1.5.1 to 1.5.3 describe how empirical studies were designed to investigate the three RQs. These sections also outline key content and critical findings in each study that inform the subsequent RQs and studies. In addition, the links between the RQs and between the studies are illustrated in Figure 1.1, as shown below.

### **1.5.1 RQ 1: Profile conference interpreting practice in China**

Conference interpreting practice in China was empirically profiled, using an exploratory-sequential MMR design. The design was operationalized by a qualitative diary study (n = 11) to initially explore English/Chinese interpreting practice, followed by a quantitative survey designed on the basis of diary findings, and distributed to a larger cohort of interpreters (n = 140).

Among other important results, it was found that a large proportion of the SI practitioners had frequently encountered fast speech rate (FSR) and strong accent (StrA) that contributed to SI difficulty. The two task characteristics, namely FSR and StrA, were therefore chosen as two independent variables to be manipulated in the design of the third study to investigate and explore RQ 2.1 How do SI tasks characterized by FSR and/or StrA affect interpreting performance quality? RQ 2.2 How do the two task characteristics engage the components of interpreting ability? and RQ 2.3 What is the relationship between interpreting ability and observed performance quality?

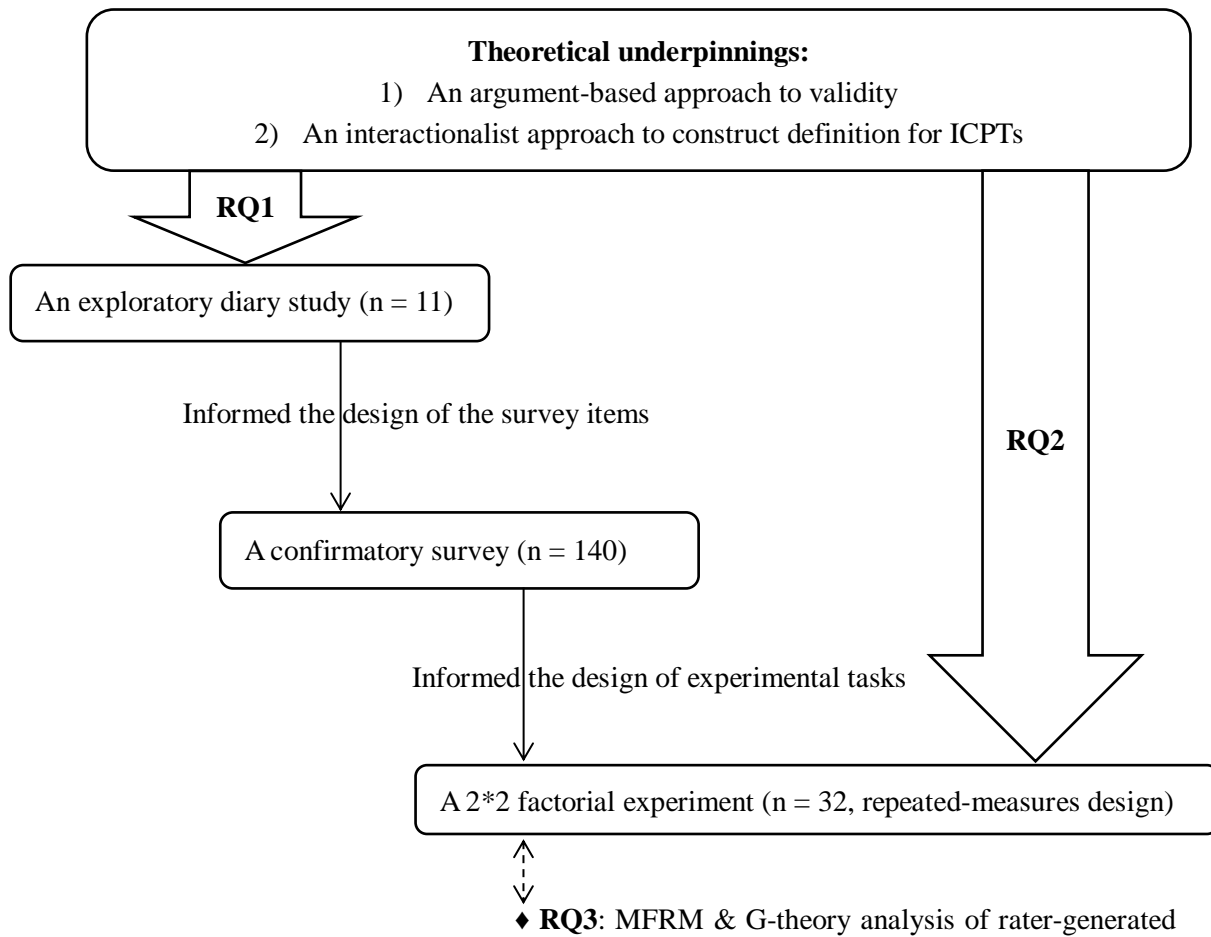


Figure 1.1 Linkages between research questions and studies

### 1.5.2 RQ 2: Exploring the interplay between task characteristics, interpreting ability and interpreting performance quality

An experiment was conducted to find evidence for the interactionalist model of construct definition. In other words, the interplay between the task characteristics represented by FSR and StrA (informed by the research results from the diary and the survey), strategy use (as a crucial component of interpreting ability), and SI performance quality were empirically investigated.

Specifically, a 2×2 factorial experimental design was used, in which 32 conference interpreters were recruited to perform English-to-Chinese SI in four tasks coded as Task<sub>SN</sub>, Task<sub>SA</sub>, Task<sub>FN</sub> and Task<sub>FA</sub>. The four tasks or treatment conditions (TCs) were produced by crossing a speed factor (consisting of two levels: fast & slow speech rates) with an accent factor (comprising two levels: accented and non-accented speeches). Four source speeches were developed and tailored to the four TCs. After completing each SI task, the interpreters were interviewed to reflect on their SI performance. After completing the four SI tasks, the

interpreters completed a questionnaire in which they used a seven-point Likert scale to rate the perceived overall difficulty of each task. All interpreters' SI performance and interviews were audio recorded with consent.

After the experiment, nine trained raters were recruited to assess each SI performance by each interpreter in each task, using three criteria including information completeness (InfoCom), fluency of delivery (FluDel) and target language quality (TLQual). Furthermore, two coders coded the interpreting strategies used during SI, based on the analysis of paralleled source- and target-language speech transcripts.

Drawing on the experiment data, the effects of the speed and the accent factors on interpreting performance quality measured by InfoCom, FluDel and TLQual was investigated based on a convergent-parallel MMR design (Creswell, 2013). Quantitative performance and perception data, and qualitative interview data were first analyzed individually, and then compared and triangulated to gain insight into the effects.

In addition, based on eight higher-achieving interpreters' SI performance in the experiment, a preliminary qualitative investigation was conducted to examine the effects of speech rate and accent on their strategy use in English-to-Chinese SI, and to explore the relationship between their strategy use and the quality of their SI performance.

### 1.5.3 RQ 3: Methodological exploration of modern measurement theory to examine rater/score reliability

Regarding the RQ 3, a methodological exploration was conducted to evaluate the utility of multifaceted Rasch measurement (MFRM) and generalizability (G) theory in investigating rater variability and score generalizability. To a large extent, the analysis of rater-generated scores in the experiment constitutes one of the important checks that are customarily taken by experimenters to ensure external validity (i.e., generalizability of experiment results, see Johnson & Christensen, 2012). Traditionally, classical test theory (CTT) has been mostly used, which is operationalized in the form of various inter-rater reliability coefficients. The application of MFRM and G theory therefore represents an extension of the conventional CTT approach. More importantly, one can regard the experiment as an interpreting performance test in which the interpreters performed SI, and their SI performance was subsequently assessed by the raters. Thus, the quantitative rater-generated data in the experiment can be regarded as simulated test scores. By demonstrating the application of

MFRM and G theory to this simulated test data, future researchers could be technically equipped to produce robust validity evidence to the generalization inference in the ICPT validity argument.

Specifically, MFRM was employed to provide detailed analysis of rater severity/leniency in the experiment. G theory was used to examine main and interaction effects of multiple assessment facets (e.g., raters, tasks) simultaneously, and to explore the effects of the number of raters/tasks used on score generalizability.

## **1.6 Thesis structure**

This thesis is written in a “thesis-by-publication” format, an encouraged and preferred practice at Macquarie University for higher degree research (HDR) students. According to *Macquarie University Higher Degree Research Thesis by Publication Guideline* (for the full guideline, please see Appendix A), a thesis by publication may include relevant papers (including conference presentations), which have been published, accepted, submitted or prepared for publication during a HDR student’s candidature. Although it is not necessary that the papers have actually been published at the time of thesis submission, each paper should be formatted in a publication-ready manner. As a general rule, theses by publication have between two and eight papers in combinations of sole and co-authored papers. The papers should form a coherent and integrated body of work, focusing on a key research question or a series of inter-related questions.

This PhD thesis consists of nine chapters, each of which reports on each component of the research project. Each chapter, except the Introduction (Chapter 1) and the Conclusion (Chapter 9), is a stand-alone and self-contained journal article. Most of the chapters have been presented in national and international conferences (Chapters 2, 3, 4, 7 and 8). In addition, some chapters have been accepted by peer-reviewed journals (Chapters 4 and 7), and the others are currently under peer review at different stages (Chapters 2, 3, 5, 6 and 8). Relevant publication details are specified in a footnote at the beginning of each chapter.

This chapter (Chapter 1) introduces and contextualizes the research, and clarifies the connections between chapters, and between studies.

Chapter 2 and Chapter 3 provide theoretical underpinnings for the subsequent empirical studies. Specifically, in Chapter 2 given the disjunction between the ubiquity of ICPTs as a

gate-keeping tool for the interpreting profession and the lack of rigorous and systematic validation of ICPTs, an argument-based approach to validity was drawn upon to build a validity argument for ICPTs. In the validity argument, a chain of six intermediate inferences scaffolded one another, along with their respective warrants and underlying assumptions, to justify and legitimate the intended score interpretations and uses. However, the explanation and the generalization inferences appeared to be the most vulnerable links based on interpreting literature. In Chapter 3, an interactionalist approach to construct definition was therefore proposed and articulated for English/Chinese ICPTs to help enhance the explanation inference. As one of three predominant approaches to construct definition in the field of language testing and assessment, the interactionalist approach accounts for performance consistency by incorporating both trait theorists' and behaviorists' points of view in a single model. As a result, the approach accommodates both trait and performance-referenced score interpretations. The theoretical construct model consists of two major components: SI task characteristics and interpreting ability, and hypothesizes interplay between SI task characteristics, interpreting ability and interpreting performance, which needs to be borne out by empirical data.

Chapters 4, 5, and 6 report on the empirical studies conducted to examine the interplay between task characteristics, interpreting ability, and quality of interpreting performance, which ultimately contributes empirical evidence to the ICPT validity argument. All the empirical studies were approved by Macquarie University Human Ethics Committee (see Appendix B). Specifically, Chapter 4 reports on the diary study and the follow-up survey that sought to address **RQ 1** (i.e., profiling the characteristics of conference interpreting practice in China). The diary study collected the following information: What are conference-related materials received by interpreters for preparation? What preparation techniques do interpreters use for conference preparation? and What are the characteristics of SI tasks (e.g., duration of an interpreting turn, directionality, difficulty factors)? Qualitative diary results were then used to inform the design of the survey. The survey obtained three general types of information on conference preparation, frequency of different SI task types performed, and contributing factors to SI task difficulty.

Chapter 5 addresses RQ 2.1, namely, the effects of task characteristics (represented by FSR and StrA) on interpreting performance quality, based on data from a factorial experiment. Quantitative and qualitative results from the experiment were described separately, and then triangulated to gain an in-depth understanding of the effects on SI performance quality.

Chapter 6 explores preliminary answers to RQ 2.2 and RQ 2.3, namely, how FSR and StrA engage and elicit strategy use, and what is the relationship between strategy use and SI performance quality, based on a subset of the experiment data.

As a methodological exploration, and as a necessary check on the external validity of the experiment results, Chapters 7 and 8 attempt to address **RQ 3**. Particularly, Chapter 7 analyzes the rater-generated scores from the experiment to identify possible rater variability (particularly rater severity) by drawing upon multifaceted Rasch measurement (MFRM).

Similarly, Chapter 8 examines the effects of rater variability on score dependability by applying generalizability (G) theory. The quantitative dataset used for analysis was also derived from the experiment. Hopefully, Chapter 7 and Chapter 8 would deepen interpreting testers' repertoire of analytic skills to investigate rater variability and score generalizability in the operational ICPT. The two chapters would also contribute to strengthening the ICPT validity argument, particularly the evaluation and the generalization inferences.

Finally, Chapter 9 summarizes and integrates findings from each individual study of this PhD research, links empirical and methodological findings to the validity argument, points out strengths and weaknesses of the research, suggests implications of the findings, and provides recommendations for further research.

## **1.7 Potential contribution of the research**

The present research aims to build preliminary theoretical and methodological foundations to scaffold ICPT development and empirical test validation. Specifically, the research has the potential to contribute to three important aspects of rater-mediated ICPT: validation of ICPTs, construct definition that informs the design of ICPTs, and analysis of the rater effects and score dependability.

Firstly, Chapter 2 represents one of the most comprehensive attempts to articulate a validity argument to guide and inform validation of ICPTs in the field of Interpreting Studies, drawing upon an approach tried and tested in language testing and assessment. The validity argument is expected to expand ICPT testers' knowledge of test validation, and to broaden the range of validity evidence to be collected.

Secondly, Chapters 3, 4, 5 and 6 articulate, flesh out and empirically test a theoretical construct model proposed for ICPT, which represents one of the first dedicated attempts in

Interpreting Studies to define the test construct for the English/Chinese ICPTs. Especially, in Chapter 4 first-hand and valuable data describing the fundamentals of the interpreting practice in China was gathered.

Thirdly, rater variability and its effects on score reliability have been investigated primarily through classical test theory (CTT) in numerous studies on interpreter performance assessment (e.g., Lee, 2008; Wu, 2010). Chapters 7 and 8 extend that body of work to apply modern measurement theory, particularly Rasch analysis and G theory, in an attempt to help interpreting testers gain in-depth understandings of rater-related measurement error and its effects on score reliability.

## 1.8 References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Angelelli, C. (2009). Using a rubric to assess translation ability: Defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp. 13-47). Amsterdam: John Benjamins.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.
- Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165-207). Amsterdam: John Benjamins.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. Doe (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41-71). Ottawa: University of Ottawa Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.
- Campbell, S., & Hale, S. (2003). Translation and Interpreting Assessment in the Context of Educational Measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing*, 20(4), 369-383.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70), Cambridge, UK: Cambridge University Press.
- Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-352). London: Routledge.
- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple...*Language Testing*, 29(1), 19-27.
- Chen, J. (2002). 从 Bachman 交际法语言测试理论模式看口译测试中的重要因素. 中国翻译. Fundamental considerations in interpreting testing. *Chinese Translator Journal*, 1, 51-53.
- Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, 3(2), 257-273.
- Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting*, 7(1), 97-13.
- Creswell, J. W. (2013). *Research design: Qualitative, Quantitative and Mixed Methods Approaches* (4th ed.). Thousand Oaks, CA: Sage.
- Cronbach, L. J. (1971). Test validation. In R.L. Thorndike (Ed.) *Educational measurement* (2nd ed.) (pp. 443-507). Washington, D.C.: American Council on Education.
- Dawrant, A., & Jiang, H. (2001). *Conference interpreting in Mainland China*. Retrieved from [http://www.aiic.net/ViewPage.cfm?page\\_id=365](http://www.aiic.net/ViewPage.cfm?page_id=365)
- Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge, UK: Cambridge
- Feng, J. Z. (2005). 论口译测试的规范化. [Towards the standardization of interpretation testing]. 外语研究, 89, 54-58.

- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385-398.
- Hale, S., Garcia, I., Hlavac, J., Kim, M., Lai, M., Turner, B., & Slatyer, H. (2012). *Development of a conceptual overview for a new model for NAATI standards, testing and assessment*. Sydney, Australia. Retrieved from: <http://www.naati.com.au/PDF/INT/INTFinalReport.pdf>
- Hlavac, J. (2013). A cross-national overview of translator and interpreter certification procedures. *The International Journal for Translation & Interpreting Research*, 5, 32-65. DOI: 10.12807/ti.105201.2013.a02
- Huang, M. (2005). 谈口译资格认证考试的规范化设计. [Toward a more standardized large-scale accreditation test for interpreters]. *中国翻译*, 6, 62-65.
- Johnson, B., & Christensen, L. (2012). *Educational Research: Quantitative, qualitative and mixed approaches* (4th ed.). Thousand Oaks: Sage.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, 29(1), 3-17.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.
- Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, 2(2), 165-184.
- Liu, M. H. (2013). Design and Analysis of Taiwan's Interpretation Certification Examination. In D. Tsagari & R. van Deemter (Eds.), *Assessment Issues in Language Translation and Interpreting* (pp. 163-178). Frankfurt am Main: Peter Lang.
- Mackintosh, J. (2006). Professionalization: Conference Interpreting - a new profession. In M. J. Chai & A. L. Zhang (Eds.), *Professionalization in Interpreting: International Experience and Developments in China* (pp. 2-14). Shanghai: Shanghai Foreign Language Education Press.
- Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, 89(3), 575-588.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.

- National Accreditation Authority for Translators and Interpreters. (2014). *Accreditation by Approved Australian Course: Information Booklet*. Retrieved from [http://www.naati.com.au/PDF/Booklets/Accreditation by Approved Australian Course booklet.pdf](http://www.naati.com.au/PDF/Booklets/Accreditation%20by%20Approved%20Australian%20Course%20booklet.pdf)
- National Accreditation Authority for Translators and Interpreters. (2015). *Accreditation by Overseas Qualification, Professional Association Membership or Advanced Standing: Information Booklet*. Retrieved from [http://www.naati.com.au/PDF/Booklets/Accreditation by Assessment OSQualification ProfessionalAssociationMembership AdvancedStanding booklet.pdf](http://www.naati.com.au/PDF/Booklets/Accreditation%20by%20Assessment%20OSQualification%20ProfessionalAssociationMembership%20AdvancedStanding%20booklet.pdf)
- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1-32.
- Roat, C. E. (2006). *Certification of Health Care Interpreters in the United States: A Primer, a Status Report and Considerations for National Certification*. Los Angeles, USA, Retrieved from: [http://www.calendow.org/uploadedFiles/certification of health care interpreters.pdf](http://www.calendow.org/uploadedFiles/certification_of_health_care_interpreters.pdf)
- Sawyer, D. B. (2000). Towards meaningful, appropriate, and useful assessment: How the false dichotomy between theory and practice undermines interpreter education. *ATA Chronicle*, 29(2), 32-40.
- Sawyer, D. B. (2004). *Fundamental aspects of interpreter education: Curriculum and Assessment*. Amsterdam & Philadelphia: John Benjamins.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Toulmin, S. E. (2003). *The uses of argument* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Wu, S. C. (2010). *Assessing Simultaneous Interpreting: A study on Test reliability and Examiners' assessment behavior*. (PhD thesis, Newcastle University upon Tyne, UK). Retrieved from <https://theses.ncl.ac.uk/dspace/bitstream/10443/1122/1/Wu%2011.pdf>
- Young, R. F. (2000). *Interactional competence: Challenges for validity*. Retrieved from [http://www.english.wisc.edu/rfyoung/IC\\_C4V.Paper.PDF](http://www.english.wisc.edu/rfyoung/IC_C4V.Paper.PDF)
- Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). London and New York: Routledge.

## **An introductory note to Chapter 2**

As can be seen in Chapter 1, interpreter certification performance tests (ICPTs) are developing rapidly in different parts of the world. Certification organizations rely on ICPT scores to make the decision of awarding a certificate to would-be interpreters. Despite the high-stakes nature of the ICPTs, it seems that there has been no systematic validation research conducted to ascertain the reliability and validity of score-based inferences and uses. One of the possible reasons for the lack of validation research is that there has been no transparent and practical guidance in the field of Interpreting Studies to inform test validation. Against this background, Chapter 2 sets out to explore an argument-based approach (ABA) to build a validity argument for ICPTs.

The ABA has gained currency in the field of educational measurement, and language testing and assessment in particular. It represents an effective response to traditional unsystematic and sometimes opportunistic collection of various types of validity evidence. The ABA also represents a principled approach to operationalizing Messick's theoretical framework of construct validity.

Specifically, Chapter 2 tracks the evolution of validity theory over the past 100 years, highlights conceptual foundations for the ABA, reviews recent applications of the ABA in language testing and assessment research, foregrounds score-based inferences and actions in ICPT, and culminates in a theoretical ABA-based roadmap that guides rigorous and systematic validity investigation into ICPTs.

## Chapter 2 Building a validity argument for interpreter certification performance testing<sup>5</sup>

***Abstract.** Over the past decade, interpreter certification performance testing has gained momentum. Interpreter certification tests are often a high-stakes test which acts as gatekeepers to professional practice. The certification decision is ordinarily based on test scores. Testing bodies make inferences about examinees' knowledge, skills and abilities, as well as their interpreting performance in a given target domain based on these test scores. To justify the appropriateness of the score-based inferences and actions, test developers need to provide validity evidence. However, a systematic approach to validation is lacking in Interpreting Studies, largely due to the absence of a methodical validation framework. In an attempt to redress this problem, this paper proposes a theoretical argument-based validation framework for interpreter certification performance tests (ICPTs) that can serve as a roadmap for interpreting testers in conducting rigorous validity investigation. Before presenting the framework, a critical literature review of validity theory and a brief examination of the argument-based approach to validation are provided. A validity argument for ICPTs is then proposed with exemplification based on the available literature. Finally, implications of the framework are discussed and future studies suggested.*

### 2.1 Introduction

The practice of interpreter certification performance testing (ICPT) has gained momentum over the past decade. This momentum can be seen not only in countries of immigration (e.g., Australia, the United States) where interpreters have traditionally played a critical role in educational, legal and medical settings (Angelelli, 2007; Ra & Napier, 2013), but also in economically emerging countries such as China where interpreters are needed to meet growing demand (e.g., Dawrant & Jiang, 2001; Setton, 2009). This momentum can be characterized by

---

<sup>5</sup> Part of this chapter was presented in the 7<sup>th</sup> American Translation and Interpreting Studies Association (ATISA) Conference at New York University, New York City, USA, April 3-5, 2014, and also at the 12th Annual Conference of the European Association for Language Testing and Assessment (EALTA), at the University of Copenhagen, Denmark, May 28-31, 2015. A revised version of the chapter is under the 3<sup>rd</sup> round of review in the journal of *Interpreting* as: Han, C., & Slatyer, H. (under review). Test validation in interpreter certification performance testing: An argument-based approach. Helen Slatyer's contribution to this paper was reviewing the writing and providing feedback on drafts.

at least three types of development. The first development pertains to building new certification systems. For example, the *Qualitas Project* in Europe (Giambruno, 2013) for the assessment of legal interpreting quality through testing and certification; the *Language and Interpreting Testing Project* (Angelelli, 2007) to evaluate medical interpreters in the United States and the project to establish an interpreter assessment standard in Taiwan (Liu, 2013). The second development has to do with the expansion of current interpreter certification testing programs. For instance, since its inception in 2003, the China Accreditation Test for Translators and Interpreters (CATTI) has grown rapidly in terms of the number of test candidates. According to the China News Service,<sup>6</sup> 50,000 candidates registered for CATTI's tests in 2012, making it one of China's largest testing programs. In the US, the National Board of Certification for Medical Interpreters' certification program has also experienced an increased demand for testing, spawning a network of testing centers across the country (Arocha & Joyce, 2013). The third development concerns the quality assessment of interpreter certification procedures. For example, Australia's National Accreditation Authority for Translators and Interpreters (NAATI) has completed its first-phase *Improvement to NAATI Testing (INT) Project*, with 17 recommendations for enhancement proposed by a panel of experts (Hale et al., 2012), and the US National Board of Certification for Medical Interpreters has provided psychometric evidence in a report to support score-based inferences for test candidates passing the National Medical Interpreter Certification Exams (PSI Services LLC, 2010). Taken together, these developments represent a growing awareness of the important role interpreter certification testing plays in the provision of quality language services in modern societies.

Interpreter certification tests are often high-stakes tests performing a gatekeeping function for professional practice. Test scores are interpreted to describe test candidates' interpreting performance as well as the knowledge, skills, abilities and strategies (KSASs) required of an interpreter,<sup>7</sup> and more importantly are used to make certification decisions.<sup>8</sup> That is, only those candidates who outscore a cut-off point will become certified interpreters. The certification decisions could produce consequential washback effects on multiple groups of relevant

---

<sup>6</sup> See [http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget\\_edu\\_jsp](http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget_edu_jsp)

<sup>7</sup> To interpret a test score is to explain meaning of the score. For example, test scores are usually interpreted as an indicator of what test takers know and can do or their KSASs. In testing and assessment literature, "test score interpretation", "score-based inferences" and "score meaning" can be used interchangeably.

<sup>8</sup> Use of test score interpretations refers to how testers and other stakeholders make decisions (e.g., admission, selection and certification) based on test scores and their interpretations. Usually, "test score use" and "score-based action" can be interchanged.

stakeholders (e.g., Vermeiren, Gucht, & De Bontridder, 2009). On the one hand, the certification decisions impact on interpreters' job prospects and potential livelihood. This is because marketplaces may demand a credible demonstration of interpreting ability such as certification from interpreter candidates before offering them employment (Hlavac, 2013). In China, although professional certification has not yet taken on a gate-keeping role in the interpreting market, there is a trend towards an interpreting certificate, particularly that of CATTI, which provides a competitive edge for its recipients, and constitutes a basic requirement in job applications (Yu, 2005). In other countries such as Australia and the US, interpreters are required to obtain recognized certification for work in government departments (e.g., police stations, immigration offices, hospitals) or legal institutions (e.g., state/federal courts). On the other hand, the certification decisions serve other stakeholders, notably consumers of interpreting services. For instance, according to Jacobs et al. (2001), the use of *ad hoc* interpreters (untrained and unqualified bilinguals) in medical settings appears to have negative clinical consequences, whereas professional interpreter services can improve the delivery of health care to limited-English-speaking patients, thus ultimately protecting public welfare.

Given the high-stakes of interpreter certification tests, certification bodies have an ethical and social obligation to ensure that score-based inferences and actions are valid and justifiable. In other words, high-stakes interpreter certification tests should be subjected to regular evaluation through a rigorous research process (Hale et al., 2012). One important component of the evaluation is "test validation" (Cronbach, 1971), in which test developers provide robust validity evidence to link candidates' test performance to test score inferences (e.g., claims about test takers' interpreting ability and predictions about their future performance in a real-life practice domain), and ultimately to certification decisions made by certifying bodies (Bachman, 1990; Kane, 1992, 2006; Kane, Crooks, & Cohen, 1999). Unless this evidential support is provided, there is no other way to sustain intended test score interpretations, and to justify final decisions.

Despite the pivotal role that validity evidence plays in justifying test score interpretations and uses, validation research involving the ICPT is still in its infancy. According to Sawyer (2004), in the context of educational interpreter assessment, an in-depth discussion of validity issues has not yet been conducted. Clifford (2005) also points to the dearth of validation

research on interpreter certification tests (i.e., analysis of psychometric properties). Notable exceptions include the work of Angelelli (2009), Clifford (2005) and PSI Services LLC (2010).

In an attempt to understand the reasons for the paucity of much-needed validation research, it is worth reflecting on possible contributing factors. One of the factors appears to be the lack of a clear and concise account of recent developments in validity theory and the related approaches to test validation. Another contributing factor could be the lack of practical guidance on how to conduct rigorous validation research, particularly in Interpreting Studies.

Against this background, in this paper we propose an assessment use argument (AUA), drawing upon an argument-based approach to validation (Bachman & Palmer, 2010; Chappelle, 2008; Kane, 1990, 2006), in the hope of informing and assisting developers of interpreter certification performance tests (ICPTs) to undertake rigorous validation research. Although the ICPTs may include a different combination of interpreting tasks (e.g., dialogue interpreting, consecutive interpreting with note-taking, sight interpreting, simultaneous interpreting), depending on the different domains of interpreter certification (e.g., international conferences, legal, medical, public services settings), the proposed AUA could be adapted to guide validation research in different contexts. In other words, the AUA represents an approach, or a strategy; it is not a tactic used for a particular ICPT.

Before describing the validity argument for the ICPTs, a concise review of major validity theories and associated validation methods from the fields of psychological and educational measurement is provided below. Recent test validation developments in the field of educational and language testing, particularly the argument-based approach to validity, are then described, followed by a brief review of some of the ICPTs for which information is publicly available, concentrating on the domains of interpreter certification, test methods, and intended score interpretations and uses. Finally, based on the review of the ICPTs, a validity argument for ICPTs is developed and described in detail.

## **2.2 The evolving nature of the validity concept and validation methods**

Validity theory has evolved gradually over the past century (e.g., Anastasi, 1986; Cronbach, 1989; Kane, 2006; Messick, 1988). The concept of validity was initially viewed as an absolute (all-or-nothing) property residing in test scores, as in the case of “criterion validity”, and was manifested in different validities (e.g., criterion, content, construct validities).

However, current theorists believe that validity is a matter of degree, and that it relates to construct validity as a unitary concept pertaining to the adequacy and appropriateness of score-based inferences and actions. As a consequence of the evolution of concepts of validity, major approaches to the validity investigation have also evolved over the years, reflecting the changing focus on different conceptualizations of validity. As a result, what is validated is not *tests or test scores*, but *score-based inferences and actions* (Bachman, 1990; Kane, 1992, 2006; Messick, 1989, 1994). In the following section, the evolving conceptualizations of validity over the previous century are outlined and summarized in Table 2.1.

### 2.2.1 Criterion validity: Correlation-based approach

As a rudimentary validity theory emerging in the early 1900s, criterion validity incorporates both concurrent and predictive validities (Lissitz & Samuels, 2007). At a time when the criterion validity model was primarily employed, validation was largely about establishing a relationship between a test (usually multiple-choice tests) and other similar criterion measures by calculating correlation coefficients between two sets of scores. However, this approach to validity posed two major problems:

- 1) The application of this model necessitates a cogently defined and demonstrably sound criterion measure (Kane, 2001, 2004). But this ideal measure is difficult to define in reality.
- 2) Even though a criterion measure is identified, there is still a need to evaluate the validity of the criterion measure, which implies a tricky loop of endless validation. (Ebel, 1961; Kane, 2001).

### 2.2.2 Content validity: Judgment-based evaluation

Between the 1940s and 1950s, influenced by behaviorism (e.g., Skinner, 1945), content validity emerged as a topic for discussion (Lissitz & Samuels, 2007). This validity model concerns two facets – content relevance and representativeness, and fared well in achievement tests that contain the behavioral domains of interest (e.g., skills, acquired knowledge). Test validation based on the content validity model was referred to as evaluation or consensual judgments by subject matter experts (SMEs) of how well test content samples different situations or subject matter about which conclusions are to be drawn (American Psychological

Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 1966; Cronbach, 1971). The most prominent weakness of relying on content validity as a sole source of validity evidence is that as a fixed property of test instruments content validity may degrade over time due to eroded test representativeness caused by an evolving target domain (Cronbach, 1971).

Table 2.1 Conceptualizations of "validity" and associated validation methods

Period	"Validity" label	Conceptualization	Validation method
<b>Early 1900s,</b> (e.g., Thurstone, 1932)	Criterion validity	1) Concurrent, 2) Predictive.	Correlation with criterion measures
<b>1940s to 1950s,</b> (e.g., Gulliksen, 1950)	Content validity	1) Content relevance, 2) Representativeness.	Expert-judgment about test content (sampling /representativeness)
<b>1954/1955,</b> (e.g., Cronbach & Meehl, 1955)	Construct validity	1) Underlying attribute, 2) Nomological theories.	Provision of internal and external evidence
<b>1950s to 1970s,</b> (e.g., Guion, 1980)	Trinitarian doctrine (validity)	1) Criterion, 2) Content, 3) Construct.	A toolkit approach (using the most available evidence)
<b>1980s to present,</b> (Messick, 1989)	Unified construct validity	1) Content, 2) Substantive, 3) Structural, 4) Generalizability, 5) External, 6) Consequential.	Accumulation of multiple lines of validity evidence; refutation of rival hypotheses.

### 2.2.3 Construct validity: An alternative approach

Although the term “construct” can be traced back to MacCorquodale and Meehl (1948), it is in Cronbach and Meehl’s (1955) seminal paper that the concept of construct validity was first elaborated. According to them, construct validity was ordinarily studied when a given trait underlying the test was of central concern, rather than test performance and test scores. This conceptualization indicates that construct validity was initially proposed as a back-up or alternative strategy to criterion and content models. In Cronbach and Meehl’s conceptualization, construct validity pertained to not only measurement of the trait, quality or construct in question, but also the development of a theory or a nomological net that relates

the latent construct of interest to its observed variables and other construct(s). Consequently, validation research needs to focus on collecting internal sources of validity evidence (e.g., studies of construct and of underlying processes) and studying external relationships (e.g., hypothesized connections of a construct with other constructs).

#### 2.2.4 Trinitarian doctrine: A toolkit approach

Between the 1950s and 1970s, criterion (predictive and concurrent), content and construct validities emerged as what Guion (1980) called a trinitarian doctrine to approach validity. In its heyday, this model provided three different paths to validity, also known as a “toolkit approach” (Kane, 2001, 2004). However, in the validation practice using this approach, a serious problem emerged: opportunistic choice of validity evidence (Guion, 1977; Kane, 2001, 2004; Messick, 1975, 1981). That is, practitioners opted for easiest and most available type of validity evidence, even though the chosen evidence may have had dubious value. In addition, it is worth noting that during this period, a major shift was under way. It was no longer a test or test scores that needed to be validated, but the interpretation of test scores (Cronbach, 1971). In other words, construct validity, reliability and predictive validity were no longer considered to reside in *tests* per se, but were instead properties of test responses (Messick, 1975).

#### 2.2.5 Construct validity as a unitary concept: An integrated approach

As early as the 1950s, some validity researchers suggested that construct validity could be a pervasive concept (Cronbach & Meehl, 1955) and should be an overriding concern in validity theory (Loevinger, 1957). These emerging ideas culminated in Messick’s seminal work – a “Validity” chapter in the third edition of *Educational Measurement* (Linn, 1989). Messick (1989) eloquently argued for construct validity as a unitary concept and as a unifying force integrating all aspects of validity.

Overall, Messick (1989, p. 13) regarded validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment”. More importantly, he highlighted six distinguishable aspects of construct validity as a means of addressing central issues implicit in the notion of validity (Messick, 1989, 1994), including

content, substantive, structural, generalizability, external, and consequential aspects of construct validity.<sup>9</sup>

Consequently, Messick (1989) treated test validation as a process of marshalling multiple lines of evidence, theoretical and/or empirical, to examine the degree of consonance between test scores and interpretive inferences, and of refuting possible rival inferences. The six aspects of unified construct validity could be gathered as different types of evidence to support test score interpretations and uses.

Despite Messick's integrated treatise on construct validity, some scholars have expressed reservations. The first source of concern relates to the impracticality of Messick's validity framework. The integrative construct validation is felt by some to be too abstract and demanding, reducing validation practice to an elusive goal (Kane, 2001, 2004, 2006; Shepard, 1993). The second source of discontent is that although Messick suggests a rich array of evidence for backing up validity claims, he does not provide detailed and concrete guidance on how to conduct validation research and how to prioritize different types of validity evidence (Bachman, 2005; Brennan, 1998; Kane, 1990, 2004, 2006). Messick's emphasis on tactics and negligence of strategies has been a stumbling block to the application of the unified validity theory to testing programs.

### **2.3 An argument-based approach to validity**

In response to problems associated with the application to validation, recent developments in validation practice focus on organizing and prioritizing validity evidence in a logical way. In the fields of educational measurement (Kane, 1992, 2006; Kane et al., 1999; Mislevy, Steinberg, & Almond, 2003) and language testing (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, 2008, 2012; Chapelle, Enright, & Jamieson, 2010), an argument-based approach to validity investigation has gained momentum. This approach explicitly draws upon a basic argument structure and formulates a conceptual template to logically link available data or evidence to score-based inferences and actions. By doing so, the plausibility and appropriateness of score interpretations and uses can be evaluated and examined, that is, "validated". At the core of the argument-based approach is a macro-structure of practical arguments such as that elucidated by Toulmin (1958, 2003). Toulmin's model of argument

---

<sup>9</sup> For details of unified construct validity, please refer to Messick (1989).

structure has provided a foundation for validation methodology and has been influential in educational and language assessment (Kunnan, 2010).

### 2.3.1 Toulmin's argument structure: Foundation building

In general, for an argument to succeed, good justification needs to be provided to support the claim of central concern. To facilitate the analysis of arguments, Toulmin (2003) proposes six key components (see Figure 2.1):

- 1) Claim: a statement whose merits must be established.
- 2) Qualifier: usually a hedge word and an expression that limits the strength of the claim or proposes a condition where the claim does not hold true.
- 3) Data: boosts the plausibility of a qualified claim, “data” (i.e. facts or evidence) needs to be collected. The arrow extending from the data to the claim represents one step of reasoning or an inference.
- 4) Warrants: statements that support the inference and authorize forward movement from the data to the claim.
- 5) Backing: in case the warrants themselves are questioned, backing evidence is accumulated to support and certify the statements expressed in the warrants.
- 6) Rebuttals: unless the qualified claim survives challenge from counter-arguments known as “rebuttals”, confidence in making the claim will be weakened and its plausibility reduced.

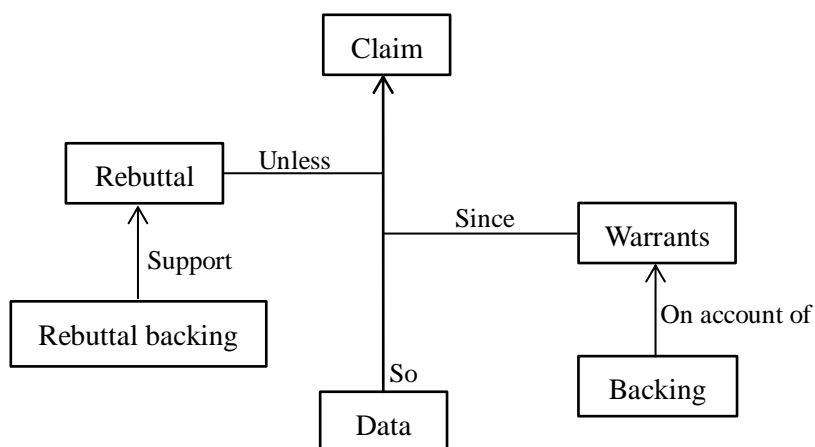


Figure 2.1 Toulmin's model of argument, based on Toulmin (2003)

An example can be used to illustrate the argument structure: according to AIIC membership admission requirements, interpreters who are members of this organization are competent professional practitioners (warrant). Since Jack was admitted to the organization as a member (data), and unless he has retired from the interpreting profession (rebuttal), it can be concluded that Jack is a competent professional interpreter (claim). In addition, concrete backing evidence (e.g., documentation of working experience, sponsorship letters by AIIC members) can also be provided to support the warrant. Because of this structure and associated components, Kunnan (2010, p. 185) observes that the argument is “expected to unfold strengths, weaknesses and limits of claims”.

### 2.3.2 An argument-based approach to validation in educational and language testing

In test validation, a claim is typically about what test candidates know and can do based on their test scores. In other words, a claim is essentially test score interpretations. To support the plausibility of the claim, pertinent evidence needs to be gleaned. Drawing upon Toulmin’s basic model, Kane (1992, 2004, 2006) proposes a network of inferences to develop interpretive and validity arguments that link observed test performance to the final use of test scores. In addition, Bachman (2003, 2005) and Bachman and Palmer (2010) have enhanced Kane’s approach by specifying and accentuating an assessment utilization argument that explicitly challenges the appropriateness of score-based actions. Furthermore, Chapelle (2008) and Chapelle et al. (2010) inherit Kane’s inferential network and Bachman’s emphasis on justification of score uses, and further introduce a new inference known as “domain description” to precede Kane’s proposed network of inferences. These additions to Toulmin’s argument structure are briefly discussed below.

#### 2.3.2.1 *Kane’s interpretive argument and validity argument*

In Toulmin’s basic argument model, there is only one step of reasoning or an inference linking a ground or available data to an ultimate claim, which serves well as a heuristic tool but fails to accommodate real-life complexity. In test validation for educational assessments, Kane (1992, 2006) points out that going from test scores (i.e., pure numbers) to meaningful descriptions of test takers (i.e., score interpretations) entails more than one step of reasoning or what he calls “a chain of inferences”. In other words, a number of “intermediate steps” (Kane, 1990, p. 14) are required to bridge what test developers observe on a test to how they

explain test scores and finally to how they use score-based inferences (i.e., decisions). This chain of inferences is illustrated in Figure 2.2. As shown in the figure, five specific inferences or steps are believed to appear regularly in test validation (Kane, 1992).

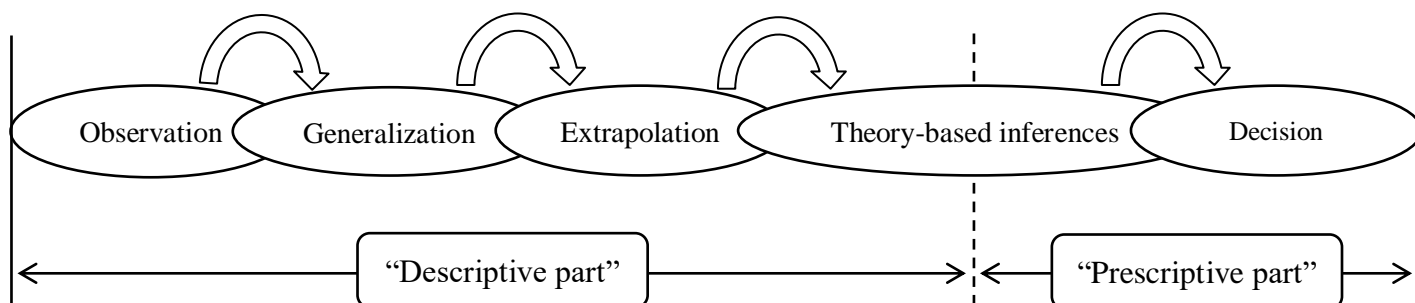


Figure 2.2 A chain of inferences linking data to score interpretations and uses, based on Kane (1994)

Each of these interlocked inferences has its underlying assumptions. Altogether they represent an *interpretive argument* (see Figure 2.3). In his later work, Kane (2001) observes that the interpretive argument can be divided into two parts (see Figure 2.2): 1) a descriptive part that links observed test performance to descriptive interpretations about test candidates' KSASs; and 2) a prescriptive part that involves decision-making based on descriptive interpretations.

To reinforce the plausibility of an *interpretive argument*, the chain of inferences and their associated assumptions need to be supported by substantial and concrete evidence (Kane, 1990, 2004). This creates a *validity argument*, as shown in Figure 2.3. While the interpretive argument aims to sanction inferences from data to claims, the validity argument tries to marshal concrete evidence to underpin the interpretive argument. Consequently, the validity argument serves to provide a rationale for accepting the interpretive argument, and ultimately, for accepting score interpretations and uses.

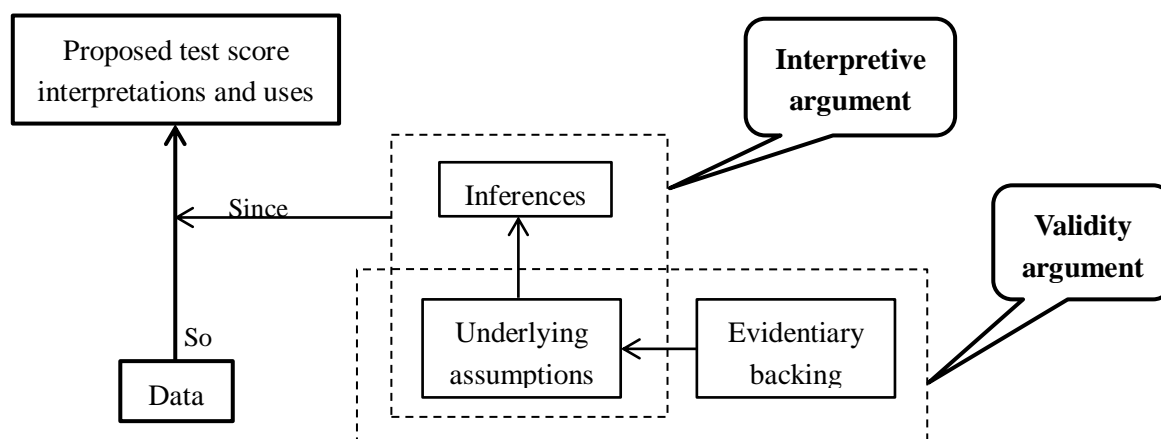


Figure 2.3 Relationship between an interpretive and a validity argument, based on Kane (2006)

### 2.3.2.2 Bachman's assessment use argument

Although Kane's validation framework represents an effective response to the lack of a methodical and systematic approach to generating and collecting validity evidence, Bachman (2005) finds that in language testing the possibility of misusing or using score-based inferences inappropriately to make decisions (e.g., certification, selection, and admission) is very real, because 1) even valid test score interpretations cannot be guaranteed to be relevant, useful and sufficient for intended uses or decisions; and 2) there is also no guarantee that these interpretations will not be subverted for other unintended uses. Given this consideration, there is a genuine need to articulate not only an argument for appropriate interpretation of test scores, but also an argument for justified use of valid score interpretations. As a result, Bachman (2003, 2005) and Bachman and Palmer (2010) develop an assessment use argument (AUA) for language tests and assessments. Particularly, as can be seen in Figure 2.4, the AUA explicitly articulates a two-part structure: 1) a *validity argument*,<sup>10</sup> which links test takers' performance on an assessment to proposed score interpretations (via a chain of inferences); and 2) an *assessment utilization argument*, which connects score interpretations to decisions.

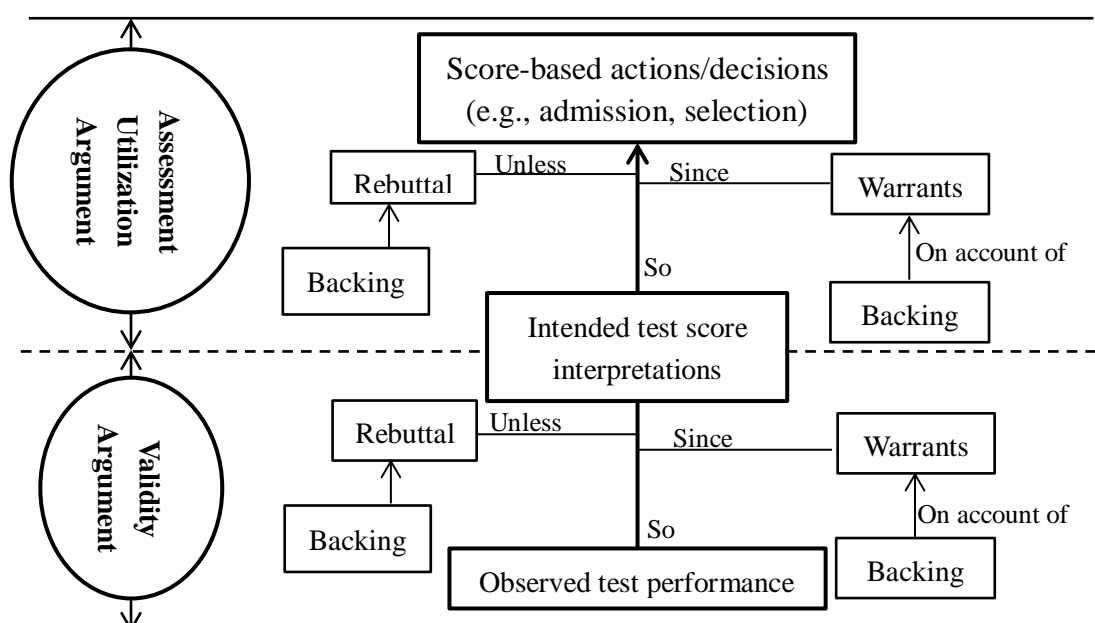


Figure 2.4 The structure of assessment use argument (AUA), based on Bachman & Palmer (2010)

<sup>10</sup> A "validity argument" in the AUA is defined differently from that of Kane's. However, the paper uses the term of "validity argument" in accordance with Bachman (2005) as well as Bachman and Palmer (2010).

In addition to benefits associated with the argument-based approach to validity, the AUA is especially explicit and pragmatic in terms of its approach to accentuating and supporting two methodologically related yet conceptually distinct arguments for score-based interpretations and actions, respectively. This double-tiered structure assists test validators in not only deliberately considering the validity of score-based inferences, but also the appropriateness of score-based uses. Given all these potential benefits, the AUA marks an important contribution to validation research,<sup>11</sup> particularly in the field of language testing and assessment.

#### 2.3.2.3 Chapelle et al.'s validity argument for the TOEFL<sup>®</sup>

In both Kane's and Bachman's validity arguments, "observed test performance" is utilized as a ground or fundamental data to proceed to score interpretations and uses. In applying the argument-based approach to evaluate the Test of English as a Foreign Language<sup>™</sup> (TOEFL<sup>®</sup>), Chapelle (2008) and Chapelle et al. (2010) use "the target language use domain" as grounds and add another inference called "*domain description*" to connect the grounds to test performance observations. According to Chapelle et al. (2010), this extension and addition is needed because since the TOEFL<sup>®</sup> is intended to be interpreted as a measure of language ability in a target language use domain, efforts to describe the domain carefully and to develop items that reflect the domain tend to support the intended interpretation (Kane, 2004). Other scholars (e.g., Briggs, 2004) have also urged test developers to empirically describe target domains and channel empirical findings into test task authoring, thus establishing "design validity" for test scores. As a result, while preserving the major inferences, warrants and assumptions in Kane's and Bachman's models, Chapelle et al. (2010) introduce one additional inference to complete the validity argument developed for the TOEFL<sup>®</sup>.

## 2.4 An argument-based approach to validation of the ICPTs

Although the approaches to interpreter certification vary across different countries (Hlavac, 2013), ICPTs are ordinarily included as a key component of certification procedures. High-stakes ICPTs should be subjected to a rigorous process of validation. According to Messick (1994, 1995), performance assessment must be evaluated by the same general validity criteria outlined by the American Educational Research Association (AERA), American

---

<sup>11</sup> Weir's (2005) socio-cognitive framework for validation represents another contribution.

Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999). More importantly, the argument-based approach can be employed to guide the validation of performance assessment (Kane, 2006). According to Bachman, the AUA can and should be used in interpreter testing and assessment, because interpreting tests are basically language tests (Chen, 2011). The present study therefore proposes an AUA to inform and guide interpreting testers to conduct a rigorous validity investigation into ICPTs.

#### 2.4.1 An overview of current practice in ICPT

This section provides a brief review of three key aspects of ICPTs before developing the validity argument: 1) target domains of interpreter certification, 2) test methods, and 3) intended score interpretations and uses.<sup>12</sup>

In terms of target domains, ICPTs may target either a range of practice domains into which score-based inferences are intended to be generalized, or be designed to assess performance in a specific domain. The CATTI level IV tests focus on the domain of general international conferences, while the Federal Court Interpreter Certification Examination (FCICE) in the USA focuses on interpreting only in the Federal Court. As a result of the different domains, the ICPTs may sample a combination of interpreting tasks so as to ensure the relevance and representativeness of test content. NAATI's professional-level test includes dialogue, sight, and consecutive interpreting tasks to model the skills and techniques in community interpreting practice in Australia in range of target domains, while the CATTI level IV tests in China use consecutive and simultaneous interpreting tasks to represent interpreting practice in conference settings.

With respect to score-based inferences, test scores have typically been explained as an indicator of 1) unobservable attributes such as knowledge, skills and abilities possessed by test candidates ("interpreting ability"), and 2) observable interpreting performance (e.g., fluency of delivery, language quality) in a given target practice domain. For example, based on the CATTI syllabus<sup>13</sup>, on the one hand, CATTI assess the basic qualities test candidates should have, including "1) using Chinese and English languages with dexterity, 2) having expansive

---

<sup>12</sup> For a detailed cross-national review of interpreter certification procedures, see Hale et al. (2012) and Hlavac (2013); for a review of certification of health care or medical interpreter in the USA, see Arocha & Joyce (2013), Roat (2006); for a review of certification of court interpreters, see Feuerle (2013); for a review of interpreter certification testing in China, see Chen (2009).

<sup>13</sup> The syllabus for the SI test: [http://bbs.catti.china.com.cn/download/syllabus\\_EN\\_SI2.pdf](http://bbs.catti.china.com.cn/download/syllabus_EN_SI2.pdf)

background knowledge of politics, economics, culture, etc., 3) applying SI skills adroitly, and 4) demonstrating sound psychological qualities and coping tactics”. On the other hand, test scores also show whether test candidates “render source-language content accurately and completely, pronounce correctly and clearly, and deliver fluently and in a natural tone” in “a various formal (conference) occasions”.<sup>14</sup>

Regarding score-based actions, test scores constitute one of the most important sources for making certification decisions. For example, the NAATI decision-makers use performance/oral test scores, coupled with written test scores (on professional ethics), to make certification decisions. Test takers must achieve a specified cut-off score for both written and oral tests to be granted certifications.

## 2.4.2 Constructing a validity argument for ICPTs’ test score interpretations

### 2.4.2.1 *A chain of inferences and data*

In this section, an assessment use argument (AUA) for ICPTs is developed, as can be seen in Figure 2.5 below. The trait- and behaviourist-based test score interpretations proposed for ICPTs are taken as a claim about test candidates. This is the ultimate claim the validity argument strives to justify (i.e., the “Destination”). To arrive at this final claim, a chain of linked inferences is needed to interweave different sets of data.

As shown in the scaffolded structure in Figure 2.5, the intermediate inferences and the related sets of data are interspersed to facilitate the logical flow from the “Ground” to the “Destination”. This represents the overall structure or major steps involved in the validation. The chain of inferences that would most likely support the AUA for ICPTs are as follows:

- 1) *Domain analysis and modeling* The fundamental set of data or the “Ground” is the real-life target domain where the interpreting practice of interest takes place. Adopted from Mislevy et al. (2003), domain analysis and modeling refers to systematic and empirical studies conducted to identify knowledge, skills, abilities and strategies (KSASs) required of interpreters, and to ascertain frequently performed interpreting tasks (with their associated characteristics). It also means that the results of the domain analysis are used to inform the development of test tasks.

---

<sup>14</sup> The syllabus is written in Mandarin Chinese. The quoted texts were translated by the author.

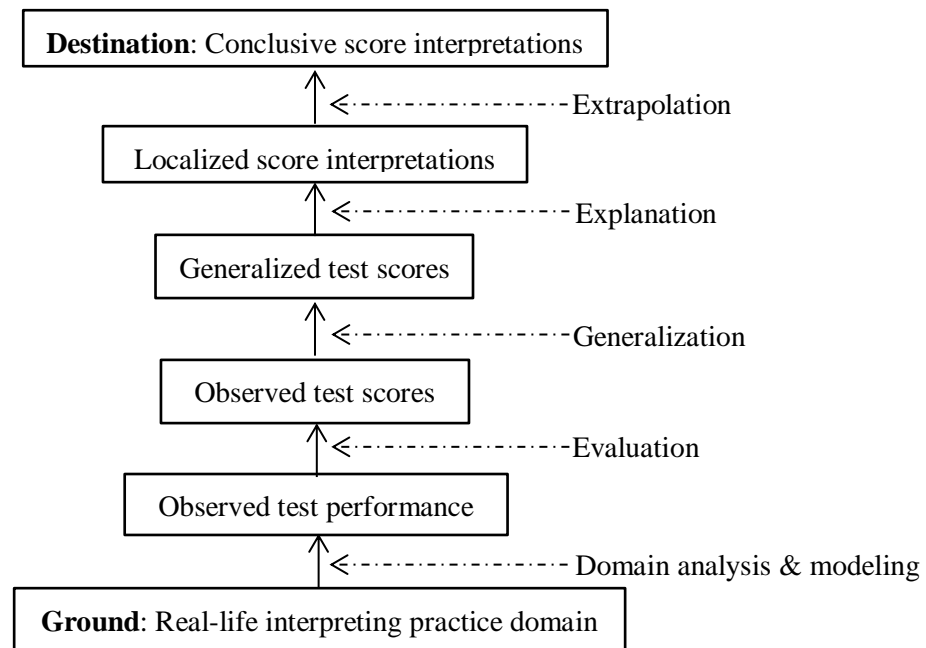


Figure 2.5 A chain of inferences and related data in the validity argument

- 2) *Evaluation* As a result of the first inference, when test candidates perform interpreting in a test situation, observations of their performance are made. This becomes the second set of data. An evaluation inference is used to ensure that appropriate scoring rules (e.g., performance assessment criteria, scoring schemes) are applied to assess performance observed in a standardized testing procedure.
- 3) *Generalization* As a result of the evaluation, test performance is quantified. Observed test scores serve as the third set of data. Typically, the observed scores are determined by specific raters on a set of specific test tasks. However, test developers and users would expect the scores to be invariant across all possible raters, test tasks and other measurement facets. This expectation could be met with a generalization inference. With this inference, the observed scores could be generalized from a specific evaluation of a specific set of performances to test takers' expected performance over a universe of a test domain (Kane, 2004), or many possible alternative tests that could be developed. Thus, the generalization inference supports the observed scores in a single test to be generalized to a much larger test domain.
- 4) *Explanation* From the generalization inference, the generalized scores become the fourth set of data. At this stage, the meaning of the numerical scores pertains to observable aspects of interpreting performance (e.g., target language quality, fluency of delivery, pronunciation). That is, the test scores are interpreted in a performance-referenced manner.

However, the test scores have also been explained in a trait-referenced manner. In other words, the generalized scores have been explained in reference to a latent trait(s) or a hypothetical construct(s) such as “interpreting ability”. Therefore, an explanation inference is needed to relate the scores on different rating dimensions to an overarching construct. For ICPTs, the construct of interest relates to a theory of interpreting ability. Via the construct theory, the generalized scores are not only related to interpreting performance per se, but also a latent trait such as interpreting ability.

5) *Extrapolation* Thanks to the explanation, the localized score interpretations are derived which becomes the fifth set of data. The score meanings are local in the sense that they are only plausible within a boundary of a test domain. In other words, the score-based interpretations only have explanatory power in an artificial testing situation. However, what test users would be most interested in are performance and ability levels in a real-life practice domain. They would expect no substantial performance and ability variation between a test situation and a real-life domain. To traverse from a test domain to a real-life practice domain, an extrapolation inference is thus needed. This inference enables the localized score interpretations to be extrapolated into the real-life domain of interpreting practice. Consequently, the score interpretations initially proposed for ICPTs can largely remain intact in real-life interpreting domains. Finally, the extrapolated score meanings represent the conclusive claim the validity argument strives to support.

As can be seen in Figure 2.5, to evaluate the adequacy and appropriateness of the conclusive test score interpretations is, therefore, to appraise the soundness and cogency of the chain of inferences.

#### 2.4.2.2 Warrants and underlying assumptions

Only when all the scaffolded inferences shown above are “warranted”, can we have confidence in the credibility of the conclusive claim, and the claim is thus “validated”. To sanction the inferences, relevant warrants and their associated underlying assumptions need to be articulated.

Table 2.2 shows the respective warrants and the associated underlying assumptions for each inference. As shown in the table, each of the five inferences in the validity argument is licensed by a warrant, and has a set of underlying assumptions. For example, to warrant the evaluation inference, four assumptions are made: 1) standardization of test conditions and administration,

2) use of appropriate and justified scoring rules, 3) consistent application of scoring rules by raters, and 4) recruitment of qualified raters. Only when these four assumptions are satisfied, can test validators move to the next inference. To support each of these assumptions, backing evidence needs to be collected, presented and assessed.

#### *2.4.2.3 Backing (validity) evidence*

To support each underlying assumption, multiple strands of concrete backing evidence (e.g., empirical, theoretical and/or judgmental), commonly known as “validity evidence”, should be generated, collected, integrated, and evaluated. Table 2.3 draws a contrast between expected validity evidence and available validity evidence (assembled from different ICPTs), and also shows further studies to generate additional validity evidence.

First, as can be seen in Table 2.3, the validity evidence underpinning the “domain analysis and modeling” inference has been generated for some ICPTs. For example, Angelelli (2007) described the development of a medical interpreter readiness test, based on an empirical job analysis conducted to profile salient features in interpreter-mediated patient-doctor encounters. For another example, Certification Commission for Healthcare Interpreter (CCHI) (2010) reported on the development of CCHI certification examinations, drawing upon a Job Task Analysis (JTA) study via a national survey of the profession. Empirical evidence from these studies could be used to show that the test content is relevant to and based on the real-life interpreting practice.

For other ICPTs that have not profiled the interpreting practice domain of interest, empirical practice analysis (Raymond, 2001), or “domain analysis” (Mislevy et al., 2003), needs to be conducted to identify what interpreting tasks (plus associated characteristics) are frequently performed in the target domain. Critical KSASs required by interpreting practice should also be identified using expert-based judgment and/or based on empirical studies (e.g., Bartłomiejczyk, 2006). Consequently, practice-analysis results should inform test design and development, following a rigorous procedure (e.g., Mislevy, Almond, & Luckas, 2004). The process of domain analysis and modeling should be carefully documented as an important strand of validity evidence.

Table 2.2 Warrants and underlying assumptions in the validity argument for ICPT score interpretations

Inference	Warrant licensing the inference	Assumptions underlying inferences
<b>Domain analysis &amp; modeling</b>	Observations of performance on ICPTs reveal KSASs in situations representative of those in real-life target domains.	<ul style="list-style-type: none"> <li>• KSASs required for real-life interpreting can be identified;</li> <li>• KSASs-engaging tasks representative of real-life practice can be identified;</li> <li>• These tasks can be replicated in a testing situation.</li> </ul>
<b>Evaluation</b>	Performance in ICPTs is consistently evaluated to produce observed scores based on appropriate scoring rules.	<ul style="list-style-type: none"> <li>• Test conditions and administration are standardized;</li> <li>• Scoring rules used are appropriate and justified;</li> <li>• Scoring rules are consistently followed by raters;</li> <li>• Raters are qualified for assessing test performance.</li> </ul>
<b>Generalization</b>	Test cores based on certain samples of performance can be generalized to expected performance over the test domain.	<ul style="list-style-type: none"> <li>• Performance on ICPTs is a random sample or a representative sample from test domains;</li> <li>• Test scores can be generalized over a test domain (across raters, occasions, etc.).</li> </ul>
<b>Explanation</b>	Generalized scores can be explained using a theory of "interpreting ability".	<ul style="list-style-type: none"> <li>• The KSASs required by interpreting vary across tasks in keeping with theoretical expectations;</li> <li>• Performance can be predicted by task characteristics as expected theoretically;</li> <li>• Test performance (does not) relate(s) to performance on the other (un)related constructs of interest as expected theoretically.</li> </ul>
<b>Extrapolation</b>	Localized score interpretations can be extrapolated to the real-life practice domain of interest.	<ul style="list-style-type: none"> <li>• Test and non-test performance dose not vary considerably;</li> <li>• Test tasks and their characteristics sampled in a test largely resemble those in the real-life practice domain.</li> </ul>

Second, for several ICPTs, there has been some partial validity evidence for the “evaluation” inference (see Table 2.3). Regarding the justifiability of scoring rules, Turner, Lai, & Huang (2010) examined the correlation of test scores generated by the error-deduction scoring method employed by NAATI and a rubrics-based rating scale used by DPSI. The results generally supported the efficacy of the error-deduction method, which could therefore be treated as a piece of validity evidence. For another example, regarding rater training, Russell and Malcolm (2009) described how raters were trained for the national certification of American Sign Language interpreters in Canada. The training procedures could be documented to partially support the “evaluation” inference.

However, robust validity evidence is still lacking to reinforce the “evaluation” inference, for instance, evidence that answers questions such as “Did rater training produce desirable effects?” “Did raters use rating scales consistently?” and “Did raters exhibit significant biased interactions?” Interpreter certifiers, therefore, should document how test administration has been standardized and implemented (e.g., test procedure, proctor behavior), and initiate at least three strands of research to generate validity evidence: 1) to rigorously develop and validate scoring rubrics and scoring schedules, as demonstrated in Fulcher, Davidson and Kemp (2011), and Knoch (2007), 2) to analyze inter-rater reliability (e.g., Slatyer, 2008), and 3) to investigate rater behavior and identify problematic raters. Specifically, rater-related analysis could follow the established practice in language testing, by applying multifaceted Rasch measurement (MFRM) to gain insight to rater internal self-consistency, differential rater functioning (DRF) and biased interactions with other measurement facets (e.g., significant rater severity/leniency), and to monitor rater behavior through inspection of rater bias assessment maps (e.g., Lumley & McNamara, 1995; Wigglesworth, 1993).

Third, as can be seen in Table 2.3, there seems to be no studies conducted to produce validity evidence that supports the “generalization” inference for ICPTs. Empirical studies should thus be initiated to show that systematic variation of test scores is largely attributed to test candidates, not to other measurement facets (e.g., raters, occasions) and/or random errors. Specifically, studies of this kind have been increasingly based on generalizability (G) theory (e.g., Brennan, 1992; Shavelson & Webb, 1991) which simultaneously investigates multiple sources of measurement error, and calculates index of generalizability or dependability (i.e., G coefficients).

Table 2.3 Expected validity evidence, available validity evidence and future studies

Inference	Expected validity evidence	Available validity evidence	Proposed studies to generate validity evidence
<b>Domain analysis &amp; modeling</b>	<ul style="list-style-type: none"> <li>• Evidence to show empirical extraction of critical interpreting tasks and KSASs from a given target interpreting domain;</li> <li>• Evidence to show principled development and sampling of KSASs-engaging tasks into an ICPT</li> </ul>	e.g., Angelelli (2007); ALTA Language Services (2007); Certification Commission for healthcare interpreters (2010), Russell & Malcolm (2009); Vermeiren et al. (2009)	<ul style="list-style-type: none"> <li>• Practice analysis to identify key tasks and KSASs required by interpreting in a given target domain;</li> <li>• Implementation of rigorous test development procedures.</li> </ul>
	<ul style="list-style-type: none"> <li>• Evidence to show standardized test administration;</li> <li>• Evidence to show justifiability, reliability &amp; validity of scoring rules;</li> <li>• Evidence to show inter-rater reliability;</li> <li>• Evidence to show recruitment of qualified raters, effective rater training &amp; monitoring.</li> </ul>	e.g., Angelelli (2007); Roat (2006); Turner et al. (2010); Russell & Malcolm (2009); Vermeiren et al., (2009)	<ul style="list-style-type: none"> <li>• Documentation of test administration;</li> <li>• Rigorous development and validation of scoring rules (rubrics, rating scales &amp; procedures);</li> <li>• Inter-rater reliability studies;</li> <li>• In-depth analysis of rater behavior in training sessions &amp; operational rating.</li> </ul>

Table 2.3 (*continued*)

<b>Inference</b>	<b>Expected validity evidence</b>	<b>Available validity evidence</b>	<b>Proposed studies to generate validity evidence</b>
<b>Generalization</b>	<ul style="list-style-type: none"> <li>• Evidence to show desirable generalizability or dependability of test scores across measurement facets.</li> </ul>	None	<ul style="list-style-type: none"> <li>• Generalizability (G) studies to estimate variance components, and to produce G coefficients, decision (D) studies to optimize measurement designs.</li> </ul>
<b>Explanation</b>	<ul style="list-style-type: none"> <li>• Evidence to show theoretically expected relationship between tasks, use of KSASs and interpreting performance;</li> <li>• Evidence to show (lack of) correlation between ICPT scores and measures of other (un)related constructs.</li> </ul>	None	<ul style="list-style-type: none"> <li>• Systematic empirical studies to investigate relationship between tasks, interpreting ability and performance;</li> <li>• Empirical studies to establish (lack of) correlation or causal relationship between a measure(s) of interpreting ability and measures of other (un)related constructs.</li> </ul>
<b>Extrapolation</b>	<ul style="list-style-type: none"> <li>• Evidence to show predictive power of ICPT performance to real-life performance;</li> <li>• Evidence to show a desirable degree of correspondence of task characteristics between testing and real-life situations.</li> </ul>	Chen (2009)	<ul style="list-style-type: none"> <li>• Correlational studies to relate ICPT scores to measures of real-life interpreting performance;</li> <li>• Comparability studies to compare characteristics of test tasks and real-life tasks.</li> </ul>

Fourth, it would appear that no evidence has been generated by ICPT testers to support the “explanation” inference (see Table 2.3). Systematic empirical studies, therefore, need to confirm the relationship between interpreting task characteristics, use of KSASs and observed

interpreting performance, as described and predicted by a construct model of interpreting ability defined *a priori*. For example, task characteristics could engage specific aspects of interpreting ability (e.g., Meuleman & Van Besien, 2009), they could also affect specific aspects of interpreting performance (e.g., Daro, Lambert, & Fabbro, 1996), and use of particular KSASs could enhance interpreting performance quality. This type of studies helps gain insight to substantive meaning of the construct, which is also known as “construct representation”<sup>15</sup> (Embretson, 1983).

In addition, the other type of studies that investigate the relationship between the construct of interpreting ability and the other (un)related constructs of interest (e.g., “translation ability”) could help demarcate the boundaries of “interpreting ability” construct, which is known as establishment of “nomothetic span”<sup>16</sup> (Embretson, 1983). For example, the ICPTs that are developed to measure the same construct of “interpreting ability” (e.g., CATTI tests and China’s National Accreditation Examination for Translators and Interpreters/NAETI) are supposed to provide invariant ability estimates on test candidates. The invariance of ability estimates across different tests can be examined, using Rasch-based common test linking (e.g., Bond & Fox, 2007). To illuminate a structural relationship between the latent construct of “interpreting ability” and other related constructs, structural equation modeling (SEM) techniques could also be applied, given the availability of a sufficient sample size and solid statistical capability of researchers (e.g., Purpura, 1997).

Fifth, to support the assumptions underpinning the “extrapolation” inference, two types of validity evidence could be provided. On the one hand, studies relating ICPT scores to predictive criterion measures such as an evaluation of examinees’ real-life interpreting performance need to show positive and meaningful correlations. On the other hand, comparability studies need to demonstrate that interpreting tasks and associated characteristics (e.g., features of source texts, task type, task conditions) sampled in ICPTs demonstrably resemble those in the real-life practice domain, as has been argued by Campbell and Hale (2003), and Chen (2009). One way to investigate the degree of correspondence, or “test authenticity” (Bachman & Palmer, 1996), is to develop a framework of interpreting task characteristics (Campbell & Hale, 2003), akin to those developed and used in language testing

---

<sup>15</sup> Put it simply, construct representation refers to a process in which construct meaning is clarified by theoretical mechanisms underlying task performance.

<sup>16</sup> A nomothetic span refers to a nomological network in which meanings of a construct can be inferred by its relationships to other related constructs.

(e.g., Bachman, 1990; Bachman, Davidson, & Milanovic, 1996; Bachman & Palmer, 1996). It seems that only Chen's (2009) study systematically investigated "test authenticity" between the real-life interpreting domain and three ICPTs in China. However, the results showed lack of authenticity, suggesting a need of further improvement of the ICPTs.

Taken together, Table 2.2 and Table 2.3 present the warrants, the underlying assumptions and the backing evidence for the respective inferences in order to evaluate the proposed score interpretations. Table 2.3 also shows that although there has been some validity evidence for several ICPTs, it is far from enough to justify the score interpretations. Further studies should be initiated to generate more validity evidence for each inference.

In the following section, the test validators need to shift their attention from the score-based inferences to assessing the appropriateness of score-based actions, guided by an assessment utilization argument.

#### 2.4.3 Constructing an assessment utilization argument for ICPT score use

To guard against unjustifiable use of score interpretations to make certification decisions, an assessment utilization argument should be developed for the ICPTs to subject the decision-making process to investigation. The argument links the score interpretations with the certification decisions. Four underlying assumptions are needed to sanction the utilization inference, including sufficiency (Bachman, 2005), values sensitivity, equitability, and beneficial consequences (Bachman & Palmer, 2010).

Table 2.4 presents the warrant, the associated assumptions and their respective backing evidence. As to the sufficiency assumption, judgmental and experiential evidence should indicate that scores from ICPTs contain sufficient information for the decision-making. Otherwise, additional indicators may be needed to warrant the decision-making. For example, in Australia the NAATI authority makes certification decisions based on three types of test scores, including scores from 1) the social and cultural awareness test, 2) the test of knowledge and application of professional ethics, and 3) the interpreting performance assessment (NAATI, 2012). These scores work together to provide complementary and non-overlapping information to inform the NAATI authority of whether certification decisions could be made.

Regarding the values sensitivity, evidentiary support needs to show that certification authorities have made deliberate efforts to engage with relevant stakeholders (e.g., test takers, educators, employers) and sought to understand their respective needs, concerns and values.

Evidence should also indicate such values are considered by certification bodies when making certification decisions. For example, in the *INT Project*, a national survey was conducted to elicit opinions from relevant stakeholders (Hale et al., 2012), the results of which are expected to help NAATI decision makers understand the ramifications of the testing decisions.

Table 2.4 Warrant, assumptions and backing evidence in the assessment utilization argument

Warrant licensing the inference	Assumptions underlying the inference	Backing evidence
The certification decision is made, based on sufficient information, and taking into account social values, equitability requirement, and possible beneficial consequences on stakeholders.	<ul style="list-style-type: none"> <li>• ICPTs provide sufficient information for decision-making;</li> <li>• Understanding social needs and values is part of the decision-making process;</li> <li>• The certification decision is made, not biased for or against a particular group of test candidates;</li> <li>• The certification decision is believed to have beneficial consequences to stakeholders and society at large.</li> </ul>	<ul style="list-style-type: none"> <li>• Evidence to show sufficiency of ICPT scores;</li> <li>• Evidence to show inclusiveness and value considerations;</li> <li>• Evidence to demonstrate test unbiasedness;</li> <li>• Evidence to show positive impacts on or benefits to stakeholders (e.g., test takers, educators, employers, services recipients)</li> </ul>

To support the third assumption of equitability, results from post-hoc studies should demonstrate the impartiality of ICPTs. Specifically, the certification decisions should be made based on test takers' performance, not on irrelevant characteristics (e.g., gender, ethnicity).

Regarding the test impacts, washback studies (e.g., Wall & Alderson, 1993; Xie & Andrew, 2013) are needed to demonstrate that overall certification decisions produce positive impacts on test takers (e.g., test preparation), on educators (e.g., efficient instruction and training), and on consumers (e.g., better interpreting services, see Jacobs et al., 2001).

Additionally, in developing the AUA for ICPT score interpretations and use, three principles deserve special attention. The first principle is that proposed test score interpretations and uses need to be articulated before the validity investigation (Bachman & Palmer, 2010; Kane, 1992, 2004). Otherwise, validation research loses its compass. Another principle is that the validation involves a program of research rather than a single study. As an on-going process, the validation research requires continual evaluation and gathering of evidence (e.g.,

Anastasi, 1986; Shepard, 1993). This on-going nature necessitates a strong program of validation to establish and maintain validity of score interpretations and uses over time and across contexts. The last principle is that multiple strands of validity evidence should be collected and triangulated to support inferences and their assumptions (Messick, 1989; Kane, 1992, 1994). Given that the AUA is a practical argument, it is impossible to prove it in a strict algorithmic fashion. But it is possible to show that the AUA is plausible and believable. Parallel lines of validity evidence contribute to greater plausibility of the AUA, especially the most vulnerable inferences and assumptions.

#### 2.4.4 Challenging ICPT score interpretations and uses

In addition to the three principles mentioned above, another critical principle is to identify rival hypotheses or alternative explanations for the score interpretations and uses. The rationale is that confidence in explaining and using scores in a particular manner accrues, when the intended score interpretations and uses survive theoretical and empirical challenges (Bachman, 2005; Kane, 1992). In fact, plausible rival hypotheses could manifest in every link between the inferences, as shown by Crooks, Kane and Cohen (1996). But more importantly, it is the weakest inference and assumption in an AUA that needs the most attention, because the overall credibility of the AUA is limited by its most questionable part (Crooks et al., 1996; Kane, 1992, 1994; Messick, 1989).

For ICPTs, two inferences appear to be the most vulnerable, as can be also seen in Table 2.3. The first one is the “generalization” inference. Typical of performance assessment, ICPTs contains a small number of tasks (i.e., ranging from one to three tasks per direction). Based on the empirical studies of educational performance tests (e.g., Mehrens, 1992; Shavelson, Baxter, & Gao, 1993), substantial variability is expected, regarding person-task interaction. ICPTs are also typical of rater-mediated performance assessment (Engelhard, 2002), where interpreting performance is evaluated by raters using a certain scoring schedule. Rater variability could contribute to unwanted systematic variance in test scores, known as *rater effects* (McNamara, 1996). This suggests that test scores may not be generalizable beyond a specific sample of interpreting tasks and raters. Therefore, empirical generalizability studies using multifaceted Rasch measurement and G theory may need to be conducted by certifying organizations to provide validity evidence for trustworthiness of rater judgment and dependability of test scores.

The other weakest inference is that of the “explanation”. This inference requires ICPT developers to explain test scores via a construct model of “interpreting ability”, be it a theory of “construct representation” or of “nomothetic span”. This appears to be a tall order, at least for now. Granted, researchers and scholars in Interpreting Studies (e.g., Gile, 1995; Moser, 1978; Setton, 1999) and other related fields (e.g., Christoffels, De Groot, & Waldorp, 2003; De Groot, 2000) have contributed significantly to the knowledge of underlying processes of interpreting. But it seems that a strong construct theory of “interpreting ability” has not yet been established, which helps organize testers’ thoughts on design and development of ICPTs and accommodates the score interpretations. Although construct definition constitutes one of the first and foremost important steps in designing an assessment instrument (Angelelli, 2009), it has not been precisely defined for interpreter testing and assessment (Sawyer, 2004). As a result, explanations based on *ad-hoc* construct models lacks due rigor and cogency.

## 2.5 Implications

The AUA for the score interpretations and uses consists of six important components: proposed score interpretations and use, ground/data, inferences, warrants/assumptions, rebuttal/rival hypotheses, and backing evidence. Figure 2.6 describes the overall AUA structure and its components. As shown in the figure, the AUA articulates the score interpretations to be made from the ground/data to a temporary claim by an inference. Each of the inferences is used to bridge a preceding body of data to a succeeding claim. Each claim in turn becomes data for a subsequent claim.

Overall, the argument-based validation framework has two implications for ICPTs. The first implication is that given the scaffolded structure of the AUA, robust validation involves a strong research program. In addition, under the AUA framework, evidence-generating studies should not be randomly conceived, but logically structured and strategically prioritized, with the most questionable inference(s) being given more attention. Considering the lack of validation research conducted for ICPTs, future test validators could initiate the studies proposed in the AUA.

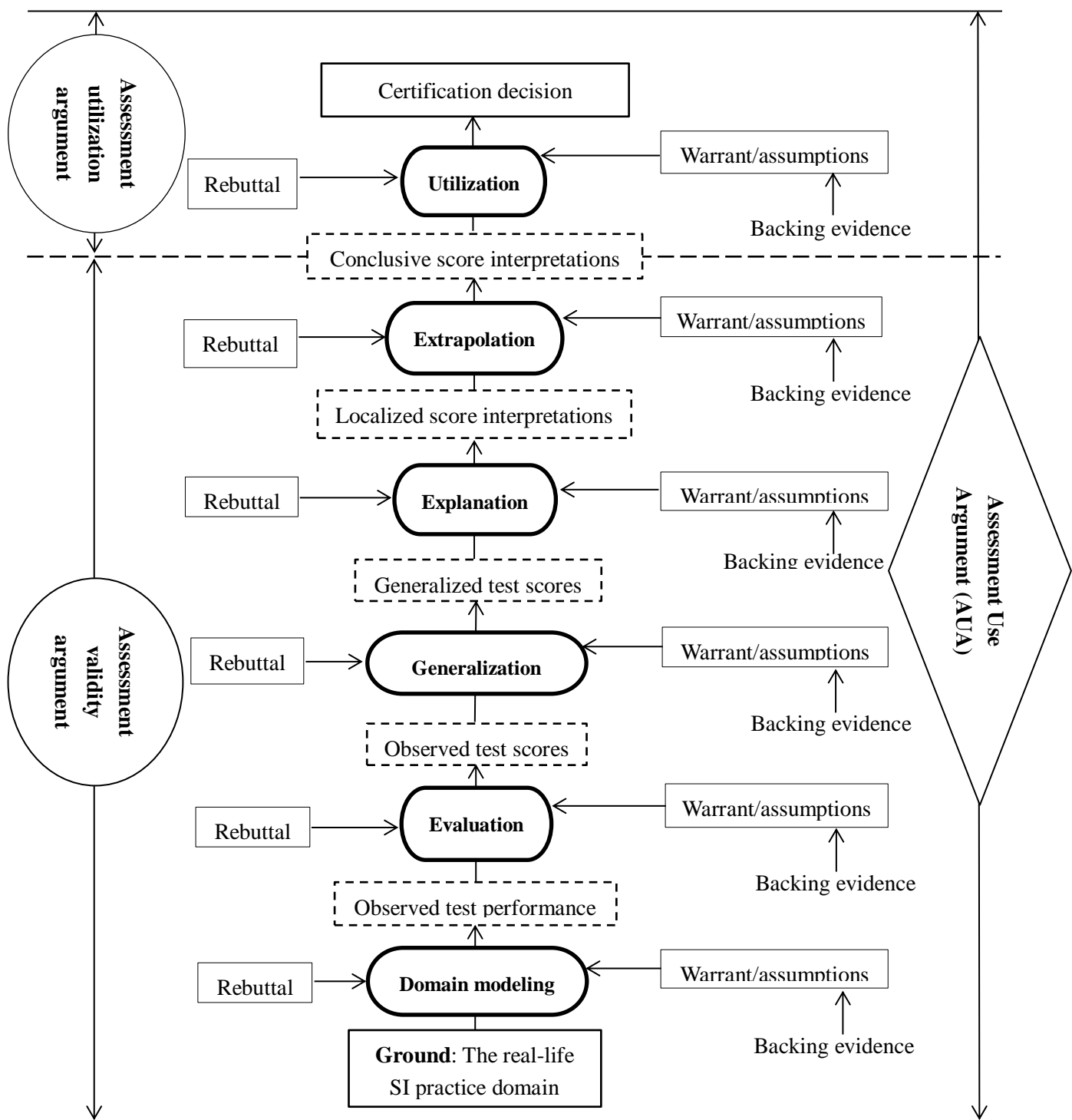


Figure 2.6 The AUA for the ICPT score interpretations and use

The second implication is that the AUA could be used to guide assessment design and development (Bachman, 2005; Crooks et al., 1996; Kane et al., 1999). As demonstrated previously, some validity evidence needed in the AUA can be collected during the test design and development stage (e.g., evidence needed for the domain analysis and modeling inference). It is actually recommended by language testers (e.g., Kane, 1990, 1994) that construction and

refinement of an AUA be synchronized with the assessment development process so as to achieve a good fit between them, and to ensure consistent score interpretations. The best example to show this interconnection would be the evidence-centered assessment design (ECD) (Mislevy et al., 2003). The ECD explicitly draws upon what is called as “an evidentiary argument”. Test developers should carry out design activities structured in a way that validity evidence emerges (Mislevy, 2007). Specifically, testers must design an assessment based on intended score interpretations and use, the practice domain of interest, and a chain of interlocking inferences.

## 2.6 Conclusion

ICPT has been growing apace around the world over the past decade. It promises to regulate interpreting markets, improve interpreting services, and deliver benefits to stakeholders. To live up to its promises, a quality control mechanism is required to allow systematic accumulation of validity evidence to support proposed test score interpretations and uses. This paper proposes a validity argument to inform and encourage future test validators to put ICPTs under robust validity investigation. Given that much discussion in the study is theoretical, it is all the more valuable to translate the proposed AUA into practice. This paper, therefore, ends by calling for ICPT testers to initiate empirical studies that have been lacking in the AUA.

## 2.7 References

- ALTA Language Services (2007). *Study of California's Court Interpreter Certification and Registration Testing*. Retrieved from <http://www.courts.ca.gov/documents/altafinalreport.pdf>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: Author.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1-16.

- Angelelli, C. (2007). Assessing medical interpreters. *The Translator*, 13(1), 63-82.
- Angelelli, C. (2009). Using a rubric to assess translation ability: defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp. 13-47). Amsterdam: John Benjamins.
- Arocha, I. S., & Joyce, L. (2013). Patient safety, professionalization, and reimbursement as primary drivers for National Medical Interpreter Certification in the United States. *The International Journal for Translation & Interpreting Research*, 5(1), 127-142. DOI: ti.105201.2013.a07
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.
- Bachman, L. F. (2003). Constructing an assessment use argument and supporting claims about test taker-assessment task interactions in evidence-centered assessment design. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 63-65.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1-34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.
- Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing*, 13(2), 125-150.
- Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting*, 8(2), 149-174.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). London: Lawrence Erlbaum.
- Brennan, R. L. (1992). *Elements of generalizability theory*. (Rev. ed.). Iowa City, IA: American College Testing.
- Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, 17(1), 5-9.

- Briggs, D. C. (2004). Comment: making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 171-174.
- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Certification Commission for Healthcare Interpreters (2010). *Job Task Analysis Study and Results*. Retrieved from <http://www.cchicertification.org/images/webinars/cchi%20jta%20report-public.pdf>
- Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319-352). London: Routledge.
- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple...*Language Testing*, 29(1), 19-27.
- Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3-13.
- Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, 3(2), 257-273.
- Chen, J. (2011). Language assessment: Its development and future – An interview with Lyle F. Bachman. *Language Assessment Quarterly*, 8(3), 277-290.
- Christoffels, I. K., De Groot, A. M. B., & Waldorp, L. J. (2003). Basic skills in a complex task: A graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition*, 6(3), 201-211.
- Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting*, 7(1), 97-13.
- Cronbach, L. J. (1971). Test validation. In R.L. Thorndike (Ed.) *Educational measurement* (2nd ed.) (pp. 443-507). Washington, D.C.: American Council on Education.
- Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

- Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice*, 3(3), 265-286.
- Daro, V., Lambert, S., & Fabbro, F. (1996). Conscious monitoring of attention during simultaneous interpretation. *Interpreting*, 1(1), 101-124.
- Dawrant, A., & Jiang, H. (2001). Conference interpreting in Mainland China. *Communicate!* Retrieved from [http://www.aiic.net/ViewPage.cfm?page\\_id=365](http://www.aiic.net/ViewPage.cfm?page_id=365)
- De Groot, A. M. B. (2000). A complex-skill approach to translation and interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting* (pp. 53-68). Amsterdam: John Benjamins.
- Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16(10), 640-647.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179-197.
- Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Mahwah, NJ: Erlbaum.
- Feuerle, L. (2013). Testing interpreters: Developing, administering, and scoring court interpreter certification exams. *The International Journal for Translation & Interpreting Research*, 5(1), 80-93. DOI: 10.12807/ti.105201.2013.a04
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing*, 28(1), 5-29.
- Giambruno, S. (2013). *EU Member States Country Profiles: The Current State of Affairs in Europe*. Retrieved from <http://www.qualitas-project.eu/sites/qualitas-project.eu/files/Leaflet.pdf>
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.
- Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological measurement*, 1(1), 1-10.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11(3), 385-398.
- Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, 5(10), 51-517.
- Hale, S., Garcia, I., Hlavac, J., Kim, M., Lai, M., Turner, B., & Slatyer, H. (2012). *Development of a conceptual overview for a new model for NAATI standards, testing and*

assessment. Sydney, Australia. Retrieved from:  
<http://www.naati.com.au/PDF/INT/INTFinalReport.pdf>

- Hlavac, J. (2013). A cross-national overview of translator and interpreter certification procedures. *The International Journal for Translation & Interpreting Research*, 5, 32-65. DOI: 10.12807/ti.105201.2013.a02
- Jacobs, E. A., Lauderdale, D. S., Meltzer, D., Shorey, J. M., Levinson, W., & Thisted, R. A. (2001). Impact of interpreter services on delivery of health care to limited-English-proficient patients. *The Journal of General Internal Medicine*, 16(7), 468-474.
- Kane, M. T. (1990). *An Argument-based Approach to Validation*. Iowa City, Iowa: American College Testing Program.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions*, 17(2), 133-159.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135-170.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Kane, M. T., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5-17.
- Knoch, U. (2007). *The development and validation of an empirically-developed rating scale for academic writing* (Unpublished doctoral thesis), University of Auckland, New Zealand.
- Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27(2), 183-189.
- Linn, R. L. (1989). *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan.
- Lissitz, R. W., & Samuels, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.

- Liu, M. H. (2013). Design and analysis of Taiwan's Interpretation Certification Examination. In D. Tsagari & R. van Deemter (Eds.), *Assessment Issues in Language Translation and Interpreting* (pp. 163-178). Frankfurt am Main: Peter Lang.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(3), 635-694.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71.
- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55, 97-105.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
- Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, 11(1), 3-9.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966.
- Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher*, 10(9), 9-20.
- Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.
- Meuleman, C., & Van Besien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting*, 11(1), 20-34.
- Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, 36(8), 463-469.
- Mislevy, R. J., Almond, R.G., & Lukas, J. (2004). *A brief introduction to evidence-centered design*. (CSE Technical Report 632). Retrieved from The National Center for Research on Evaluation, Standards, Student Testing (CRESST) website: <http://www.cresst.org/reports/r632.pdf>

- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-66.
- Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver & H. W. Sinaiko (Eds.), *Language, Interpretation and Communication* (pp. 353-368). New York/London: Plenum Press.
- National Accreditation Authority for Translators and Interpreters. (2012). *Accreditation by Testing: Information Booklet*. Retrieved from [http://www.naati.com.au/PDF/Booklets/Accreditation by Testing booklet.pdf](http://www.naati.com.au/PDF/Booklets/Accreditation_by_Testing_booklet.pdf)
- PSI Services LLC (2010). *Development and Validation of Oral and Written Examinations for Medical Interpreter Certification: Technical Report*. Burbank, California, USA. Retrieved from: <http://www.certifiedmedicalinterpreters.org/sites/default/files/oral-and-written-medical-i-nterpreter-technical-report-final.pdf>.
- Purpura, J. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, 47(2), 289-325.
- Ra, S., & Napier, J. (2013). Community interpreting: Asian language interpreters' perspectives. *The International Journal for Translation & Interpreting Research*, 5(2), 45-61.
- Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education*, 14(4), 369-415.
- Roat, C. E. (2006). *Certification of Health Care Interpreters in the United States: A Primer, a Status Report and Considerations for National Certification*. Los Angeles, USA, Retrieved from: [http://www.calendow.org/uploadedFiles/certification\\_of\\_health\\_care\\_interpretors.pdf](http://www.calendow.org/uploadedFiles/certification_of_health_care_interpretors.pdf)
- Russell, D., & Malcolm, K. (2009). Assessing ASL–English interpreters: The Canadian model of national certification. In Angelelli, C. V. & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 331-376). Amsterdam: John Benjamins.
- Sawyer, D. B. (2004). *Fundamental aspects of interpreter education: Curriculum and Assessment*. Amsterdam & Philadelphia: John Benjamins.

- Setton, R. (1999). *Simultaneous Interpretation: A Cognitive and Pragmatic Analysis*. Amsterdam and Philadelphia: John Benjamins.
- Setton, R. (2009). Introduction: Interpreting China, interpreting Chinese. *Interpreting*, 11(2), 109-117.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215-232.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review*, 52(5), 270-277.
- Slatyer, H. (2008). *Proposal for rater training workshops* (Research report), Sydney: Macquarie University.
- Thurstone, L. L. (1932). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Toulmin, S. E. (2003). *The uses of argument* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Turner, B., Lai, M., & Huang, N. (2010). Error deduction and descriptors – A comparison of two methods of translation test assessment. *The International Journal for Translation and Interpreting Research*, 2(1), 11-23. Retrieved from <http://trans-int.org/index.php/transint/article/view/42/66>
- Vermeiren, H., Gucht, J. V., & De Bontridder, L. (2009). Standards as critical success factors in assessments: Certifying social interpreters in Flanders, Belgium. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 291-330). Amsterdam: John Benjamins.
- Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lanka impact study. *Language Testing*, 10(1), 41-69.
- Weir, C. (2005). *Language Testing and Validation: An evidence-based approach*. Houndgrave, UK: Palgrave-Macmillan.

- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-319.
- Xie, Q., & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, 30(1), 49-70.
- Yu, D. R. (2005). *T&I Labor Market in China*. Sydney, Australia. Retrieved from: [http://www.ling.mq.edu.au/translation/lmtip\\_china.htm](http://www.ling.mq.edu.au/translation/lmtip_china.htm)

### **An introductory note to Chapter 3**

In Chapter 2, a theoretical validity argument has been proposed for the ICPTs. The validity argument is designed to help interpreting researchers and testers collect multiple strands of validity evidence to link interpreting performance observed in a testing situation to actual performance in a real-life practice domain and the certification decisions. Particularly, it is emphasized that researchers and testers need to focus on the most vulnerable links in a validity argument, as the weakest link(s) may jeopardize the whole validation enterprise. In the proposed ICPT validity argument, the explanation inference represents one of the weakest links, which should be strengthened and enhanced by using a strong construct theory to impute substantive meanings to ICPT scores.

Again this background, Chapter 3 proposes an interactionalist approach to construct definition for English/Chinese ICPTs. Typically, in English/Chinese ICPTs, test scores are not only an indicator of unobservable traits such as knowledge, skills, strategies and abilities required of an interpreter, but also are used to describe observable aspects of performance in a given practice domain. In other words, ICPT scores are explained in both trait- and performance-referenced manner. An examination of three traditional approaches (i.e., trait, behaviourist, and interactionalist) to construct definition in language testing and assessment research indicates that the interactionalist approach can accommodate the score interpretations intended by ICPTs. A theoretical construct model is therefore proposed and described for ICPTs, and its implications are also discussed.

## Chapter 3 An interactionalist approach to construct definition for English/Chinese interpreter certification performance testing<sup>17, 18</sup>

***Abstract.** In the field of language testing and assessment, construct definition is generally approached in three perspectives: trait, behaviorist and interactionalist. These approaches are based on different ontological views and lead to different test score interpretations. In interpreter certification performance testing, test developers often explain test scores in relation to both test takers' unobservable attributes and observable dimensions of interpreting performance within a certain practice domain. This approach to score interpretation presupposes an appropriate construct theory being used to guide test design and development. However, it would appear that current approaches to construct definition in Interpreting Studies are either trait- or behaviorist-based, which does not enable test scores to be interpreted as intended. As an initial step to redress the disjunction, an interactionalist approach is proposed and a theoretical model developed. The model contends that interpreting performance consistency can be primarily attributed to interpreting ability, characteristics of interpreting tasks and interaction between the two in a various and arguable proportion. A further step is taken to specify components of interpreting ability and describe a framework of interpreting task characteristics. Implications of the model are also discussed.*

### 3.1 Introduction

In psychological, educational and language testing, a test is ordinarily used to measure “something” that is of central concern to psychologists, educators and language testers. The “something” is commonly referred to as a “construct”. A construct is a verbal surrogate for a phenomenon of interest. It is a means of ordering observations (Stenner, Smith, & Burdick,

---

<sup>17</sup> This chapter was presented at the 9<sup>th</sup> China National Conference and International Forum on Interpreting at Beijing Languages and Culture University, Beijing, China, 1-2 June, 2012, and also at the 7<sup>th</sup> International Conference on English Language Teaching in China, at Nanjing University, Nanjing, China, 23-26 October, 2014. A revised version of the chapter is under the 1<sup>st</sup> round of peer review in the journal of *Translation and Interpreting Studies* as: Han, C. (under review). An interactionalist approach to construct definition for English/Chinese interpreter certification performance testing..

<sup>18</sup> This chapter limits its discussion to simultaneous interpreting (SI). Accordingly, interpreting ability is discussed from the perspective of SI, and interpreter certification performance testing concerns assessment of SI.

1983) and “a meaningful interpretation of observed behaviour” (Chapelle, 1998, p. 33). To operationalize a construct in a test, the construct needs to be defined clearly in the first place to guide test design and development (Angelelli, 2009; Bachman, 1990; Messick, 1994). Essentially, construct definition refers to a process in which construct meaning is clarified by theoretical mechanisms underlying task performance, and could also be inferred by its relationships to other related constructs. In other words, the meaning of a construct is explicated by “construct representation” and could be informed by a “nomothetic span” (Whitely, 1983). Defining a construct is critical, because a different perspective on construct definition “encompasses beliefs about what can and should be defined, how tests should be designed, and what the priorities for validation should be” (Chapelle, 1998, p. 50). Particularly, it plays a central role in providing a working conception for test scores to be meaningfully interpreted and explained (Read & Chapelle, 2001).

For English/Chinese interpreter certification performance testing (ICPT) (see Chapter 1), test scores are usually used to infer examinees’ knowledge, skills, abilities and strategies (KSASs), or in nutshell “interpreting ability”, and also to describe simultaneous interpreting (SI) performance in a certain target practice domain.<sup>19</sup> That is, score interpretations are made, with reference to both internal, unobservable traits, and observable dimensions of interpreting performance in the target domains of generalization (See Figure 3.1).

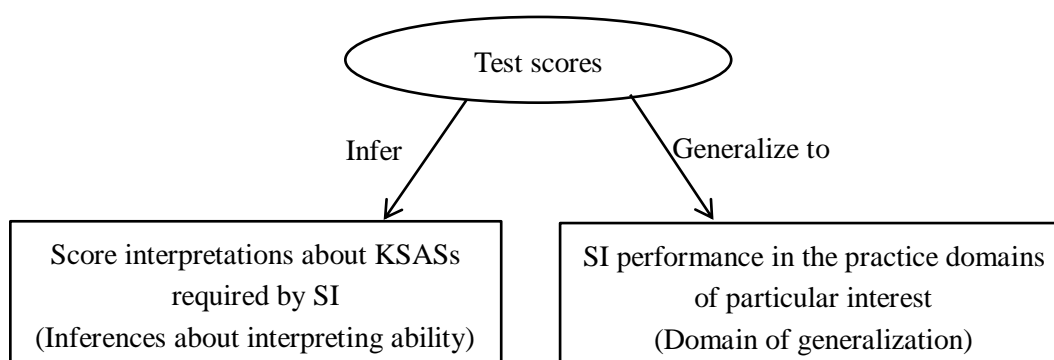


Figure 3.1 Two types of score-based inferences

There are some concrete examples of how scores from English/Chinese interpreter certification performance tests (ICPTs) are explained. For instance, based on the syllabus of China

<sup>19</sup> To interpret a test score is to explain meaning of the score. In testing and assessment literature, “test score interpretation”, “score-based inferences” and “score meaning” can be used interchangeably.

Accreditation Tests for Translators and Interpreters (CATTI),<sup>20</sup> test scores are used to indicate whether test candidates are able to “use Chinese and English languages with dexterity, have expansive background knowledge of politics, economics, culture, etc., apply SI skills adroitly, and demonstrate sound psychological qualities and coping tactics”. Test scores are also treated as indicators of whether test takers have “rendered source-language content accurately and completely, pronounced correctly and clearly, delivered fluently and in a natural tone” in “various formal (conference) occasions”. Another example is from English Interpreting Certificate (EIC) test. It seems that test scores can be related to whether test takers “have high-level English and Chinese bilingual language proficiency, and possess professional skills in consecutive interpreting (CI) and simultaneous interpreting (SI)”, and whether test candidates can perform CI and SI in various settings including “large international conferences, diplomatic occasions, business negotiations, court hearing and other high-level meetings”.<sup>21</sup>

Given that how a test construct is defined informs how score-based inferences could be made, and also given the way how test scores are explained by the interpreter certifying organizations in China, it is logical to speculate that the test construct of central concern must have been properly defined and articulated by test developers of ICPTs. However, it would appear that little credible evidence has been provided for the current approach to score interpretation for ICPTs. Critical information concerning test design could not be found in testing manuals and related publications (Wu, 2010). In addition, as the literature review below would indicate, the current approaches to construct definition in Interpreting Studies may not justify the score interpretations made by ICPTs.

Against the backdrop, the chapter aims at proposing and articulating a theoretical construct model for English/Chinese ICPTs, a model that is intended primarily to help organize interpreting testers’ thoughts on test design, identify potentially useful empirical research to inform better test development, and ultimately (and hopefully) enable test scores to be explained in an ability- and performance-referenced manner. The chapter first reviews and evaluates different approaches to construct definition. The chapter then provides rationales to and proposes an interactionalist approach to construct definition for ICPTs. After that, the interactionalist construct model for ICPTs is specified. Finally, implications of the approach are discussed.

---

<sup>20</sup> The syllabus for the SI test: [http://bbs.catti.china.com.cn/download/syllabus\\_EN\\_SI2.pdf](http://bbs.catti.china.com.cn/download/syllabus_EN_SI2.pdf)

<sup>21</sup> See full descriptions: <http://www.xiadakouyi.com/1154.htm>

### **3.2 Literature review**

This section reviews different approaches to construct definition, particularly a behaviorist and a trait approach. It also evaluates the representation of each approach in Interpreting Studies.

#### **3.2.1 Construct definition**

Generally, observed behavior can be interpreted as and generalized to a construct, provided that the behavior reflects consistency across a range of assessment tasks (Bachman, 2007; Chapelle, 1998). In other words, only consistent and stable performance allows for measurement of discrete behaviors or isolated observations (Messick, 1989). While inconsistent behavior is an interesting topic that has its own value, it is impossible for it to be treated as a dependable indicator of a construct (Chapelle, 2006). As a result, the centerpiece of construct definition is to “hypothesize the source of performance consistency” (Chapelle, 1998, p. 34).

Historically, theorists hold different perspectives of defining a construct and impute performance consistency to different sources (Messick, 1981, 1989). Three general approaches to construct definition have been proposed for the testing and assessment purpose: a behaviorist, a trait and an interactionalist approaches. These approaches are based on different ontological stances and postulate different relationships between entity and context (Bachman, 2006). As a result, test score interpretations vary as a function of the approach to construct definition.

#### **3.2.2 A behaviorist approach**

Behaviorists contend that behaviors are interrelated, primarily because they are elicited and maintained by similar environmental conditions (Messick, 1981). These related behaviors form a response class, or “a set of behaviors all of which change in the same or related ways as a function of stimulus contingencies” (Messick, 1989, p. 15). As a result, the consistent performance is related in a principled way to the context in which the behavior is observed (Young, 2011). The performance is also viewed as a sample of related contexts. From this perspective, a construct comprises the contextual characteristics, and the meaning of the construct is synonymous with the operations or procedures that are used to elicit the

phenomenon (Bachman, 2006). If a test is developed using the behaviorist approach, test scores cannot be explained in a trait-referenced manner, and are only generalizable to those contexts similar to the context where the performance is observed. In other words, score-based inferences should stick strictly to behaviorist language (Messick, 1975).

From the behaviorist perspective, interpreting performance consistency is primarily attributed to contextual characteristics, especially characteristics of interpreting tasks. Figure 3.2 shows the possible relationship between SI task characteristics and SI performance consistency from a behaviourist perspective. As shown in the figure, SI performance consistency could be largely elicited and maintained by a proper contextual configuration.

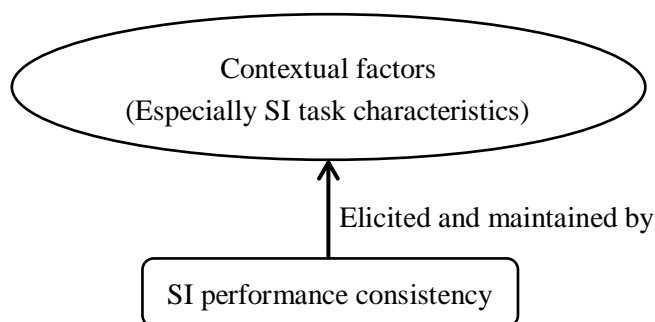


Figure 3.2 A behaviorist approach to interpreting performance consistency

Consequently, if an ICPT is developed based on the behaviorist approach, relevant and representative SI task types and their associated characteristics (e.g., speech rate, accent, register, syntactic features) should be sampled in the ICPT, based on empirical profiling of practice domains (e.g., international conferences, legal, medical settings), so that test scores could be generalized to the target domains of interest.

However, for English/Chinese ICPTs in China, test content may not adequately reflect and represent the target domains of generalization. For example, the CATTI SI performance test only samples one task type (i.e., SI with text). However, the real-life SI practice in China encompasses several frequently performed task types, including (but not limited to) SI with text (Wang & Lin, 2006), SI with PPT (e.g., Wan, 2004) and SI for Q&A session (Chang & Wu, 2009). In addition, the CATTI test primarily samples source-language (SL) speeches characterized by well-controlled speech rates and oral presentations of written texts (Chen, 2009). In the real-life practice, SL speeches for SI are characterized by varying speech rates, different accents, registers, modes of presentation (Chen, 2009; Huang, 2005). As a result, there seems to be a disjunction between the desired score interpretations (i.e., performance in

“various formal conference occasions”) and the interpretations the operational CATTI test actually supports. It also seems that operations or stimulus materials used in the ICPTs would better represent and relate to the real-life practice, if they could be developed consistently based on an empirical profiling of relevant practice domains.

### 3.2.3 A trait approach

According to Messick (1989, p. 15), a trait is “a relatively stable characteristic of a person – an attribute, enduring process, or disposition – which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances”. That is, a person’s consistent performance is taken to index a stable configuration of knowledge, skills, processes and attributes inherent in the person and can be observed across contexts (Young, 2011). Performance is thus treated as a sign of underlying traits. From this perspective, a construct resides in and is internalized by individuals (Bachman, 2006; Deville & Chalhoub-Deville, 2006). If a test is developed based on this approach, score inferences are supposed to maintain across various contexts, since a trait is a relatively stable attribute of a person.

Based on the trait approach, interpreting performance consistency is primarily influenced by internal attributes of interpreters. The attributes could be described by a cognitive model for the testing and assessment purpose (Embretson, 1983; Gorin, 2006). There have been two types of the cognitive model. One type specifies people’s representation of a domain in terms of knowledge, skills, processes and strategies (KSPSs), the other is basically a processing model that describes how problems are represented and how information is processed. As a result, either KSPSs or underlying cognitive processes underlies the performance consistency.

In Interpreting Studies, models of both types have been developed. One type is generally referred to as “models of interpreting ability” (see Angelelli & Degueudre, 2002; Wang, 2007; Yuan, 2007), the other is known as information-processing (IP) models of SI or the IP paradigm (see Bacigalupe, 2010; Daro & Fabbro, 1994; Gerver, 1975; Gile, 1995; Mizuno, 2005; Moser, 1978). The former theorizes the KSPSs required by SI, the latter hypothesizes the internal cognitive processes to sustain SI. Figure 3.3 shows the trait approach to the SI performance consistency. As described in the figure, although the two models have a different focus, the substantive components they represent contribute equally to the SI performance consistency.

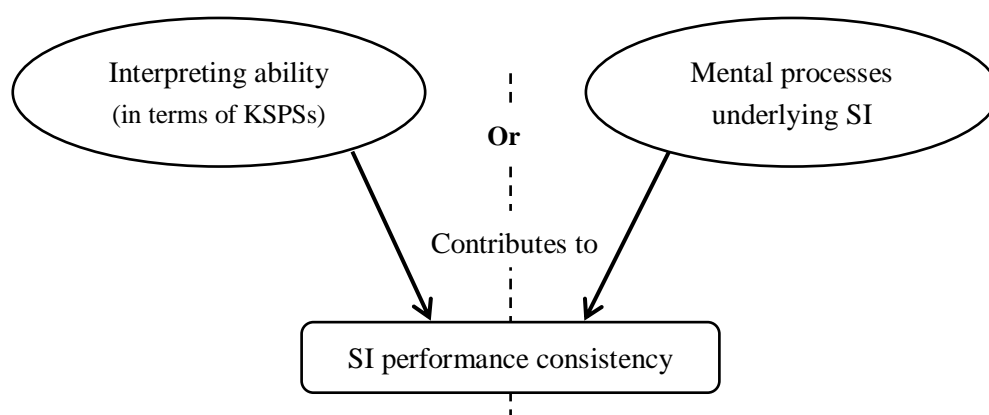


Figure 3.3 A trait approach to interpreting performance consistency

If an ICPT is developed using the trait approach, the sampled test materials would be those that could elicit and engage desired KSPSs or cognitive processes required by SI. Consequently, test scores should only be explained as test takers' possession of the KSPSs required by SI, or being capable of SI-related information processing. For example, the CATTI authority explains the SI test scores in a trait-referenced manner, as shown previously. However, Setton (1997, p. 2) observes that the IP paradigm makes "simplifying assumptions about language in communication" and "hardly address the question of context" (p. 5). It seems that the IP paradigm over-emphasizes the contribution of internal processes to the performance consistency.

As the reviews shows, either of the two approaches to construct definition is unable to accommodate the trait- and behaviourist-based score interpretations for the ICPTs. To redress this problem, an interactionalist approach is therefore proposed.

### 3.3 Proposing and articulating an interactionalist construct model for ICPTs

This section justifies the interactionalist approach to construct definition, proposes a general model, and articulates and specifies each component in the model.

#### 3.3.1 Rationales

Three reasons are provided to justify the approach. The first reason is that previous research supports the interactionalist approach to construct definition in the field of language testing

and assessment (Bachman, 1990, 2007; Bachman & Palmer, 1996, 2010; Chapelle, 1998; Read & Chapelle, 2001). Specifically, Bachman and Palmer (1996, p. 62) state: “Language use involves complex and multiple interactions among the various individual characteristics of language users, on the one hand, and between these characteristics and the characteristics of the language use or testing situation, on the other”. Since SI is a special case of language use and processing (De Groot & Christoffles, 2006; Englund-Dimitrova & Hyltenstam, 2000), it also involves interaction between interpreters’ attributes and SI task conditions. The second reason is that an emerging body of interpreting literature (e.g., Chang & Schallert, 2007; Setton, 1999) indicates that SI performance consistency could be attributed to internal processes and abilities, contextual characteristics, as well as the interaction between the two. For instance, Setton’s cognitive-pragmatic analysis of SI (1997, 1999) fuses together the IP and the IT paradigms, thus taking into account both cognitive-linguistic processes entailed by SI and contextual-pragmatic influences on SI. The third reason is a pragmatic one. Using the interactionalist approach could accommodate the score interpretations for ICPTs.

### 3.3.2 An interactionalist approach to construct definition

The interactionalist approach postulates that some performance or behavioral consistency could be referenced to traits, some to situational factors, and some to interactions between the two, in a various and arguable proportion (Bachman, 2006, 2007; Chapelle, 1998, 2006; Messick, 1981, 1989; Read & Chapelle, 2001). That is, “performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts” (Chapelle, 1998, p. 43). This approach to construct definition offers a way to infer from performance something about both a practice-specific behavior and a practice-independent, person-specific trait (Young, 2000, 2011).

Building on the language ability models in language testing research (Bachman, 1990; Bachman & Palmer, 1996, 2010), the interactionalist approach is adapted for English/Chinese ICPTs. Figure 3.4 describes different components of the model. As shown in the figure, the proposed model is not an IP model, but is complementary to a SI processing model such as Setton’s (1999). The model also does not intend to include traditional components such as different skills and competences, because the model treats them as interpreters’ internal mental processes. The model is multi-componential. But the multi-componentiality does not strive for a complete representation of every aspect of SI, but only for those aspects that are

critical. This is because “no model is meant to correspond exactly to the phenomena” (Moser-Mercer, 1997, p. 159). If it did, the model would no longer be a surrogate for the phenomena, but reality itself.

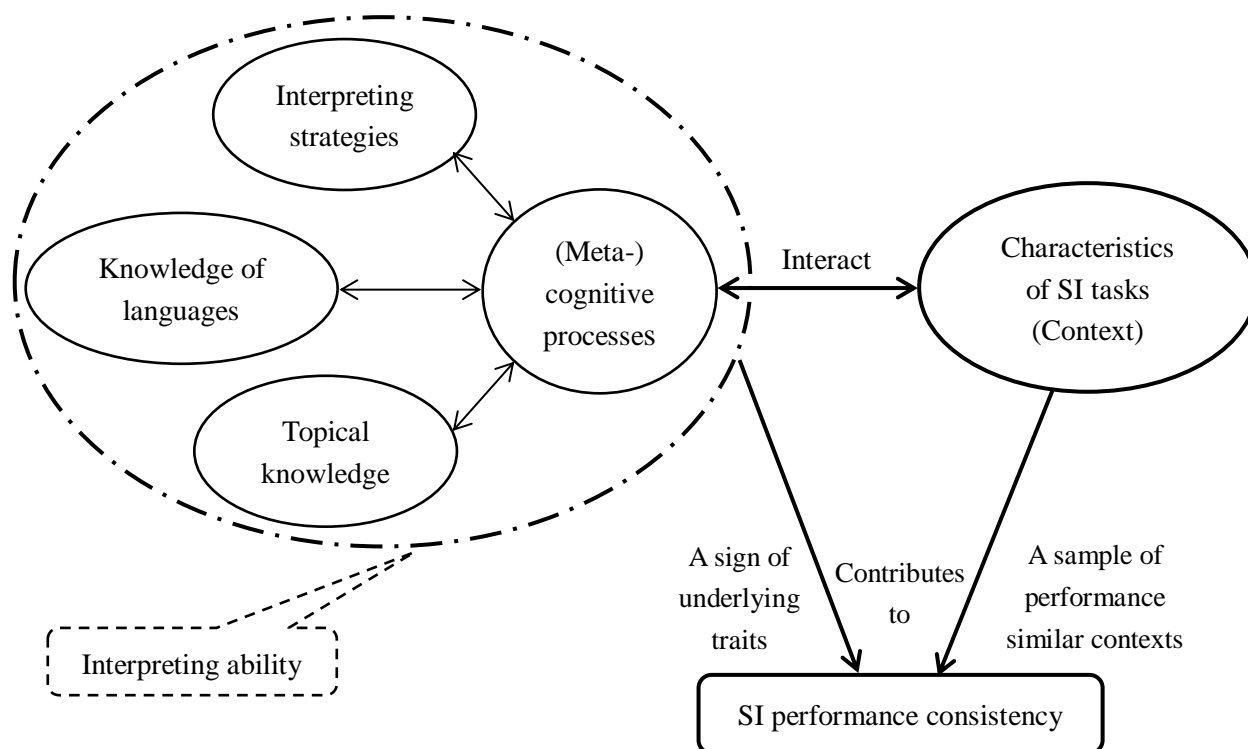


Figure 3.4 An interactionalist approach to interpreting performance consistency

### 3.3.3 Components of interpreting ability

This section specifies components of interpreting ability, which includes knowledge of languages, interpreting strategies, topical knowledge, and (meta-)cognitive processes. The first of three components are involved and engaged in (meta-)cognitive processes.

#### 3.3.3.1 Knowledge of languages

The most widely and unanimously agreed component of interpreting ability is bilingual knowledge (Gile, 1995; Kopczynski, 1980), which is a prerequisite for interpreting (e.g., Angelelli & Degueldre, 2002). In the model, knowledge of languages (both source and target languages) is assumed to be identical with language knowledge in the communicative language ability model (Bachman & Palmer, 1996). Figure 3.5 shows different aspects of language knowledge. As can be seen, it is hierarchical in nature, and consists of two major types of knowledge: organizational and pragmatic knowledge.

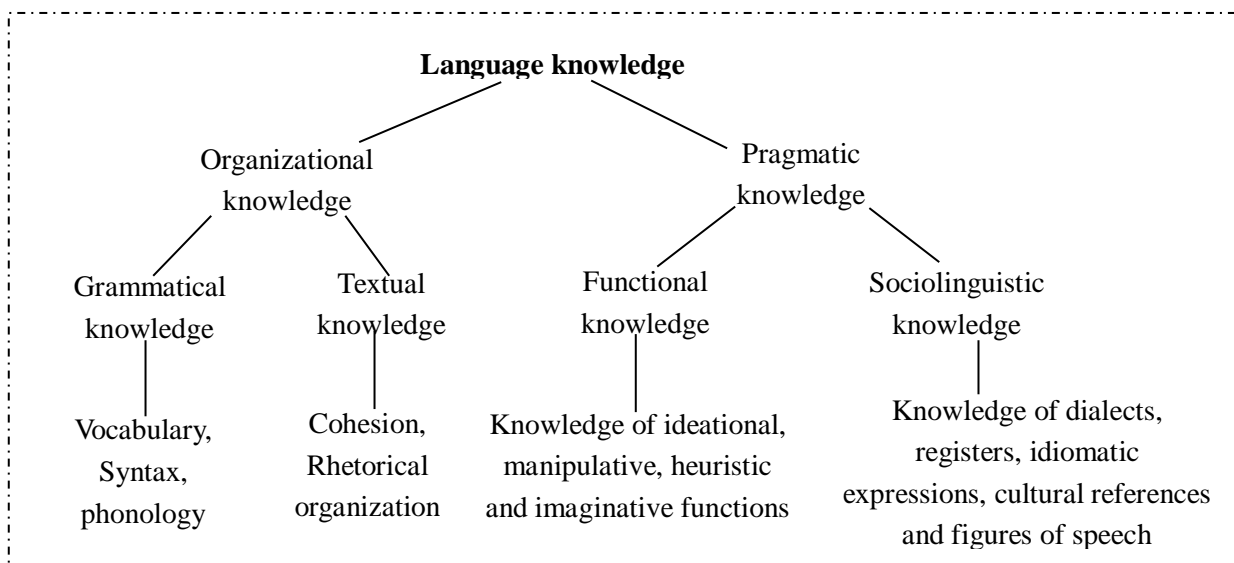


Figure 3.5 Structure and components of language knowledge (Bachman & Palmer, 1996)

### 3.3.3.2 Interpreting strategies

Good knowledge of interpreting strategies is an important part of the interpreting ability construct. It is important, because SI is a strategic activity (Chang & Schallert, 2007; Kohn & Kalina, 1996), and requires constant decision making by appropriately employing strategies (Pym, 2008; Riccardi, 2005). In addition, strategy use is typical of and crucial for the result of SI (Kalina, 2000). Shlesinger (2000, p. 7) even observes that “a strategy which is used regularly by competent professionals tends to acquire normative force”. Interpreting strategies have been categorized using different schemes (e.g., Bartłomiejczyk, 2006; Donato, 2003; Gile, 1995; Riccardi, 2005), and some categorization schemes are empirically derived based on parallel text analysis (Donato, 2003) or a retrospective verbalization technique (Bartłomiejczyk, 2006).

### 3.3.3.3 Topical knowledge

Topical knowledge is included as part of the construct for two important reasons. One reason is that previous interpreting ability theories (e.g., Gile, 1995; Kalina, 2000; Kopczynski, 1980; Wang, 2007; Yuan, 2007) recognize the role of topical knowledge in SI. The other reason is that results from empirical studies (e.g., Díaz-Galaz, 2011) indicate that topical knowledge could potentially influence interpreting performance, especially when interpreting for a technical subject matter.

#### 3.3.3.4 (Meta-)cognitive processes

Based on Bachman and Palmer (1996), cognitive processes refer to procedures and mechanisms whereby all kinds of information are processed and manipulated, and meta-cognitive processes pertain to higher-order executive procedures that assess, monitor and plan cognitive processes and interpreting performance.

Regarding the cognitive processes, they usually take place in three stages outlined in SI processing models (e.g., Gerver, 1975; Moser, 1978; Setton, 1999) including 1) input reception and comprehension stage, 2) SL-TL conversion stage, and 3) output production stage. In the first stage, visual and auditory SL information is temporarily stored in “operational memory” (Gerver, 1975) or “working memory” (Daro & Fabbro, 1994), and processed by “parser” and “assembler” with the help of TL knowledge and contextual characteristics (Setton, 1999). In the second stage, the processed SL input exists as concepts, and are further processed into TL concepts via a “conceptual base/network” (Moser, 1978) or a “formulator” (Setton, 1999). In the third stage, the TL concepts are encoded with the help of TL knowledge via the “parser” (Setton, 1999), and finally articulated.

In terms of the meta-cognitive processes, they are supposed to oversee all lower-order processes and performance. For instance, some meta-cognitive processes are involved in the first stage to assess adequacy of comprehension, others to appraise information reformulation and monitor final performance (e.g., Setton, 1999, “self-monitoring”). Once potential problems are forecasted and/or detected, the meta-cognitive processes are involved in planning a response, or strategy to overcome hindrances.

During SI, language knowledge, knowledge of interpreting strategies, topical knowledge and contextual characteristics interact with these (meta-)cognitive processes, feeding them necessary informative materials. The (meta-)cognitive processes also respond to and interact with characteristics of SI tasks (e.g., delivery speed, linguistic features of speech, characteristics of situational context).

#### 3.3.4 A framework of SI task characteristics

In addition to the components that directly constitute “interpreting ability”, another crucial part of the interactionalist model is a framework of SI task characteristics. The framework provides a principled way to describe an SI task that could be sampled in an ICPT. Campbell

and Hale (2003) call for building such a framework, drawing upon language testing and assessment literature. The framework of SI task characteristics presented in the chapter is thus modeled against Bachman and Palmer (1996), and draws upon relevant literature in Interpreting Studies (e.g., Kalina, 2002) to extend a previous similar framework from Chen (2009). Table 3.1 presents the framework. As shown, there are four dimensions: 1) a situational context that provides a background for SI tasks; 2) a physical context (the booth) where SI tasks are performed; 3) characteristics of SI tasks; and 4) expected SI performance. The descriptions in the table are not intended to be either exhaustive or definitive; hopefully it can serve as a heuristic tool to characterize different facets of SI tasks.

#### *3.3.4.1 Situational context*

The situational context situates SI tasks in a meaningful environment, giving them purpose and background. It is similar to what is called the “hypertext skopos” (Pöchhacker, 1995). That is, it sets the scene on which focal context supersedes.

#### *3.3.4.2 Physical (booth) condition*

A booth is the immediate environment where SI tasks are performed. Based on the literature, the booth environment is characterized by acoustic separation, visibility, ergonomic design (Jumpelt, 1985), and physical parameters (e.g., air quality, ventilation and lighting) (AIIC, 2002). Adverse booth conditions not only cause physiological exhaustion and post-work stress for interpreters (AIIC, 2002), but also reduce interpreting quality (e.g., Gerver, 1974).

#### *3.3.4.3 Characteristics of SI tasks*

This section describes characteristics of SI tasks, with respect to task-level characteristics, linguistic and paralinguistic characteristics of source speeches, and kinesic characteristics of speakers.

Table 3.1 A framework of SI task characteristics

<b>Situational context</b>	SC 1. Conference setting (e.g., theme, time, place)	
	SC 2. Participants (e.g., speaker, audience, relationship, status)	
	SC 3. Purpose (e.g., purpose of communication, speaker's motive)	
<b>Physical condition</b>	PC 1. Acoustics, visibility and ergonomic design (e.g., sound insulation)	
	PC 2. Physical parameters (e.g., air quality, temperature, ventilation)	
<b>SI tasks</b>	CST 1. Task	CST 1.1 Conference-related materials received (e.g., PPT)
		CST 1.2 When received?
		CST 1.3 Task type (e.g., SI with PPT, SI with Text)
		CST 1.4 Topical characteristics (e.g., economics, politics)
		CST 1.5 Directionality (e.g., English to Chinese)
		CST 1.6 Duration of an interpreting turn
	CST 2. Linguistic	CST 2.1 Lexical characteristics (e.g., word/vocabulary)
		CST 2.2 Semantic and propositional characteristics
		CST 2.3 Syntactic structure (e.g., subordination)
		CST 2.4 Textual organization (e.g., cohesion, logic, genre)
		CST 2.5 Pragmatic characteristics (e.g., illocutionary force)
		CST 2.6 Sociolinguistic characteristics (e.g., register)
	CST 3. Paralinguistic	CST 3.1 Delivery speed
		CST 3.2 Non-native speaker (NNS) (e.g., intonation)
	CST 4. Kinesic	CST 4.1 Gestures, manners and postures
<b>Expected response</b>	ER 1. Accuracy of interpretation (e.g., fidelity, sense consistency)	
	ER 2. Language quality of interpretation (e.g., grammaticality)	
	ER 3. Delivery of interpretation (e.g., fluency, pronunciation)	

### 3.3.4.3.1 Task-level characteristics

*Conference-related materials* In the real-life interpreting practice, a wide range of materials could be received (e.g., Kalina, 2002). The materials, if received prior to a conference, are of great value in helping interpreters prepare for SI tasks (e.g., Gile, 1995, 2002). Gile (1995) observes that as an important part of working conditions, conference organizers should provide in advance a full of set of relevant documents to interpreters.

*When received* Another important issue is when interpreters received the conference-related materials. Timing is of significance because it dictates available time for preparation (AIIC, 2004; Donavan, 2001; Kalina, 2002). The timing also affects how interpreters prepare (AIIC, 2004; Gile, 1995, 2002).

*Task type* Interpreters are said to be “frequently confronted with the task of interpreting on the basis of written manuscripts or overhead transparencies” (Kalina, 2002, p. 17). Interpreting researchers have also mentioned other types of SI tasks, for example, SI with PPT (Wan,

2004), SI with texts (Anderson, 1979; Wang & Lin, 2006), SI with other supplementary materials (e.g., a summary or abstract) (Anderson, 1979), and SI without any materials (Barik, 1973).

*Topical characteristics* This aspect pertains to specialized knowledge involved in SI tasks. Presumably, the more technical and jargon-laden a conference is, the more time and effort should be set aside for preparation. AIIC (2002) mentions “interpreters’ information deficit” and “frequent change of subject matter” as sources of stress for interpreters.

*Directionality* It is recommended that interpreters work into their A language(s) to ensure interpretation quality (e.g., Seleskovitch & Lederer, 1989). However, some researchers have discussed “interpreting into B” (e.g., Godijns & Hinderdael, 2005). In addition, some empirical studies suggest reasons for interpreting both into and from A language (Pan, Sun, & Wang, 2009)

*Duration of one interpreting turn* As a cognitive venture, SI is taxing, even for professionals. It is recommended that an optimal duration of an interpreting turn be approximately 20 minutes and a maximum 30 minutes (Chmiel, 2008; Moser-Mercer, Künzli, & Korac, 1998). Working too long in one turn could have negative effects on interpreters’ well-being (Klonowicz, 1991) and on interpretation quality (Kalina, 2002).

#### 3.3.4.3.2 Linguistic characteristics of source speeches

*Lexical characteristics* This aspect primarily deals with vocabulary. It is suggested that some types of vocabulary cause more cognitive load, for example, proper names, figures, acronyms and terminology (Gile, 1995, 2008; Kalina, 2005; Vianna, 2005). These words or word-like elements are meaning-laden, and of specific domains, thus requiring interpreters to summon more efforts.

*Semantic and propositional characteristics* This aspect pertains to semantic complexity and propositional density. They usually take the form of a cascade of quickly presented enumerations, meaning units and culturally-loaded expressions (e.g., puns) (Gile, 1995, 2008; Kalina, 2005; Vianna, 2005), which may cause cognitive saturation. Irrealis, counterfactuals and attributed beliefs are also supposed to cause processing difficulties for SI (Bülow-Møller, 1999; Setton, 2002a, 2002b).

*Syntactic structure* Syntactic organization in source texts can affect SI processing. For example, in a case where sentence structure is multi-layered and complicated, it hinders

comprehension in SI (Gile, 2008; Meuleman & Besien, 2009). However, the theme-rheme structure is said to be a “path indicator” in guiding correct understating of source texts in SI (Consorte, 1999; Torsello, 1996).

*Textual organization* This aspect generally refers to text cohesion and rhetorical organization. On the one hand, the illogical thread of or “tortuous logic” of source texts triggers processing problems in SI (Gile, 2008; Kalina, 2005; Setton, 2001). On the other hand, if interpreters are familiar with typical structures of source texts, they can work with more confidence and even predict what is to be presented in a speech (Bao, 1998).

*Pragmatic characteristics* The pragmatic dimension of source speeches primarily pertains to underlying message speakers convey. Application of relevance theory in SI pertains to the pragmatic dimension (see Vianna, 2005). Specifically, with the aid of pragmatic resources, a mental model of SI helps analyze and derive attitude, intentionality and implicatures from source speeches (Setton, 1999, 2001).

*Sociolinguistic characteristics* This aspect relates to language variety (i.e., dialect) and register. The former is relevant to SI because conference speakers may belong to a similar culture, but speak different language varieties (e.g., Arabic, Chinese). Interpreters are even encouraged to use the same language variety as the participants do (Gold, 1973). The latter aspect, register, is also closely related to SI. This is because in different contexts speakers tend to use different registers or speech levels (Ardito, 1999; Setton, 2001; Vianna, 2005), which may impact the way interpreters work.

#### 3.3.4.3.3 *Paralinguistic characteristics of source speeches*

*Delivery speed* Delivery speed is often cited as one of major factors contributing to cognitive load during SI (e.g., Cooper, Davis, & Tung, 1982; Meuleman & Van Besien, 2009; Pio, 2003). Specifically, interpretation quality varies as a function of speech rate, with fast delivery associated with poor quality (Gerver, 1969/2002).

*Accent and non-native speaker (NNS)* As a lingua franca, English is used by non-native speakers in conference settings, which has consequences for interpreting (Albl-Mikasa, 2010). Interpreters’ performance could be influenced by speakers’ accented speech or non-native speakers (NNSs) due to phonemic and intonational deviation and degradation (e.g., Gile, 2008). Survey findings also report NNSs as one of work stressors (AIIC, 2002).

#### *3.3.4.3.4 Kinesic characteristics of speakers*

Referred to as conscious or unconscious psycho-muscularly-based body movements (Poyatos, 1987), kinesics includes gestures, manners that are mainly learned and socially ritualized, and postures that are codified by social norms. Kinesics could also influence interpreting processes (Poyatos, 1987; Anderson, 1979; Rennert, 2008).

#### *3.3.4.3.5 Expected response from interpreters*

Expected interpretation quality is included as one dimension of the framework, because it contributes to a meaningful context in which interpreters are expected of their performance. Expectations on interpreting performance may vary as a function of heterogeneous groups of interpretation users (Pöchhacker, 1995). Therefore, knowing different expectations in advance helps create a meaningful context and even has normative effects on interpreters. For example, Shlesinger (1997) observes that the Chinese delegation to the UN demands for a rather literal rendering instead of style and fluency.

Although the expected interpretations may differ among groups, studies have revealed a number of consistently mentioned quality criteria (e.g., Pöchhacker, 2001, 2002, 2005). Pöchhacker (2001) states that considerable agreement emerges in the community of interpreting studies as to what common criteria should be used for interpretation quality assessment. Overall, three general criteria have been persistently mentioned including accuracy, language quality and delivery of interpretation.

### **3.4 Implications**

Given that the interactionalist construct model and its relationship to interpreting performance is theoretical, it does not therefore automatically lead to trait- and performance-referenced score interpretations. However, the model provides a heuristic tool to help interpreting testers organize their thoughts on test design. For the theoretical model to be useful to design of ICPTs, the framework of SI task characteristics needs to be fleshed out with empirical data, and the interactions between SI task characteristics, interpreting ability and interpreting performance needs to be investigated to gain an in-depth understanding. At least four lines of research could thus be conducted.

The first line of research is to profile the real-life English/Chinese conference interpreting practice in China, or to obtain empirical data that describe characteristics of the practice domains, which is also known as “domain analysis” research in the testing and assessment literature (Mislevy, Steinberg, & Almond, 2003). By doing so, interpreting researchers would obtain detailed information on characteristics of SI tasks performed in the real-life practice domains of interest. For example, researchers would know what kind of SI tasks interpreters frequently undertake in the target practice domain, and what characteristics of the SI tasks are (e.g., any materials received, duration, features of source texts). Based on the empirical profiling, ICPT testers would be better informed to define the real-life practice domains (e.g., international conferences, legal, court settings) to which test scores are to be generalized. More importantly, informed by the empirically-derived framework, ICPT developers would sample into a test relevant and representative SI tasks (with their respective characteristics). Ultimately, as can be seen in Figure 3.6, the degree of correspondence of test tasks and real-life tasks in terms of task structure and associated characteristics pertains to test “authenticity” (Bachman & Palmer, 1996). In this PhD thesis, the first line of research represents the first major research question (RQ): **RQ 1** What are the characteristics of the real-life English/Chinese conference interpreting practice in China?

The second line of research is to examine how SI task characteristics affect interpreting performance (e.g., Pio, 2003; Shlesinger, 2003). Identifying particular task characteristics or “input variables” (Pöchhacker, 2004) that affect task difficulty could help testers develop multiple versions of a test with similar difficulty *a priori*. Keeping the difficulty level consistent across different forms of a test not only relates to parallel-form reliability (Bachman, 1990), but also to test fairness (Kunnan, 2000).

The third line of research is to systematically investigate the interactive features between SI task characteristics and components of interpreting ability. For example, empirical studies could be conducted to explore how SI task characteristics (e.g., speech rate, syntactic complexity, propositional density) affect the use of interpreting strategies (e.g., Meuleman & Van Besien, 2009) or underlying cognitive processes (e.g., Liu, Schallert, & Carroll, 2004). This line of research is of great importance, in that it would produce evidence that speaks to how SI tasks and associated characteristics elicit and engage knowledge, skills and abilities that ICPT developers are interested in. In other words, it is related to the crucial test quality of “interactiveness” and construct validity (Bachman & Palmer, 1996).

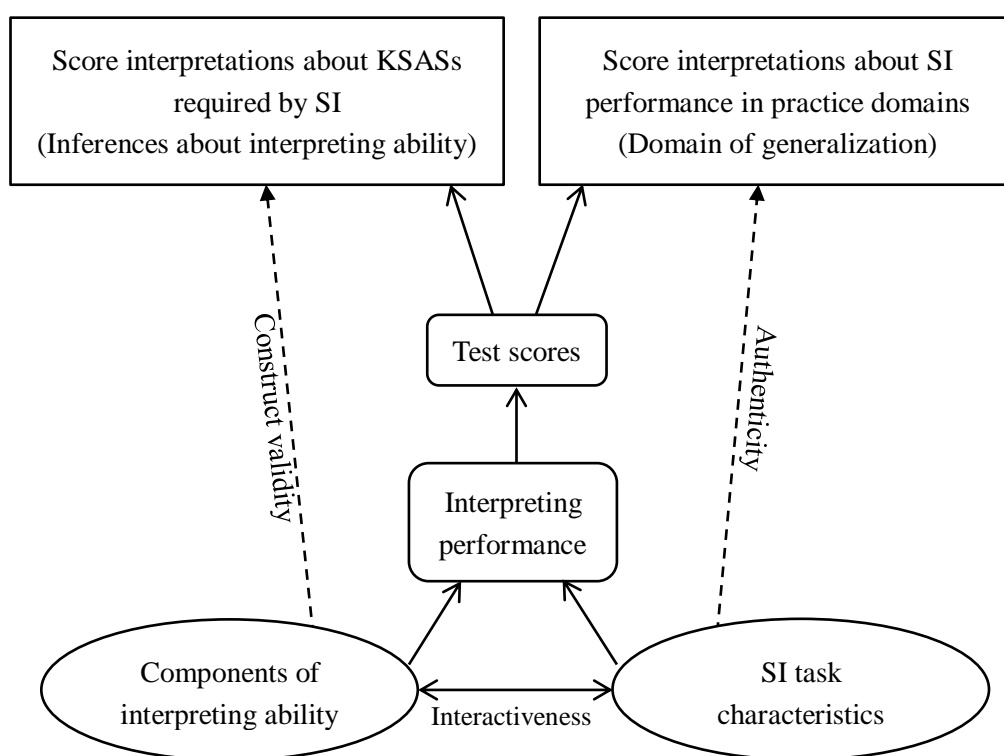


Figure 3.6 The interactionalist construct model and score interpretations

The last line of research is to explore how the use of certain knowledge, skills and strategies affects interpreting performance. For example, Meuleman and Van Besien (2009) found that in interpreting a fast-paced source text, almost all interpreters who utilized the tailing strategy yielded an acceptable translation, while an absence of strategies resulted in failure.

Based on the results from these lines of research, ICPT developers are able to 1) sample adequate SI tasks and their characteristics to establish a domain of generalization, and 2) to determine whether certain SI tasks can effectively engage and elicit desirable components of interpreting ability. Following rigorous test design and development procedures (Bachman & Palmer, 1996), score-based inferences can be made in reference to interpreting ability (i.e., the trait approach), and performance levels in a certain practice domain (i.e., the behaviorist approach) (see Figure 3.6, adapted from Bachman & Palmer, 1996).

In this PhD thesis, the last three lines of research represent the second major research questions: **RQ 2** What is the possible interplay between SI task characteristics, interpreting ability and SI performance quality? Logically, subsumed under RQ 2 are three sub-RQs: **RQ 2.1** What are the effects of SI task characteristics on SI performance quality? **RQ 2.2** What

are the effects of SI task characteristics on use of interpreting ability? **RQ 2.3** What is the relationship between strategy use and SI performance quality?

### 3.5 Conclusion

ICPT is developing with good momentum. This rapid development requires a solid evidentiary support to underpin score-based inferences and actions. This paper identifies a disjunction between the current approaches to construct definition and the way test scores are explained. As the initial step to narrow this gap, an interactionalist approach to interpreting ability is proposed, which postulates that SI performance consistency is a result of interpreting ability, contextual characteristics and interaction between the two. Given that the interactionalist model is largely theoretical, empirical studies are needed to test these hypotheses. In addition, certification testing has been extensively discussed in other occupations, and a large body of literature has been accumulated. In order not to reinvent the wheel, the paper ends by calling for importing of proven techniques and methodologies from the broader assessment and testing community into ICPT.

### 3.6 References

- AIIC. (2002). *Interpreter workload study – Full report*. Retrieved from <http://aiic.net/page/657/interpreter-workload-study-full-report/lang/1>
- AIIC. (2004). *Practical guide for professional conference interpreters*. Retrieved from <http://aiic.net/ViewPage.cfm/article21.htm>.
- Albl-Mikasa, M. (2010). Global English and English as a lingua franca (ELF): Implications for the interpreting profession. *Trans-kom*, 3(2), 126-148.
- Anderson, L. (1979). *Simultaneous Interpretation: Contextual and Translation Aspects* (Master's thesis, Concordia University, Canada). Retrieved from <http://spectrum.library.concordia.ca/5/1/MK43196.pdf>
- Angelelli, C. (2009). Using a rubric to assess translation ability: defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp. 13-47). Amsterdam: John Benjamins.
- Angelelli, C., & Degueudre, C. (2002). Bridging the gap between language for general purposes and language for work: An intensive superior level language/skill course for

- teachers, translators, and interpreters. In B. L. Leaver & B. Shekhtman (Eds.), *Developing professional-level language proficiency* (pp. 91-110), Cambridge, UK: Cambridge University Press.
- Ardito, G. (1999). The systematic use of impromptu speeches in training interpreting students. *The Interpreters' Newsletter*, 9, 177-89.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.
- Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville, C. A. Chapelle & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165-207). Amsterdam: John Benjamins.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. Doe (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41-71). Ottawa: University of Ottawa Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, UK: Oxford University Press.
- Bacigalupe, L. A. (2010). Information processing during simultaneous interpretation: a three-tier approach. *Perspectives*, 18(1), 39-58.
- Bao, G. (1998). 口译理论概述. 旅游教育出版社. [An Overview of Interpreting Theories]. Beijing: Tourism Education Press.
- Barik, H. C. (1973). Simultaneous interpretation: Temporal and quantitative data. *Language and Speech*, 16(3), 237-270.
- Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting*, 8(2), 149-174.
- Bülow-Møller, A. M. (1999). Existential problems: On the processing of irrealis in simultaneous interpreting. *Interpreting*, 4(2), 145-168.

- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Chang, C., & Schallert, D. L. (2007). The impact of directionality on Chinese/English simultaneous interpreting. *Interpreting*, 9(2), 137-176.
- Chang, C., & Wu, M. M. (2009). Address form shifts in interpreted Q&A sessions. *Interpreting*, 11(2), 164-189.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70), Cambridge, UK: Cambridge University Press.
- Chapelle, C. A. (2006). L2 vocabulary acquisition theory: The role of inference, dependability and generalizability in assessment. In M. Chalhoub-Deville, C. A. Chapelle & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspective* (pp. 47-64). Amsterdam: John Benjamins.
- Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, 3(2), 257-273.
- Chmiel, A. (2008). Boothmates forever? - On teamwork in a simultaneous interpreting booth. *Across Languages and Cultures*, 9(2), 261-276.
- Consorte, C. (1999). Thematic structure and simultaneous interpretation. Some experimental evidence. *Interpreters' Newsletter*, 9, 99-124.
- Cooper, C. L., Davis, R., & Tung, R. L. (1982). Interpreting stress: Sources of job stress among conference interpreters. *Multilingua*, 1(2), 97-108.
- Daro, V., & Fabbro, F. (1994). Verbal memory during simultaneous interpretation: Effects of phonological interference. *Applied Linguistics*, 15(4), 365-381.
- De Groot, A. M. B., & Christoffels, I. K. (2006). Language control in bilinguals: Monolingual tasks and simultaneous interpreting. *Bilingualism: Language and Cognition*, 9(2), 189-201.
- Deville, C., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C. A. Chapelle & P. A. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspective* (pp. 9-25). Amsterdam: John Benjamins.

- Díaz-Galaz, S. (2011). The effect of previous preparation in simultaneous interpreting: Preliminary results. *Across Languages and Cultures*, 12(2), 173-191.
- Donato, V. (2003). Strategies adopted by student interpreters in SI: A comparison between the English-Italian and the German-Italian language pairs. *The Interpreters' Newsletter*, 12, 101-134.
- Donovan, C. (2001). Interpretation of technical conferences. *Conference Interpretation and Translation*, 3, 4-22.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179-197.
- Englund-Dimitrova, B., & Hyltensta, K. (Eds.). (2000). *Language Processing and Simultaneous Interpreting: Interdisciplinary Perspectives*. Amsterdam: John Benjamins.
- Gerver, D. (1969/2002). The effects of source language presentation rate on the Performance of simultaneous conference interpreters, Republished in F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 52-66). London and New York: Routledge.
- Gerver, D. (1974). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica*, 38(3), 159-167.
- Gerver, D. (1975). A psychological approach to simultaneous interpretation. *Meta*, 20(2), 119-128.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.
- Gile, D. (2002). The Interpreter's preparation for technical conferences: Methodological questions in investigating the topic. *Conference Interpretation and Translation*, 4(2), 7-27.
- Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum*, 6(2), 59-77.
- Godijns, R, & Hinderdael, M. (Eds.). (2005). *Directionality in Interpreting. The 'Retour' or the Native?* Gent: Communication and Cognition.
- Gold, D. L. (1973). On quality in interpretation. *Babel*, 19(4), 154-155.
- Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, 25(4), 21-35.
- Huang, M. (2005). 谈口译资格认证考试的规范化设计. [Toward a more standardized large-scale accreditation test for interpreters]. *中国翻译*, 6, 62-65.

- Jumpelt, W. R. (1985). The conference interpreters' working environment under the new ISO and IEC standards. *Meta*, 30(1), 82-90.
- Kalina, S. (2000). Interpreting competences as a basis and a goal for teaching. *The Interpreters' Newsletter*, 10, 3-32.
- Kalina, S. (2002). Quality in interpreting and its prerequisites A framework for a comprehensive view. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century: Challenges and Opportunities* (pp. 121-130). Amsterdam/Philadelphia: John Benjamins.
- Kalina, S. (2005). Quality assurance for interpreting processes. *Meta*, 50(2), 769-784.
- Klonowicz, T. (1991). The effort of simultaneous interpretation: It's been a hard day... *FIT Newsletter*, 9(4), 446-457.
- Kohn, K., & Kalina, S. (1996). The strategic dimension of interpreting. *Meta*, 41(1), 118-138.
- Kopczynski, A. (1980). *Conference Interpreting: Some Linguistic and Communicative Problems*. Poznan: Wydawn.
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1-14). Cambridge, UK: Cambridge University Press.
- Liu, M. H., Schallert, D., & Carroll, P. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting*, 6(1), 19-42.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966.
- Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, 89(3), 575-588.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Meuleman, C., & Van Besien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting*, 11(1), 20-34.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-66.
- Mizuno, A. (2005). Process model for simultaneous interpreting and working memory. *Meta*, 50(2), 739-752.

- Moser, B. (1978). Simultaneous interpretation: A hypothetical model and its practical application. In D. Gerver & W. Sinaiko (Eds.), *Language, Interpretation and Communication* (pp. 353-368). New York/London: Plenum Press.
- Moser-Mercer, B. (1997). Skill components in simultaneous interpreting. In Y. Gambier, D. Gile & C. Taylor (Eds.), *Conference interpreting: Current trends in research* (pp. 133-148). Amsterdam: John Benjamins.
- Moser-Mercer, B., Künzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (pilot study). *Interpreting*, 3(1), 47-64.
- Pan, J., Sun, Z. X., & Wang, H. H. (2009). 口译的职业化与职业化发展 – 上海及江苏地区口译现状调查研究. [Professionalization in interpreting: current development of interpreting in Shanghai and Jiangsu Province]. *解放军外国语学院学报*, 6, 81-85.
- Pio, S. (2003). The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter*, 12, 69-100.
- Pöchhacker, F. (1995). Simultaneous interpreting: A functionalist approach. *Hermes*, 14, 31-53.
- Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta*, 46(2), 410-425.
- Pöchhacker, F. (2002). Researching interpreting quality: Models and methods. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century: Challenges and Opportunities* (pp. 95-106). Amsterdam/Philadelphia: John Benjamins.
- Pöchhacker, F. (2005). Quality research revisited. *The Interpreters' Newsletter*, 13, 143-166.
- Poyatos, F. (1987). Nonverbal communication in simultaneous and consecutive interpretation: A theoretical model and new perspectives. *TEXTconTEXT*, 3(2), 73-108.
- Pym, A. D. (2008). On omission in simultaneous interpreting: Risk analysis of a hidden effort. In G. Hansen, A. Chesterman & H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile* (pp. 83-105). Amsterdam: John Benjamins.
- Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1-32.
- Rennert, S. (2008). Visual input in simultaneous interpreting. *Meta*, 53(1), 204-217.
- Riccardi, A. (2005). On the evolution of interpreting strategies in simultaneous interpreting. *Meta*, 50(2), 753-767.

- Seleskovitch, D., & Lederer, M. (1989). *Pédagogie raisonnée de l'interprétation* [A reasoned pedagogy of interpreting]. Bruxelles : Didier Érudition.
- Setton, R. (1997) *A Pragmatic Theory of Simultaneous Interpretation* (Unpublished doctoral thesis) The Chinese University of Hong Kong, China.
- Setton, R. (1999). *Simultaneous Interpretation: A Cognitive and Pragmatic Analysis*. Amsterdam and Philadelphia: John Benjamins.
- Setton, R. (2001). Deconstructing SI: a contribution to the debate on component processes. *The Interpreters' Newsletter*, 11, 1-26.
- Setton, R. (2002a). Pragmatic analysis as a methodology: A reply to Gile's review of Setton (1999), *Target*, 14(2), 353-360.
- Setton, R. (2002b). Seleskovitch: A radical pragmatist before her time. *The Translator*, 8(1), 117-124
- Shlesinger, M. (1997). Quality in simultaneous interpreting. In Y. Gambier, D. Gile & C. Taylor (Eds.), *Conference Interpreting: Current Trends in Research* (pp. 123-132). Amsterdam / Philadelphia: John Benjamins.
- Shlesinger, M. (2000). Interpreting as a cognitive process. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of Translation and Interpreting: Outlooks on empirical research* (pp. 3-15). Amsterdam/Philadelphia: John Benjamins.
- Shlesinger, M. (2003). Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreters' Newsletter*, 12, 37-50.
- Stenner, A. J., Smith III M., & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, 20(4), 305-316.
- Torsello, C. T. (1996). Theme as the interpreter's path indicator through the unfolding text. *The Interpreter's Newsletter*, 7, 113-149.
- Vianna, B. (2005). Simultaneous interpreting: A relevance-theoretic approach. *Intercultural Pragmatics*, 2(2), 169-190.
- Wan, H. Y. (2004). 解读图表:另一项重要的口译技能. [Interpreting graphics: An important skill for interpreters]. 中国翻译, 2, 83-86.
- Wang, B. H. (2007). 口译能力评估和译员能力评估-口译的客观评估模式初. [From interpreting competence to interpreter competence - A tentative model for objective assessment of interpreting]. 外语界, 3, 44-50.
- Wang, L., & Lin, W. (2006). Interpretation training: SI with text. In M. J. Cai & A. L. Zhang

- (Eds.), *Professionalization in interpreting: International experience and developments in China* (pp. 237-244). Shanghai: Shanghai Foreign Language Education Press.
- Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- Wu, S. C. (2010). *Assessing Simultaneous Interpreting: A study on Test reliability and Examiners' assessment behavior*. (PhD thesis, Newcastle University upon Tyne, UK). Retrieved from <https://theses.ncl.ac.uk/dspace/bitstream/10443/1122/1/Wu%2011.pdf>
- Young, R. F. (2000). *Interactional competence: Challenges for validity*. Retrieved from [http://www.english.wisc.edu/rfyoung/IC\\_C4V.Paper.PDF](http://www.english.wisc.edu/rfyoung/IC_C4V.Paper.PDF)
- Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). London and New York: Routledge.
- Yuan, X. L. (2007). 口译能力与口译测试有用性之关系研究. [On the relationship between interpretation ability and interpretation testing usefulness]. *外语教育*, 28(5), 87-90.

## **An introductory note to Chapter 4**

In Chapter 3, an interactionalist approach to construct definition for ICPTs has been proposed. This approach offers a way to infer from interpreting performance something about both a practice-specific behavior and a practice-independent, person-specific trait, which is generally consistent with the score interpretations proposed by ICPTs. In this interactionalist construct model, an emphasis is placed on interpreting ability, characteristics of SI tasks and interpreting performance, and possible interactions between them. Particularly, the theoretical interactionalist construct model gives rise to four lines of research and two major research questions (RQs) the thesis tries to answer.

In Chapter 4, the author intends to provide a preliminary answer to **RQ 1** what are the characteristics of the real-life English/Chinese conference interpreting practice in China? Answers to this RQ provide empirical and real-life data to flesh out the framework of SI task characteristics, which will in turn influence interpreting testers on how to design authentic test tasks to be used in ICPTs.

As will be seen in Chapter 4, an exploratory diary study is first conducted, based on 11 English/Chinese interpreters. The diary study aims to collect initial empirical data on characteristics of interpreting practice. Informed by the diary results, a survey is subsequently designed and administered to a larger sample of 140 interpreters to profile interpreting practice in China. Results are expected to provide preliminary information on how test tasks could be designed.

## Chapter 4 Profiling conference interpreting practice in China: A sequential-exploratory mixed-methods design study<sup>22</sup>

**Abstract.** *An empirical and detailed description of real-life interpreting practices is potentially beneficial to interpreting researchers, students, educators and testers. However, few empirical studies have been conducted to generate data that account for the real-life practice in China. The present study therefore was conducted to provide an initial empirical description of the interpreting practice in China, based on a diary study (n=11) followed by a survey (n=140). Main findings are 1) even though the interpreters received conference-related materials in advance, they did not have sufficient time to prepare; 2) the interpreters performed a much wider variety of simultaneous interpreting (SI) tasks than previously thought, with some tasks occurring appreciably more frequently than others; and 3) the interpreters constantly encountered an array of factors contributing to SI difficulty, such as strong accent and fast delivery speed. In addition, a series of ordered probit regression analyses indicate that among three demographic variables, only SI experience had the potential to predict how frequently SI tasks were performed. The findings are discussed in relation to previous relevant studies in China and in other countries.*

### 4.1 Introduction

Conference interpreting has evolved in China since its emergence in 1979 when a China-UN joint training program was established in Beijing (Dawrant & Jiang, 2001). From 1979 to 1993, a total of 103 first-generation conference interpreters were trained in the program, and the majority of them worked for government-related agencies (Wang, 2006). Over the past 20 years (approximately from 1994 to 2014), driven by China's increasing economic, social and

---

<sup>22</sup> This chapter was presented at the 10<sup>th</sup> China National Conference and International Forum on Interpreting at Xiamen University, Xiamen, Fujian province, China, 17-18 October 2014. The sequential-exploratory mixed-methods design (MMR) was operationalized by an exploratory qualitative diary study, followed by a quantitative survey. Specifically, the diary study is published as: Han, C. (2015). Lacunae, myths and legends about conference interpreters: A diary study to explore conference interpreting practice in China. *Perspectives: Studies in Translatology*, 23(3), 440-457. The survey findings were reported in another article, which has been accepted by the journal of *Interpreting* as a report: Han, C., (forthcoming). A survey on the profile of conference interpreting practice in China.

cultural exchanges with other parts of the world, conference interpreting has developed into a profession (Wang, 2005). Particularly, the development has gained momentum in the recent decade, boosted by China's growing language services market (Guo, 2010), the creation of the Master of Translators and Interpreters (MTI) program in 2007, and the launch of the national interpreter certification system in 2003 (i.e., the China Accreditation Test for Translators and Interpreters/CATTI). As the market develops, conference interpreting practice is becoming more dynamic, diverse and challenging, particularly in the private market sector (Dawrant & Jiang, 2001), and the make-up of the profession could also be re-shaped by a recent influx of new interpreting graduates from both at home and overseas.

Given the exciting changes, providing detailed empirical descriptions of the real-life interpreting practice in China would produce potential benefits to interpreting researchers, practitioners, educators and testers. Firstly, there is a need for such descriptions, as the interpreting practice in China may differ from the other countries. China's interpreting market is dominated by English/Chinese interpreting in conference settings (Dawrant & Jiang, 2001), despite other emerging types of interpreting such as community interpreting (Su, 2009). In the other countries, particularly immigration countries such as Australia, the UK and the US, interpreting practice occurs in various settings (e.g., court, hospital, police station, conference), and takes on different forms of interpreting (e.g., dialogue interpreting, sight interpreting). Detailed descriptions of real-life practices would thus contribute to an enhanced understanding of similarities and differences in interpreting across the countries. Secondly, detailed descriptions of the interpreting practice in China would inform future interpreters, especially the increasing number of postgraduate students enrolled in the MTI program, of the professional practice so that they could form appropriate career expectations. Thirdly, detailed descriptions of the interpreting practice would also inform interpreting educators of designing up-to-date instructional and training materials that align with the real-life practice (e.g., Wang & Lin, 2006). Lastly, detailed descriptions of the interpreting practice would inform interpreter certification bodies and testing specialists of developing authentic assessment tasks that tap relevant knowledge, skills and abilities required of a conference interpreter (e.g., Campbell & Hale, 2003; Chen, 2002, 2009; Feng, 2005; Huang, 2005). Particularly, although interpreter certification performance testing (ICPT) is developing rapidly in China, there have been proposals to enrich test content based on characteristics of the real-life practice (Chen, 2009; Feng, 2005; Huang, 2005).

Particularly, regarding the ICPT, there has been no empirical information on the real-life interpreting practices in China on which design and development of English/Chinese ICPTs could be based. It would appear that much of ICPT content, especially the characteristics of interpreting tasks, is primarily based on experts' intuition and personal experience. As a result, ICPT content may not accurately reflect the real-life interpreting practice. For instance, based on Bachman and Palmer's (1996) framework of task characteristics, Chen (2009) compares four certification tests in China, and finds that only 37% of the test task characteristics correspond to those of the real-life interpreting domains, leaving much room for improvement.

Given the potential benefits outlined above and lack of empirical data describing the real-life interpreting practice in China, the present study builds on the results from a previous qualitative diary study and the other relevant literature (reviewed below) to generate quantitative empirical data describing fundamentals of the interpreting practice in China.

## **4.2 Descriptions of the interpreting practice in China and in other countries**

This section first provides an overview of three surveys that have produced general information on the interpreting profession in China, and then reviews the relevant literature that primarily describes interpreting practice in European countries.

### **4.2.1 Surveys on the interpreting profession in China**

Three surveys have been conducted, providing an initial empirical description of the interpreting profession in China (see Table 4.1). Overall, the surveys provide much information on the profession-related issues such as the interpreting market, the level of professionalization, and interpreter training and certification. The information is valuable, but does not describe systematically how the interpreters practice their occupation. Nevertheless, the available descriptions on the practice are presented below.

In Wang's (2005) survey, he found that a large proportion (employers: 45% & interpreters: 36%) of the interpreter-mediated events were related to economy and trade issues; 49% of the employers and 46% of the interpreters reported that the conferences on biomedical science were the most difficult to interpret; 40% of the employers would provide conference-related materials (e.g., bio sketch, glossary, draft speech) to interpreters prior to

the conferences; and 34% of the interpreters would take the initiative to ask for relevant materials.

The second survey was jointly conducted by the Science and Technology Translators Association of the Chinese Academy of Sciences (STTACAS) and a private translation agency (STTACAS & TRANSN, 2007). The survey recruited both translators and interpreters of any language pairs, with a valid sample size of 14,600, 19% of whom reported to be more competent in interpreting than translation. Specifically, about 71% of the respondents were based in China's economically dynamic regions: Beijing (26.4%), Guangdong (14.9%), Shanghai (12.3%), Jiangsu (7.3%), Zhejiang (4.9%), Sichuan (3.2%) and Chongqing (2.4%). In addition, the survey found that the topics most familiar to the translators and interpreters had to do with finance, investment, electronics and government-led foreign exchanges.

Table 4.1 General information on the three surveys

	Wang (2005)	STTACAS & TRANSN (2007)	Pan et al. (2009)
◆ Location	Beijing	Greater China area	Shanghai & Jiangsu province
◆ Target population	Intp. & Emp.	Trans. & Intp.	Intp. & Emp.
◆ Sample size	Intp. (n = 34); Emp. (n = 39)	Trans. (n = 11,826); Intp. (n = 2,774)	Intp. (n = 64); Emp. (n = 59)
◆ Survey content	<ul style="list-style-type: none"> <li>• International conferences;</li> <li>• Supply &amp; demand;</li> <li>• Intp. assessment criteria;</li> <li>• Intp. training &amp; certification.</li> </ul>	<ul style="list-style-type: none"> <li>• Translation-related issues;</li> <li>• Work intensity;</li> <li>• Health problems;</li> <li>• Work-related pressure;</li> <li>• Income level;</li> <li>• Leisure time;</li> <li>• Training.</li> </ul>	<ul style="list-style-type: none"> <li>• Level of professionalization;</li> <li>• Use of professional skills;</li> <li>• Market development;</li> <li>• Intp. certification;</li> <li>• Intp. education.</li> </ul>

Notes: Intp. = Interpreters; Emp. = Employers; Trans. = Translators;

In Pan et al.'s (2009) survey, they found that the interpreters worked both into and from their A language (i.e., Chinese), although the portion of interpreting from Chinese to foreign languages was slightly higher than that of the opposite direction; and almost 40% of the interpreting services were related to commerce and trade (24%), and finance (14%).

Based on the three surveys, the interpreting practice in China is generally related to economic activities and government-led foreign exchanges. But more specific aspects of the practice need to be investigated.

#### 4.2.2 Detailed descriptions of conference interpreting practice

There has been an abundant amount of scholarly literature on conference interpreting practice, but most of it is based on researchers and practitioners primarily in Europe. They typically describe what interpreters do in a “conference cycle” (AIIC, 2004). For example, describing the pre-conference preparation, AIIC (2004), Gile (1995), and Kalina (2002) numerate many possible types of materials for preparation: visual materials (e.g., transparencies), relevant documents (e.g., background papers) and information on speakers (e.g., bio sketch), which can be also categorized into human and textual sources (Gile, 2002). In addition, depending on when interpreters receive the materials, different kinds of preparation can be performed: “advance preparation”, “last-minute preparation” and “in-conference preparation” or “online preparation” (Gile, 1995, 2002). Whereas the advance preparation features systematic study of conference materials such as “long, meticulous reading of background documents and conference documents” (Gile, 2002, p. 9), the last-minute preparation refers to preparation on the premise just before the conference. In practice, it is said that interpreters usually do not have sufficient time to prepare (AIIC, 2002; Donavan, 2001).

When performing SI, interpreters have reported undertaking different types of SI task. For example, Kalina (2002, p. 17) observes that “interpreters are frequently confronted with the task of interpreting on the basis of written manuscripts or overhead transparencies”. In addition, commenting on SI practice in China, Setton (2009, p. 109) points out that interpreters perform SI from “fast, recited formal or ceremonial speeches with little or no preparation”. Similarly, Wang and Lin (2006) believe that SI with text is frequently performed in China. Furthermore, Wan (2004) and Wu (2007) claim that presentation with PowerPoint slides (PPT) in conferences has become a norm in China, which necessitates SI with PPT.

When it comes to characteristics of SI tasks, directionality, duration of an interpreting turn and factors contributing to SI difficulty have been most discussed in the literature. According to Seleskovitch and Lederer (1989), interpreters can only work into their A language(s) to maintain quality. However, in their surveys, Pavlović (2007) and Szabari (2002) found that the interpreters in Europe performed SI between their A and B languages; and in China, Pan et al. (2009) also found that the interpreters in greater Shanghai area worked between Chinese and foreign languages.

For how long an interpreter typically works during one turn in the real-life practice, Moser-Mercer, Kunzli and Korac (1998) recommend an optimal duration of about 20 minutes and Chmiel (2008) claims that the maximum duration should be around 30 minutes. However, anecdotes suggest conference interpreters in China sometimes do interpret for longer period of time in one turn than recommended, due to unknown reasons.

In the literature, many factors have been identified to contribute to SI difficulty (e.g., AIIC, 2002; Gile, 1995, 2008; Kalina, 2005). For example, Setton (2009) regards SI for fast, recited speeches as a hazard, which (he believes) is probably more common in China than elsewhere. In addition, SI difficulty is found to increase with non-native speakers/NNS (e.g., Albl-Mikasa, 2010; Kurz, 2009), complex syntactic structures (e.g., Tommola & Helevä, 1998), background noise (e.g., Gerver, 1971), and propositional complexity (Dillinger, 1990).

To sum up, the scholarly literature discusses the specific aspects of the interpreting practice. But most of it is provided by scholars who describe the practice in European countries. Only a small amount of the literature concerns the practice in China, and most of it is based on individual experience.

### **4.3 An exploratory qualitative diary study**

Given the review results, a diary study was conducted by the author to empirically explore the Chinese/English conference interpreting. Based on the above literature, a PDF electronic event-contingent diary was designed, piloted, revised and finally sent to the interpreters who were scheduled to provide SI for international conferences. The diary required the participants to record information about a target event that occurred during a conference cycle. It examined two major areas of SI practice: 1) conference preparation, and 2) characteristics of an SI task. For part of a completed diary, please see Appendix C. Eleven interpreters participated in the

study and kept diaries for 11 conferences. Diary entries were then coded and analyzed, using NVivo 10. Main findings are 1) the interpreters received conference-related materials such as PPT and draft speech texts, but had insufficient time to prepare; 2) the interpreters performed a much greater variety of interpreting tasks than previously thought; and 3) the interpreters needed to work bi-directionally, and frequently confronted an array of factors underlying SI difficulty such as fast delivery and dense information. For more details, please refer to Han (forthcoming). Although the study generated initial data, they lacked generalizability, due to the exploratory nature of the study.

In summary, based on the literature reviewed, it is found: 1) the empirical descriptions of the practice in China either lack desired generalizability or are not specific enough to be useful for interpreting students, educators and testers; 2) most of the specific descriptions are based on the experience of individual English/Chinese interpreters; and 3) an abundant literature have been generated by European scholars based on individual experience, but the literature may not adequately reflect the Chinese interpreting practice.

Against the backdrop, the diary study was followed by a survey, based on a larger cohort of Chinese/English simultaneous interpreters ( $n = 140$ ). The primary purpose of the survey was thus to provide quantitative and empirical data that describe the conference interpreting practice in China. Survey results were also expected to provide some useful information for interpreting students, educators and testers.

## **4.4 Method**

### **4.4.1 Survey design**

Using the online tool *SurveyMonkey*,<sup>23</sup> the survey was designed to have four sections. Section I introduced the study briefly; Section II profiled demographic information (Question 3 – 9, including gender, age, education, interpreting training, employment status, SI experience, and working location); Section III was the core of the survey and consisted of three parts examining three specific aspects of the practice outlined below; and Section IV was a “Thank you” page.

Specifically, Section III investigated: Part I. What conference-related materials were received by interpreters in advance? (Question 10 – 11); Part II. How frequently 18 varieties of

---

<sup>23</sup> The SurveyMonkey keeps track of the number of participants who entered the survey and who actually completed it, respectively.

SI task were performed by practitioners?<sup>24</sup> (Question 13 – 16); and Part III. What were the characteristics of these SI tasks? (Question 17 – 19, including directionality, duration of an interpreting turn, and factors underlying task difficulty).

More specifically, in Part II, to investigate how frequently the 18 varieties of SI task were performed, 18 Likert-type items were constructed with frequency descriptors attached to each item. These items used a seven-point frequency rating scale and each frequency descriptor was quantified by assigning an arbitrary number (e.g., “Always” - 7, “Never” - 1).

All survey questions were designed based on three strands of sources: 1) the empirical findings from the exploratory diary study, 2) the scholarly literature discussed above, and 3) feedback from interpreters in a small-scale pilot. For a view of the survey, please visit the web link.<sup>25</sup>

#### 4.4.2 Sampling

To participate, three criteria must be met: 1) participants practiced SI; 2) they interpreted between Chinese and English; and 3) they were working in China. In order to boost sample size, a multi-pronged approach was taken. One of the methods was to seek external help. The researcher collaborated with a Chinese website devoted to promoting interpreting profession. The website administrators sent the survey web link and relevant information to its *Weibo* (the Chinese version of Twitter). It was hoped that eligible interpreters would self-select to participate. Another method was to send an invitation email to each of Chinese/English interpreters affiliated with two professional organizations: the AIIC, and the Shanghai Interpreter Association (SIA), the only professional society for conference interpreters in China. The last method was to distribute the survey web link to the interpreters within the researcher’s professional networks. Consequently, non-probability sampling was employed.

#### 4.4.3 Procedure

A draft survey was designed and revised before a pilot involving six interpreters. The interpreters were asked to trial the survey and pay attention to four aspects: 1) the appropriateness and the legitimacy of question stems and response categories, 2) clarity of

---

<sup>24</sup> A majority of the 18 task varieties was identified and categorized in the exploratory diary study. For detailed descriptions of these tasks, please refer to Appendix D.

<sup>25</sup> Online survey: [https://www.surveymonkey.com/s/A\\_profile\\_of\\_conference\\_interpreting\\_practice](https://www.surveymonkey.com/s/A_profile_of_conference_interpreting_practice)

wording, 3) logic of question order, and 4) the amount of time used to complete the survey. Pilot feedback helped further revision. Finally, the online survey was opened to potential respondents. In order to ensure sample quality, participants were asked in the survey introduction page to confirm that they met relevant recruitment criteria (i.e., Question 1-2) before proceeding to the main survey. The survey took 15-20 minutes to complete. At the end of the data collection, 15 respondents were randomly selected to win an electronic gift voucher worth 300RMB (approximately 50 Australian dollars).

#### 4.4.4 Data analysis

A total of 232 hits on the survey web link were recorded by the SurveyMonkey. A meticulous review of each response helped filter 92 invalid responses in which 62 responses were totally blank, 26 responses were incomplete, and four completed responses were also deleted because of apparently contradictory information provided. The reason for the large number of totally blank responses is probably because the interpreters found that they did not meet the recruitment criteria after entering the survey, and exited without answering any questions. As a result, a total number of 140 valid responses were stored and prepared for further analysis.

The organized survey dataset was analyzed using NVivo 10 and Stata 10. While the NVivo 10 was used to analyze the qualitative data (e.g., verbal comments) and produce descriptive and cross-tabulation results, Stata 10 was utilized to perform inferential statistical analyses of the quantitative data (e.g., task frequency ratings). Specifically, an ordered probit regression model was used to investigate the degree and the direction of relationship between the three demographic variables of interest and task frequency ratings. The model is customarily applied to estimate the statistical significance and direction of the relationship each predictor variable has to more than two outcomes of an ordinal dependent variable (Boes & Winkelmann, 2006). Therefore, the model is suited to the study. In addition, a one-way repeated-measures ANOVA was performed to detect significant difference of frequency ratings between the nine SI tasks of particular interest and a baseline task, respectively.

## 4.5 Results

### 4.5.1 Demographic information

Table 4.2 shows the basic demographic information of the respondents including gender, age, education, and interpreting training. As can be seen, the proportion of the female interpreters was nearly 30% higher than that of their male counterparts; the interpreters were between 22 and 64 years old, with the majority (64.3%) belonging to the age group of 26-35 years; and 71.4% of the respondents had a Master's degree. One respondent reported attending courses provided by Technical and Further Education (TAFE) institutions, Australia's vocational tertiary education provider. In terms of the interpreting training, 59.3% of the respondents held a postgraduate-level interpreting degree. In the category of "Other", three respondents indicated that they received on-job or pre-job training, and two respondents received training from the EU Directorate-General for Interpretation.

Table 4.2 Basic demographic information

Demographic variables	No.	Percent (%)
<b>Gender</b>		
Male	51	36.4
Female	89	63.6
<b>Age</b>		
22-25 years old	30	21.4
26-35 years old	90	64.3
Over 35 years old	20	14.3
<b>Education</b>		
High school	4	2.9
Bachelor	26	18.6
Master	100	71.4
Doctorate	9	6.4
Other	1	0.7
<b>Interpreting training and education</b>		
Self-taught	15	10.7
Intensive interpreting training course	30	21.4
Interpreting diploma	7	5.0
Postgraduate-level interpreting degree	83	59.3
Other	5	3.6

Table 4.3 shows the three demographic variables of particular interest including respondents' employment status, SI experience and working location. As shown in the table, 45.0% identified themselves as part-timers who hold a formal job, and only interpret part-time; 39.3% were freelancers who are not committed to a particular employer long term and usually work on a piecemeal basis; and 15.7% were in-house who are employed staff and work within an organization.

Table 4.3 Descriptive statistics of the three demographic variables

Demographic variables	No.	Percent (%)
<b>Employment status</b>		
Part-time	63	45.0
Freelance	55	39.3
In-house	22	15.7
<b>SI experience</b>		
Less than 3 years ( $\leq 3$ years)	82	58.6
More than 4 years ( $\geq 4$ years)	58	41.4
<b>Location</b>		
Bohai Economic Rim	38	27.1
Yangtze River Delta	46	32.9
Pearl River Delta	42	30.0
Cheng-Yu Economic Zone	14	10.0

SI experience ranged from half a year to 30 years and was averaged at about five years. It is noteworthy that 58.6% of the respondents had no more than three years' experience, constituting the less experienced group, and the rest of the interpreters (41.4%) were thus categorized as the more experienced group (see Table 4.3).

To obtain information about geographical location, the participants were asked to report where they frequently worked as interpreters. Twenty-eight Chinese cities were reported with Shanghai, Beijing, Guangzhou and Shenzhen being mentioned most frequently. The 28 cities can be categorized into four geographical areas where China's economic activities are concentrated: the Bohai Economic Rim centering on Beijing, the Yangtze River Delta

surrounding Shanghai, the Pearl River Delta revolving around Guangzhou, Shenzhen and Hong Kong, and the Cheng-Yu Economic Zone linking two mega-cities in China's mid-west. Accordingly, the respondents were conveniently categorized into one of the four areas based on their reported location.<sup>26</sup> As shown in Table 4.3, the proportion of the interpreters was similar across the first three areas, accounting for roughly one third of total respondents, respectively; whereas for the Cheng-Yu Economic Zone, the number of interpreters recruited was almost three times fewer than that of the other areas. This difference could be explained by the facts that 1) demand for interpreting services in China is related with economic activities, and 2) compared with the other three areas the Cheng-Yu Economic Zone is the least economically dynamic and internationally oriented one.

#### 4.5.2 Part I: Results

Figure 4.1 summarizes the answers to Question 10 which asks what kind of materials interpreters usually receive in advance for preparation. As shown, the majority of the interpreters usually received draft speech text (51.4%), bio sketch/C.V. (52.1%), conference agenda (87.1%) and PPT (91.4%). In addition, it is interesting to know that some interpreters usually received audio/video materials for conference preparation. In the "Other" category, the respondents (i.e., R015 & R023) mentioned that they received "warm-up advertorials" and "company's website".

---

<sup>26</sup> For those respondents who identified two cities/places, the first identified city/place was used to represent their location.

**As a conference interpreter, I usually obtained the following conference-related materials and information in advance:**

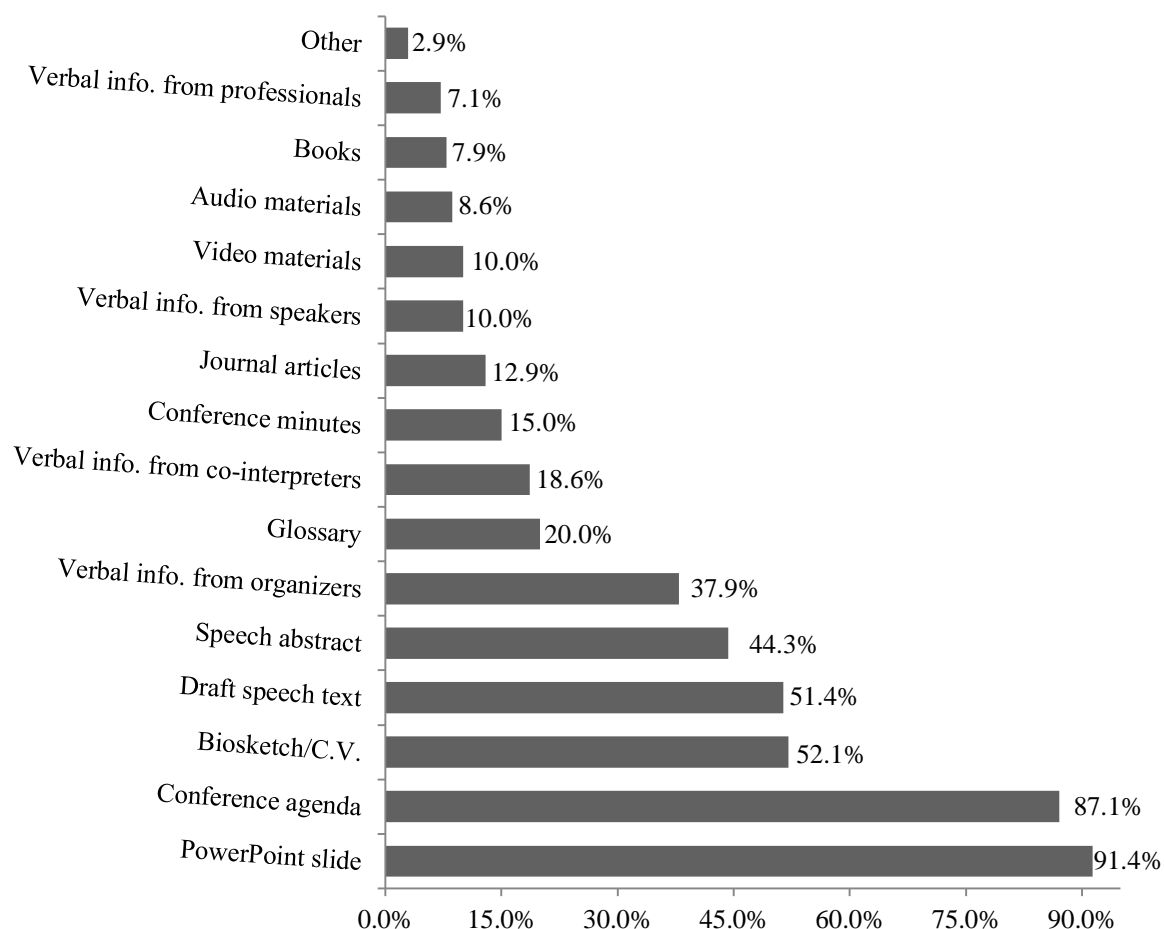


Figure 4.1 Conference-related materials or information obtained in advance

Table 4.4 presents the number of different materials (and associated percentages) different types of interpreters received, which helps shed a light on the relationship between the three demographic variables (i.e., location, employment status, and SI experience) and the top four most received materials (i.e., PPT, conference agenda, bio sketch/C.V., and draft speech text). Due to the relatively small samples, the “Cheng-Yu Economic Zone” group ( $n = 14$ ) and the “In-house” group ( $n = 22$ ) were dropped from the current analysis and similar analyses below. In addition, a percentage difference equal to or larger than 10% was arbitrarily treated as a large (appreciable) difference in the study (indicated by the bold percentages in Table 4.4). As can be seen in Table 4.4, while there were no considerable percentage differences between the interpreters working in the Yangtze and the Pearl River Delta areas across the received materials, the proportions were much larger than those of the interpreters working in the Bohai Economic Rim, particularly for “Conference agenda” and “Draft speech text”. Similarly, a considerably larger proportion of the freelancers received “Conference agenda”

than the part-timers (i.e.,  $92.7\% - 82.5\% = 10.2\%$ ). Furthermore, an appreciably larger percentage of the more experienced interpreters received “bio sketch/C.V.” than their less experienced counterparts (i.e.,  $60.3\% - 45.1\% = 15.2\%$ ).

Table 4.4 The number of different materials received by different types of the interpreters

Type of materials	Location			Employment status		SI experience	
	(n=126)			(n=118)		(n=140)	
	Bohai (n=38)	Yangtze (n=46)	Pearl (n=42)	Freelance (n=55)	Part-time (n=63)	More exp. (n=58)	Less exp. (n=82)
PPT	35 <sup>a</sup> (92.1) <sup>b</sup>	40 (87.0)	40 (95.2)	52 (94.5)	56 (88.9)	56 (96.6)	73 (89.0)
Conference agenda	<b>30 (78.9)</b>	41 (89.1)	38 (90.5)	<b>51 (92.7)</b>	<b>52 (82.5)</b>	52 (89.7)	70 (85.4)
Bio sketch/ C.V.	20 (52.6)	21 (45.7)	23 (54.8)	30 (54.5)	33 (52.4)	<b>35 (60.3)</b>	<b>37 (45.1)</b>
Draft speech text	<b>22 (57.9)</b>	21 (45.7)	19 (45.2)	28 (50.9)	31 (49.2)	32 (55.2)	39 (47.6)

Notes: <sup>a</sup> The number of materials received; <sup>b</sup> Percentage (%) was calculated by dividing the number of materials by the number of interpreters in a given sub-group.

Figure 4.2 shows when the interpreters received the three materials of PPT, conference agenda and draft speech text (i.e., Question 11). As can be seen in the figure, although these materials were received at different times, most of the PPTs and the draft speech texts were received close to the actual SI, while conference agenda was distributed more or less equally across the different times. In addition, five respondents (i.e., R009, 011, 014, 020 and 078) reported that they usually did not receive these materials; R013 commented that some PPTs came early, others late.

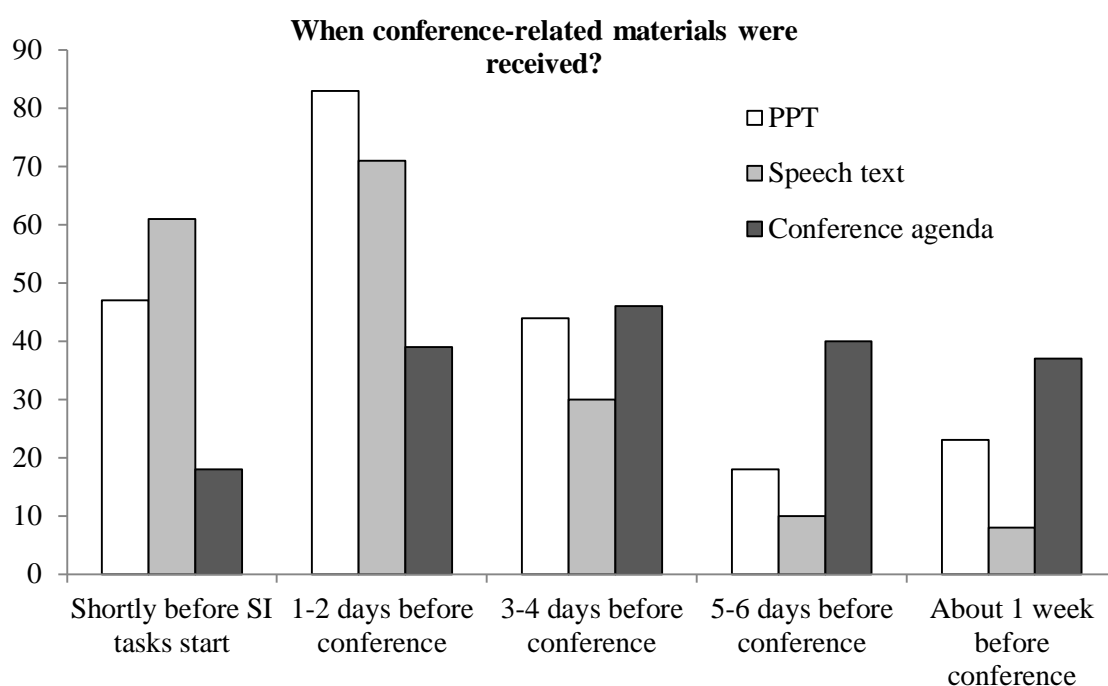


Figure 4.2 When conference-related materials were received

#### 4.5.3 Part II: Results

Table 4.5 presents the descriptive statistics of the frequency counts for the 18 SI tasks (i.e., Question 13 – 16). SI with Text (ShortMod) was taken as a baseline task,<sup>27</sup> because its average score and standard deviation ( $M = 3.97$ ,  $SD = 1.47$ ) were very similar to those of the grand average ( $M = 3.95$ ,  $SD = 1.43$ ). As shown in the table, nine tasks were performed more frequently by interpreters than the baseline task.

To examine whether a real difference exists between the frequency scores of the top nine SI tasks and that of the baseline, a one-way repeated-measures ANOVA was performed. Mauchly's test indicated that the assumption of sphericity had been violated,  $\chi^2(44) = 366.16$ ,  $p < 0.05$ , therefore degree of freedom was corrected using Greenhouse-Geisser estimate of sphericity ( $\epsilon = 0.65$ ). The main ANOVA showed that frequency scores were significantly affected by the type of SI task performed,  $F(5.8, 32.17) = 13.77$ ,  $p < 0.05$ . Following the significant main effect of task type, simple contrasts (taking the baseline task as the reference) indicated that the frequency scores of the SI task types were statistically significantly higher than that of the baseline ( $p < 0.05$ ), with the exception of SI with PPT (LongMod),  $F(1, 0.46) =$

<sup>27</sup> A type of SI task in which the interpreters received the speaker's draft speech text shortly before the task started, and found it to be moderately matched with the speech actually delivered.

0.18,  $p = 0.67$ , and SI with PPT (ShortMod),  $F(1, 5.6) = 3.60$ ,  $p = 0.06$ . The results indicate that the top seven SI tasks were performed significantly more frequently than the baseline task.

Table 4.5 Mean frequency scores and standard deviations for the 18 varieties of SI task

SI task variety <sup>a</sup>	Mean	SD	Rank
SI (DiaIntr)	5.09	1.34	1
SI with PPT (ShortAbun)	4.96	1.32	2
SI (MonoImprm)	4.75	1.45	3
SI with Text (ShortAbun)	4.66	1.52	4
SI with PPT (LongAbun)	4.42	1.55	5
SI with no Materials (NoPPT)	4.35	1.27	6
SI with no Materials (NoText)	4.26	1.24	7
SI with PPT (ShortMod)	4.17	1.40	8
SI with PPT (LongMod)	4.03	1.30	9
<b>SI with Text (ShortMod)</b>	<b>3.97</b>	<b>1.47</b>	<b>10</b>
SI with Text (LongMod)	3.77	1.36	11
SI with Text (LongAbun)	3.65	1.66	12
SI with no Materials (NoText&PPT)	3.56	1.39	13
SI (Audio/video)	3.39	1.50	14
SI with Text (LongBar)	3.21	1.39	15
SI with PPT (LongBar)	3.09	1.48	16
SI with PPT (ShortBar)	2.90	1.56	17
SI with Text (ShortBar)	2.85	1.62	18
<b>Grand Average Score</b>	<b>3.95</b>	<b>1.43</b>	<b>n/a</b>

Note: <sup>a</sup>. Definitions for each task variety can be found in Appendix D; n/a = not applicable

In addition, to investigate relationships between the three demographic variables (i.e., employment status, SI experience and working location) and the frequency scores of the top seven task varieties, a series of ordered probit regression was conducted. Specifically, for each variety of SI task, an ordered probit regression model was used to analyze the relationship between the three predictor variables and the task frequency scores, respectively. In other words, three predictor variables were simultaneously included in a model. A significant

coefficient indicates that compared to a reference group, a given demographic group has a significant impact (either positive or negative) on frequency scores of a particular SI task. To facilitate the understanding of the regression results, only useful findings were presented.

#### 4.5.3.1 Employment status

Table 4.6 shows the degree and the direction of relationship the employment status has on the frequency scores of the top seven task varieties, respectively.

Table 4.6 Relationship between employment status and SI task varieties

SI task variety	In-house ( <b>Ref.</b> ) (n=22)		Part-time( <b>Ref.</b> ) (n=63)
	v.s.		v.s.
	Freelance (n=55)	Part-time (n=63)	Freelance (n=55)
SI (DiaIntr)	n.s.	n.s.	n.s.
SI with PPT (ShortAbun)	n.s.	n.s.	n.s.
SI (MonoImprm)	n.s.	n.s.	n.s.
SI with Text (ShortAbun)	n.s.	n.s.	−0.33*; (0.20)
SI with PPT (LongAbun)	n.s.	0.77***; (0.28)	−0.61***; (0.19)
SI with no Materials (NoPPT)	0.61**; (0.30)	n.s.	0.39**; (0.20)
SI with no Materials (NoText)	n.s.	n.s.	n.s.

Note: n.s. = not significant; Ref. = reference group; Coefficients; (Standard errors) are results of ordered probit regression; \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

As shown in Table 4.6, a limited number of the cases achieved significance. For example, in the second column of the table, compared to the in-house interpreters, the freelancers had a significantly positive impact on the frequency rating for SI with no Materials (NoPPT) at the 0.05 level. It means that the freelancers performed this task more frequently than the in-house interpreters. In the fourth column, with the reference to the part-timers, the freelancers had a significantly negative effect on the frequency scores of SI with Text (ShortAbun) at the 0.1 level, and of SI with PPT (LongAbun) at the 0.01 level, but had a substantially positive impact on the frequency rating for SI with no Materials (NoPPT) at the 0.05 level. It indicates that in comparison with the part-timers, the freelancers performed the former two tasks less frequently, but the latter task more frequently. However, it would appear that in general employment status

did not exert a sweeping impact on the frequency scores. In other words, different types of the interpreters did not deviate substantially from one another in terms of how frequently they performed the SI tasks.

#### 4.5.3.2 *SI experience*

Table 4.7 shows the regression results for the degree and the direction of relationship SI experience has on the frequency scores of the top seven task varieties, respectively. As can be seen in Table 4.7, most cases achieved statistical significance. This result indicates that the “experience” variable had the potential to predict how frequently interpreters performed the SI tasks. Specifically, compared to their less experienced counterparts, the more experienced interpreters performed more frequently the tasks in which relevant supplementary materials (i.e., draft speech text, PPT) were made available to them shortly before SI, such as SI with Text (ShortAbun) and SI with PPT (ShortAbun). In addition, they also performed more frequently the tasks in which relevant materials were even not provided in advance, such as SI with no Materials (NoText), SI with no Materials (NoPPT) and SI (MonoImprm).

Table 4.7 Relationship between interpreting experience and SI task varieties

SI task variety	Less experienced ( <b>Ref.</b> ) (n=82) v.s. More experienced (n=58)
SI (DiaIntr)	n.s.
SI with PPT (ShortAbun)	0.48***; (0.17)
SI (MonoImprm)	0.39**; (0.17)
SI with Text (ShortAbun)	0.33*; (0.17)
SI with PPT (LongAbun)	n.s.
SI with no Materials (NoPPT)	0.41**; (0.17)
SI with no Materials (NoText)	0.57***; (0.18)

Note: n.s. = not significant; Ref. = reference group; Coefficients; (Standard errors) are results of ordered probit regression; \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

#### 4.5.3.3 *Working location*

Table 4.8 shows the regression results for the degree and the direction of relationship the “location” variable has on the frequency scores of the top seven task varieties, respectively. As shown in the table, there were only a small number of significant cases, which suggests that the

location as a predictor variable had a negligible effect on how frequently SI tasks were performed. Additionally, the significant cases occurred only in the comparison with the Cheng-Yu Economic Zone. Considering the small sample size (n=14), the results were likely to be unstable.

Table 4.8 Relationship between location and SI task varieties

SI task variety	Cheng-Yu ( <b>Ref.</b> ) (n=14)			Bohai ( <b>Ref.</b> )		Pearl ( <b>Ref.</b> )
	v.s.			(n=38) v.s.		(n=42) v.s.
	Bohai (n=38)	Pearl (n=42)	Yangtze (n=46)	Pearl (n=42)	Yangtze (n=46)	Yangtze (n=46)
SI (DiaIntr)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SI with PPT (ShortAbun)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SI (MonoImprm)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SI with Text (ShortAbun)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SI with PPT (LongAbun)	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
SI with no Materials (NoPPT)	0.64**; (0.30)	0.44*; (0.25)	0.51*; (0.30)	n.s.	n.s.	n.s.
SI with no Materials (NoText)	0.69***; (0.25)	n.s.	0.53***; (0.27)	n.s.	n.s.	n.s.

Note: n.s. = not significant; Ref. = reference group; Coefficients; (Standard errors) are results of ordered probit regression; \* $p < 0.1$ , \*\* $p < 0.05$ , \*\*\* $p < 0.01$ .

#### 4.5.4 Part III: Results

Figure 4.3 presents the results of Question 17 which explores SI directionality. Overall, there are three findings: 1) all the respondents interpreted bi-directionally (i.e., E-to-C & C-to-E), though to a different degree; 2) nearly half of the respondents reported interpreting in both directions for an approximately equal amount of time and 3) it seems that C-to-E SI was performed more than the opposite direction.

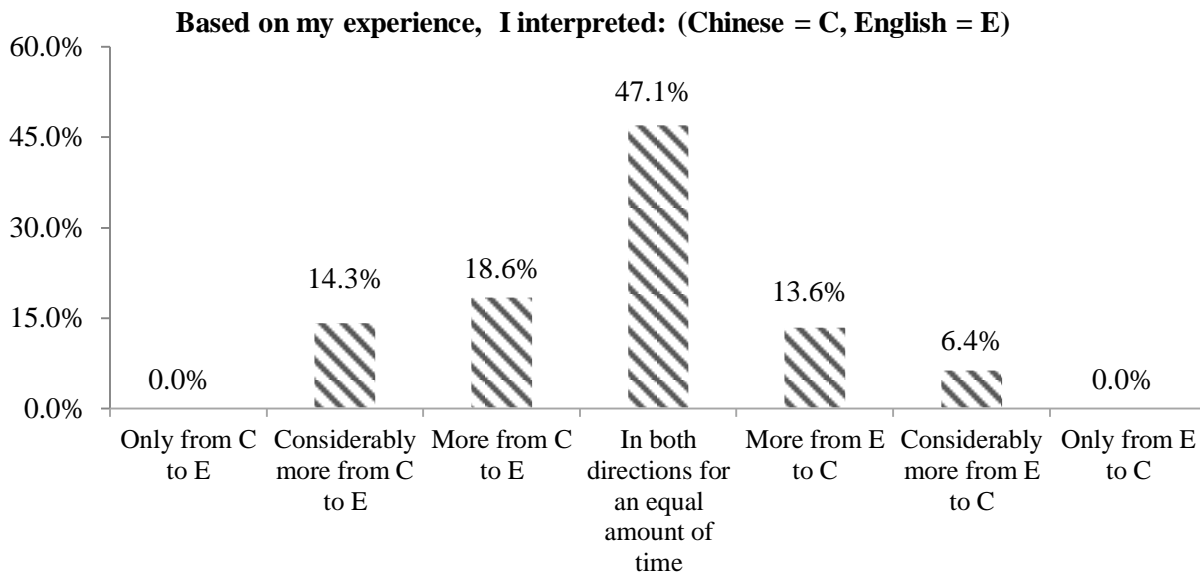


Figure 4.3 Directionality of SI

Figure 4.4 displays the results concerning the duration of a single interpreting turn (i.e., Question 18). The trend line shows that most of the interpreters chose “11min-20min” and “21min-30min” categories. Specifically, 85 out of the 140 respondents (60.7%) reported usually performing SI for 11 to 20 minutes, and 42.1% of them for 21 to 30 minutes. It is also worth noting that more than one-fifth of the respondents (14.3%) reported usually interpreting for over 31 minutes in one turn.

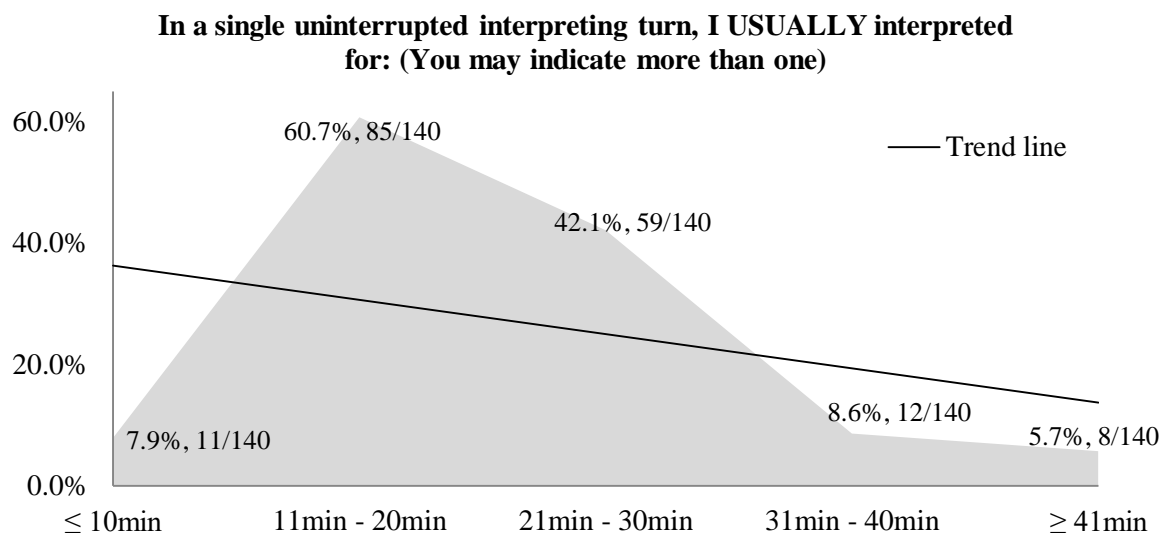


Figure 4.4 Duration of an interpreting turn

To examine which group of interpreters was more likely to work over 31 minutes in one turn, the survey data were re-structured (see Table 4.9).

Table 4.9 Who interpreted for over 31 minutes in one interpreting turn?

Duration	Location			Employment status		SI experience	
	(n=126)			(n=118)		(n=140)	
	Bohai (n=38)	Yangtze (n=46)	Pearl (n=42)	Freelance (n=55)	Part-time (n=63)	More exp. (n=58)	Less exp. (n=82)
$\geq 31$ min	<b>2.63%</b> <b>(1/38)</b>	<b>15.21%</b> <b>(7/46)</b>	<b>21.43%</b> <b>(9/42)</b>	10.91% (6/55)	11.11% (7/63)	12.07% (7/58)	14.63% (12/82)

Note: exp. = experienced.

As shown in Table 4.9, the largest difference occurred in the interpreters categorized by the “location” variable, whereas the differences between the freelancers and the part-timers, as well as between the more and the less experienced interpreters were small. Specifically, an appreciably larger proportion of the interpreters working in the Yangtze and the Pearl River Delta areas reported interpreting for over 31 minutes in one turn than their counterparts in the Bohai Rim area. Exact reasons are unknown, based on the survey data. However, one respondent (i.e., R129) who identified him/her as a Shanghai-based freelancer with 4-year SI experience commented that “in the market, there is a kind of SI called marketing research SI which normally requires one interpreter to interpret 1-2 hours.” This comment suggests that one of the reasons for the excessively long duration could be that only one interpreter is employed for a particular type of SI.

Figure 4.5 shows respective percentage of the interpreters who attributed difficulty of SI tasks to a variety of factors (i.e., Question 19). The factors were categorized into six major groups including “interpreter factor”, “working condition”, “SI task dimension”, “linguistic” and “paralinguistic dimensions” of input materials, and “other”.

**Which of the following could be FREQUENT contributing factors to SI tasks difficulty?**

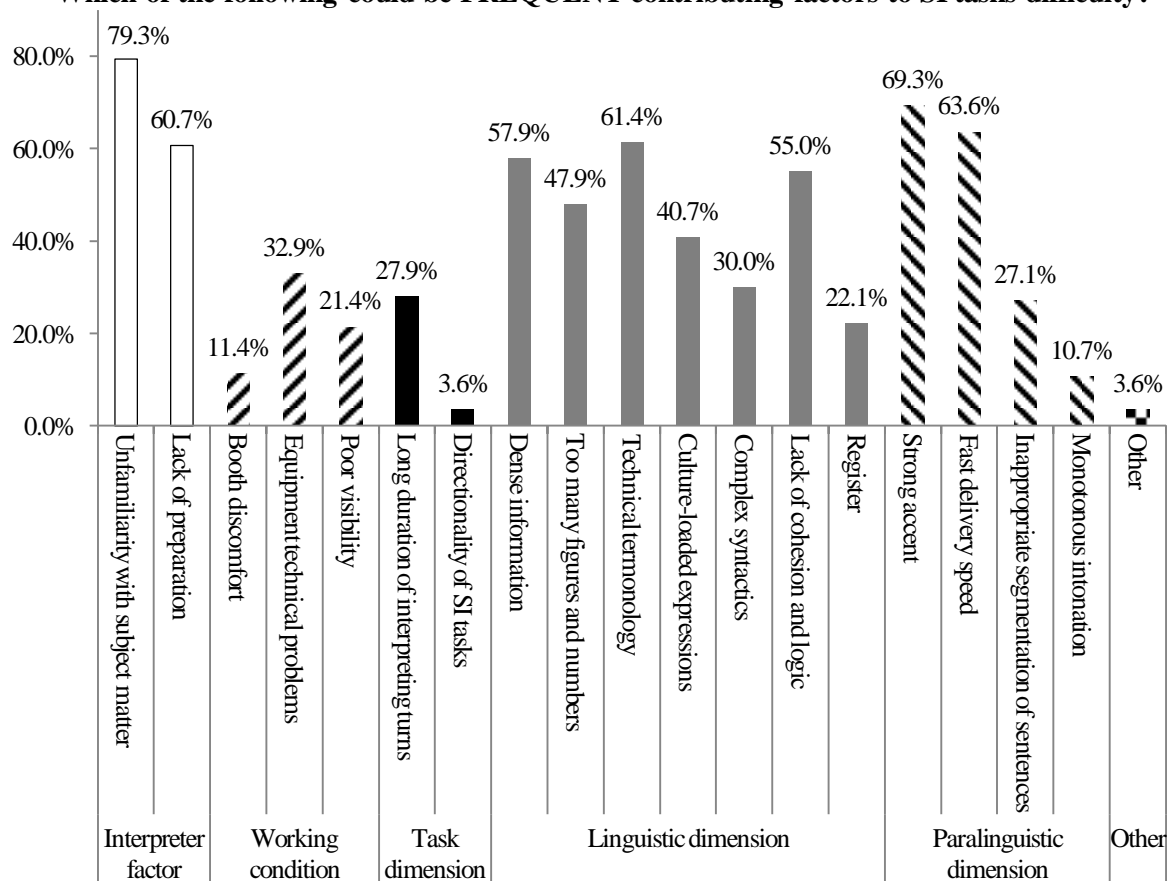


Figure 4.5 Factors contributing to SI task difficulty

As shown in Figure 4.5, seven factors were chosen by more than 50% of the respondents. The least chosen factor was “directionality of SI tasks” (3.6%). In the “other” category, R001 mentioned “visual distraction from either booth mates or people moving around in front of the booth”; R020 reported “speaker’s mixing code”; and R118 cited “physical ailment”.

Table 4.10 presents the results from a further analysis of the relationship between the two demographic variables (i.e., employment status and SI experience) and the top five perceived difficulty factors (i.e., unfamiliarity with subject matter, lack of preparation, technical terminology, strong accent and fast delivery speed). As can be seen in Table 4.10, considerably larger proportions of the part-timers regarded “Lack of preparation”, “Technical terminology” and “Strong accent” as the difficulty factors than the freelancers. In addition, quite surprisingly, markedly larger percentages of the experienced interpreters regarded “Strong accent” and “Fast delivery speed” as the frequent factors underlying SI difficulty than the inexperienced practitioners.

Table 4.10 The frequency (and percentage) of a certain difficulty factor identified by different types of the interpreters

Difficulty factors	Employment status (n=118)		SI experience (n=140)	
	Freelance (n=55)	Part-time (n=63)	More exp. (n=58)	Less exp. (n=82)
Unfamiliarity with subject matter	45, (81.8)	49, (77.8)	49, (84.5)	63, (76.8)
Lack of preparation	<b>30, (54.6)</b>	<b>43, (68.3)</b>	37, (63.8)	48, (58.5)
Technical terminology	<b>31, (56.4)</b>	<b>42, (66.7)</b>	38, (65.5)	48, (58.5)
Strong accent	<b>34, (61.8)</b>	<b>47, (74.6)</b>	<b>45, (77.6)</b>	<b>53, (64.6)</b>
Fast delivery speed	40, (72.7)	42, (66.7)	<b>42, (72.4)</b>	<b>48, (58.5)</b>

Notes: <sup>a</sup> The number of materials received; <sup>b</sup> Percentage (%) was calculated by dividing the number of materials by the number of interpreters in a given sub-group.

## 4.6 Discussion

### 4.6.1 Demographic information

Based on the data, the majority of the sample was female interpreters (63.6%), received good education (Master's degree: 71.4% & Postgraduate-level interpreting degree: 59.3%), and aged between 26-35 years old (64.3%). In addition, most of the respondents were part-timers (45%) or freelancers (39.3%), had less than 3 years of SI experience (58.6%), and worked in the more economically dynamic regions (90%). However, due to the non-probability sampling, the respondents may not represent the population adequately.

Interestingly, the demographic profile from the previous surveys (i.e., Pan et al., 2009; STTACAS & TRANSN, 2007; Wang, 2005) is similar to the current one: a group of relatively young interpreters who are well-educated, have a moderate amount of experience, and work as freelancers or part-timers. For example, in the national survey (STTACAS & TRANSN, 2007), 70% of the respondents are between 20 to 35 years old (i.e., 20 – 25 yr: 20%; 26 – 30 yr: 27%; 31 – 35 yr: 23%); 54% of the respondents have working experience of only 1-5 years; nearly 70% of them work part-time. In Pan et al.'s (2009) survey, 60% of the interpreters receive Master's or higher degrees; 41% major in conference interpreting; and according to the employers surveyed, 63% of the interpreters employed are not in-house. In addition, university

lecturers and students constitute a large source of interpreters. In Wang's survey (2005), 56% of the interpreters receive professional interpreting training before entering the market.

The interpreter profile shared by the four surveys could be explained in two ways. One less plausible scenario is that the profile happens to reflect the make-up of the profession in China. Over the past years, the influx of new graduates from domestic and overseas programs probably has re-shaped the composition of the profession, resulting in a younger and less experienced workforce who is willing to work more flexibly than the first-generation interpreters. The alternative and more plausible explanation is that the sampling method in all four surveys are based on non-probabilistic selection, particularly convenience sampling. As a result, due to unknown reasons, the more experienced senior interpreters could be less willing to participate. The same sampling error has been repeated, leading to the biased samples.

However, the sampling problem is not uncommon to social sciences (e.g., Gideon, 2012), particularly in Interpreting Studies (Pöchhacker, 2009). Pöchhacker identifies two obstacles to sound survey sampling, based on a meta-analysis of 40 surveys worldwide on conference interpreting. One obstacle is the lack of reliable information on population of conference interpreters, with China as a case in point (Pöchhacker, 2009). Currently, there are no national organizations for conference interpreters in China, despite the proposals to establish one (Wang, 2005). Unlike the AIIC members, conference interpreters in China cannot be found on any publicly available list, and their population is very difficult to estimate (A. Dawrant, personal communication, June 2, 2013). In addition, unlike other countries such as the US and Australia where a specific practice domain/setting is well-defined (e.g., courts, hospitals and police stations), conference interpreting in China has been conceived as an umbrella term for SI performed in any international conferences. Consequently, the vague conceptualization may also make it difficult to pinpoint specific practice domains and register practicing interpreters.

The other obstacle is small and unsystematic samples that give sparse coverage of the wider population. This problem is frequently encountered by survey researchers (e.g., Johnson & Christensen, 2012), due to practical constraints (e.g., lack of sources, inaccessibility to part of target population). One possible way to overcome the obstacle is replication. That is, researchers conduct multiple small-scale surveys for later comparison and meta-analysis.

In summary, despite the relatively large sample in the present survey ( $n = 140$ ) compared to the surveys reviewed in Pöchhacker (2009), and despite the similar demographic information to that of the previous surveys (e.g., STTACAS & TRANSN, 2007), the non-probability sampling

necessitates a cautious and discreet approach to interpreting and generalizing the survey findings.

#### 4.6.2 Discussion of the results from Part I, II and III

Results from Part I confirm the wide-range of materials numerated in the scholarly literature (e.g., Gile, 1995; 2002, Kalina, 2002), and corroborate the diary findings that PPT and draft speech text were more likely to be received. Part I also provides some additional observations that certain types of interpreters are more likely to receive a certain type of conference-related material, although the findings are preliminary. Furthermore, given that most of the PPTs and draft speech texts become available only one or two days before conferences, interpreters may not have sufficient time for preparation, and probably need to use what Gile (1995) calls “last-minute” or “in-conference preparation”. The time constraint on interpreters’ preparation is also echoed by Donovan (2001, p. 12) who claims that “conference interpreters often have little time to prepare a meeting”, and is supported by the AIIC’s (2002) Workload Study that cites “too little time to prepare” as one of stressors.

Results from Part II show that a wide-range of SI tasks was performed by the English/Chinese conference interpreters, albeit to a varying degree of frequency, which confirms the exploratory diary findings. Especially, the finer-grained categorization of SI task variety extends and enriches the three general types of SI tasks that have been traditionally discussed in the literature: SI with Text (Kalina, 2002; Setton, 2009; Wang & Lin, 2006), SI with PPT (Kalina, 2002; Wan, 2004; Wu, 2007) and SI for Q&A (Chang & Wu, 2009). In addition, based on the frequency scores and the ANOVA results (particularly the top seven task varieties), it could be said that the interpreters performed significantly more frequently the tasks in which they were inadequately informed with the speech content, and lacked sufficient time for thorough preparation, as the relevant materials were not received in advance or received only shortly before SI. The findings indirectly support Donovan’s (2001) and AIIC’s (2002) comments on the pre-conference preparation, and lend credence to Gile’s (1995) categorization of the “last-minute” and the “in-conference preparation”. Furthermore, Part II generates some new findings in terms of the relationship between the three demographic variables of interest and the SI task frequency ratings. It was found that employment status and working location as predictors did not produce across-the-board impacts on the frequency ratings of the top seven task varieties, whereas SI experience had the potential. Specifically, the more experienced

interpreters were more likely to perform SI tasks in which systematic preparation was almost impossible. This is probably because when working in pairs the more experienced interpreters tend to take on more difficult, adversely conditioned or unexpected SI tasks. These results could be used to help testers re-analyze the tasks sampled in the interpreter certification performance tests (ICPTs). Currently, the operational English/Chinese ICPTs that focus on assessing SI in China include one or two types of SI task, especially SI with no Materials (NoText). Given the variety of SI tasks identified in the survey, the current ICPTs may risk under-representing the real-life interpreting practice domain, and the testers therefore probably need to consider broadening the range of SI tasks to be included in the tests. For example, Huang (2005) suggests sampling such tasks as SI (DiaIntr) and SI for uni-directional presentations of different styles.

The results on SI directionality in Part III support Setton's (2009, p. 109) observation that "practice is fully-bidirectional" in China. The results also corroborate the diary findings, and echo Pan et al.'s (2009), Pavlović's (2007) and Szabari's (2002) survey findings that real-life conference interpreters work between their A language(s) and other less dominant language(s).

The finding on the duration of an interpreting turn in Part III is generally consistent with the recommended optimal duration of about 20 minutes (Chmiel, 2008) and the maximum amount of approximately 30 minutes (Moser-Mercer et al., 1998). However, the data also show that due to some practical constraints some interpreters worked for an excessively long period of time in one turn, which corresponds with the diary finding. The heavy workload in one turn required of an interpreter may indicate the lack of recognition of the demanding nature of SI on the part of employers, and the lack of standardization on the part of the regulators (Feng, 2005).

Furthermore, the results on "difficulty factors" in Part III generally accord with the diary findings and support the claims made by researchers and scholars (e.g., Gile, 1995, 2008; Kalina, 2005) that a number of factors contribute to SI difficulty such as lack of preparation, fast speech rate, strong accent/non-native speakers (NNS), and lack of cohesion and logic. More importantly, several factors are identified as frequent contributors in the real-life practice, such as fast speech rate and lack of preparation. In fact, Setton (2009) believes that SI from fast-delivered, recited speeches with little or no preparation is probably more common in China than elsewhere. The surprising result that the substantially larger percentages of the more experienced interpreters regarded "Strong accent" and "Fast delivery speed" as the difficulty factors than their less experienced counterparts could be explained in relation to the results

from Part II. Given that the more experienced interpreters were likely to perform SI tasks without adequate preparation (indicated by Part II results), they could be more vulnerable and sensitive to additional external cognitive-loading factors such as “Fast delivery speed” and “Strong accent” than the less experienced interpreters. The surprising result could also be accounted for by the possibility that the experienced interpreters were more conscious of and thus more responsive to such difficulty factors and their negative effects on SI performance.

#### 4.7 Conclusion

The study reports a survey of 140 interpreters to explore the conference interpreting practice in China. The survey findings support and extend the experiential accounts based on individual practitioners, and the previous results from empirical studies. However, the non-probability sampling makes it difficult to generalize the findings to the practice domain across China. Nevertheless, the empirical data are valuable in providing insight to the real-life practices. To gain more understandings of the practice, interpreting researchers, practitioners and regulators need to make concerted efforts in order to overcome barriers that impede flow of and access to relevant information. All parties will benefit long term from an accurate and reliable profile of the interpreting practice. For the moment, multiple samplings of interpreters by different researchers represent a viable approach to obtaining a clear picture of the profession.

#### 4.8 References

- AIIC. (2002). *Interpreter workload study*. Retrieved from <http://aiic.net/page/657/interpreter-workload-study-full-report/lang/1>
- AIIC. (2004). *Practical guide for professional conference interpreters*. Retrieved from <http://aiic.net/ViewPage.cfm/article21.htm>
- Albl-Mikasa, M. (2010). Global English and English as a lingua franca (ELF): Implications for the interpreting profession. *Trans-kom*, 2, 126-148.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Boes, S., & Winkelmann, R. (2006). Ordered response models, *AStA Advances in Statistical Analysis, Springer*, 90(1), 167-181.

- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Chang, C., & Wu, M. M. (2009). Address form shifts in interpreted Q&A sessions. *Interpreting*, 11(2), 164-189.
- Chen, J. (2002). 从 Bachman 交际法语言测试理论模式看口译测试中的重要因素. [Fundamental considerations in interpreting testing]. 中国翻译, 1, 51-53.
- Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, 3(2), 257-273.
- Chmiel, A. (2008). Boothmates forever? - On teamwork in a simultaneous interpreting booth. *Across Languages and Cultures*, 9(2), 261-276.
- Dawrant, A., & Jiang, H. (2001). *Conference interpreting in Mainland China*. Retrieved from [http://www.aiic.net/ViewPage.cfm?page\\_id=365](http://www.aiic.net/ViewPage.cfm?page_id=365)
- Dillinger, M. (1990). What do interpreters know that bilinguals don't? *The Interpreters' Newsletter*, 3, 41-58.
- Donovan, C. (2001). Interpretation of technical conferences. *Conference Interpretation and Translation*, 3, 4-22.
- Feng, J. Z. (2005). 论口译测试的规范化. [Towards the standardization of interpretation testing]. 外语研究, 89, 54-58.
- Gerver, D. (1971). *Simultaneous and Consecutive Interpretation and Human Information Processing*. (Research report), Retrieved from ERIC database.
- Gideon, L. (2012). *Handbook of survey methodology for the social sciences*. Springer: New York.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.
- Gile, D. (2002). The interpreter's preparation for technical conferences: Methodological questions in investigating the topic. *Conference Interpretation and Translation*, 4(2), 7-27.
- Gile, D. (2008). Local cognitive load in simultaneous interpreting and its implications for empirical research. *Forum*, 6(2), 59-77.
- Guo, X. Y. (2010). 中国语言服务行业发展状况, 问题及对策 – 在 2010 中国国际语言服务行业大会上的主旨发言. [Development of China's language services industry:

- Problems and solutions – A keynote speech at 2010 China International Language Industry Conference]. 中国翻译, 6, 34-37.
- Han, C. (forthcoming). Lacunae, myths and legends about conference interpreters: a diary study to explore conference interpreting practice in China. *Studies in Translatology: Perspectives*.
- Huang, M. (2005). 谈口译资格认证考试的规范化设计. [Toward a more standardized large-scale accreditation test for interpreters]. 中国翻译, 6, 62-65.
- Johnson, B., & Christensen, L. (2012). *Educational Research: Quantitative, qualitative and mixed approaches* (4th ed.). Thousand Oaks: Sage.
- Kalina, S. (2002). Quality in interpreting and its prerequisites A framework for a comprehensive view. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century: Challenges and Opportunities* (pp. 121-130). Amsterdam/Philadelphia: John Benjamins.
- Kalina, S. (2005). Quality assurance for interpreting processes. *Meta*, 50(2), 769-784.
- Kurz, I. (2009). The impact of non-native English on students' interpreting performance. In G. Hansen, A. Chesterman & H. Gerzymisch-Arbogast (Eds.), *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile* (pp. 179-192). Amsterdam: Benjamins.
- Moser-Mercer, B., Kunzli, A., & Korac, M. (1998). Prolonged turns in interpreting: Effects on quality, physiological and psychological stress (pilot study). *Interpreting*, 3(1), 47-64.
- Pan, J., Sun, Z. X., & Wang, H. H. (2009). 口译的职业化与职业化发展 – 上海及江苏地区口译现状调查研究. [Professionalization in interpreting: current development of interpreting in Shanghai and Jiangsu Province]. 解放军外国语学院学报, 6, 81-85.
- Pavlović, N. (2007). Directionality in translation and interpreting practice. Report on a questionnaire survey in Croatia. *Forum*, 5(2), 79-99.
- Pöchhacker, F. (2009). Conference interpreting: Surveying the profession. *Translation and Interpreting Studies*, 4(2). 172-186.
- Seleskovitch, D. and Lederer, M. (1989). *A systematic approach to teaching interpretation* (Trans. J. Harmer). Silver Spring, MD: Registry of Interpreters for the Deaf.
- Setton, R. (2009). Interpreting China, interpreting Chinese – Introduction. *Interpreting*, 11(2), 109-117.
- Science and Technology Translators Association of the Chinese Academy of Sciences & TRANSN. (2007). 中国地区译员生存状况调查报告. [An investigation into working

- and living conditions of translators and interpreters in China: A report]. Retrieved from <http://wenku.baidu.com/view/3ff59c88d0d233d4b14e6964.html>
- Su, W. (2009). 社区口译在中国. [Community interpreting in China]. 上海翻译, 4, 42-45.
- Szabari, K. (2002). Interpreting into the B language. In EMCI, *Teaching Simultaneous Interpretation into a "B" Language*. Retrieved from <http://www.emcinterpreting.org/?q=system/files/EMCI-TeachingSimultaneousIntoB-voll.pdf>
- Tommola, J., & Helevä, M. (1998). Language direction and source text complexity: Effects on trainee performance in simultaneous interpreting. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), *Unity in diversity? Current trends in translation studies* (pp. 177-186). Manchester: St. Jerome.
- Wan, H. Y. (2004). 解读图表:另一项重要的口译技能. [Interpreting graphics: An important skill for interpreters]. 中国翻译, 2, 83-86.
- Wang, E. M. (2005). 口译在中国 调查报告. [Interpretation as a profession in China: A survey]. 中国翻译, 2, 57-60.
- Wang, L. & Lin, W. (2006). Interpretation training: SI with text. In M.J. Cai & A. L. Zhang (Eds.), *Professionalization in Interpreting: International Experience and Developments in China* (pp. 237-244). Shanghai: Shanghai Foreign Language Education Press.
- Wang, R. J. (2006). Simultaneous interpretation and its professionalization in China. In M.J. Cai & A. L. Zhang (Eds.), *Professionalization in Interpreting: International Experience and Developments in China* (pp. 144-154). Shanghai: Shanghai Foreign Language Education Press.
- Wu, A. H. (2007). PPT 演讲口译: 视觉辅助还是视觉干扰? [Interpretation for presentations with PowerPoint slides: A visual aid or distraction?]. (Unpublished Master's thesis), Xiamen University, China.

## **An introductory note to Chapter 5**

In Chapter 4, an exploratory-sequential mixed-methods study is conducted to provide a preliminary answer to **RQ 1**: what are the characteristics of the real-life English/Chinese conference interpreting practice in China? It is found that the interpreters encountered a wider and finer-grained variety of simultaneous interpreting (SI) tasks in real-life practice domains than the three general types of SI tasks: SI with texts, SI with PPT and SI with no materials. It is also found that among many factors underlying SI difficulty, fast speech rate (FSR) and strong accent (StrA) figure prominently.

Against this backdrop, Chapter 5 sets out to address the first part of **RQ 2**: RQ 2.1 What are the effects of SI task characteristics on SI performance quality? Specifically, what are the effects of speech rate and accent of source-language speakers on SI quality?

The investigation reported in Chapter 5 is based on a convergent-parallel mixed-methods approach. As will be seen, a quantitative experiment and a qualitative analysis of retrospection data are implemented in a paralleled manner. Informed by the results reported in Chapter 4, fast speech rate (FSR) and strong accent (StrA) are chosen to characterize simultaneous interpreting (SI) tasks. 32 interpreters are recruited to perform SI in four different tasks, and also interviewed to provide self-reflection on their performance after each task. SI performance quality is subsequently assessed by trained raters on three dimensions: information completeness (InfoCom), fluency of delivery (FluDel) and target language quality (TLQual). Quantitative performance data and qualitative interview data are analyzed separately, and also compared and triangulated to shed insight to RQ 2.1.

## Chapter 5 The effects of interpreting task characteristics on the quality of simultaneous interpreting: A convergent parallel mixed-methods approach<sup>28</sup>

**Abstract.** *Simultaneous interpreting (SI) performance is affected by a variety of parameters. One of these parameters is known as “difficulty factors” or input variables of source speeches. Effects of the difficulty factors on SI performance have been investigated, because an in-depth understanding of the effects of source speech characteristics could benefit interpreting researchers, educators and testers. Two factors that have been studied are speech rate and accent of source speeches. Although a number of empirical studies have modeled the causal effects of fast speech rate (FSR) and strong accent (StrA) on SI performance, research results are inconsistent and even contradictory. This study was therefore initiated to investigate the effects of these factors in more depth, using a mixed-methods research design. The quantitative component of the study indicates that FSR produced mixed effects on three performance quality measures, while StrA exerted a consistent detrimental effect across the measures. In addition, the qualitative component provides the interpreters’ reflections on their performance under different SI conditions. Results of the quantitative and qualitative phases of the study are discussed, and the inconsistency between the quantitative and qualitative findings is accounted for by drawing a meta-inference.*

### 5.1 Introduction

In Interpreting Studies, variability of simultaneous interpreting (SI) performance is a complex phenomenon affected by many variables (e.g., interpreter ability, acoustic quality, fatigue, etc.). One of the parameters that affect SI performance relates to the characteristics of a given SI task, also known as “input variables” or “difficulty factors”. Such factors may include speaker’s accent, intonation and delivery speed (see Pöchhacker, 2004). Using experimental or experiment-like methods, interpreting researchers have examined the causal effects of difficulty factors on interpreting performance quality, such as “lexical and syntactic

---

<sup>28</sup> A revised version of this chapter is under the 1<sup>st</sup> round of peer review in the journal of *Target* as: Han, C., & Riazi, M. (under review). A partial replication study to investigate the effects of speech rate and accent on simultaneous interpretation: A mixed-methods approach. Mehdi Riazi’s contribution to the chapter was as PhD supervisor (i.e., reviewing writing and providing feedback on drafts).

complexity” (Tommola & Helevä, 1998), “source text types” (Dillinger, 1989), “noise levels”, (Gerver, 1974), and “visual input” (Jesse, Vrignaud, Cohen, & Massaro, 2000/01).

Results from the difficulty factors studies have had three benefits to interpreting researchers, curriculum developers and interpreting performance assessors. First, the accumulated knowledge of how input variables affect SI performance (see Dillinger, 1989) has improved the understanding of interpreters’ cognitive load during SI (see Gile, 1995, 1999). Second, curriculum designers and course content developers have been able to design SI training materials that represent different levels of cognitive complexity to suit trainee interpreters at different developmental stages (Wang & Lin, 2006). Third, designers of interpreter performance assessment could manipulate “difficulty levels” of input materials *a priori*, based on research findings, to ensure consistency of task complexity across test forms (Liu & Chiu, 2009).

To corroborate the current knowledge base, the present study aims to examine the causal effects of two task characteristics on the quality of English-to-Chinese SI performance: speech rate and accent. The two factors are chosen, because of two reasons:

- 1) In a previous survey on English/Chinese conference interpreting practice in China (see Chapter 4), the majority of the respondents cited fast speech rate (FSR) and strong accent (StrA) as two main difficulty factors in SI. This is while several empirical studies did not find negative effects of these two factors on SI performance, as discussed in the literature review below. The disjunction between the interpreters’ perception of SI difficulty factors and non-negative effects of these factors on SI quality warrants further investigation.
- 2) Despite the proposals to use (para-)linguistically diverse speech samples in interpreter certification performance testing (ICPT) (Chen, 2009; Huang, 2005), it seems that the speed and the accent factor have not been operationalized in high-stakes certification tests in China, which may run the risk of making assessment tasks less authentic (Angelelli, 2009; Campbell & Hale, 2003). An in-depth empirical analysis would therefore inform test developers of possible consequences of operationalizing these two factors in SI tests.

To shed insight to the effects, the study uses a mixed-methods research (MMR) design by combining quantitative experimental and qualitative analysis of retrospection data. The use of MMR design enhances the study, because it aims at collecting multiple strands of data using different research methods, so that the resulting mixture has complementary strengths and non-overlapping weaknesses (Johnson, 2009). Specifically, a quantitative experimental research is the strongest research method for generating evidence of a causal relationship between two variables (Johnson & Christensen, 2012). Qualitative analysis of interpreters' retrospection data also provide an emic or insider's viewpoint and understanding of personal experience, in this case interpreters' experience of performing SI.

## **5.2 Literature review**

The literature review is organized into two sections. In the first section, the empirical studies related to the speed and the accent factors are reviewed. In the second section, a discussion is conducted to account for study results.

### **5.2.1 Review of empirical studies: Speed factor**

Six empirical studies of different methodological paradigms have investigated the effects of the speed factor on SI quality. These studies include Chang (2005), Gerver (1969/2002), Liu, Schallert and Carroll (2004), Meuleman and Van Besien (2009), Pio (2003), and Shlesinger (2003). Of the six studies, three studies (Gerver, 1969/2002; Meuleman & Van Besien, 2009; Pio, 2003) have found negative effects of fast speech rate (FSR) on SI performance. Gerver (1969/2002), for example, concluded that the principal effect of increasing presentation rate for interpreters is the increase of the number of incorrectly translated words, longer ear-voice span, fewer target-language (TL) utterances, and higher pause-to-speech ratio. Two other studies produced mixed results. Chang (2005) reported that the accuracy of renditions was negatively affected by FSR in a statistically significant manner, while there was no apparent effect of FSR on TL quality. Liu et al. (2004) found that among three different source-language (SL) speeches, SI performance was significantly lower for the fast delivery than for the slower delivery only for one SL speech, and overall speech rates did not differentially affect the interpreters. In contrast with the previous findings, Shlesinger (2003) reported the

“counter-intuitive” result that the SI performance measured by correct rendition of successive adjective modifiers was consistently better at a faster rate than at a slower rate.

### 5.2.2 Review of empirical studies: Accent factor

Another six empirical studies (Cheung, 2013; Kurz, 2009; Lin, Chang, & Kuo 2013; Mazzetti, 1999; Proffitt, 1997; Sabatini, 2000/01) based on different methodological paradigms examined the effects of the accent factor (i.e., native speaker/NS & non-native speaker/NNS) on SI performance. In general, the studies produced different results. On the one hand, four studies indicate that a NNS produced detrimental effects on SI performance: Kurz (2009), Lin et al. (2013), Mazzetti (1999), and Sabatini (2000/01). For example, Kurz (2009) found a markedly higher loss of information in the interpretation for a NNS than for a NS. On the other hand, Proffitt (1997) reported that interpreters achieved better results when working from NNS texts. In addition, Cheung (2013) found that in English-to-Cantonese SI, the quality of the interpretations produced by a native Cantonese interpreter was perceived to be better than that by non-native Cantonese speakers.

### 5.2.3 Discussion of the empirical studies

The inconsistent results from the empirical studies are discussed in light of six aspects including methodological paradigm, sampling, control of extraneous variables (EVs), assessment criteria used, measurement error, and statistical conclusion validity (SCV).

First, regarding the methodological paradigms, most of the studies use a quantitative approach (Gerver, 1969/2002; Kurz, 2009; Liu et al., 2004; Mazzetti, 1999; Meuleman & Van Besien, 2009; Pio, 2003; Sabatini, 2000/01; Shlesinger, 2003), usually embodied by an experiment. However, some studies are based on weak experimental research designs (Mazzetti, 1999; Meuleman & Van Besien, 2009), with no control groups, and without random assignment. In some strong experiments using repeated-measures designs (Pio, 2003; Proffitt, 1997), sample sizes are not balanced across groups due to limited samples and/or missing data; in the factorial designs (Liu et al., 2004), the independent variables (IVs) are not fully crossed, but nested, which limits the partition of variances. Moreover, in a few MMR designs (Kurz, 2009; Lin et al., 2013), it would appear that the quantitative and the qualitative data are not fully merged, drawing explicitly on MMR data analytic procedures (see Creswell, 2013). As a result,

future studies could employ strong MMR designs for in-depth understandings of the relationships.

Second, sampling has always been an issue in Interpreting Studies, given the difficulty to recruit a sufficient number of representative participants (Gile, 1998; Liu, 2011). In the studies reviewed, sample sizes range from 3 to 16 per group, and non-probability sampling is typically used. To make best of a small sample size, appropriate experimental designs should be chosen. Single-case experimental designs are a practical choice, which requires one to three carefully-chosen participants/cases. In addition, a repeated-measures design is an alternative, as it works well with smaller sample sizes (from 15 to 30) and uses participants as their own controls. To improve the generalizability of research results, replication or multiple studies using parallel samples represent a pragmatic method at the moment.

Third, although confounding effects of EVs on experimental results are widely recognized (see Johnson & Christensen, 2012; Marczyk, DeMatteo, & Festinger, 2005), the rigor and stringency in controlling potential EVs differs in the studies. For example, to standardize input materials for SI, some researchers use both qualitative and quantitative indicators to maximize consistency across SL texts (e.g., Liu et al., 2004), while others provide qualitative descriptions (e.g., Pio, 2003). More efforts should be invested to better control potential EVs. An exemplar is provided by Dillinger (1989) who applies stringent methods to improve comparability of experimental SI tasks.

Fourth, the divergent findings could be partly attributed to the use of different quality assessment criteria and quantification methods. Although the academia has generally agreed on a common set of SI quality criteria (see Pöchhacker, 2001), there seems to lack an empirically-driven and validated rating scale (with rubrics) tailored to SI quality assessment (see Lee, 2008). Using a common rating scale to assess SI performance can be efficient and reliable, and contribute to greater transparency and easier communication, although the alternative qualitative analysis of linguistic features (e.g., errors, omissions, pauses) as a mean of quality assessment also has its value (e.g., Gerver, 1969/2002). A rubrics-based rating scale could therefore be developed and validated for performance assessment purposes.

Fifth, although all the experiments incorporate a measurement procedure of some sort to assign quantitative indicators to performance, measurement errors (systematic & random) are not unanimously estimated and controlled for. Only Chang (2005) and Liu et al. (2004) estimate measurement errors due to rater/coder variability, and provide rater training. Given

that a reliable measurement underpins validity of research results, a robust measurement procedure should be used to reduce errors and generate reliable scores.

Last, regarding the statistical conclusion validity (SCV, see Maxwell & Delaney, 2004), majority of the studies using inferential statistics forget to report results of statistical assumption testing (e.g., normality, equality of variances), and to provide effect size indicators for the relationships of interest. Results from assumption testing and effect size estimates could be reported to help audience better understand the nature of collected data and the magnitude of the strength of an effect. Particularly, appropriate indicators of effect size should be calculated. Although statistical programs such as SPSS conventionally produce partial eta squared ( $\eta_p^2$ ), it has a number of limitations such as biasedness and over-estimation (see Pierce, Block, & Aguinis, 2004). For instance, based on the  $\eta_p^2$  statistics in Lin et al. (2013), the combined effects of phonemic and prosodic deviations accounted for 111.1% of the total model variation, which is practically impossible.

### **5.3 Research purpose and questions**

In light of the reviewed literature, the present study used a convergent parallel MMR design (see Creswell, 2013; Onwuegbuzie, Slate, Leech, & Collins, 2007) to investigate the effects of the speed and the accent factors on SI performance quality on the one hand, and to gain an emic view of how the two factors influence interpreters' perception of their performance, on the other. Specifically, the study endeavors to answer the three questions listed below. Given the inconsistent results in the previous studies, no specific and directional hypotheses were formulated *a priori*.

- 1) Does variation in speakers' delivery rates produce changes in SI performance quality measured by information completeness (InfoCom), fluency of delivery (FluDel) and target language quality (TLQual)?
- 2) Does native and non-native speakers' accent produce changes in SI performance quality measured by InfoCom, FluDel and TLQual?
- 3) How do interpreters perceive the variation of delivery speed and the native/non-native accent in SI tasks may affect their performance?

## 5.4 Methods

As indicated before, the study used a MMR design drawing on both quantitative and qualitative data and analysis. In the quantitative phase of the study, a factorial experiment was conducted, and in the qualitative phase interpreters' retrospection data was analyzed. In what follows, each of the two phases of the study is explained.

### 5.4.1 Mixed sampling design: Participants

Two groups of interpreters were recruited for a pilot and the experiment. A pilot was run in which 11 student interpreters participated. The pilot aimed at trialing relevant materials and data-collection procedures so as to identify potential problems and to streamline formal administration.

To participate in the experiment, interpreters must satisfy three criteria: 1) They were active Chinese/English simultaneous interpreters working in China; 2) They had received formal interpreting training, preferably obtained a postgraduate interpreting degree; and 3) They had practiced SI for at least two years. Using snowball sampling, 32 Beijing-based active interpreters were recruited. Averaged at 31 years old, they all had Mandarin Chinese as their L1 and English their L2. In addition, they had an average amount of 56-month SI experience and annual workload of about 47 conferences. For more demographic information, please see Table 5.1.

An identical concurrent mixed sampling design was used (see Collins, Onwuegbuzie, & Jiao, 2007). In other words, the same interpreters participated in both quantitative and qualitative phases of the study.

Table 5.1 Demographic information of the participants in the main study

Demographics	No.	Percent (%)
<b>Gender</b>		
Male	13	40.6
Female	19	59.4
<b>Education</b>		
Bachelor	5	15.6
Master	26	81.3
Doctorate	1	3.1
<b>Interpreter training</b>		
Intensive training course	9	28.1
Postgraduate interpreting degree	23	71.9
<b>Type of interpreter</b>		
Part-time	10	31.2
Freelance	17	53.1
In-house	5	15.6

#### 5.4.2 Experimental design

The quantitative part of the study used a 2×2 within-subjects factorial design (see Table 5.2). Two IVs were speech rate and accent, with each having two levels. The IVs were crossed to produce four treatment conditions (TCs). A repeated-measures design allowed us to observe all participants performing English-to-Chinese SI in all the TCs.

Table 5.2 The 2×2 factorial design

Independent variables (IVs)		IV A: Speech rate	
		a <sub>1</sub> : Slow	a <sub>2</sub> : Fast
IV B: Accent	b <sub>1</sub> : Native English	TC <sub>1</sub> : a <sub>1</sub> b <sub>1</sub> (Task <sub>SN</sub> )	TC <sub>3</sub> : a <sub>2</sub> b <sub>1</sub> (Task <sub>FN</sub> )
	b <sub>2</sub> : Accented English	TC <sub>2</sub> : a <sub>1</sub> b <sub>2</sub> (Task <sub>SA</sub> )	TC <sub>4</sub> : a <sub>2</sub> b <sub>2</sub> (Task <sub>FA</sub> )

#### 5.4.3 SI materials

##### 5.4.3.1 Development of SI tasks

Four SI tasks coded as Task<sub>SN</sub>, Task<sub>SA</sub>, Task<sub>FN</sub> and Task<sub>FA</sub> were developed to comply with TC<sub>1-4</sub>, respectively. The four tasks were carefully calibrated so that they could be comparable to each other, except speech rate and accent. To ensure task comparability, a multi-pronged approach was taken. First, eight SI tasks were developed in synchronicity from the outset, based on eight different authentic English speeches on the general topic of the Australia-China relationship (see Appendix E). Second, a framework of SI task characteristics (see Chapter 3) was used to ensure that all characteristics (excluding speed and accent) were maintained as consistently as possible across the eight tasks. Third, based on the eight tasks, four best aligned tasks were chosen for the experiment. Inspired by Dillinger (1989), a diverse array of indices was calculated to quantify linguistic features of the four source speeches for SI. Particularly, lexical, propositional and syntactic characteristics of the source texts were kept as similar as possible (see Appendix F). To compute these indices, a variety of computer programs were utilized. For example, linguistic characteristics were quantified using *Lexical Complexity Analyzer* (Lu, 2012) and *L2 Syntactical Complexity Analyzer* (Lu, 2010). Propositional density was calculated using *Computerized Propositional Idea Density Rater* (CPIDR) (Brown, Snodgrass, Kemper, Herman, & Covington, 2008). Nucleus and satellite elementary discourse units (EDUs) were coded, based on Carlson and Marcu (2001) (intra-coder agreement index > 90% at a two-month interval). Following Liu and Chiu (2009) and Liu et al. (2004), overall readability indices were also computed.

Even though the efforts were made to maximize comparability across the tasks, it is impossible to guarantee a perfect alignment. The point is to control the EVs as much as the research sources allow, so that research findings are defensible.

#### 5.4.3.2 *Manipulating the IV: Accent*

A presence/absence technique (see Johnson & Christensen, 2012) was used to distinguish native accent from non-native accent. Specifically, two speakers were recruited to record the four source texts: a native English speaker who was a human sciences PhD candidate, and an accented English speaker from India who had a doctorate in linguistics. Each speaker recorded one fast and one slow speech in a sound-proof studio with consent.

#### 5.4.3.3 *Manipulating the IV: Speech rate*

To properly define a fast and a slow speech rate, relevant literature was consulted (e.g., Chang, 2005; Liu et al., 2004). The guiding principle was that the speech rates should manifest a discernable difference, but not represent two extremes. An amount technique (see Johnson & Christensen, 2012) was applied such that a fast speed was defined as approximately 155 wpm; and a slow one about 105 wpm.

In addition, to maintain a uniform “text-internal rate of delivery” (Dillinger, 1989), the source texts for fast and slow delivery were divided into 50-word and 35-word segments, respectively. The speakers finished each segment at a 20-second interval while monitoring a timer. They practiced the source speeches until their delivery was smooth and natural. The speech rates were further fine-tuned digitally, using *Amazing Slow Downer*, which manipulates speech rates, without changing pitch levels.

#### 5.4.4 Other instruments

##### 5.4.4.1 *Post-task interview*

After each SI task, the participants were interviewed, which represents the qualitative phase of the study. Four questions were asked (see Appendix G), but only the first two questions are of concern in the study: 1) “What are the *prominent factor(s)* do you think contribute(s) to the difficulty of this SI task?” and 2) “How did these factors affect your SI performance?”

For the first question, if the participants correctly identified the manipulated IVs but nothing else as *prominent* difficulty factors, then the task development and the IVs manipulation had been successful. For the second question, the participants were expected to provide reflective descriptions on the effects of the TCs on SI quality. The interviews were conducted in Chinese to prevent any language barrier, and were audio-recorded with participants’ permission. The interviews were later transcribed for analysis. The transcripts were also translated into English when used to exemplify a certain phenomenon. The translation was done by the first author and double-checked by a NAATI-certified Chinese/English translator.<sup>29</sup>

##### 5.4.4.2 *Post-hoc questionnaire*

---

<sup>29</sup> NAATI: Australia’s National Accreditation Authority for Translators and Interpreters. NAATI is the national standards and accreditation body for translators and interpreters in Australia. It is the only agency that issues accreditations for practitioners who wish to work in this profession in Australia

After the participants completed all tasks, they filled out a questionnaire, providing demographic information on gender, age, education, interpreting training, type of interpreter, SI experience, and annual workload. They also rated the overall difficulty of each task, using a seven-point Likert scale (i.e., Very easy – 1, Very difficult – 7) (see Appendix H).

#### 5.4.5 Data collection procedure

One day before the experiment, a three-page background reading material covering all topics in the four source speeches was sent to the participants by email (see Appendix I). The researcher also contacted each participant by phone to ensure they read the material in advance; and they were asked to use only this material to prepare. On the experiment day, the data collection was conducted on an individual basis. Table 5.3 shows the experiment procedure.

Table 5.3 An overview of the experiment procedure

<b>Data collection procedure</b>
<b>1.</b> Introduction (Participant consent, evidence of ethics clearance, rapport-building)
<b>2.</b> Equipment training & practice session, followed by a <i>3-min short break</i>
<b>3.</b> Experimental session
<b>3.1</b> The 1 <sup>st</sup> round, followed by a <i>6-min short break</i>
<b>a.</b> Contextualizing interpreters (Background Information Sheet)
<b>b.</b> Performing SI for a given task
<b>c.</b> Post-task interview
<b>3.2</b> The 2 <sup>nd</sup> round (a, b & c), followed by a <i>8-min short break</i>
<b>3.3</b> The 3 <sup>rd</sup> round (a, b & c), followed by a <i>10-min short break</i>
<b>3.4</b> The 4 <sup>th</sup> round (a, b & c)
<b>4.</b> Wrapping-up (Questionnaire & compensation)

As shown in the table, written consent was first sought for all the participants. To offset practice effects, the participants were given sufficient time to warm up. To reduce fatigue effects, multiple short breaks of different lengths were provided. To counterbalance order effects of the tasks, a Latin square design was used. That is, the participants were randomly selected into four groups, with each group taking a different order of the tasks. In addition, at

the beginning of each round of SI, a *Background Information Sheet* was provided to contextualize the participants (see Appendix J). All performances were audio-recorded with consent. The experiment took approximately three hours to complete. By the end of the experiment, each participant was compensated with 1000 RMB (about US\$ 170) to offset impacts of potential work time lost.

#### 5.4.6 Performance assessment

Nine raters were recruited and normed in a 5-hour training session with consent. Raters used a descriptor-based rating scale to assess SI (see Appendix K). The scale consisted of three 8-point subscales. The subscales were InfoCom, FluDel and TLQual. In addition, each subscale was divided into four 2-point bands with descriptors provided for each band. The rating scale was constructed, piloted and revised prior to the operational use, and functioned properly based on Rasch-generated fit statistics (see Chapter 7). In addition, a fully-crossed rating design was employed in which each rater assessed all recorded interpretations.

Generalizability (G) theory was then used to calculate standard error of measurement (SEM) and a reliability-like G coefficient ( $\rho^2$ ) for the design of four tasks and nine raters (Table 5.4, also see Chapter 8). As can be seen in the table, all  $\rho^2$  values were greater than minimally accepted level of 0.80 for all three rating dimensions, suggesting reliable measurement.

Table 5.4 SEM &  $\rho^2$  for the scores of the three rating dimensions

Indices/Criteria	InfoCom	FluDel	TLQual
<b>SEM</b>	0.36	0.29	0.28
<b><math>\rho^2</math></b>	0.92	0.89	0.90

Given the high G coefficients, multiple scores provided by the nine raters were averaged to represent quantitative measures of the three criteria in each TC for each interpreter.

#### 5.4.7 Data analysis

Overall, the study used a balanced design, with no missing data. Three strands of data were collected (Table 5.5), including the rater-generated performance scores (i.e., performance data), the perceived task difficulty ratings (i.e., perception data) and the interpreters' interview recordings (i.e., interview data).

Table 5.5 Data source &amp; data type matrix

Data source	Data type	
	Qualitative	Quantitative
Interpreter	Interview data	Perception data
Rater	N/A	Performance data

Note: N/A = not applicable

For the performance data, a two-way repeated-measures multivariate analysis of variance (MANOVA) was performed to simultaneously investigate how the two IVs affected the overall SI performance. In addition, given that each task represented a unique combination of the IVs, investigating inter-task score differences sheds light on the effects of the two IVs, and also speaks to relative difficulty of the tasks. Therefore, treating the tasks/TCs as a new IV, a one-way repeated-measures MANOVA was also run to examine how the TCs affected the overall performance. Following significant MANOVA effects, univariate ANOVAs were conducted to examine the effects of the IVs on InfoCom, FluDel and TLQual, respectively. For significant main effects of ANOVAs, *post-hoc* comparisons were also carried out.

Similarly, for the perception data, a two-way repeated-measures ANOVA was run to explore the effects of the two IVs on the perceived difficulty of the tasks. A one-way repeated-measures ANOVA was also performed to examine effects of the TCs on the perceived difficulty ratings. SPSS 21 was used for all statistical analyses.

For the interview data, the software of NVivo 10 was employed to code interview recordings, and a series of matrix coding queries was then run to reveal how participants' self-identified "difficulty factors" affected the performance. Specifically, the codings categorized in a given theme were counted and transformed to quantitative data (i.e., data transformation/quantitizing, see Johnson & Christensen, 2012). In addition, transcripts of the interview recordings were used to exemplify a certain phenomenon.

Finally, to merge data in a convergent parallel MMR design, three data analytic procedures were taken following Creswell (2013): 1) data transformation (i.e., quantitizing), 2) side-by-side comparison of quantitative and qualitative findings, and 3) a joint display of data, which displays and merges both data types in a single visual, as provided in *Summary discussion: Triangulation and meta-inference*.

## 5.5 Results and discussions

### 5.5.1 Results I and discussion I

#### 5.5.1.1 Quantitative results: Performance data

*Effects of FSR and StrA* Table 5.6 shows the descriptive statistics for the performance scores. Although Shapiro-Wilk test of normality revealed a violation: FluDel scores in Task<sub>SA</sub>,  $W(32) = 0.93$ ,  $p = 0.05$ , the overall results satisfied the univariate normality assumption. Sphericity (i.e., homoscedasticity across all possible pairs of levels of an IV) was not an issue in the study.<sup>30</sup>

Table 5.6 Descriptive statistics for the performance data

Criteria/Task ID	Descriptive statistics M (SD)			
	Task <sub>SN</sub>	Task <sub>SA</sub>	Task <sub>FN</sub>	Task <sub>FA</sub>
<b>InfoCom</b>	4.99 (1.53)	4.35 (1.49)	4.66 (1.31)	3.73 (1.24)
<b>FluDel</b>	4.74 (1.05)	4.07 (0.99)	5.10 (0.95)	4.34 (0.98)
<b>TLQual</b>	4.95 (1.11)	4.65 (0.97)	5.04 (0.94)	4.71 (0.79)

Notes: M = mean, SD = standard deviation

To investigate whether the speed and the accent factors had significant overall impacts on performance scores, a MANOVA was first performed. Using Pillai's trace, there were statistically significant effects of speech rate,  $V = 0.71$ ,  $F(3, 29) = 24.17$ ,  $p < 0.01$ , and of accent,  $V = 0.69$ ,  $F(3, 29) = 21.82$ ,  $p < 0.01$ , on performance scores. However, the interaction effect was not significant,  $V = 0.15$ ,  $F(3, 29) = 1.73$ ,  $p = 0.18$ .

Following the significant MANOVA results, a series of univariate ANOVA was conducted to examine how the two IVs affected InfoCom, FluDel and TLQual. As shown in Table 5.7, the speed factor had a statistically significant main effect on InfoCom,  $F(1, 31) = 14.68$ ,  $p < 0.01$ , and on FluDel,  $F(1, 31) = 11.53$ ,  $p < 0.01$ , but not on TLQual,  $F(1, 31) = 0.65$ ,  $p = 0.43$ . In addition, the accent factor significantly affected all the three measures: InfoCom,  $F(1, 31) = 42.01$ ,  $p < 0.01$ ; FluDel,  $F(1, 31) = 61.28$ ,  $p < 0.01$ , and TLQual,  $F(1, 31) = 18.25$ ,  $p < 0.01$ . However, a significant effect for the speed-by-accent interaction was not observed across the measures.

<sup>30</sup> Sphericity becomes an issue when an independent variable has three or more than three levels.

Table 5.7 Univariate ANOVA effects

ANOVA effect	Measures	<i>df</i>	MS	<i>F</i>	$\rho$
<b>Speech rate</b>	InfoCom	1	7.14	14.68	***
	FluDel	1	3.05	11.53	***
	TLQual	1	0.16	0.65	0.43
<b>Accent</b>	InfoCom	1	19.53	42.01	***
	FluDel	1	16.35	61.28	***
	TLQual	1	3.19	18.25	***
<b>Speech rate <math>\times</math> Accent</b>	InfoCom	1	0.68	2.58	0.12
	FluDel	1	0.07	0.32	0.57
	TLQual	1	0.007	0.03	0.86

Note: \*\*\*  $\rho < 0.01$ ; \*\*  $p < 0.05$ ; \*  $\rho < 0.1$ ; MS = mean squares

As shown in Figure 5.1(a), regarding the speed factor, the contrast indicated that the average InfoCom score for the fast-speech-rate SI tasks ( $M = 4.20$ ,  $SD = 1.35$ ) was significantly lower than that of the slow-speech-rate tasks ( $M = 4.67$ ,  $SD = 1.53$ ),  $r = 0.57$ .<sup>31</sup> Additionally, on average, the interpreters performed significantly better on FluDel, when they interpreted for the fast speeches ( $M = 4.72$ ,  $SD = 1.03$ ) than the slow speeches ( $M = 4.41$ ,  $SD = 1.07$ ),  $r = 0.52$ . Finally, there was no statistically substantial difference of the average TLQual scores between the fast ( $M = 4.87$ ,  $SD = 0.88$ ) and the slow speeches ( $M = 4.80$ ,  $SD = 1.05$ ),  $r = 0.14$ . In Figure 5.1(b), the contrasts revealed that the interpreters performed significantly worse in the accented speeches than the non-accented speeches for all the three measures: InfoCom,  $r = 0.76$ ; FluDel,  $r = 0.66$ ; TLQual,  $r = 0.61$ .

<sup>31</sup> Effect size  $r$  was calculated for the contrasts in the factorial repeated-measures ANOVAs after Field (2009). Conventions for interpreting  $r$  as an effect size indicator are:  $\leq 0.10$  (small effect),  $\geq 0.30$  (medium effect) and  $\geq 0.50$  (large effect) (Murphy & Myers, 2004).

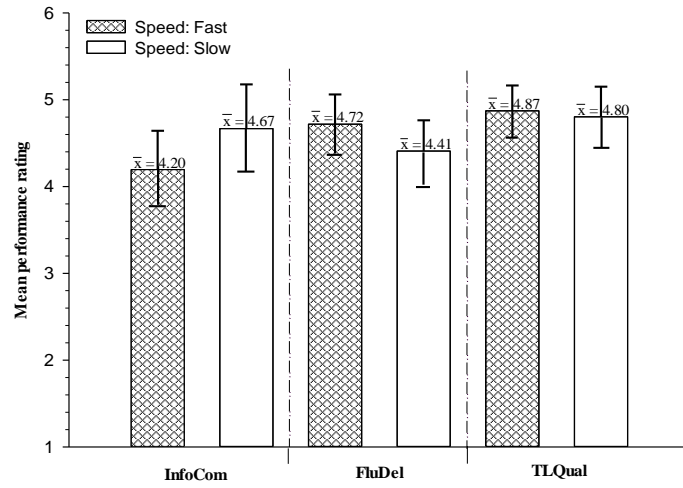


Figure 5.1(a) Performance scores (Speed)

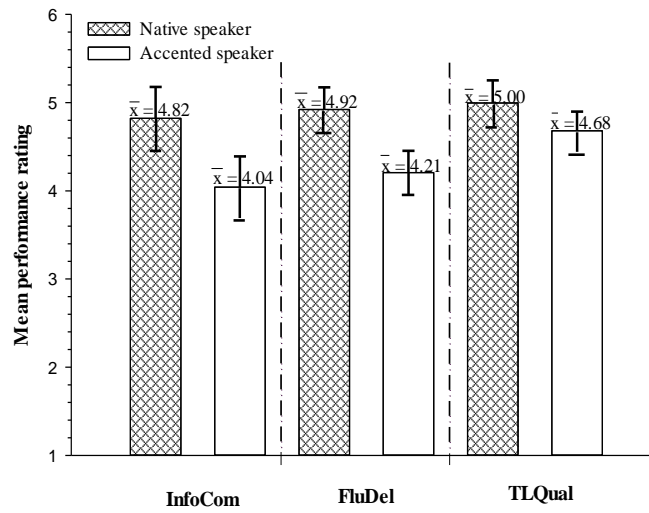


Figure 5.1(b) Performance scores (Accent)

*Effects of TCs* To investigate how the TCs affected the performance scores, a one-way repeated-measures MANOVA was carried out. Using Mauchly's test, the sphericity assumption was not violated: for InfoCom,  $W = 0.85$ ,  $\chi^2(5) = 4.94$ ,  $p = 0.42$ ; for FluDel,  $W = 0.90$ ,  $\chi^2(5) = 2.99$ ,  $p = 0.70$ ; and for TLQual,  $W = 0.79$ ,  $\chi^2(5) = 6.90$ ,  $p = 0.23$ , therefore degrees of freedom ( $df$ ) did not need correction. Using Pillai's trace, there was a statistically significant effect of the TCs on the overall performance scores,  $V = 1.03$ ,  $F(9, 279) = 16.18$ ,  $p < 0.01$ .

Following the significant MANOVA results, univariate ANOVAs were conducted to examine how the TCs affected the three measures. As displayed in Table 5.8, the TCs had statistically significant effects on each measure. This result suggests that for each measure at least two TCs differed significantly regarding the average scores.

Following the significant ANOVA effects, a series of *post-hoc* multiple comparisons were conducted for each measure, using the Bonferroni correction method. Results show that all pairwise comparisons turned out to be statistically significant (albeit at different  $\rho$  levels), except for Task<sub>SN</sub> – Task<sub>FN</sub> (mean difference /MD = -0.08,  $\rho = 0.52$ ) and Task<sub>SA</sub> – Task<sub>FA</sub> (MD = -0.06,  $\rho = 0.63$ ).

Table 5.8 ANOVA results for the effects of the TCs on the quality criteria

ANOVA effect	Measures	df	MS	F	$\rho$	Omega squared <sup>32</sup> ( $\omega^2$ )
Treatment conditions	InfoCom	3	9.12	22.53	***	0.05
	FluDel	3	6.49	25.62	***	0.03
	TLQual	3	1.12	5.11	***	0.00

Note: \*\*\*  $\rho < 0.01$ ; \*\*  $\rho < 0.05$ ; \*  $\rho < 0.1$ ; SS = sum of squares; MS = mean squares

#### 5.5.1.2 Discussion I: Performance data

Regarding the causal effects of the speed factor on SI quality, the analysis shows a pattern of mixed relationships: 1) Compared with the performance in the slow-speech-rate conditions, the FSR had detrimental effects on InfoCom, which generally concurs with Chang (2005), Gerver (1969/2002), and Pio (2003); 2) The FSR contributed appreciably to higher FluDel scores, running counter to Gerver (1969/2002) and Pio (2003); and 3) The FSR did not exert substantial impacts on TLQual, similar to Chang's (2005) results. Regarding the effects of the accent factor on SI quality, the presence of StrA had considerable impairing effects on all the three measures, which corroborates the findings from Kurz (2009), Lin et al. (2013), and Mazzetti (1999).

The inconsistent results between the present study and the previous studies, and among the previous studies is probably due to the operationalization of SI quality criteria. For example, in assessing FluDel, descriptors in a rating scale could be written only in relation to delivery rate of interpretations, while speech analysis of renditions could only focus on filled/unfilled pauses, repetitions, and other disfluencies. The disjunction of conceptualizations and definitions of FluDel probably led to divergent results. In addition, even if the content of the

<sup>32</sup> Omega squared ( $\omega^2$ ) is calculated as a measure of effect size for the one-way repeated-measures ANOVAs after Field (2009). Conventions for interpreting  $\omega^2$  are:  $\leq 0.01$  (small effect),  $\geq 0.06$  (medium effect) and  $\geq 0.15$  (large effect) (Murphy & Myers, 2004).

descriptors corresponds to that of speech analysis, raters may have different internalized representations of the rating scale, thus resulting in inconsistent use of the scale between the raters. It is therefore worthwhile to compare data derived from a rating scale and speech analysis to examine their fit.

Regarding the effects of the TCs on the SI quality, a mixed pattern is also identified: 1) The pattern of the InfoCom scores was  $\text{Task}_{\text{SN}} > \text{Task}_{\text{FN}} > \text{Task}_{\text{SA}} > \text{Task}_{\text{FA}}$ , indicating the increasing task difficulty levels; 2) The pattern of the FluDel scores was  $\text{Task}_{\text{FN}} > \text{Task}_{\text{SN}} > \text{Task}_{\text{FA}} > \text{Task}_{\text{SA}}$ , suggesting the increasing difficulty levels in terms of FluDel. 3) The pattern of the TLQual scores was  $\text{Task}_{\text{FN}} \approx \text{Task}_{\text{SN}} > \text{Task}_{\text{FA}} \approx \text{Task}_{\text{SA}}$ .

## 5.5.2 Results II and discussion II

### 5.5.2.1 Quantitative results: Perception data

*Effects of FSR and StrA* Table 5.9 provides the descriptive statistics for the perception data. Normality testing revealed violation of the assumption in all the tasks. However, given the robustness of the  $F$  test, the violation should not be a major problem.

Table 5.9 Descriptive statistics for the perception data

Difficulty/Task ID	Descriptive statistics M (SD)			
	$\text{Task}_{\text{SN}}$	$\text{Task}_{\text{SA}}$	$\text{Task}_{\text{FN}}$	$\text{Task}_{\text{FA}}$
<b>Perceived difficulty</b>	2.81 (1.03)	4.06 (1.08)	4.59 (1.39)	5.34 (1.29)

Notes: M = mean, SD = standard deviation

To examine the effects of speech rate and accent on the perceived task difficulty, a two-way repeated-measures ANOVA was performed. The ANOVA shows that there were statistically significant main effects of speech rate,  $F(1, 31) = 37.54, \rho < 0.01$ , and of accent,  $F(1, 31) = 28.34, \rho < 0.01$ . It indicates that FSR and StrA substantially altered the interpreters' perception of task difficulty. No significant interaction effect was found,  $F(1, 31) = 2.70, \rho = 0.11$ .

As shown in Figure 5.2, the contrasts revealed that the interpreters perceived the fast-speech-rate tasks ( $M = 4.97, SD = 1.38$ ) to be more difficult than the slow-speech rate tasks ( $M = 2.44, SD = 1.22$ ),  $r = 0.74$ , and the non-accented tasks ( $M = 3.70, SD = 1.51$ ) to be much easier than the accented tasks ( $M = 4.70, SD = 1.34$ ),  $r = 0.69$ .

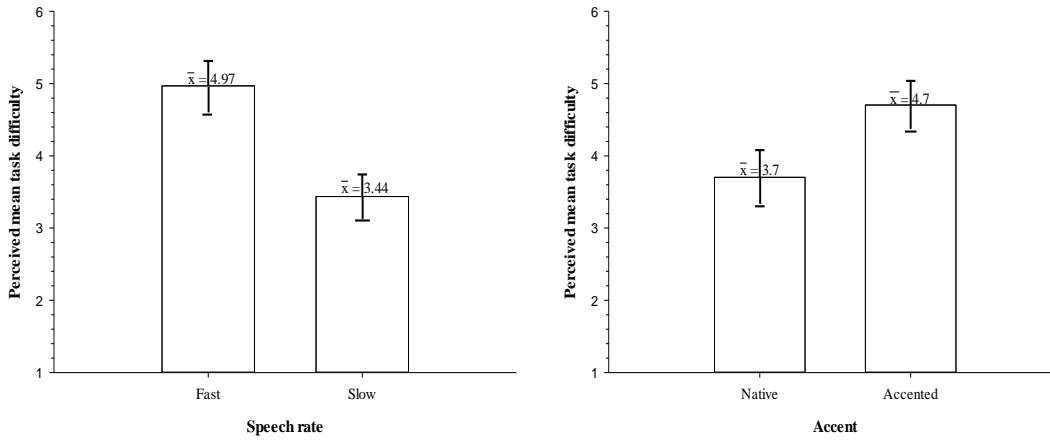


Figure 5.2 Averaged perceived difficulty ratings

*Effects of the TCs* To investigate the effects of the TCs on the perceived difficulty, a one-way repeated-measures ANOVA was carried out. Given the four levels of the TCs, the sphericity assumption was examined, using Mauchly's test. The result showed a violation of the sphericity assumption at the  $\rho$  level of 0.05,  $W = 0.69$ ,  $\chi^2(5) = 11.07$ . Degrees of freedom were thus corrected using Huynh-Feldt estimates of sphericity ( $\varepsilon = 0.87$ ). Based on the correction, the ANOVA shows a statistically significant effect of the TCs on the perceived difficulty ratings,  $F(2.61, 31) = 28.17$ ,  $\rho < 0.01$ ,  $\omega^2 = 0.37$ .

Following the significant main effect, *post-hoc* multiple comparisons were conducted, using the Bonferroni method. Results show that all pairwise comparisons were statistically significant ( $\rho < 0.01$ ), except  $\text{Task}_{\text{SA}} - \text{Task}_{\text{FN}}$  (mean difference = -0.53,  $\rho = 0.14$ ).

#### 5.5.2.2 Discussion II: Perception data

Based on the interpreters' perceptions, FSR made the tasks significantly more difficult; and on average the SI tasks involving the NS were regarded significantly easier than those involving the NNS. These results confirm the survey findings that FSR and StrA are cited as a prominent difficulty factor in SI (see Chapter 4). The results also imply that the performance scores could be significantly lower in FSR and StrA conditions.

In terms of the perceived difficulty of the tasks, the following observations can be made: 1)  $\text{Task}_{\text{SN}}$  was perceived to be significantly less difficult than both  $\text{Task}_{\text{SA}}$  and  $\text{Task}_{\text{FN}}$ , 2)  $\text{Task}_{\text{SA}}$  and  $\text{Task}_{\text{FN}}$  did not differ substantially, and 3) both  $\text{Task}_{\text{SA}}$  and  $\text{Task}_{\text{FN}}$  were significantly less difficult than  $\text{Task}_{\text{FA}}$ . That is, the pattern of the perceived difficulty was  $\text{Task}_{\text{SN}} < \text{Task}_{\text{SA}} \approx$

$\text{Task}_{\text{FN}} < \text{Task}_{\text{FA}}$ . This result implies that overall the SI performance scores could be highest in  $\text{Task}_{\text{SN}}$  and lowest in  $\text{Task}_{\text{FA}}$ .

### 5.5.3 Results III and discussion III

#### 5.5.3.1 Qualitative results: Task development and IVs manipulation

Based on the analysis of the answers to the first interview question, prominent difficulty factors were identified for each SI task (see Table 5.10). For  $\text{Task}_{\text{SN}}$ , difficulty factors were sparsely distributed, and the participants felt comfortable, interpreting for  $\text{Task}_{\text{SN}}$ . For example:

**P08:** 这一个好像是 4 个里面最合适的一个。语速和发音可能都会比较舒服。

It seems that this speech is the most suitable one for (SI) among the four speeches. (I feel) comfortable with its delivery speed and accent.

**P21:** 这篇应该没有什么难点, 几乎都没有。他说得也比较清楚一点吧。

There should be no difficulty factors in this speech. He delivered (messages) clearly.

Given no outstanding difficulty factors in  $\text{Task}_{\text{SN}}$ , some interpreters were even able to focus on language refinement. For instance:

**P16:** .....说话的方式, 用词更注意一些。对自己译语更多的期待。

(I) paid more attention to my way of delivering and diction. I had more expectation on my target language.

**P22:** 信息传达方面不是大问题, ....., 但更多的就是怎么把语言说得更漂亮一些。

It's not a problem to communicate information. ..., the important thing is to deliver your speech more elegantly.

For  $\text{Task}_{\text{SA}}$ , it seems that the task difficulty primarily stemmed from the strong accent (StrA) of the Indian speaker, attested by the fact that 84% of the 32 interpreters cited StrA. For example:

**P14:** 有点口音, 有几个关键词反应不过来。

(The speaker spoke with) a bit of accent. I couldn't understand several key words.

For Task<sub>FN</sub>, the task was felt difficult, largely because of the fast speech rate (FSR) of the speaker, since all the 32 interpreters attributed difficulty to FSR. For example:

**P20:** .....它语速特别，特别快，我一下子没反应过来。然后语速对我造成很大的干扰。

... the delivery rate was very, very fast. I couldn't respond immediately. The delivery rate greatly hampered my (SI).

**P31:** 这篇我觉得语速比较快，跟的时候比较难，有心里负担。

I think the delivery rate of this speech was relatively fast and difficult to follow, which was a burden psychologically.

Table 5.10 Difficulty factors mentioned by the interpreters

Difficulty factors	Task ID			
	Task <sub>SN</sub>	Task <sub>SA</sub>	Task <sub>FN</sub>	Task <sub>FA</sub>
Bad sound quality	†	2 (6%)	†	17 (53%) ♠
Dense information	2 (6%)	1 (3%)	†	†
Fast speech rate (FSR)	†	1 (3%)	32 (100%) ♠	22 (69%) ♠
Illogic text structure	†	1 (3%)	†	†
Lack of background	2 (6%)	2 (6%)	†	3 (9.4%)
Lexical complexity	1 (3%)	1 (3%)	†	†
Numbers & figures	4 (13%)	1 (3%)	3 (9%)	5 (16%)
Strong accent (StrA)	1 (3%)	27 (84%) ♠	†	23 (72%) ♠
Syntactic complexity	4 (13%)	6 (19%)	†	†
Too slow delivery	1 (3%)	5 (16%)	†	†
<b>Total</b>	15 (-)	47 (-)	35 (-)	70 (-)

Note: † no coding; numbers outside of parenthesis = the No. of codings, percent (%) is calculated by dividing the No. of codings by 32 (i.e., the No. of the interpreters); ♠ = possible prominent difficulty factor. - “not applicable”.

Regarding Task<sub>FA</sub>, Table 5.10 presents three possible difficulty factors including bad sound quality (17, 53%), FSR (22, 69%) and StrA (23, 72%). For example:

**P24:** 难点是他说的太快了。正常人的 delivery 不会是这样子的，太快了。

The difficulty is that he spoke too fast. Delivery by an average speaker wouldn't be like this.

It's too fast.

**P22:** 我没有想到她是印度人说话，所以我觉得心理上有点意外。一开始的时候没有适应过来。

I didn't expect her to be an Indian speaker. I was a bit surprised, psychologically. At the beginning, I couldn't adapt myself to (her accent).

#### 5.5.3.2 Discussion III: Prominent difficulty factors

Given “bad sound quality” was not intended to be a difficulty factor, it requires special attention. An inspection of the codings and the original recordings reveal that there was a sudden decrease of sound level in the last minute for Task<sub>FA</sub>, probably due to technical problems at the recording stage. Although “bad sound quality” is identified, it is argued that it does not constitute a *prominent* difficulty factor, for three reasons. First, the decrease of sound occurred at the very end of the task and accounted for only 12% of the recording length. If it occurred at the beginning of the recording, it would impair interpreters' comprehension throughout. Second, the interpreters reported raising the sound level when they encountered the problem. In fact, prior to the experiment, all the participants were trained on how to adjust input sound level according to their own needs. Third, almost half of the participants did not report this sound problem, which may indicate its limited impacts. As a result, only FSR and StrA were regarded as the *prominent* factors contributing to the difficulty of Task<sub>FA</sub>. Overall, the qualitative analysis indicates that the task development and the IVs manipulation had been largely successful, and the tasks functioned as expected.

#### 5.5.4 Results IV and discussion IV

##### 5.5.4.1 Qualitative results: Interpreters' reflections on SI performance

The qualitative data (i.e., answers to the second interview question) were first analyzed and coded into two categories for each task-by-criterion condition. The first category was about the interpreters' comments on the negative effects of a TC on a certain criterion (i.e., presence of negative effects), while the second category was about the positive and/or the non-negative effects (i.e., absence of negative effects). For example:

Presence of negative effects:

**P02:** 他的语速和口音交叠起来, 会影响我的信息完整度。

His (the speaker's) delivery speed and accent overlapped with each other, which (negatively) affected the completeness of information.

Absence of negative effects:

**P19:** .....因为他很流畅, 不想前面一个一点一点地说, 比前一个的流畅度要稍微好一些。

... because he (the speaker) spoke fluently, unlike the previous speaker who uttered bit by bit, (my output) is slightly more fluent than my previous performances.

The qualitative data were then transformed into quantitative data. Table 5.11 shows the interpreters' self-reflections on how the way speakers delivered speeches affected the three criteria. Based on Table 5.11, five observations can be made:

- 1) There were much fewer comments provided on Task<sub>SN</sub> than the other tasks. This may indicate that given an opportunity to comment, the interpreters tended to focus on Task<sub>SA</sub>, Task<sub>FN</sub> and Task<sub>FA</sub>, and were less concerned with Task<sub>SN</sub>. Task<sub>SN</sub> was then regarded as a benchmark or a reference task.
- 2) Task<sub>SA</sub> differs from Task<sub>SN</sub> primarily in terms of StrA. For Task<sub>SA</sub>, across the measures, the interpreters made more comments on the negative effects of the StrA than the non-negative effects. In other words, they believed that the StrA was detrimental to overall SI performance. This is largely because the StrA caused problems to the interpreters' listening and comprehension. For example:

1)

**P10:** .....因为我要把精力放在听上面, 然后在说的这一块儿, 就不能好好的监听我说的东西。对信息完整度会有影响, 准确度会有影响。

... I need to direct my attention on listening (due to the accent). I couldn't monitor what I had said effectively. (As a result,) it (negatively) affected information completeness and accuracy.

**P32:** ..... 因为你的思维就不这么连贯了, 你要去想她到底在说什么东西, 会分一部分神去做那个事情。

... your train of thoughts was no longer smooth. You have to think about what she

tried to express, and you need to split some attention to attending to that.

Table 5.11 Number of codings & percentage for each task-by-criterion condition

Criteria	Effects	Task ID				
		Task <sub>SN</sub>	Task <sub>SA</sub>	Task <sub>FN</sub>	Task <sub>FA</sub>	Average
InfoCom	▲	2 (6%)	17 (53%)	25 (78%)	23 (72%)	17 (52%)
	△	1 (3%)	4 (13%)	4 (13%)	0 (0%)	2 (7%)
FluDel	▲	1 (3%)	13 (41%)	9 (28%)	14 (44%)	9 (29%)
	△	1 (3%)	6 (19%)	13 (41%)	5 (16%)	6 (20%)
TLQual	▲	1 (3%)	8 (25%)	16 (50%)	10 (31%)	9 (27%)
	△	0 (0%)	5 (16%)	1 (3%)	2 (6%)	2 (6%)
<b>Average</b>	<b>▲/△</b>	1(4%)/2 (2%)	13(40%)/5(16%)	17(52%)/6(19%)	16(49%)/2(7%)	N/A

Notes: ▲ = presence of negative effects, △ = absence of negative effects; numbers outside parenthesis = No. of codings; % was calculated by dividing the No. of codings by 32. N/A = not applicable.

- 3) Task<sub>FN</sub> differs from Task<sub>SN</sub> primarily because of FSR. For Task<sub>FN</sub>, when it comes to InfoCom and TLQual, the interpreters commented more on the negative effects of FSR than the non-negative effects. However, regarding FluDel more comments were made on the non-negative effects of FSR. It could suggest that overall the interpreters believed they had performed reasonably well on the FluDel criterion even under the FSR condition. This is probably because in keeping with the FSR of the source speeches, they speeded up SI, making renditions sound more fluent. For example:

**P03:** 流畅还好, 他要快, 你就跟着快。

It is OK with fluency of delivery, (because) if he (the speaker) speeds up, you pace up as well.

**P16:** 我觉得语速太慢的话, 对流畅度有影响。那这个说得很快, 译员也需要翻得很快, 听上去可能是更加流畅一些。因为如果他说得很慢, 你需要的等很久才出一句, 他的语言反而更加支离破碎。

I think if (a speaker) speaks too slowly, it will affect (my) delivery. This (speaker)

speaks fast, the interpreter (I) need to respond and interpret fast as well. (As a result,) (my output) may sound more fluent. If a speaker speaks slowly, you need to wait for a long time before producing a sentence, then the output will become more fragmented instead.”

- 4) Task<sub>FA</sub> is different from Task<sub>SN</sub> largely because of the presence of the StrA and FSR. In Table 5.11, the interpreters expressed far more opinions on the negative effects of Task<sub>FA</sub> on the SI performance across the three criteria than its non-negative impacts. For example:

**P29:** 对信息完整度, 和流畅度都有影响。整个表现会受到很大的折扣。

(the way speaker delivers the speech) had (negative) impacts on both information completeness and fluency of delivery. The overall performance quality was reduced appreciably.

- 5) In Table 5.11, the InfoCom criterion was most commented on by the interpreters. Specifically, summarized over the tasks, proportionally more comments of the negative effects were associated with InfoCom than the other two criteria: 52% for InfoCom, 29% for FluDel, 27% for TLQual. This result may suggest that InfoCom was perceived to be more vulnerable to the adverse speech conditions.

#### *5.5.4.2 Discussion IV: Perceived effects on SI*

The qualitative data provides an insider’s viewpoint of how the interpreters perceive the prominent difficulty factors to affect SI performance. Overall, FSR and StrA forced the interpreters to step out of a “cognitive” comfort zone, where the pattern of assigning cognitive capacity to multiple parallel processes required by SI has been more or less habituated in normal speech conditions. The unexpected conditions compelled the interpreters to break the old pattern and re-distribute cognitive capacity in order to establish a new attention-sharing mechanism (Gile, 1995). Such a mechanism seems to be unstable and unnatural to the interpreters.

Specifically, the FSR led to more information units expressed in a unit time span, and tended to overload interpreters' cognitive capacity during SI. As a result of cognitive saturation, some information was lost, reducing InfoCom scores. For FluDel, since SI is an externally (speaker-) paced activity, the FSR prompted the interpreters to comprehend speech segments in a relatively short period of time, and catapulted them to produce renditions as quickly as possible, thus increasing the general flow of speech.

Regarding the accent factor, one of the biggest problems of working with the StrA was that the interpreters had to exert more efforts on listening and lagged long behind the source speeches for better comprehension, which leads to decreased attention on TL production and induces more pauses. Consequently, some information was lost due to difficulty in comprehension; fluency impaired because of intermittent pauses; and language quality suffered owing to incomprehension.

#### 5.5.5 Summary discussion: Triangulation and meta-inference

Table 5.12 presents all the results from the qualitative and the quantitative components on the same panel. As can be seen in the table, to colligate the results, three strands of triangulation were conducted: A (i.e.  $a_1$ ,  $a_2$ ,  $a_3$  &  $a_4$ ), B (i.e.  $b_1$ ,  $b_2$ ,  $b_3$  &  $b_4$ ) and C (i.e.  $c_1$ ,  $c_2$  &  $c_3$ ).

Strand A concerns the causal effects of the FSR on the SI performance. Specifically,  $a_1$ ,  $a_2$  and  $a_4$  point to the similar results that the FSR had a negative impact on the SI quality, especially InfoCom. However,  $a_3$  indicates a generally positive effect of the FSR on FluDel. This is primarily because the interpreters speeded up SI to keep up with the FSR speakers, making the renditions sound fluent.

Strand B relates to the causal effects of the StrA on the SI performance. An analysis of  $b_1$ ,  $b_2$ ,  $b_3$  and  $b_4$  confirms a sweeping detrimental effect on all the quality measures, primarily because the comprehension phase of SI was hampered.

Strand C has to do with the effects of the TCs on the SI performance and the task difficulty. Specifically,  $c_3$  demonstrates that Task<sub>SN</sub> was a benchmark or baseline, and Task<sub>FA</sub> was the most difficult. Additionally,  $c_1$  shows that Task<sub>SN</sub> and Task<sub>FA</sub> were the easiest and most difficult tasks, respectively; while Task<sub>SA</sub> and Task<sub>FN</sub> lied somewhere in the continuum, which is generally consistent with the results based on  $c_3$ .

However, an inconsistency stands out. That is, as shown in Table 5.12, while the perception-based data inference of "StrA → Overall SI" is confirmed by the performance data

across the three measures, the inference of “FSR → Overall SI” is only supported by the InfoCom criterion.

The inconsistency could be explained by a meta-inference (see Teddlie & Tashakkori, 2009). The interpreters as a group may have internalized InfoCom as a foremost criterion to assess their SI performance. Consequently, task difficulty was evaluated primarily based on how a given task would affect the InfoCom criterion. That is, if a task is viewed as difficult, it is primarily because the task involves factors that are perceived to do disservice to InfoCom. The inference is supported by the qualitative results that the interpreters provided far more comments on InfoCom than the others, suggesting a dominant status of InfoCom as a quality criterion. The interpreters talked much about InfoCom, because the criterion was important to them.

Table 5.12 Comparison and triangulation of the results from the quantitative and the qualitative components

Quantitative: Performance data	Quantitative: Perception data	Qualitative: Interview data
<ul style="list-style-type: none"> <li>• FSR → InfoCom, (-) <sup>***</sup>, <math>r = 0.57</math></li> <li>• FSR → TLQual, (-), <math>r = 0.14</math></li> <li>• FSR → FluDel, (+) <sup>***</sup>, <math>r = 0.52</math></li> <li>• StrA → InfoCom, (-) <sup>***</sup>, <math>r = 0.76</math></li> <li>• StrA → TLQual, (-) <sup>***</sup>, <math>r = 0.61</math></li> <li>• StrA → FluDel, (-) <sup>***</sup>, <math>r = 0.66</math></li> </ul>	<ul style="list-style-type: none"> <li>• FSR → Overall difficulty, (-) <sup>***</sup>, <math>r = 0.74</math></li> <li>♦ Inference: FSR → Overall SI, (-), significant</li> <li>• StrA → Overall difficulty, (-) <sup>***</sup>, <math>r = 0.69</math></li> <li>♦ Inference: StrA → Overall SI, (-), significant</li> <li>• Overall difficulty: <math>T_{SN} &lt;^{***} T_{SA} \approx T_{FN} &lt;^{***} T_{FA}</math></li> <li>♦ Inference: Overall SI: <math>T_{SN} &gt; T_{SA} \approx T_{FN} &gt; T_{FA}</math></li> </ul>	<ul style="list-style-type: none"> <li>• Prominent difficulty factors identified</li> <li>• <math>T_{SN}</math> is a benchmark</li> <li>• FSR → InfoCom &amp; TLQual (-)</li> <li>• FSR → FluDel (+)</li> <li>• StrA → InfoCom, FluDel &amp; TLQual (-)</li> <li>• StrA &amp; FSR (-)</li> <li>• InfoCom: most commented on</li> </ul>
<p>Notes: FSR = fast speech rate, StrA = strong accent, → Affect, (-) Negative effect, (+) Positive effect, <sup>***</sup> <math>\rho &lt; 0.01</math>, <sup>*</sup> <math>\rho &lt; 0.1</math>, <math>r</math> = effect size, T = Task, &lt;---&gt; Inter-relationship.</p>		

## 5.6 Limitations of the study

The study has three limitations: 1) The snowball sampling reduces the generalizability of the results. Future studies could recruit a different cohort of interpreters to confirm the results. 2) “Bad sound quality” could confound the results. Preventative measures should have been implemented prior to the recording. 3) The rating scale used is not fully validated, which could affect measurement accuracy. Future studies could compare quantitative indicators based on the rating scale and speech analysis of the renditions to examine how the rating scale functions in relation to text-based assessment of the performance.

## 5.7 Conclusion

Despite the limitations above, the present study has some main contributions. Specifically, it uses a convergent parallel MMR design to investigate how FSR and StrA affect SI performance. The results show a pattern of mixed impacts of the speed factor on InfoCom, FluDel and TLQual dimensions of SI performance, and a consistent pattern of detrimental impacts of the accent factor across the rating dimensions. The data triangulation also reveals that InfoCom could have been internalized by the interpreters as a key criterion to assess SI performance quality.

## 5.8 References

- Angelelli, C. (2009). Using a rubric to assess translation ability: defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp. 13-47). Amsterdam: John Benjamins.
- Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540-545.
- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.

- Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual* (ISI Technical Report ISI-TR-545). Retrieved from <http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf>
- Chang, C. C. (2005). *Directionality in Chinese/English simultaneous interpreting: Impact on performance and strategy use* (Doctoral thesis, University of Texas at Austin, USA). Retrieved from <https://www.lib.utexas.edu/etd/d/2005/changc71804/changc71804.pdf>
- Chen, J. (2009). Authenticity in accreditation tests for interpreters in China. *The Interpreter and Translator Trainer*, 3(2), 257-273.
- Cheung A. K. F. (2013). Non-native accents and simultaneous interpreting quality perceptions. *Interpreting*, 15(1), 25-47.
- Creswell, J. W. (2013). *Research design: Qualitative, Quantitative and Mixed Methods Approaches* (4th ed.). Thousand Oaks: Sage.
- Dillinger, M. L. (1989). *Component Processes in Simultaneous Interpreting* (Doctoral thesis, McGill University, Canada). Retrieved from [http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object\\_id=39215&local\\_base=GEN01-MCG02](http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object_id=39215&local_base=GEN01-MCG02)
- Gerver, D. (1969/2002). The effects of source language presentation rate on the Performance of simultaneous conference interpreters, Republished in F. Pöchhacker & M. Shlesinger (Eds.), *The Interpreting Studies Reader* (pp. 52-66). London and New York: Routledge.
- Gerver, D. (1974). The effects of noise on the performance of simultaneous interpreters: Accuracy of performance. *Acta Psychologica*, 38(3), 159-167.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.
- Gile, D. (1998). Observational studies and experimental studies in the investigation of conference interpreting. *Target*, 10(1), 69-93.
- Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting – a contribution. *Hermes*, 23, 153-172.
- Huang, M. (2005). 谈口译资格认证考试的规范化设计. [Toward a more standardized large-scale accreditation test for interpreters]. *中国翻译*, 6, 62-65.
- Jesse, A., Vrignaud, N., & Massaro, D. W. (2000/01). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, 5(2), 95-115.
- Johnson, B. R. (2009). Toward a more inclusive “scientific research in education”.

*Educational Researcher*, 38(6), 449-457.

- Johnson, B., & Christensen, L. (2012). *Educational Research: Quantitative, qualitative and mixed approaches* (4th ed.). Thousand Oaks: Sage.
- Kurz, I. (2009). The impact of non-native English on students' interpreting performance. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (pp. 179-192). Amsterdam: John Benjamins.
- Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, 2(2), 165-184.
- Lin, I. I., Chang, F. A., & Kuo, F. (2013). The impact of non-native accented English on rendition accuracy in simultaneous interpreting. *The International Journal for Translation and Interpreting Research*, 5(2), 30-44. DOI: ti.105202.2013.a03
- Liu, M. H. (2011). Methodology in interpreting studies: A methodological review of evidence-based research. In B. Nichoemus & L. Swabey (Eds.), *Advances in Interpreting Research: Inquiry in action* (pp. 85-119). Amsterdam: John Benjamins.
- Liu, M. H., & Chiu, Y. H. (2009). Assessing source material difficulty for consecutive interpreting: Quatifiable measures and holistic judgement. *Interpreting*, 11(2), 244-266.
- Liu, M. H., Schallert, D., & Carroll, P. (2004). Working memory and expertise in simultaneous interpreting. *Interpreting*, 6(1), 19-42.
- Lu, X. F. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474-496.
- Lu, X. F. (2012). The relationship of lexical richness to the quality of ESL learners's oral narratives. *The Modern Language Journal*, 96(2), 190-208.
- Marczyk, G., DeMatteo, D., & Festinger, D. (2005). *Essentials of research design and methodology*. Hoboken, New Jersey: John Wiley & Sons.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.
- Mazzetti, A. (1999). The influence of segmental and prosodic deviations on source-text comprehension in simultaneous interpretation. *The Interpreters' Newslette*, 9, 125-147.
- Meuleman, C., & Van Besien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting*, 11(1), 20-34.

- Murphy, K. R., & Myers, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Onwuegbuzie, A. J., Slate, J. R., Leech, N. L., & Collins, K. M. T. (2007). Conducting mixed analyses: A general typology. *International Journal of Multiple Research Approaches*, 1(1), 4-17.
- Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, 64(6), 916-924.
- Pio, S. (2003). The relation between ST delivery rate and quality in simultaneous interpretation. *The Interpreters' Newsletter*, 12, 69-100.
- Pöchhacker, F. (2001). Quality assessment in conference and community interpreting. *Meta*, 46(2), 410-425.
- Pöchhacker, F. (2004). *Introducing Interpreting Studies*. Shanghai: Shanghai Foreign Language Education Press.
- Proffitt, L. (1997). *Simultaneous Interpretation of the Non-Native Speaker of English – Perceptions and Performance*, (Unpublished Master's thesis), University of London, UK.
- Sabatini, E. (2000/01). Listening comprehension, shadowing and simultaneous interpreting of two “non-standard” English speeches. *Interpreting*, 5(1), 25-48.
- Shlesinger, M. (2003). Effects of presentation rate on working memory in simultaneous interpreting. *The Interpreters' Newsletter*, 12, 37-50.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating qualitative and quantitative approaches in the social and behavioral sciences*. Thousand Oaks: Sage.
- Tommola, J., & Helevä, M. (1998). Language direction and source text complexity: Effects on trainee performance in simultaneous interpreting. In L. Bowker, M. Cronin, D. Kenny & J. Pearson (Eds.), *Unity in diversity? Current trends in translation studies* (pp. 177-186). Manchester: St. Jerome.
- Wang, L., & Lin, W. (2006). Interpretation training: SI with text. In M. J. Cai & A. L. Zhang (Eds.), *Professionalization in interpreting: International experience and developments in China* (pp. 237-244). Shanghai: Shanghai Foreign Language Education Press.

## **An introductory note to Chapter 6**

In Chapter 5, a convergent parallel mixed-methods study is conducted to ascertain the effects of speech rate and accent on the quality of English-to-Chinese simultaneous interpretation (SI) (i.e., the first part of **RQ 2**). It is found that although strong accent consistently exerted a negative influence across the three quality measures of SI (i.e., information completeness/InfoCom, fluency of delivery/FluDel, and target language quality/TLQual), fast speech rate actually had a positive impact on FluDel. This pattern is also supported by qualitative analysis of the interpreters' retrospective data. However, it is still unknown whether there are any impacts of speech rate and accent on strategy use, and whether there is any relationship between use of interpreting strategy and SI quality.

In the next chapter (i.e., Chapter 6), the remaining parts of **RQ 2**, namely RQ 2.2 and RQ 2.3 are thus explored. The two sub-RQs are: What are the effects of SI task characteristics (i.e., speech rate and accent) on strategy use (as a crucial part of interpreting ability)? and What is the relationship between strategy use and SI performance quality?

The investigation reported in Chapter 6 represents a preliminary exploration of RQ 2.2 and RQ 2.3, based on a further analysis of a subset of the data derived from the experiment in Chapter 5. Specifically, a contrastive analysis of source-language scripts and target-language renditions is first conducted to identify interpreting strategies used. Strategy use is then analyzed in relation to SI task characteristics (i.e., speech rate and accent), and SI quality.

## Chapter 6 Exploring the relationship between task characteristics, strategy use and performance in English-to-Chinese simultaneous interpreting<sup>33</sup>

**Abstract.** *Strategy use in simultaneous interpreting has long been a topic of interest for interpreting researchers. Empirical studies have been carried out to investigate the effect of source-language variables on strategy use, and the relationship between strategy use and interpreting performance. In the study, an experiment was conducted to explore strategy use in English-to-Chinese simultaneous interpreting, searching for possible patterns, examining the effect of source-language speech rate and accent on strategy use, and exploring the relationship between strategy use and overall performance quality. The preliminary results show that the interpreters developed a deep repertoire of interpreting strategies, but utilized strategies of syntactic transformation and substitution most frequently across the tasks. They also employed strategy clusters, a sequential combination of strategies, to cope with complex source-language segments. In addition, while the speech rate affected the use of the two prominent strategies (i.e., syntactic transformation and substitution) considerably, the accent did not produce the same effect. Moreover, the results show that overall the more strategies were used, the better the performance was. But this positive correlation did not hold across all the strategies. These results are explained and accounted for, and limitations of the study are discussed to inspire further research.*

### 6.1 Introduction

Interpreting strategies have long been a topic of interest for researchers (e.g., Bartłomiejczyk, 2006; Gile, 1995; Kohn & Kalina, 1996; Wang, 2012). The importance of strategy use in simultaneous interpreting (SI) cannot be emphasized more. For instance, strategy use is believed to be all-pervasive (Kohn & Kalina, 1996), and plays a vital role in interpreting (Donato, 2003). Kalina (2000) observes that the interpreting-specific use and interaction of

---

<sup>33</sup> A revised version of the chapter is under the 1<sup>st</sup> round of peer review in *the International Journal of Interpreter Education* as: Han, C. & Chen, S. J. (under review). The relationship between source-text characteristics, strategy use and performance of English-to-Chinese simultaneous interpreting: An exploratory study. Sijia Chen's contribution to the chapter was a coder who worked with the first author for one month on piloting an interpreting strategy categorization scheme, coded strategy use for eight interpreters' simultaneous interpreting performance in the operational coding, and proof-read the chapter.

strategies is typical of and crucial for the result of simultaneous interpreting (SI). Shlesinger (2000, p. 7) even claims that “a strategy which is used regularly by competent professionals tends to acquire normative force”.

Given the importance of strategies in SI, many researchers have investigated the impact of source speech characteristics on strategy use (e.g., Al-Salman & Al-Khanji, 2002; Bartłomiejczyk, 2006; Donato, 2003; Kim 2005; Liontou, 2011; Van Besien, 1999), and a few have looked into the relationship between strategy use and SI performance (e.g., Al-Salman & Al-Khanji, 2002; Meuleman & Van Besien, 2009). To enrich the literature and to gain further perspectives, the present study contributes some empirical data for exploring the effect of two source speech characteristics, namely speech rate and accent, on strategy use in English-to-Chinese SI. It also aims to examine the relationship between strategy use and SI performance quality. Speech rate and accent are chosen, primarily because in a previous survey on English/Chinese SI practice in China (see Chapter 4), fast speech rate and strong accent were cited as two frequently occurring factors underlying SI difficulty.

## **6.2 Literature review**

This section first provides an overview of the attempts made by interpreting researchers to define, identify and categorize various interpreting strategies, and then reviews the empirical studies that explore the effect of task characteristics on strategy use, and the relationship between strategy use and SI quality.

### **6.2.1 Defining, identifying and categorizing interpreting strategies**

In Interpreting Studies, a plethora of definitions have been proposed for interpreting strategies (e.g., Bartłomiejczyk, 2006; Chang, 2005; Gile, 1995; Kalina, 2000; Kirchhoff, 1976/2002; Riccardi, 2005; Wang, 2012). For example, Gile (1995) refers interpreting strategies or “coping tactics” as conscious solutions implemented by interpreters to solve problems during the interpreting process.

Based on the various definitions, two characteristics stand out. One is that strategy use is problem-directed or goal-oriented. That is, interpreting strategies are applied to address potential problems and to achieve certain goals. The other characteristic is that strategy use is

non-automatic and is consciously planned. What is worth noting here, however, is that automaticity and consciousness is perhaps more of a continuum than “either-or”.

In order to identify strategy use, two methods are used in the literature. One is called paralleled text analysis. It refers to identifying strategy use through segment-by-segment analysis of the source text (ST) and the target text (TT). The underlying rationale is that cognitive processes and strategy use in SI leave traces in interpreted texts, and therefore can be detected (Ivanova, 2000; Riccardi, 2005). Text analysis has been widely used by researchers to identify individual strategies of particular interest (e.g., Vandepitte, 2001; Petite, 2005) and to explore all possible strategies (e.g., Donato, 2003). The other method is called immediate retrospection (Ericsson & Simon, 1980, 1993). It requires interpreters to verbally report their thinking processes involved in SI immediately after performing the task. A number of studies have collected retrospective data to illuminate the interpreting process and strategy use during SI (e.g., Bartłomiejczyk, 2006; Chang, 2005; Vik-Tuovinen, 2002).

With a great variety of interpreting strategies identified, interpreting researchers have made efforts to categorize these strategies. For example, some strategies are used to address ST comprehension problems, others to solve difficulties associated with TT production, still others are known as emergency strategies that are deemed last-resort solutions. In general, interpreting researchers have been largely consistent in categorizing and defining each specific strategy, with only a few variations across different categorization schemes (see Bartłomiejczyk, 2006; Donato, 2003; Gile, 1995; Jones, 2002; Riccardi, 2005).

#### 6.2.2 The effect of SI task characteristics on strategy use

This section provides a literature review of empirical investigations into how SI task characteristics affect the use of interpreting strategies, with a special focus on the effect of speech rate and accent on strategy use in English-to-Chinese SI.

Most relevant to the present study is Setton (1999), Chang (2005), and Meuleman and Van Besien (2009). The former two studies involve strategy use in Chinese/English SI, the latter examines the effect of fast delivery rate on strategy use. Setton (1999) discussed strategy use in Chinese-to-English SI by looking at the differences between the two languages. For example, structural and syntactic differences (e.g., word order asymmetry) between Chinese and English

may create problems for SI. Given that Chinese is a left-branching language,<sup>34</sup> strategies suggested for Chinese-to-English SI include waiting for more information, stalling by using neutral fillers, chunking the content of the left-branching phrase as a constituent, and anticipating.

Using stimulated retrospective interview, Chang (2005) investigated strategy use as a function of interpreting directionality (between Chinese and English). It was found that apparent differences emerged when the interpreters translated into and from their A language. They tended to omit parts of the information when having difficulty expressing in their B language or resort to meaning-based strategies such as generalization.

Meuleman and Van Besien (2009) studied the effect of fast delivery speed on strategy use in French-to-Dutch SI. They found that to cope with fast speech rate the interpreters preferred lagging behind the source speech (i.e., “tailing”), and a few of them also segmented a source-language sentence into different parts (i.e., “segmentation”).

### 6.2.3 Relationship between strategy use and SI performance

Two empirical studies have examined the possible relationship between strategy use and interpreting performance in SI. Meuleman and Van Besien (2009) conducted an experiment to investigate whether strategy use has an enhancing effect on French-to-Dutch SI performance. It was found that the segmentation strategy was preferred by the interpreters working with a syntactically complex ST, and that the tailing strategy was most frequently used for a high delivery speed ST. When they examined strategy use together with the performance as measured by a 3-point acceptability scale, they found that 1) by employing the segmentation strategy, two thirds of the interpreters produced an acceptable rendition, 2) by utilizing the tailing strategy, almost all interpreters yielded an acceptable translation, while an absence of the strategy resulted in failure.

Al-Salman and Al-Khanji (2002) found that on average more achievement strategies (e.g., approximation, skipping and summarizing) were used in Arabic-to-English SI, and more reduction strategies (e.g., message abandonment, literal interpretation) were used in

---

<sup>34</sup> Left-branching structures are attached to the left of the constituent which governs them. That is, the structures occur in the speech string before the item they qualify or modify. For example, in Chinese a whole “participial” relative clause may precede and modify a noun, and may itself contain a left-branching phrase (Setton, 1999, p. 132).

English-to-Arabic SI.<sup>35</sup> In addition, the Arabic-native interpreters felt more comfortable in Arabic-to-English SI. It could be inferred that using achievement strategies more frequently would result in better interpretations, while frequent use of reduction strategies reduced performance quality.

### **6.3 Research questions**

Given the limited number of empirical studies, the study aims to seek answers to the following three questions:

- 1) What are the characteristics and patterns of strategy use in English-to-Chinese SI?
- 2) How do speech rate and accent affect strategy use in English-to-Chinese SI?
- 3) Is there a relationship between strategy use and SI performance quality?

### **6.4 Method**

#### **6.4.1 Participant recruitment**

A total of 32 Beijing-based English/Chinese interpreters were recruited to participate in the experiment. All of them had Mandarin Chinese as their L1, and English as their L2. In addition, a panel of nine trained raters evaluated all interpreters' performance, giving each performance three scores on information completeness (InfoCom), fluency of delivery (FluDel), and target language quality (TLQual) (for more details, please see Chapter 5 and Chapter 7).

#### **6.4.2 Experimental design**

In the experiment, there were two independent variables (IVs), namely speech rate and accent. The speed variable had two levels: a fast speech rate (FSR) of 155 words per minute (wpm), and a slow speech rate (SSR) of 105 wpm. The accent variable also had two levels: a native speech (NS) and an accented speech (AS). The two IVs were fully-crossed (i.e., a  $2 \times 2$  factorial design), producing four conditions: a slow and native speech (SN), a slow and accented speech

---

<sup>35</sup> According to Al-Khanji, El-Shiyab and Hussein (2000), achievement strategies refers to those with which speakers achieve a solution to communicative problem they face, while reduction strategies are those that are used to avoid a communicative problem and characterized by absence of an alternative plan to address the problem head-on.

(SA), a fast and native speech (FN), and a fast and accented speech (FA). A repeated-measures design was also implemented so that each interpreter performed SI in all conditions.

#### 6.4.3 SI tasks

Four SI tasks were carefully developed, namely Task<sub>SN</sub>, Task<sub>SA</sub>, Task<sub>FN</sub>, and Task<sub>FA</sub>. Each task included a tailor-made ST based on an authentic speech. All four STs centered on the general topic of Australia-China relationship, and were calibrated to be comparable with each other regarding length, lexical complexity, propositional density, syntactic structure, and text readability. For a detailed and complete description, please refer to Chapter 5.

To operationalize the accent variable, two English speakers were recruited: a native Australian English speaker and an Indian English speaker, each of whom recorded two source speeches. To operationalize the speed variable, each speaker recorded one fast speech (155wpm) and one slow speech (105wpm). For more details, please see Chapter 5. As a result, despite the similar text length across the four STs (i.e., approximately 1250 words), the duration of Task<sub>SN</sub> and Task<sub>SA</sub> was about 12 minutes, while that of Task<sub>FN</sub> and Task<sub>FA</sub> was about 8 minutes.

#### 6.4.4 Experiment procedure

One day prior to the experiment, all participants were provided with a three-page background reading material that covers the general topic of Australia-China relationship in the four STs, and they were asked to use the material only for preparation.

On the experiment day, there were four rounds of data collection for each participant. In each round, the participants performed SI in one of the tasks and completed other related activities. For more details, please see Chapter 5. Overall, the experiment took approximately three hours to complete, and SI performance was audio-recorded with consent. By the end of the experiment, each participant was compensated with 1000 RMB (about US\$ 170).

#### 6.4.5 Strategy coding

To prepare for strategy coding, all audio recordings were transcribed verbatim. In addition, a coding guideline was developed by the authors, based on Donato (2003), Gile (1995), Jones (2002), and Riccardi (2005). Strategy categories and their definitions are provided in Appendix L.

The strategy coding was performed by the two authors. Before operational coding, the two coders spent a total of about 50 hours within a month to discuss and practice coding until acceptable inter-coder agreement was reached. Specifically, the coders first discussed thoroughly each strategy category in order to establish a common conceptual ground. Then based on the coding guideline the coders worked together to code strategies, using random samples of paralleled STs and TTs. Next, the coders worked independently on random samples selected from different tasks. After this, inter-coder agreement index was calculated, and differences in coding were discussed. This process was repeated until the inter-coder agreement index plateaued and reached acceptable levels. For the four tasks, the final percent agreement indices ranged from 64% to 70%, meaning that 64% to 70% of the two coders' decisions were the same. For the operational coding, the texts were imported into NVivo 10 for qualitative analysis. The two coders worked independently to code strategy use.

#### 6.4.6 English glosses for the target language interpretations

English glosses are provided for the target language (TL) (i.e., Chinese) interpretations. Following Setton (1999), the glosses include a few TL particles and other language-specific features which have no apparent equivalent in English (see Table 6.1).

Table 6.1 Glosses for the Chinese interpretations

Chinese	Description	Gloss
呢	modal particle	<i>ne</i>
的	attributive resultative/adverbial particle	<i>de</i>
是	copula/affirmative predicate	<i>shi</i>
了	aspect particle	<i>le</i>
着	aspect particle	<i>zhe</i>
亿	numeral: one hundred million	<i>yi</i>
万	numeral: ten thousand	<i>wan</i>
呃, 嗯, 啊, 这个	filled pauses	♣

#### 6.4.6 Samples selected for analysis

Based on the overall performance scores, the interpretations from the top eight interpreters were chosen for analysis in the study. On an overall scale of one to eight, each of the interpreters scored five points or higher.

The choice of selecting part of the data for analysis was made, because the authors agreed that inferring strategy use from lower-quality interpretations was problematic and unreliable. In the pilot coding, the coders found that higher-performing interpreters usually rendered the majority of the source language (SL) segments correctly, thus leaving traceable evidence to infer strategy use. The higher-performing interpreters were also more likely to employ a wide range of strategies. But when working on lower-quality interpretations, the coders found it difficult to infer strategy use with sufficient confidence and in a consistent manner. For example, when lower-performing interpreters omitted something, the omissions were often important constituent parts of a sentence. It was therefore difficult to decide whether omissions of such type were strategically motivated or just because of incompetence. Categorizing these behaviors as use of strategic “omission” would over-inflate its importance. Consequently, a focused analysis of the higher-quality interpretations was believed to be more helpful in answering the research questions.

#### 6.4.7 Data analysis

Using NVivo 10, four types of nodes were created: interpreter, task, sentence in each task, and strategy category.<sup>36</sup> In each node, sub-nodes were also created. Coded text segments were stored in relevant nodes and sub-nodes. For example, all coded text segments concerning a given interpreter were stored in that interpreter’s sub-node. Consequently, using NVivo 10’s matrix coding query, quantitative data were generated, including frequency count for each node and sub-node, and cross-tabulation of the nodes and the sub-nodes. Qualitative data were also produced. That is, codings for a given interpreter in a given task were generated for each of the strategy categories.

---

<sup>36</sup> In NVivo 10, a node is a collection of references about a specific theme, place, person or other area of interest. Nodes make it possible to code one’s materials or sources, and thereby organize them.

## 6.5 Results

In this section, illustrated examples for each strategy category are provided first so that readers could gain a concrete understanding of each strategy in Appendix L. Then, characteristics and patterns of strategy use are identified, followed by the analysis of the effect of speech rate and accent on strategy use. Finally, the relationship between strategy use and SI performance quality is explored.

### 6.5.1 Strategy use: Illustrated examples

A typical example includes an SL segment (usually an SL sentence), its corresponding TL interpretations, and an English gloss for the TL renditions. The source of the example can also be identified. For example, the content in Example 1 is based on Participant 16's interpretations for Sentence 68 in Task<sub>SN</sub>, thus SN-S68-P16.

*Stalling by using neutral material* In Example 1, the neutral materials were included in the curly bracket { }.

#### Example 1: SN-S68-P16

**SL:** The Global Financial Crisis showed that we were so close to collective disaster, ...

**TL:** 呃，金融危机呢也告诉我们，{现在实际上呢}，对于，我们对于共同的这个，呃，灾难是如此的接近，.....

**Gloss:** ♣, financial crisis *ne* also inform us, {at present as a matter of fact *ne*}, regarding, we regarding common ♣, ♣, disaster is so close, ...

*Syntactic transformation* In Example 2, two segments in the SL sentence was underlined and numbered with superscripts (1 → 2). As can be seen in the TL renditions, the original syntactic order was re-structured to form the TL sentence (2 → 1).

#### Example 2: FA-S05-P01

**SL:** As you know, it has been forty years<sup>1</sup> since Australia and China established diplomatic relations<sup>2</sup>.

**TL:** 那么大家知道呢，中澳之间建立外交关系<sup>2</sup>已经有 40 年的历史了<sup>1</sup>。

**Gloss:** So all of us know *ne*, China and Australia establish foreign diplomatic relations<sup>2</sup> already

have 40 years' history *le*<sup>1</sup>.

*Syntactic segmentation* In Example 3, the parallel || was used to indicate where the SL sentence was segmented into two independent TL segments.

Example 3: SA-S44-P09

**SL:** A productive relationship with China, based on mutual interest and mutual respect, is in Australia's national interest.

**TL:** 我们和中国之间富有成果的合作伙伴关系是建立在互利共赢的基础之上的。|| 这是符合澳大利亚的根本利益的。

**Gloss:** Our and China's productive cooperative partnership is built on mutual interest and mutual benefit *de* basis above *de*. || This is serve Australia's fundamental interest.

*Changing the order of elements* In Example 4, the SL adjectives were underlined and numbered with superscripts (1 → 2 → 3). The order was reversed in the TL interpretations (3 → 2 → 1).

Example 4: SA-S11-P01

**SL:** The Government is committed to building a mature<sup>1</sup>, balanced<sup>2</sup> and sustainable<sup>3</sup> relationship with China.

**TL:** 澳洲政府是承诺要建立一个可持续性的<sup>3</sup> 和平衡的<sup>2</sup> 成熟的<sup>1</sup> 关系。

**Gloss:** Australian government promise establish a sustainable<sup>3</sup> and balanced<sup>2</sup> mature<sup>1</sup> relationship.

*Generalization* In Example 5, the SL parts that were generalized in TL were placed in the round brackets.

Example 5: SN-S56-P14

**SL:** I am glad to announce that Monash University, in collaboration with (Southeast University), is establishing (a joint research institute) in the Suzhou Industrial Park.

**TL:** 我非常高兴的宣布, Monash University 和 (中国的一个大学合作), 在苏州工业园区开展 (一个项目)。

**Gloss:** I very glad announce, Monash University and (China's one university) cooperate, in Suzhou Industrial Park conduct (a project).

*Simplification* In Example 6, the two interpreters did not follow the original structure, but performed stylistic simplifications for the SL sentence.

Example 6: FA-S15-P06/P14

**SL:** So, China matters greatly to Australia, but Australia also matters greatly to China.

**TL<sub>1</sub>:** 呃，两国相互的重要性是非常高的。(P06)

**Gloss<sub>1</sub>:** ♣, two countries mutual importance is very high.

**TL<sub>2</sub>:** 所以，中国对澳大利亚非常重要，反过来也是一样的。(P14)

**Gloss<sub>2</sub>:** Therefore, China to Australia very important, vice versa.

*Omission* In Example 7, the underlined SL text was strategically omitted. The omission did not affect the transfer of information, but could ease the interpreter's cognitive load to better deal with the figure that follows.

Example 7: SN-S18-P09

**SL:** At present, Australia has seventy two resource projects at an advanced stage of development, the total value of which is worth about one hundred and thirty billion dollars.

**TL:** 现在澳大利亚已经有七十二个项，资源的项目。这些项目的总价值大约是，一千三百亿澳元。

**Gloss:** At present Australia already have seventy two pro..., resource project. These project(s) *de* total value approximately is, one thousand three hundred *yi* Australian dollar(s).

*Explanatory additions* In Example 8, the double-underlined TL segment could not find its SL equivalents, but represented an addition to further explain the previous utterance.

Example 8: FN-S18-P06

**SL:** The rise of China and Australia simultaneously is no coincidence.

**TL:** 澳大利亚和中国的经济发展并不是偶然。中国的经济发展，嗯，这是，与澳大利亚紧密相关的。

**Gloss:** Australia's and China's economic development is no coincidence. China's economic development, ♣, this is, Australia closely related.

*Addition to maintain coherence* In Example 9, the wavy line under the TL segment represented an addition to maintain textual coherence.

Example 9: SN-S05-P14

**SL:** China's development into an important nation in the twenty-first century is absolutely an extraordinary achievement.

**TL:** 中国发展已经成为二十一世纪的一个重要国家。这样的一个变化呢本身就是一个巨大的成绩。

**Gloss:** China develop already become twenty first century *de* an important country. Such a transformation *ne* itself is a huge achievement.

*Repetition* In Example 10, the underlined TL segment was a semantic repetition of the previous utterance, but enhanced lexical and semantic accuracy.

Example 10: SA-S45-P14

**SL:** It is hard to think of a single international issue of importance to Australia where China is not a key player on the world stage.

**TL:** 可以说，任何重大的国际事务，额，都要求中国，澳大利亚的共同的参与，对澳大利亚影响的这些国家事务呢都离不开中国的贡献和参与。

**Gloss:** Can say, any important international issue, ♣, require China, Australia *de* common participation, regarding Australia affect *de* these national issue *ne* all leave not China's contribution and participation.

*Paraphrase* In Example 11, the underlined SL segment “more growth potential” was not rendered directly. Instead, the interpreter paraphrased the segment, as can be seen in the underlined TL segment.

Example 11: FN-S49-P11

**TL:** As I said earlier, a large number of Chinese students have been enrolled in Australian courses, and more growth potential is expected in the future.

**GL:** 现在有很多的中国学生在澳大利亚读书，而且呢我认为越来越多的中国学生将会选择，

呃，选择澳大利亚。

**Gloss:** At present have many Chinese students in Australia study, and *ne* I think more and more Chinese would choose, ♣, choose Australia.

*Substitution* In Example 12, the square-bracketed TL content seemed to be related to the SL sentence. But it was actually a mis-representation of the original information. This is probably because the interpreter did not comprehend the SL sentence, but produced something contextually plausible to cover his/her otherwise silence.

Example 12: SN-S02-P01

**SL:** I am here as the leader of an economically confident nation whose current and future prosperity is connected with China's.

**TL:** [我今天来呢其实呢是来谈一下澳大利亚和中国之间的经济关系。]

**Gloss:** [I today come *ne* in fact *ne shi* come discuss Australia and China *de* economic relation.]

*Reproduction* In Example 13, the interpreter probably did not know exactly the Chinese translation of the Prime Minister, and had to reproduce the name in English, as shown by the underlined TL content.

Example 13: FN-S09-P11

**SL:** The following year, Gough Whitlam, as an Australian Prime Minister, made the first official visit to China.

**TL:** 在第二年，Whitlam作为澳大利亚的首相就第一次访华。

**Gloss:** In the second year, Whitlam as Australia's Prime Minister the first time visit China.

*Repair* In Example 14, the TL content in the angel bracket was added to repair the incorrect renditions uttered previously.

Example 14: FN-S52-P13

**SL:** Prior to nineteen seventy two, only a small number of Australians had visited China, but today almost half a million visit annually, which is predicted to grow in years to come.

**TL:** 那么，92 年的时候只有少数的中国人来澳大利亚，现在已经超过了 50 万，未来还会

增长，啊，<刚才说的呢是澳大利亚到中国的出访的人数>。

**Gloss:** So, nine two year *de* time only have limited Chinese come to Australia, at present already surpass fifty *wan*, future would increase, ♣, <just say *de ne* is Australia go to China *de* visit *de* person number>.

*Transcoding* In Example 15, based on the speech context, the underlined SL phrase means “a stake or involvement in an undertaking”. Instead, the interpreter translated the phrase literally into a TL term that means “the feeling of wanting to know or learn about something or someone”.

Example 15: SA-S60-P01

**SL:** It also means that China has an interest in the stability of world markets and how the markets function.

**TL:** 同时它也意味着，中国呢，对于全球市场的稳定，以及市场的正常运行，呃，有自己的很大的这样一个兴趣。

**Gloss:** Meanwhile it also means *zhe*, China *ne*, regarding global market *de* stability, and market *de* normal function, ♣, have itself *de* very big this an interest.

## 6.5.2 Characteristics of strategy use

### 6.5.2.1 Patterns of strategy use

Table 6.2 summarizes the descriptive statistics for strategy use. Based on the table, four major observations can be made on the pattern of strategy use. First, a total of 1998 strategies were identified for the performance by the eight interpreters in the four tasks. Second, the total number of strategies used in each task decreased from Task<sub>SN</sub> (n = 588) to Task<sub>FA</sub> (n = 420). This pattern may have to do with the characteristics of the tasks. Third, syntactic transformation and substitution figured prominently, accounting for 37.5% and 30.4% of all strategies used. The interpreters also used the following strategies frequently: syntactic segmentation (4.9%), generalization (5%), and repair (5.3%). Fourth, while the interpreters used a decreasing number of syntactic transformation from Task<sub>SN</sub> (n = 230) to Task<sub>FA</sub> (n = 134), they resorted to substitution more frequently from Task<sub>SN</sub> (n = 126) to Task<sub>FA</sub> (n = 195).

Table 6.2 Frequency count of strategy use in the interpreting tasks

Task / Strategy	Task <sub>SN</sub>	Task <sub>SA</sub>	Task <sub>FN</sub>	Task <sub>FA</sub>	Total (%)
Stalling	8	1	11	4	24 (1.2%)
Syntactic transformation	230	208	178	134	<b>750 (37.5%)</b>
Syntactic segmentation	44	31	11	11	<b>97 (4.9%)</b>
Changing the order of phrases	17	17	17	5	56 (2.8%)
Generalization	27	28	18	26	<b>99 (5.0%)</b>
Simplification	19	11	16	9	55 (2.8%)
Omission	16	6	2	3	27 (1.4%)
Explanatory addition	20	8	7	10	45 (2.3%)
Addition to maintain coherence	25	25	20	4	74 (3.7%)
Repetition	11	7	2	1	21 (1.1%)
Paraphrase	4	1	7	3	15 (0.8%)
Substitution	126	131	155	195	<b>607 (30.4%)</b>
Reproduction	4	6	6	2	18 (0.9%)
Repair	35	38	20	12	<b>105 (5.3%)</b>
Transcoding	2	2	0	1	5 (0.3%)
<b>Total (%)</b>	588 (29.4%)	520 (26.0%)	470 (23.5%)	420 (21.0%)	1998 (100.0%)

### 6.5.2.2 Strategy clusters

For an interpreting strategy to be effective in coping with cognitively taxing SL segments, it is sometimes combined with additional strategies in sequence, thus forming strategy clusters. Based on the corpus, there seemed to be two types of strategy cluster. One type concerns the use of one strategy (particularly syntactic transformation) multiple times within a SL segment. Example 17 shows how syntactic transformation was utilized twice to re-organize the syntactic structure of an SL segment. As can be seen, the interpreter (i.e., P01) reversed the syntactic order of 1 → 2 and 3 → 4 → 5 to 2 → 1 and 5 → 4 → 3, respectively.

Example 17: SN-S16-P01

**SL:** Chinese investment is welcomed<sup>1</sup> in Australia<sup>2</sup>, as is shown by the steady stream of proposals<sup>3</sup> already approved by<sup>4</sup> our Foreign Investment Review Board<sup>5</sup>.

**TL:** 中国的投资在澳大利亚<sup>2</sup> 很受欢迎<sup>1</sup>, 而这也体现在已经由我们的外资审核委员会<sup>5</sup> 所

批准<sup>4</sup>的这样一些投资的项目<sup>3</sup>。

**Gloss:** Chinese investment in Australia<sup>2</sup> very much welcomed<sup>1</sup>, and this is also shown by our Foreign Investment Review Board<sup>5</sup> approved<sup>4</sup> *de* these some investment projects<sup>3</sup>.

The other type concerns the use of different strategies within an SL segment. Example 18 shows that the interpreter (P01) used a chain of strategies within an SL sentence: syntactic transformation → addition to maintain coherence → syntactic segmentation → addition to maintain coherence → syntactic transformation. That is, the interpreter first reversed the order of 1 → 2 to 2 → 1, and added “providing to” to maintain coherence. Then s/he started a new TL sentence, and inserted “So this is ...” to refer back to the previous TL sentence. Finally, the interpreter reformulated once again the syntactic order of 3 → 4 to 4 → 3.

Example 18: SN-S08-P01

**TL:** Australia’s role as a stable, reliable and high quality supplier<sup>1</sup> of energy and mineral resources<sup>2</sup> to China is the bedrock<sup>3</sup> of a comprehensive economic partnership<sup>4</sup>.

**SL:** 我们是一个可靠的, 高质量的能源和矿产资源<sup>2</sup>的提供方<sup>1</sup>, 提供给中国。|| 那么这是我们全面经济关系<sup>4</sup>的基础<sup>3</sup>。

**Gloss:** We are a reliable, high-quality energy and mineral sources<sup>2</sup> *de* provider<sup>1</sup>, providing to China. || So this is our comprehensive economic relationship<sup>4</sup> *de* basis<sup>3</sup>.

Example 19 also shows how the interpreter (i.e., P13) strategically approached an SL segment by using a chain of strategies. As can be seen, in the beginning s/he stalled by using a neutral utterance “We can see”. The strategy of “stalling” was used probably because the interpreter tried to avoid an awkward silence while waiting for more information. S/he began her/his interpretation by reversing the syntactic order from 1 → 2 to 2 → 1, and quickly repaired her/his first rendition of the proper name (i.e., National Development and Reform Commission) when s/he realized the rendition was inaccurate. However, it seems that s/he could not remember exactly “Vice-chairman” because too much cognitive effort may have been spent on the repair. As a result, a superordinate term “leader” was opted for. For the remaining part of the sentence, the interpreter transformed the syntactic structure from 4 → 3 to 3 → 4, and replaced “Ministerial Climate Change Dialogue” with a general term “bilateral dialogue”.

Example 19: SN-S41-P13

**SL:** We recently welcomed the Vice-Chairman<sup>1</sup> of the National Development and Reform Commission<sup>2</sup> to take part in our Ministerial Climate Change Dialogue<sup>3</sup> in Australia<sup>4</sup>.

**SL:** {我们可以看到了}, 这个, 国家发展委, <发改委><sup>2</sup> 的, 嗯, (领导人)<sup>1</sup> 参加了我们在澳大利亚<sup>4</sup> 进行的(双边对话)<sup>3</sup>。

**Gloss:** We can see, ♣, National Development Commission, <Development and Reform Commission><sup>2</sup> *de*, ♣, (leader)<sup>1</sup> participate in our in Australia<sup>4</sup> on-going (bilateral dialogue)<sup>3</sup>.

### 6.5.3 Effect of speech rate and accent on strategy use

Table 6.3 compares the strategy use between the fast speech rate (FSR) and the slow speech rate (SSR) conditions. As can be seen, the comparison was only made for those strategies whose total frequency count equals or exceeds 50 ( $n \geq 50$ ). In addition, for a given strategy, only when the absolute difference equals or exceeds 15% ( $n_{\%} \geq 15\%$ ), can the strategy use be said to differ substantially between the FSR and SSR conditions. Consequently, six strategies qualified for the analysis, as shown in Table 6.3.

Table 6.3 A comparison of strategy use between the fast and the slow conditions

Strategy	Total No.	Speech rate		Absolute % difference
		FSR	SSR	
Syntactic transformation	750	312 (41.6%)	438 (58.4%)	16.8%
Syntactic segmentation	97	22 (22.7%)	75 (77.3%)	54.6%
Changing the order of phrases	56	22 (39.3%)	34 (60.7%)	21.4%
Addition to maintain coherence	74	24 (32.4%)	50 (67.6%)	35.1%
Substitution	607	350 (57.7%)	257 (42.3%)	15.3%
Repair	105	32 (30.5%)	73 (69.5%)	39.0%

Note: To calculate the percentage in the parenthesis, the frequency count for syntactic transformation under the FSR conditions (i.e., 312) was divided by the total count (i.e., 750).

It turns out that syntactic transformation, syntactic segmentation, changing the order of phrases, addition to maintain coherence and repair were used more frequently in the SSR than the FSR condition. However, much more substitution was employed in the FSR than the SSR condition. In addition, the largest percentage difference was for syntactic segmentation (54.6%),

indicating that the SSR and the FSR conditions differed most substantially regarding the use of that strategy.

Table 6.4 compares the strategy use between the native speech (NS) and the accented speech (AS) conditions. The same analysis criteria ( $n \geq 50$  and  $n\% \geq 15\%$ ) were applied. As shown in the table, the strategies of changing the order of phrases, simplification and addition to maintain coherence were used much more in the NS than the AS conditions.

Statistics for two strategies, namely syntactic transformation and segmentation, were also presented in Table 6.4. As can be seen, by sheer number they remained prominent strategies in both conditions, but they were applied similarly across the conditions.

Table 6.4 A comparison of strategy use between the native and the accented speech conditions

Strategy	Total No.	Accent		Absolute % difference
		NS	AS	
Changing of the order of phrases	56	34 (60.7%)	22 (39.3%)	21.4%
Simplification	55	35 (63.6%)	20 (36.4%)	27.3%
Addition to maintain coherence	74	45 (60.8%)	29 (39.2%)	21.6%
Syntactic transformation	750	408 (54.4%)	342 (45.6%)	8.8%
Substitution	607	281 (46.3%)	326 (53.7%)	7.4%

#### 6.5.4 Relationship between strategy use and SI performance

To explore the relationship between strategy use and SI performance, three participants were selected from the eight interpreters to represent high- (i.e., P01), medium- (i.e., P13), and low-performance (i.e., P14) interpreters, based on their overall performance scores. Table 6.5 presents a profile of strategy use for each interpreter. As can be seen, a total of 293, 240 and 187 strategies were employed in high, medium, and low performance, respectively. There seems to be a positive correlation between strategy use and performance quality. In other words, the number of strategies used was positively related to performance quality. Or it could also be interpreted as: better-skilled interpreters use more strategies than less-skilled ones.

However, the positive relationship does not seem to hold across the strategies. For example, the performance quality seemed to improve, with the increasing use of the first seven strategies in Table 6.5, from syntactic transformation to omission. It would appear that the overall quality

impaired when more substitution was employed. In addition, the performance was not clearly related with the last seven strategies, from stalling to simplification.

Table 6.5 A comparison of strategy use between the three interpreters

<b>Participant / Strategy</b>	<b>P01</b> (High)	<b>P13</b> (Medium)	<b>P14</b> (Low)
Total No.	293	240	187
<b>Syntactic transformation</b>	<b>140 (47.8%)</b>	<b>81 (33.8%)</b>	<b>71 (38.0%)</b>
Syntactic segmentation	22 (7.5%)	14 (5.8%)	4 (2.1%)
Changing of the order of phrases	17 (5.8%)	6 (2.5%)	1 (0.5%)
Explanatory addition	9 (3.1%)	5 (2.1%)	2 (1.1%)
Addition to maintain coherence	17 (5.8%)	7 (2.9%)	4 (2.1%)
Repair	11 (3.8%)	6 (2.5%)	3 (1.6%)
Omission	10 (3.4%)	5 (2.1%)	1 (0.5%)
<b>Substitution</b>	<b>37 (12.6%)</b>	<b>79 (32.9%)</b>	<b>84 (44.9%)</b>
Stalling	1 (0.3%)	1 (0.4%)	0 (0.0%)
Generalization	13 (4.4%)	15 (6.3%)	10 (5.3%)
Repetition	1 (0.3%)	3 (1.3%)	0 (0.0%)
Paraphrase	1 (0.3%)	2 (0.8%)	3 (1.6%)
Reproduction	5 (1.7%)	2 (0.8%)	1 (0.5%)
Transcoding	2 (0.7%)	0 (0.0%)	0 (0.0%)
Simplification	7 (2.4%)	14 (5.8%)	3 (1.6%)

Particularly, for the two prominent strategies of syntactic transformation and substitution, there seems to be a countervailing effect between them. For example, in the strategy profile for the higher performance, nearly half of the strategies used were syntactic transformation (47.8%), overshadowing the use of substitution (12.6%). For the medium performance, the proportion of syntactic transformation decreased to 33.8%, almost equal to that of substitution (32.9%). For the lower performance, the use of substitution accounted for 44.9% of the strategies, surpassing the use of syntactic transformation (38.0%). Therefore, the performance quality seems to be related with the relationship between syntactic transformation and substitution.

## 6.6 Discussion

### 6.6.1 Characteristics of strategy use

The results show that syntactic transformation (37.5%) and substitution (30.4%) accounted for a large proportion in the strategy use profile. These large percentages may reveal the dominant strategic behaviors present in English-to-Chinese SI. But the large percentages could also be inflated by the method used to identify strategy use.

The prominent position of syntactic transformation may reflect an important aspect of the strategic decision-making process in English-to-Chinese SI. Working in the SI mode (i.e., sequential processing), interpreters need to handle structural asymmetry between English and Chinese. One method is to transform the syntactic order (syntactic transformation) to form left-branching Chinese sentences. An alternative is to segment original information into individual meaning units, thus preserving linearity (chunking or syntactic segmentation) (Zhong, 1984; Zhuang, 1991). As has been illustrated in Example 19, while syntactic transformation helped P13 produce well-formed Chinese sentences, s/he had to wait for the left-branching structures that are attached to the left of a constituent, thus increasing cognitive load. On the other hand, syntactic segmentation helped P03 ease cognitive load by immediately recasting the content of left-branching phrases as a constituent. The strategy isolated the compact sentence into separate units, weakening the integrity of the original message. Based on the results, the interpreters preferred syntactic transformation in handling syntactic differences.

The frequent use of substitution indicates that overall SI is a cognitively taxing activity, even for the experienced interpreters. Regarded as an emergency strategy (Donato, 2003; Riccardi, 2005), substitution is used when interpreters encounter comprehension failures, and have to produce a contextually plausible TL segment in order to maintain the smooth flow of SI. As can be seen in Table 6.2, when the tasks are characterized by strong accent (Task<sub>SA</sub>), fast speech rate (Task<sub>FN</sub>) or both (Task<sub>FA</sub>), they become progressively challenging compared to the baseline Task<sub>SN</sub>. Consequently, the number of the substitution strategy used increased from Task<sub>SN</sub> to Task<sub>FA</sub>.

However, the method used to identify strategy use (i.e., paralleled text analysis) may contribute to the prominent status of syntactic transformation and substitution. Since an SL-TL contrastive study immediately reveals syntactic and semantic differences, the coders found it easier to infer strategic behaviors such as syntactic transformation and substitution. However, other strategies are difficult to infer. For example, an interpreter may omit an SL segment

strategically to avoid redundancy, or merely because of incompetence. While strategic omissions should be a form of strategy use, other types of omission represent a performance error (Napier, 2004). As a result the coders had to take into account the immediate textual context to make an educated inference, which inevitably complicates strategy coding.

In addition, the identification of effective strategy clusters suggests that in handling complex SL segments interpreters should proactively employ a combination of strategies. More importantly, as demonstrated in Example 19 that P13 and P03 produced adequate interpretations using different sets of strategy cluster, interpreters are encouraged to explore an optimal solution for themselves, based on their cognitive capacity and interpreting style. Riccardi (1998) observes that interpreting performance will be creative whenever interpreters succeed in employing a combination of strategies in a flexible way.

#### 6.6.2 The effect of speech rate and accent on strategy use

For the effect of speech rate, while the strategies of syntactic transformation, syntactic segmentation, changing the order of phrases, addition to maintain coherence and repair were used much less frequently (by at least 15%) in the FSR than the SSR conditions, the substitution strategy was utilized much more in the FSR conditions. This pattern of strategy use could be attributed to the fact that SI is an externally paced activity. Particularly, in a FSR condition, the primary challenge facing interpreters is to keep up with speakers. The interpreters must produce TL renditions as quickly as possible to empty working memory for new information. It leaves little time for the interpreters to re-think, re-plan and re-structure. Consequently, strategies that require extra reformulation (such as syntactic transformation, changing the order of phrases and repair), and that tend to expand the length of TL productions (such as syntactic segmentation and addition to maintain coherence) would be used less frequently. For example, using syntactic segmentation in English-to-Chinese SI usually entails addition of connective phrases to link separate meaning units, as shown in Example 19 (P03). Additionally, given an SL text, the increase of delivery speed leads to higher information density per unit time. As a result of the potentially heavy cognitive load, the interpreters may encounter comprehension failures. In order to avoid awkward silence, they may resort to the strategy of substitution, producing seemingly plausible renditions based on a few words they have heard. This could explain why substitution was used more frequently in the FSR conditions.

For the effect of accent, the study shows that overall the strategies of changing the order of phrases, simplification and addition to maintain coherence were used more frequently in the NS than the AS conditions. When interpreting for an accented speech, interpreters may encounter comprehension difficulties because of the unfamiliar accent. Presumably, they would use such strategies as tailing or lagging behind to help comprehension, and substitution to maintain fluent production. Although the results show that substitution was used more often in the AS than the NS conditions, the difference was not substantial (less than 15%). In addition, other strategies such as tailing cannot be identified based purely on the text analysis.

#### 6.6.3 Relationship between strategy use and SI performance quality

The results seem to echo the previous findings that if successfully implemented certain strategies have the potential to enhance SI performance quality (Al-Salman & Al-Khanji, 2002; Meuleman & Van Besien, 2009). In addition, the distinction made between achievement and reduction strategies (Al-Khanji, El-Shiyab, & Hussein, 2000) indicates that not all strategies play an equal role in contributing to higher SI quality. For instance, the use of the first seven strategies in Table 6.5 seems to have a positive relationship with SI quality. The nature of these strategies is either to reformulate syntactic structure, or to elucidate TL renditions. If properly implemented, they would help interpreters retain SL information in TL interpretations. However, as an emergency strategy, substitution is drawn upon primarily to maintain the interpretation flow and to avoid embarrassing moments of silence. In other words, the use of substitution is not intended to enhance SI quality, particularly information completeness.

As for the counteracting effect of substitution on syntactic transformation, it probably could be explained by their different roles in SI. In English-to-Chinese SI, syntactic transformation is a strategy that is constantly needed due to syntactic asymmetry. As discussed above, successful implementation of syntactic transformation implies two things: 1) structural differences have been addressed, and 2) left-branching structures and their associated main constituent have been reproduced in TL. Consequently, the more syntactic transformation is used, the more SL content is adequately translated, contributing to better overall quality, particularly higher degree of information completeness. However, when substitution is used, SL content is not rendered accurately in SL. In other words, the use of substitution produces semantically incorrect TL segments that seem to be plausible in a given context. As a result, larger

percentage of syntactic transformation was present in the higher performance, while greater proportion of substitution was in the lower performance.

## **6.7 Limitations**

The study has two limitations. One limitation has to do with the method used to identify strategy use. Not all strategies used can be accurately and reliably identified based on comparative SL-TL text analysis. This is because 1) the method can only reveal those strategies that leave detectable traces in interpreted texts, and 2) the coders have to rely on their experience to infer strategy use through the SL-TL analysis, which could be unreliable. As a result, the reported strategies at best represent only a part of the interpreters' strategic behaviors. The other limitation is the purposive selection of a small sample size. As an exploratory study, only eight relatively high-performing interpreters were selected for the analysis. Consequently, the findings, especially the observed relationships, are not generalizable.

## **6.8 Conclusion**

Based on the eight high-performing interpreters' performance in the four tasks, the study explores strategy use in English-to-Chinese SI, focusing on patterns of strategy use, the effect of speech rate and accent on strategy use, and the relationship between strategy use and SI performance quality. Preliminary results show that the interpreters utilized a variety of interpreting strategies, but resorted to syntactic transformation and substitution most frequently, and employed strategy clusters to cope with complex SL segments. In addition, SL speech rate substantially affected how syntactic transformation and substitution were used in English-to-Chinese SI, but accent did not produce the same effect. Furthermore, the results show that the more strategies were used, the better performance. But the relationship did not hold across all strategies. Particularly, the relationship was reversed for the strategy of substitution. Given the exploratory nature of the study, future research could be conducted to ascertain the preliminary findings, using enhanced methods to identify strategies and larger high-quality samples for the purpose of generalizability.

## 6.9 References

- Al-Khanji, R., El-Shiyab, S., & Hussein, R. (2000). On the use of compensatory strategies in simultaneous interpretation. *Meta*, 45(3), 548-557.
- Al-Salman, S., & Al-Khanji, R. (2002). The native language factor in simultaneous interpretation in an Arabic/English context, *Meta*, 47(4), 607-626.
- Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting*, 8(2), 149-174.
- Chang, C. C. (2005). *Directionality in Chinese/English simultaneous interpreting: Impact on performance and strategy use* (Doctoral thesis, University of Texas at Austin, USA). Retrieved from <https://www.lib.utexas.edu/etd/d/2005/changc71804/changc71804.pdf>
- Donato, V. (2003). Strategies adopted by student interpreters in SI: A comparison between the English-Italian and the German-Italian language-pairs. *The Interpreters' Newsletter*, 12, 101-134.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215-251.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.
- Ivanova, A. (2000). The use of retrospection in research on simultaneous interpreting. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and Mapping the Processes of Translation and Interpreting: Outlooks on empirical research* (pp. 27-52), Amsterdam/Philadelphia: John Benjamins.
- Jones, R. (1998). *Conference Interpreting Explained*. (2nd ed.). Manchester: St Jerome Publishing.
- Kalina, S. (2000). Interpreting competences as a basis and a goal for teaching, *The Interpreters' Newsletter*, 10, 3-32.
- Kim, H. R. (2005). Linguistic characteristics and interpretation strategy based on EVS analysis of Korean-Chinese, Korean-Japanese interpretation, *Meta*, 50(4), 1492-1421.
- Kirchhoff, H. (1976/2002). Simultaneous interpreting: Interdependence of variables in the interpreting process, interpreting models and interpreting strategies. In F. Pöchhacker & M. Shlesinger (Eds.), *Interpreting Studies Reader* (pp. 110-119). London: Routledge.

- Kohn, K., & Kalina, S. (1996). The strategic dimension of interpreting. *Meta*, 41(1), 118-138.
- Liontou, K. (2011). Strategies in German-to-Greek simultaneous interpreting: A corpus-based approach. *Gramma*, 19, 37-56.
- Meuleman, C., & Van Besien, F. (2009). Coping with extreme speech conditions in simultaneous interpreting. *Interpreting*, 11(1), 20-34.
- Napier, J. (2004). Interpreting omissions: A new perspective. *Interpreting*, 6(2), 117-142.
- Petite, C. (2005). Evidence of repair mechanisms in simultaneous interpreting. *Interpreting*, 7(1), 27-49.
- Riccardi, A. (1998). Interpreting strategies and creativity. In A. Beylard-Ozeroff, J. Králová, B. Moser-Mercer (Eds.), *Translators' Strategies and Creativity* (pp. 171-179). Amsterdam/Philadelphia: John Benjamins.
- Riccardi, A. (2005). On the evolution of interpreting strategies in simultaneous interpreting. *Meta*, 50(2), 753-767.
- Setton, R. (1999). *Simultaneous Interpretation: A Cognitive and Pragmatic Analysis*. Amsterdam and Philadelphia: John Benjamins.
- Shlesinger, M. (2000). Interpreting as a cognitive process. In S. Tirkkonen-Condit & R. Jääskeläinen (Eds.), *Tapping and mapping the processes of Translation and Interpreting: Outlooks on empirical research* (pp. 3-15). Amsterdam/Philadelphia: John Benjamins.
- Van Besien, F. (1999). Anticipation in simultaneous interpretation. *Meta*, 44(2), 250-259.
- Vandepitte, S. (2001). Anticipation in conference interpreting: A cognitive process. *Revista Alicantina de Estudios Ingleses*, 14, 323-335.
- Vik-Tuovinen, G. V. (2002). Retrospection as a method of studying the process of simultaneous interpreting. In G. Garzone & M. Viezzi (Eds.), *Interpreting in the 21st Century: Challenges and opportunities* (pp. 63-71). Amsterdam: John Benjamins.
- Wang, B. H. (2012). Interpreting strategies in real-life interpreting – Corpus-based description of seven professional interpreters' performance. *Translation Journal*, 16(2). Retrieved from <http://translationjournal.net/journal/60interpreting.htm>
- Zhuang, M. L. (1991). 汉英同声传译的技巧. [Techniques in Chinese-English simultaneous interpreting]. 中国翻译, 2, 24-27.
- Zhong, S. K. (1984). 实用口译手册. [A Practical Handbook of Interpretation]. Beijing: China Foreign Language Translation Publishing Corporation.

## **An introductory note to Chapter 7**

As has been pointed out in Chapter 2, apart from the explanation inference, the other weakest link in the proposed ICPT validity argument is the generalization inference. To enhance this inference, one piece of validity evidence, among others, should show that ICPT scores are generalizable across raters, tasks and other interested assessment facets. In particular, given that ICPTs are typically rater-mediated performance tests, and that rater variability figures prominently in performance assessment, achieving a desirable level of rater consistency constitutes an important task for ICPT testers. The attempt to reduce rater variability to obtain reliable scores necessitates an in-depth understanding of rater behavior in an operational rating context.

Given that nine raters are employed to assess and evaluate a large amount of audio-recorded interpreting performance in the experiment (see Chapter 5), and also that to a large extent, experimental conclusions are made on the basis of the rater-generated scores, the quality of these scores need to be investigated, and raters' rating behavior and pattern needs to be carefully examined.

Chapter 7 intends to use a sophisticated psychometric model, known as multifaceted Rasch measurement, to investigate rater variability, particularly rater severity/leniency, in the rater-mediated assesement of SI performance in this PhD research (i.e., part of **RQ 3**). Hopefully, it will contribute to strengthening the generalization inference methodologically.

## Chapter 7 Investigating rater severity/leniency in interpreter performance testing: Using multifaceted Rasch measurement<sup>37</sup>

**Abstract.** *Rater-mediated performance assessment (RMPA) constitutes a critical component of interpreter certification testing systems worldwide. Given the acknowledged rater variability in RMPA and the high-stakes nature of certification testing, it is crucial to ensure a desirable level of rater reliability in interpreter certification performance testing (ICPT). A review of current ICPT practice shows that although interpreter certifying bodies organize rater training to minimize rater variability, the complexities associated with rater behavior such as variable rater severity/leniency and raters' biased interactions with other assessment facets have not been examined with due rigor. It is also unclear to what extent the current correlation-based approach to rater reliability estimation benefits rater training and helps diagnose problematic raters. Against this background, the present study reports on an application of multifaceted Rasch measurement (MFRM), a sophisticated psychometric model, to investigate potential rater severity/leniency displayed in a rater-mediated assessment of interpreting performance. Through the application of MFRM, it is hoped that interpreting testers and researchers would benefit from alternative pathways and new perspectives proffered by MFRM to detect and inspect nuances of rater behaviour in ICPT. Implications for practical rating designs in ICPT and rater training are also discussed.*

### 7.1 Introduction

The need to professionalize interpreting and the desire to ensure the quality of interpreting services have given birth to interpreter certification testing programs worldwide (see Hlavac, 2013). Central to all interpreter certification tests is rater-mediated performance assessment

---

<sup>37</sup> Revisions of this chapter benefit from the author's attendance in the *Language Testing at Lancaster* course, at Lancaster University, UK, 28 July - 8 August 2014, and a Rasch analysis workshop organized by Language Testing Research Center (LTRC) at the University of Melbourne, Australia, 12 July 2014. Part of this chapter was presented at the Australian Institute of Interpreter and Translators Biennial Conference at University of Queensland, Brisbane, Australia, 1-2 November 2014, and at the Biennial Conference of Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), at the University of Queensland, Brisbane, Australia, 27-29 November 2014. A revised version of the chapter will be published as: Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting*, 17(2), 255-283. Page numbers are subject to changes.

(RMPA), a type of assessment in which a candidate's individual performance in response to a stimulus is evaluated by raters using a rating scale or other types of scoring schedules (McNamara, 1996). In rater-mediated assessment of interpreting performance, four typical assessment facets figure prominently: test candidates, raters, interpreting tasks, and assessment criteria. Holding the other facets constant, raters play a critical role in assigning and determining test scores that subsequently constitute one of the most important bases for making certification decisions by certifying authorities. Like other types of score-based actions (e.g., selection, admission, placement), the certification decisions could have consequential impacts or washback effects on relevant stakeholders, particularly test takers (e.g., Yu, 2005) and potential users of interpreting services (e.g., Jacobs et al., 2001).

Given the unanimous use of RMPA in the interpreter certification testing and the potential washback effects, it is an obligation to investigate whether RMPA functions appropriately as intended in interpreting performance assessment. More importantly, considering the pivotal role played by raters in RMPA, it is necessary for test developers and users to be well informed of whether rater-generated scores are vulnerable to both intentional and inadvertent rater variability. One of the rater-related variations is known as *rater severity/leniency*. That is, raters may have the tendency to consistently give ratings that are substantially lower or higher than is warranted by examinees' performance (e.g., McNamara, 1996). Rater severity/leniency has been acknowledged by language testers as a perennial and prominent problem (e.g., Eckes, 2005; McNamara, 1996), and contributes to variability of scores, undermining the utility of test scores as a valid basis for subsequent inferences and actions (e.g., Messick, 1989).

In particular, rater severity/leniency could exist for an individual rater, known as internal self-inconsistency. In other words, a rater may alternate between severe and lenient interpretation and use of a rating scale in the course of performance assessment. For example, in educational and language testing, Cason and Cason (1984, cited in Lumley & McNamara, 1995) found that differences in judge severity can explain as much variance in ratings as differences in examinee ability. In addition, raters may display a pronounced pattern of harshness or leniency in relation to a particular facet of an assessment context such as examinees, assessment criteria, occasions, etc. In other words, raters can be consistently biased (i.e., harsh or lenient) toward a particular group of examinees (Kondo-Brown, 2002; Lynch & McNamara, 1998), a particular rating criterion (Wigglesworth, 1993), and a particular occasion of rating (Lumley & McNamara, 1995; Lunz & Stahl, 1990). Such systematic interaction

between raters and other facets of an assessment context is referred to as *rater bias* (e.g., McNamara, 1996). As a result, it is difficult, if not impossible, to estimate a test candidate's true ability (represented by a true score or fair score), given the existence of rater bias. Examinees' ability estimates depend on, to some extent, a particular harsh or lenient rater and it is a matter of luck for examinees whether they are assigned a particular group of raters. Therefore, rater severity/leniency needs to be considered and accounted for in estimating candidates' ability.

However, it appears that rater severity/leniency has not been investigated with due rigor for interpreter certification performance testing (ICPT). As the following literature review would show, rater variability in ICPT has been primarily encapsulated in inter-rater correlation coefficients modeled through classical test theory (CTT).

## **7.2 A review of rater training and rater variability in ICPT**

In interpreter performance assessment, rater variability is a cause for concern (Angelelli, 2009; Arjona-Tseng, 1993; Hale & Campbell, 2003; Feng, 2005; Wu, 2010), because rater inconsistency in awarding scores undermines the validity of measurement outcomes. To ensure rater reliability, interpreting researchers and testers have called for instituting a rigorous program for rater selection, training and monitoring (Hale, Garcia, Hlavac, Kim, Lai, Turner & Slatyer, 2012; Feng, 2005; Roat, 2006). In a recent report submitted to Australia's National Accreditation Authority for Translators and Interpreters (NAATI), Hale et al. (2012, p. 89) recommend that "examiners undertake compulsory training before being accepted on the (examiners') panel, and continuous training while on the panel". Feng (2005) also recommends strict recruitment, training, calibration and monitoring of raters for large-scale ICPT in China.

To gain a better understanding of how rater training has been conducted to reduce rater variability, 13 national-level interpreter certification performance tests (ICPTs) from eight countries were reviewed (see Appendix M). Four major sources of literature were consulted: 1) ICPT testing manuals, 2) websites of certification organizations, 3) officially published documents, and 4) academic reports, book chapters and journal articles that discuss a certain ICPT. To provide transparency, the literature reviewed is provided in Appendix M. Specifically, answers to three questions were searched: 1) What models have been proposed for rater reliability estimation in ICPTs? 2) Whether rigorous rater training has been

conducted? 3) Have detailed descriptions of rater training been provided? Appendix M summarizes information bearing on each question.

The review shows that, firstly, inter-rater reliability model has been emphasized by the majority of the certifying bodies. This suggests that interpreter certifiers regard rater effects as a primary contributing factor to measurement error. Hale and Campbell (2003, p. 221) observe that “the practice of ‘second marker’, ‘trial marking’, etc. indicates a focus on the marker rather than items as a source of information about reliability”.

Secondly, most of the certifying organizations have conducted rater training of some sort, but only the US-based certifiers have provided detailed training procedures.

Thirdly, rater variability is primarily investigated via models of inter-rater reliability, that is, correlation between two raters (e.g., Pearson’s correlation coefficient) or consistency among a group of raters (e.g., intraclass correlation).

However, the review reveals more questions than it answers. For instance, on the one hand, given the complexity of rater behaviors (i.e., possible rater severity/leniency, biased interactions with other assessment facets), the reliability models used in current ICPTs do not provide an in-depth analysis of rater variability. Although Pearson’s correlation coefficient is useful with two raters, with the increasing number of raters involved, correlation matrix gets progressively complex, making rating patterns difficult to identify. More importantly, under the correlation framework, nuances of rater behavior such as biased rater interactions with other assessment facets (e.g., test takers, tasks, criteria) cannot be detected. Two raters could be highly correlated, because both of them are significantly severe or lenient to a particular test candidate than others. Moreover, the correlation framework is unable to generate diagnostic indicators of an individual rater’s behavior, because correlation always involves at least two raters. On the other hand, it is still unclear whether the rater training provided in ICPTs is effective, how problematic raters are identified, and what remedial procedures can be taken to deal with problematic raters.

In general, it seems that there lacks a methodology that is capable of revealing complexities and nuances associated with rater variability to benefit and facilitate rater training, and ultimately contribute to reliable measurement in ICPTs. To fill this methodological gap, this paper reports on an application of multifaceted Rasch measurement (MFRM), a sophisticated psychometric model, to explore potential rater severity/leniency displayed in a rater-mediated assessment of interpreting performance.

### 7.3 A brief introduction to multifaceted Rasch measurement

Multifaceted Rasch measurement (MFRM) has been used by language testers and researchers to investigate rater behavior in high-stakes language tests (Eckes, 2011; Green, 2013; McNamara, 1996; McNamara & Knoch, 2012), based on the computer program FACETS (Linacre, 2013).<sup>38</sup> As an extension of basic Rasch models, MFRM can analyze both dichotomous and polytomous data, and incorporate multiple facets of an assessment context (e.g., examinees, raters, tasks). The facets under examination and their associated parameters are simultaneously but statistically independently analyzed using joint maximum likelihood estimation procedures. As a result, estimates for the facets of interest are calibrated onto a single linear scale with log odds units (i.e., logits) as its standard metric unit. In other words, MFRM analysis produces calibrated estimates for relevant facets of an assessment context in a common equal-interval metric, thus creating a single frame of reference for interpreting results. For example, for a three-facet measurement model involving examinees, raters and tasks, calibrated estimates for overall ability of each examinee, overall severity/leniency of each rater, and relative difficulty of each task are displayed on a common logit scale. Therefore, measures for elements of one facet specified in a Rasch model are estimated as independent as is statistically possible of the particularities of elements in other facets.<sup>39</sup> In addition, systematic analysis of sub-pattern of rater behavior, known as *bias analysis*, can be conducted in MFRM implemented through the FACETS program. In other words, FACETS can be used to analyze interaction between a rater and other facets specified in a model. For instance, whether a particular rater shows a tendency to give higher-than-warranted scores to an examinee or applies a certain rating scale more leniently.

### 7.4 Context of the present study

This study represents an extension of a full factorial experiment conducted by the researcher to quantitatively investigate how presence and absence of fast speech rate and strong accent

---

<sup>38</sup> To gain in-depth knowledge of the multifaceted Rasch measurement, please refer to Bond & Fox (2007), Eckes (2011), as well as McNamara (1996).

<sup>39</sup> In MFRM, an element is subsumed under a facet. For example, if there are 30 raters in an assessment context, the rater facet includes 30 elements.

would affect the quality of English-to-Chinese simultaneous interpreting (SI) performance (for more details, please refer to Chapter 5).

In the experiment, two independent variables (IVs), namely speech rate and accent, were manipulated. The speed variable had two levels: a fast speech rate (FSR) of approximately 155 wpm, and a slow speech rate (SSR) of about 105 wpm. The accent variable also had two levels: a native speaker (NS) and a non-native speaker (NNS). These two IVs were fully-crossed (i.e., a  $2 \times 2$  factorial design), producing four treatment conditions (TCs) or four differently conditioned tasks (see Table 7.1). Four source-language speeches used in the tasks were manipulated and calibrated to be comparable on multiple dimensions (e.g., length, lexical complexity), except for speech rate and accent (see Appendix F).

Table 7.1 A  $2 \times 2$  factorial design used in the experiment

Independent variables (IVs)		IV A: Speech rate	
		a <sub>1</sub> : Slow	a <sub>2</sub> : Fast
IV B: Accent	b <sub>1</sub> : Native English	TC <sub>1</sub> : a <sub>1</sub> b <sub>1</sub> (Task <sub>SN</sub> )	TC <sub>3</sub> : a <sub>2</sub> b <sub>1</sub> (Task <sub>FN</sub> )
	b <sub>2</sub> : Accented English	TC <sub>2</sub> : a <sub>1</sub> b <sub>2</sub> (Task <sub>SA</sub> )	TC <sub>4</sub> : a <sub>2</sub> b <sub>2</sub> (Task <sub>FA</sub> )

A total of 32 Beijing-based interpreters were recruited to perform English-to-Chinese SI in the four tasks. Their performance was audio-recorded, and 128 interpretation recordings were collected (i.e., 32 interpreters  $\times$  4 tasks).

In order to use inferential statistical procedures (primarily multivariate analysis of variance or MANOVA) to analyze how speech rate and accent would affect SI performance quality, the recorded interpretations need to be evaluated by raters to generate quantitative data. Given that accurate statistical analysis and defensible experimental results hinge on high-quality raw data (in this case, rater-generated scores or ratings), the researcher decided to recruit multiple raters for assessment, and conduct rigorous analysis of rater-generated scores, using MFRM, to identify potentially problematic rating behavior. These considerations underlie the present study.

## 7.5 Foci of MFRM analysis in the study

The Rasch analysis reported in the study focused on three specific and local questions relating to potential rater severity/leniency displayed in the assessment of the recorded interpretations:

- 1) Did the raters recruited in the study differ in overall severity/leniency when assessing the recorded interpreting performance?
- 2) Did the recruited raters consistently use rating scales overall in operational rating? In other words, how internally self-consistent were these raters?
- 3) Did the recruited raters maintain a uniform level of severity/leniency across the interpreters, the SI tasks, and rating criteria in operational rating? That is, did the raters display significant biased interaction with a particular interpreter, task and criterion?

By exploring these questions, it is hoped that an in-depth understanding of the rating behaviors displayed in the study could be gained. Such understanding would also inform subsequent actions to be taken to generate high-quality raw data for statistical analysis to be performed in the experiment. For instance, if a rater is found to be significantly inconsistent and biased, remedial measures could be taken to mathematically correct for her/his severity/leniency, or even drop all her/his scores from subsequent analysis.

In the process of examining rater severity/leniency, it is also hoped that potential audience such as interpreting researchers and testers would gain a concrete and better understanding of procedures involved in MFRM, and new possibilities and rich information provided by MFRM. Implications of MFRM for operational ICPTs are also discussed.

## **7.6 Method**

### **7.6.1 Raters**

To evaluate the interpretations, twelve raters were recruited and trained, but nine of them participated in operational rating.<sup>40</sup> The nine raters were university postgraduate students majoring in English/Chinese interpreting (male:  $n = 3$ , female:  $n = 6$ ; 3rd-year students:  $n = 3$ , 2nd-year students:  $n = 6$ ; average age of about 24 years old). They had experience of assessing interpreting performance for a regional certification test in China. They were also trained by the researcher before operational rating (see 7.6.3 *Procedure*).

### **7.6.2 Materials**

#### **7.6.2.1 Data source: Interpretation recordings**

---

<sup>40</sup> Three raters (R02, R10, R11) did not attend operational rating due to clashing schedules.

The interpretation recordings were derived from the experiment, as is described earlier. A total number of 128 audio recordings of SI performance were used for assessment.

#### 7.6.2.2 *Rating scale*

A rubrics-based rating scale was used to evaluate the interpretations. The rating scale consisted of three 8-point subscales. The subscales represented three rating criteria focusing on three dimensions of interpreting performance: information completeness (InfoCom), fluency of delivery (FluDel) and target language quality (TLQual). In addition, each sub-scale was divided into four 2-point bands with descriptors for each band (see Appendix K). The rating scale was trialed in a small-scale study and revised prior to operational use. Overall, the three subscales functioned properly, as can be seen in the section of 7.7.2 *Effectiveness of the rating scale*.

#### 7.6.3 Procedure

To situate the raters in the current assessment context, rater preparation and rater training were organized before operational rating. Four days preceding rater training, all source-language texts (in English) were sent by the researcher to 12 recruited raters. They were required to familiarize themselves with the content of all source-language texts.

Subsequently, the 12 raters participated in a 5-hour rater training. Each rater was assigned a code (e.g., R01) which stayed with them for the rest of the rating. Specifically, the raters went through five stages of training including 1) an introduction (i.e., an overall introduction to the purpose of the SI performance rating); 2) a familiarization session in which rating assessment criteria and a rating sheet (see Appendix K) were introduced and explained in detail to the raters; 3) a practice session during which the raters were asked to familiarize with using the rating sheet to assess two random recording samples; 4) a norming session in which the raters assessed five different pre-anchored recordings and compared notes with one another; 5) a mini-pilot session in which the raters trialed the rating procedure by assessing a bundle of four recordings in a row before having a short break.

On the next day after the training, all nine raters gathered together in a quiet room for the operational rating. Specifically, each rater was asked to assess interpreting performance independently. The sequence of the recordings to be assessed was randomly generated using a

random number generator.<sup>41</sup> In addition, each rater was provided source-language texts to help them check the original information against the renditions. Furthermore, raters were asked to rate a batch of four recordings before having a short break. By the end of the study, each rater was compensated with 1000RMB or approximately 170 US dollars.

#### 7.6.4 Measurement design

Four facets were specified in the design including interpreters (32 elements), raters (9 elements), SI tasks (4 elements) and assessment criteria (3 elements). These four facets and their associated elements were fully crossed with each other to produce an optimum design from a measurement point of view, thus meeting the connectedness requirement of MFRM (Eckes, 2011; Schumacker, 1999). In other words, each rater assessed each interpreter in each SI task, using the rating scale. As a result of this design, a total of 3456 data points were generated (i.e., 32 interpreters  $\times$  4 tasks  $\times$  3 criteria  $\times$  9 raters), and this empirical dataset constituted raw data for MFRM analysis.

#### 7.6.5 Rasch models

##### 7.6.5.1 Rasch model variants

The dataset was first entered into both a Rating Scale Model (RSM) (Andrich, 1978) and a Partial Credit Model (PCM) (Masters, 1982) for analysis. The former assumed that all three rating subscales shared the same structure, while the latter premised that each subscale had its own distinctive structure for each rating dimension. Both Rasch model variants incorporated such facets as interpreters (*i*), raters (*r*), tasks (*t*) and criteria (*c*).

##### 7.6.5.2 Unidimensionality assumption

To examine psychometric unidimensionality assumed by Rasch models (Henning, 1992; McNamara, 1996), fit statistics, particularly those of rating criterion facet, were used, since departures from the assumption are observable in fit values (Bonk & Ockey, 2003; Eckes, 2005, 2008). Table 7.2 presents the fit statistics of criterion facet produced by FACETS. As can be seen, all infit indices were between the limits of 0.7 and 1.3 (Linacre, 2002), thus providing initial evidence of unidimensionality.

---

<sup>41</sup> Each recording was coded by a number from 001 to 128. A random number sequence was produced by The Random Sequence Generator ([www.random.org](http://www.random.org)) to guide the selection of recordings to be rated.

Table 7.2 Infit statistics of the rating criteria to support unidimensionality

Rating criteria	RSM			PCM		
	Measure	SE	Infit	Measure	SE	Infit
InfoCom	0.14	0.03	1.24	0.21	0.02	1.03
FluDel	0.04	0.03	0.84	-0.02	0.03	0.95
TLQual	-0.17	0.03	0.90	-0.19	0.03	1.02

Note: Measure is in logit; SE = standard error; Infit = mean-squared fit statistics.

### 7.6.5.3 Choice of Rasch model

Given the evidence of unidimensionality, an effort was made to select an appropriate Rasch model between RSM and PCM. The criterion used was separation values (i.e., separation ratio) associated with each facet in a model (Bonk & Ockey, 2003; Fox & Jones, 1998), because greater separation indicates more reliable division of elements in a facet into discernible levels when compared to noise in a model. Table 7.3 compares the separation values for each facet in RSM and PCM, as well as shows variances and standard errors in PCM.

Table 7.3 Separation values and variance statistics in RSM and PCM

Facets	Separation value		Variance in the PCM		
	RSM	PCM	Measure (logit)	Percentage	SE
Interpreter	8.89	8.93	0.59	74.7%	*
Rater	6.07	6.13	0.08	10.1%	*
SI task	9.01	9.11	0.08	10.1%	*
Criterion	6.08	7.33	0.04	5.1%	*
<b>Total</b>	n/a	n/a	0.79	100%	*

Note: Variance is calculated after Bonk & Ockey (2003); Percentage = % of total variance; SE = standard error; \* = SE is negligible; n/a = not applicable.

As can be seen in Table 7.3, separation values were slightly higher in PCM than those of RSM for all facets. That is, the PCM model was potentially superior and subsequently used for the rest of the analysis.

### 7.6.6 Data analysis

To implement MFRM, FACETS 3.71.0 (Linacre, 2013) was used. By convention (Linacre, 2013), mean logit measures of raters, tasks and criteria facets were arbitrarily centered to zero, while the facet of candidates (i.e., interpreters) was made non-centered facet, with mean logit measure of candidate abilities varying according to samples analyzed. Apart from calibrating measures of the four facets, particularly the rater facet, bias analyses were also run to investigate interaction between rater severity/leniency and interpreters, tasks and criteria, respectively.

## 7.7 Results

### 7.7.1 Global model fit

When empirical data are fitted to a Rasch model, there is a need to investigate practical utility of the model (Eckes, 2011). One simple way to assess overall data-model fit is to examine data points that are unexpected given the assumptions of the model (Eckes, 2005, 2008, 2011). Specifically, satisfactory model fit is indicated when approximately 5% or less of (i.e.,  $\leq 5\%$ ) (absolute) standardized residuals are larger than or equal to 2; and about 1% or less of (i.e.,  $\leq 1\%$ ) are larger than or equal to 3 (Linacre, 2013). Table 7.4 presents the overall model fit statistics. As shown, only 2.9% and 0.4% of valid responses were associated with (absolute) standardized residuals  $\geq 2$  and  $\geq 3$ , respectively, suggesting satisfactory overall fit.

Table 7.4 Overall data-model fit statistics

	No.	Percentage
Total valid responses used for estimating model parameters	3456	100%
Valid responses with (absolute) standardized residuals $\geq 2$	100	2.9%
Valid responses with (absolute) standardized residuals $\geq 3$	14	0.4%

### 7.7.2 Effectiveness of the rating scale

Used separately, the three subscales functioned as intended. Due to limited space, the statistics associated with the FluDel subscale was presented in Table 7.5. As shown, the distribution of ratings was spread out across the rating scale scores (indicated by frequency percentage). This

result suggested that the interpreters had a relatively high probability of being correctly given a score that best described their performance. In addition, the thresholds advanced monotonically as expected, which corresponds with their respective rating score. Furthermore, the mean-squared outfit values for each rating score were reasonably close to the expected value of 1 (Eckes, 2008). Similar patterns were also observed for the other two subscales. Therefore, in general, these three subscales functioned reliably with the raters.

Table 7.5 Statistics relating to FluDel rating subscale

Rating criterion	Rating score	Frequency count	Frequency percentage	Threshold (SE)	Outfit statistics
FluDel	1	10	1%	-	0.8
	2	78	7%	-3.02 (0.32)	0.9
	3	176	15%	-1.47 (0.12)	0.9
	4	298	26%	-0.83 (0.08)	0.9
	5	311	27%	0.02 (0.07)	0.8
	6	160	14%	1.10 (0.08)	1.0
	7	95	8%	1.39 (0.11)	1.0
	8	24	2%	2.80 (0.22)	1.3
<b>Total</b>	n/a	1152	100%	n/a	n/a

Note: Threshold = Rasch-Andrich threshold; SE = standard error; n/a = not applicable.

### 7.7.3 Rater calibration reports: Rater severity and internal self-consistency

Table 7.6 shows the estimates of overall rater severity for each of the nine raters in a descending order. As shown, with 0.39 logits R12 was most severe, while R06 (i.e., -0.35 logits) was most lenient. That is, there was a 0.74 logit spread regarding rater severity.

To better understand the differences of the overall rater severity, three statistical indices were computed including 1) homogeneity statistic, 2) separation index, and 3) separation reliability statistic. These indices are shown under Table 7.6. As can be seen, the result of Chi-square test for the rater facet was significant ( $\chi^2 = 307.6$ ,  $df = 8$ ) at  $p < 0.01$ , thus rejecting the null hypothesis that all raters were equally severe. In other words, at least two raters were significantly different. In addition, a high separation index of 6.13 (much higher than zero) meant that rater severity differences were about six times greater than the estimate error.

Furthermore, a very high separation reliability index of 0.97 indicated that the raters could be reliably distinguished from one another. In general, despite the less-than-one logit difference of the overall severity measures, the variations turned out to be statistically significant.

Table 7.6 Logit estimates for overall rater severity

Rater ID	Severity measure (in logit)	Model error	Infit mean square
R12	0.39	0.05	1.03
R05	0.32	0.05	0.80
R04	0.19	0.05	0.71
R07	0.14	0.05	1.14
R03	0.00	0.05	1.09
R01	-0.09	0.05	0.84
R09	-0.27	0.05	0.83
R08	-0.34	0.05	1.96
R06	-0.35	0.05	0.63

Homogeneity statistic: Chi-square ( $\chi^2$ ) = 307.6\*\*,  $df = 8$ ;

Separation index = 6.13;

Separation reliability index = 0.97

\*\* $p < 0.01$

In terms of an individual rater's internal self-consistency, the mean-squared infit statistic in Table 7.6 can be used to gauge whether raters use rating scales consistently across candidates, tasks and criteria (Brown, 1995; Eckes, 2005; Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998).<sup>42</sup> To determine an acceptable level of infit statistics, both a loose fit control (Linacre, 2002) and a tight one (Bond & Fox, 2007; McNamara, 1996) were used. The loose control range suggests that infit values (in logit) between the lower limit of 0.5 and the upper limit of 1.5 (i.e., 0.5-1.5 logits) are useful and productive for measurement, while the tight one ranges from 0.7 to 1.3 logits. In addition, raters with infit values less than the lower limits show less variation than expected by Rasch models, thus constituting an *overfit*, while raters with infit

<sup>42</sup> For each element of each facet, the MFRM provides two fit indices including infit and outfit mean square, both of which indicate the degree of match between observed scores and expected scores generated by Rasch model. For the purpose of this study, only the mean-square infit index is used for all facets, except in 7.7.2 *Effectiveness of the rating scale* where outfit statistic was used.

values greater than the upper limits display more variation than predicted, thus resulting in a *misfit* or *underfit* (Eckes, 2011; McNamara, 1996). Overall, misfit is believed to be more problematic than overfit (Myford & Wolfe, 2003).

Table 7.7 summarizes the distribution of infit statistics for the rater facet. As can be seen in the table, using the tight fit control, only two raters were not internally self-consistent. The two raters were R08 (infit = 1.96) and R06 (infit = 0.63). While R08 as a misfit showed a high degree of inconsistency in his/her ratings, R06 as an overfit were generally more consistent than what the Rasch model expected. Therefore, R08 was more problematic.

Table 7.7 Tight and loose fit ranges to determine misfit and overfit of a rater-facet element

Facet	Fit range					
	Tight control			Loose control		
Rater	fit < 0.7	$0.7 \leq \text{fit} \leq 1.3$	fit > 1.3	fit < 0.5	$0.5 \leq \text{fit} \leq 1.5$	fit > 1.5
No. (%)	(overfit)	(acceptable)	(misfit)	(overfit)	(acceptable)	(misfit)
	1 (11.1%)	7 (77.8%)	1 (11.1%)	0 (0%)	8 (88.9%)	1 (11.1%)

Note: Numbers may not total exactly 100% due to rounding.

Taken together, it can be concluded that although the logit difference of rater severity was only 0.74, the analyses found a statistically significant variation of harshness among the raters. In addition, despite variable rater severity, most raters exercised their respective degree of harshness consistently across the interpreters, the tasks and the criteria, based on either the tight or the loose fit control

#### 7.7.4 Bias analysis

##### 7.7.4.1 Summary statistics on two-way interaction

Apart from the overall severity, raters may have tendency to provide more severe/lenient ratings to a particular interpreter than others (i.e., rater-by-interpreter interaction), assess interpreters more harshly/leniently in a particular task (i.e., rater-by-task interaction) and use a particular criterion more severely/leniently (i.e., rater-by-criterion interaction). Table 7.8 presents the summary statistics for three two-way interactions including Rater  $\times$  Interpreter, Rater  $\times$  Task and Rater  $\times$  Criterion. In particular, a large (absolute) standardized Z score (i.e., an absolute Z score equal to or greater than 2) indicates a significant interaction for a given rater

with a given element in a given facet (Eckes, 2005, 2011; McNamara, 1996). As shown in Table 7.8, whereas the large Z-score percentage for Rater  $\times$  Task interaction was fairly low (5.6%), about 1/6 of Rater  $\times$  Interpreter combinations and about 1/3 of Rater  $\times$  Criterion combinations were found to be significant biased interactions. That is, they were associated with statistically significant differences between observed ratings and expected ratings ( $p = 0.05$ ). Therefore, a further attempt was made to investigate how rater severity interacted with particular interpreters and rating criteria.

Table 7.8 Summary statistics for three two-way interactions

Statistics	Bias analyses: Type of interaction		
	Rater $\times$ Interpreter	Rater $\times$ Task	Rater $\times$ Criterion
No. of combination	288	36	27
Large Z scores (No.; %)	45; <b>15.6%</b>	2; 5.6%	8; <b>29.6%</b>
Minimum Z; Maximum Z	-3.85; 7.19	-2.50; 2.80	-9.42; 7.94
Mean; SD	-0.01; 1.50	0.00; 1.04	0.00; 2.99

#### 7.7.4.2 Rater $\times$ Interpreter interaction

Table 7.9 summarizes the statistics for significant rater-by-interpreter interaction for all the raters. As displayed in the table, altogether, the raters and the interpreters produced 288 different combinations, 45 of which showed significant interaction. That is, a particular rater consistently awarded harsher or more lenient ratings to a particular interpreter than would be predicted by the model, given the rater's rating behavior with other interpreters.

Considering that 13 significant interactions involved R08, accounting for almost 30% of all 45 significant interactions and nearly 40% of 32 R8-by-interpreter combinations, respectively (see Table 7.9), significant Rater  $\times$  Interpreter interactions concerning R08 were further examined with relevant statistical information represented in Table 7.10.

Table 7.9 Statistical information on all significant Rater × Interpreter interactions

<b>Rater ID</b>	<b>Total No. of R × Intp combination</b>	<b>No. of sig. int. (absolute Z &gt; 2)</b>	<b>% of sig. int. in respective R × Intp combinations</b>	<b>% in total No. of sig. int.</b>
R01	32	1	3.10%	2.20%
R03	32	7	21.90%	15.60%
R04	32	2	6.30%	4.40%
R05	32	5	15.60%	11.10%
R06	32	3	9.40%	6.70%
R07	32	6	18.80%	13.30%
R08	32	<b>13</b>	<b>40.60%</b>	<b>28.90%</b>
R09	32	4	12.50%	8.90%
R12	32	4	12.50%	8.90%
<b>Total</b>	288	45	n/a	100%

Note: sig. int. = significant interaction; R × Intp = Rater × Interpreter; n/a = not applicable

As shown in Table 7.10, the absolute Z scores for all the interactions were equal to and greater than 2, indicating significant rater-by-interpreter interaction. Specifically, when a Z-score is below -2.0, a rater gives a more lenient rating to an interpreter given her/his ratings provided to other interpreters; when a Z-score is above 2.0, a rater gives a more severe rating to an interpreter, given all relevant information about the rater. Accordingly, R08 had significant bias, being unexpectedly severe on Interpreter 14, 25, 10, 16, 24, 30 and 31, as well as being unpredictably lenient on Interpreter 08, 19, 23, 21, 05 and 03. Additionally, more information is available from the infit statistics in Table 7.10 as to how consistent is the bias pattern for R8 to evaluate interpreters' ability across all the tasks and the criteria. Using the tight fit control range (i.e., 0.7-1.3 logits), R08 was found to provide both harsher ratings (on Interpreter 16, 30 and 31) and more lenient ratings (on Interpreter 08, 19, 23, 21 and 03) in an inconsistent manner (i.e., all abnormal infit values were above upper limit of 1.3 logits), summarized over all the tasks and the criteria.

Table 7.10 Statistical information on significant R08 × Interpreter interactions

Rater ID	Interpreter ID	Average		SE	Z-score	Bias fit: Infit
		observed-expected	Bias (logit)			
		raw ratings				
R08	Intp14	1.05	1.11	0.34	3.28	1.1
R08	Intp25	1.12	0.93	0.28	3.35	0.7
R08	Intp10	1.03	0.92	0.29	3.15	1.2
R08	Intp16	0.76	0.88	0.35	2.52	<b>2.1</b>
R08	Intp24	0.91	0.82	0.29	2.81	1.0
R08	Intp30	0.89	0.69	0.26	2.66	<b>1.7</b>
R08	Intp31	0.82	0.66	0.27	2.47	<b>2.8</b>
R08	Intp08	-0.67	-0.52	0.25	-2.09	<b>2.5</b>
R08	Intp19	-0.73	-0.57	0.27	-2.13	<b>1.6</b>
R08	Intp23	-0.83	-0.65	0.27	-2.42	<b>1.6</b>
R08	Intp21	-0.95	-0.69	0.25	-2.82	<b>1.8</b>
R08	Intp05	-1.02	-0.86	0.29	-2.97	0.9
R08	Intp03	-1.24	-0.95	0.25	-3.85	<b>2.5</b>

Note: SE = standard error; infit = mean-squared infit statistics.

Moreover, Z scores associated with all 32 R08 × Interpreter interactions were plotted onto a graph, providing a clear representation of any bias for R08, also known as “assessment map” (Wigglesworth, 1993). Figure 7.1 shows the bias assessment map for R08, with Z scores on the Y axis and ability estimates for the 32 interpreters on the X axis. As can be seen in the figure, the significant rater-by-interpreter interactions (represented by a cross) were outside the “safe zone” (absolute Z score  $\geq 2$ ).

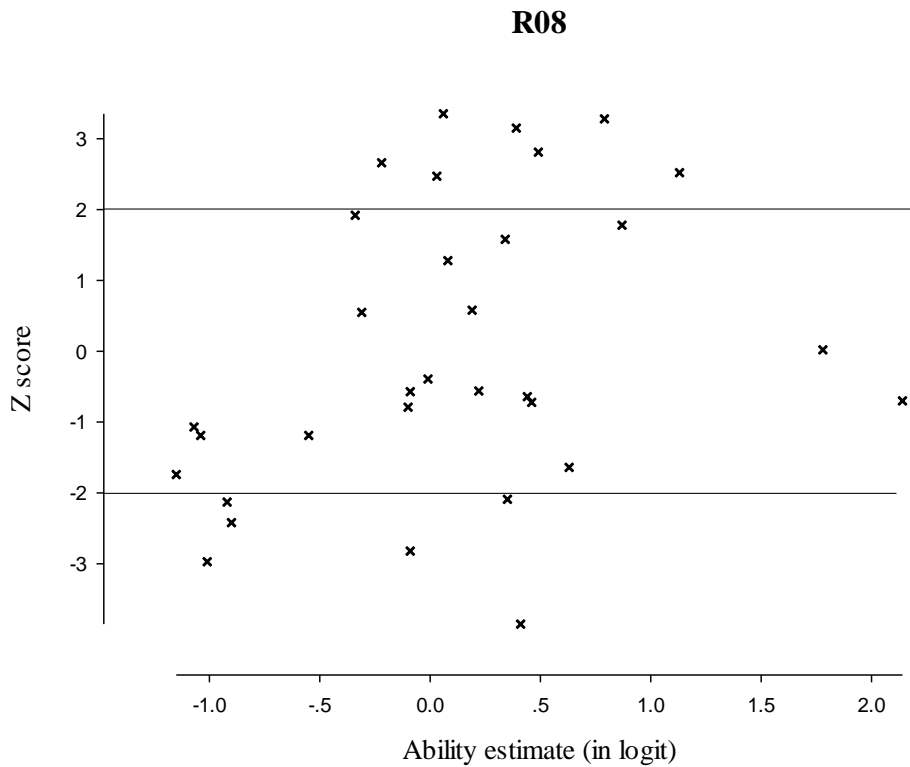


Figure 7.1 Bias assessment map for R08

Furthermore, to gain a panoramic view of significant rater-by-interpreter interactions for all the raters, an assessment map for each rater was created and put together in Figure 7.2 for a direct comparison. By eyeballing, R01 was found to have the least significant biased interactions with the interpreters, while R08 had the largest number of biased interactions. To summarize, all the nine raters displayed significant rater-by-interpreter bias, albeit to a varying degree. In particular, R08 was the most problematic.

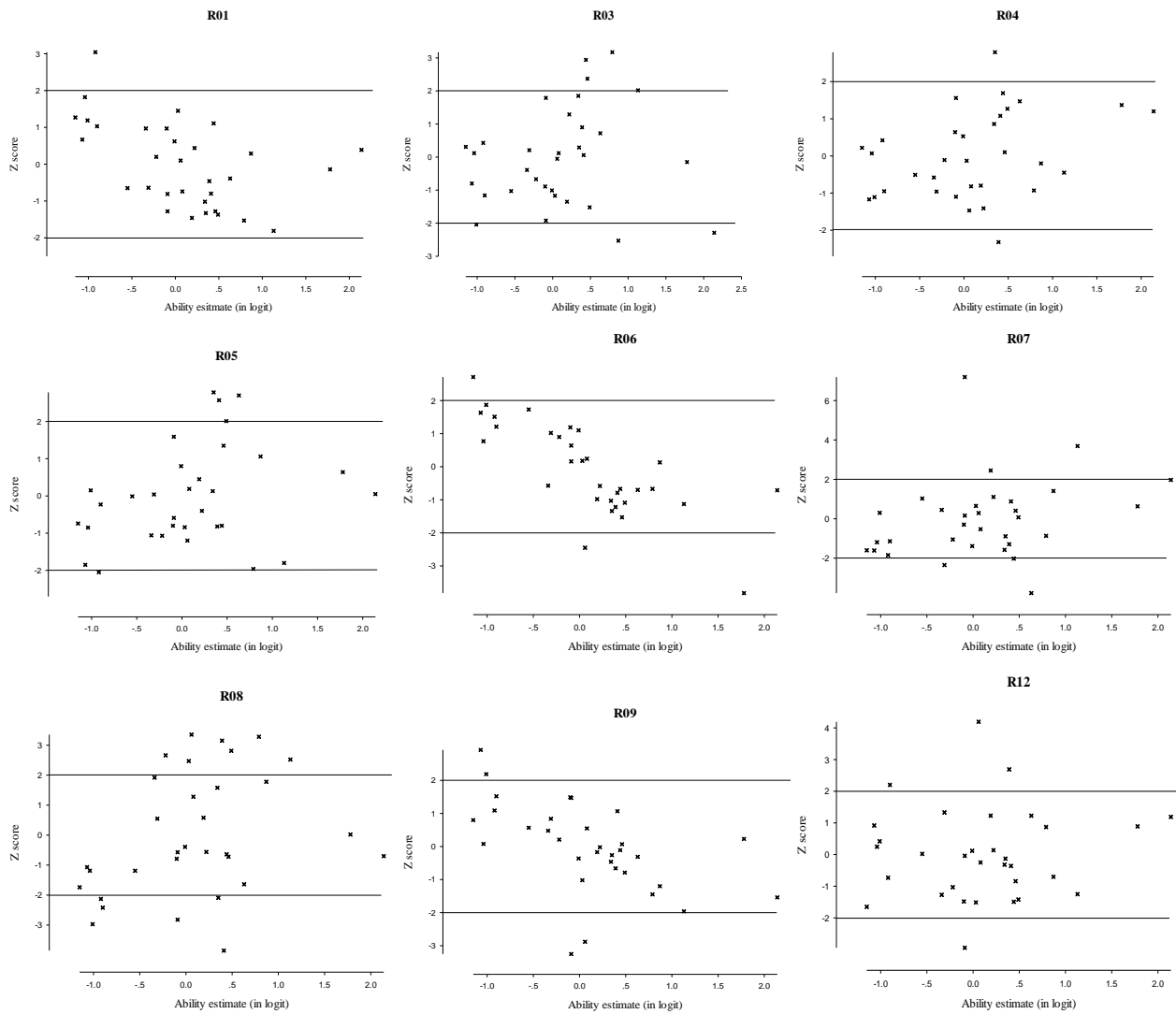


Figure 7.2 Bias assessment maps for all the raters

#### 7.7.4.3 Rater $\times$ Criterion interaction

Overall, rater-by-criterion interaction would test for patterns of unexpected ratings related to a particular rating criterion. Altogether, the raters and the criteria had 27 combinations, eight of which turned out to be significant interactions. Table 7.11 summarizes the statistics relating to the eight significant rater-by-criterion interactions. As can be seen, five out of the eight raters displayed biased interaction with the rating criteria including R01, 04, 06, 08 and 09. In particular, with extremely large absolute Z scores, R08 exercised a variable level of severity/leniency across all the three rating criteria. In other words, R08 tended to alternate between harsher ratings on one criterion and more lenient ratings on other criteria. This result indicates unwanted rater variability in construing the importance of the three criteria. With the infit values larger than 1.3, it also seems that R08's degree of harshness in using rating criteria

was rather inconsistent across the interpreters and the tasks. Therefore, R08 was once again found to be a problematic rater.

Table 7.11 Statistical information on all significant Rater  $\times$  Criterion interactions

Rater ID	Rating criterion	Average		SE	Z-score	Bias fit: Infit
		observed-expected raw ratings	Bias (logit)			
R01	FluDel	-0.20	-0.17	0.08	-2.08	0.7
R04	InfoCom	0.31	0.20	0.07	2.83	<b>0.6</b>
R06	TLQual	0.23	0.22	0.09	2.55	0.7
R08	FluDel	0.76	0.67	0.08	7.94	1.3
R08	TLQual	0.27	0.26	0.09	2.96	<b>1.7</b>
R08	InfoCom	-1.03	-0.67	0.07	-9.42	<b>1.5</b>
R09	InfoCom	0.38	0.26	0.07	3.50	<b>0.5</b>
R09	TLQual	-0.21	-0.19	0.08	-2.24	0.9

Note: SE = standard error; infit = mean-squared infit statistics.

## 7.8 Discussion and implications

### 7.8.1 Sample size issue in the present study

Like any other statistical analysis, MFRM generally requires sufficiently large samples to produce stable and precise estimates with small standard errors (Linacre 1994). In this study, it seems that the sample size was not particularly large, with only 32 interpreters and 9 raters. However, it can be argued that the sample size used in the analysis was useful, given the scope and the purpose of the study.

Firstly, the application of MFRM in this study was an *ex post facto* exploration of rater severity/leniency displayed in a one-time specific assessment context. It was not intended to design and validate a scale or a test usually used in psychological and educational testing.

Secondly, with 3456 data points generated in the study, the total size of the empirical dataset in the MFRM analysis compares well to that in published MFRM-based research (e.g., Elder, Barkhuizen, Knoch, & von Rnadow 2007; Kondo-Brown 2002; Sudweeks, Reeve, &

Bradshaw 2005; Weigle 1994), although there are Rasch-based empirical studies using huge sample sizes (e.g., Bonk & Ockey 2003; Eckes, 2005).

Thirdly, to calibrate rater estimates, the more data points there are to provide information to a given rater, the more stable rater estimates will be. That is, standard errors associated with rater estimates will be very small. With 3456 data points generated in this study, as many as 348 data points contributed information to each rater, which compares favourably to published studies (e.g., Elder et al. 2007; Weigle 1994). In addition, by inspecting standard error (SE) associated with the raters in Table 7.6, the error term at 0.05 was relatively small compared to the severity measures. Given the *ex post facto* and the exploratory nature of this study, the level of precision is acceptable.

However, on a cautionary note, it is worth pointing out that the MFRM results should not be interpreted out of this specific exploratory context. The Rasch-based rater estimates may not be stable with a larger sample of interpreters and tasks. For operational high-stakes ICPTs, MFRM analysis should be based on a sufficiently large sample size (i.e., large numbers of examinees, tasks and raters), so that stable and precise logit measures for each facet can be estimated.

#### 7.8.2 Rater severity/leniency in the present assessment context

In the study, the recruited raters differed in terms of overall severity/leniency, when assessing the SI performance. This finding generally concurs with those found in second language testing literature (Brown, 1995; Eckes, 2005; Kondo-Brown, 2002; Wigglesworth, 1993). In addition, some raters interacted significantly with the interpreters, the SI tasks, and the assessment criteria by providing harsher or more lenient ratings than expected. For example, R08 interacted significantly with over 2/5 of interpreters, giving them considerably lower or higher ratings than what the Rasch model predicted. Other raters (e.g., R03, R09) also display substantial biased interactions, albeit less severe than R08. Such problematic raters need to be identified, analyzed, and even subjected to individualized training, if high-stakes decisions are to be made based on rater-generated measures.

In hindsight, the rater variability and the biased interactions exhibited in the study could be attributed to lack of rating experience on the part of the student interpreters/raters recruited, ineffectiveness of the rater training to some raters, lack of rigorous oversight of the

operational rating procedure, and inherently challenging nature of assessing interpreting performance.

In particular, interpreting researchers believe that assessing interpretations is cognitively taxing and challenging, especially for real-time assessment (Gile, 1995; Vermeiren et al., 2009; Wu, 2010). In assessing SI, raters not only need to split attention on different aspects of interpreting performance (target-language output), but also should attend to source-language input to determine fidelity, accuracy and appropriateness of target-language output. In other words, raters typically need to juggle with several mental activities simultaneously: 1) listening to target-language recordings, 2) reading source-language input, 3) comparing and analyzing the fit between source- and target-language materials, using a multi-dimensional rating scale, and 4) making judgment and awarding scores. According to Gile (1995), this complicated multitasking process could saturate short-term memory capacity of most if not all assessors. As a result, the complexity of raters' mental activity may be no less than that of simultaneous interpreters (Wu, 2010). Given this cognitive complexity involved in assessing interpreting performance, raters with their own idiosyncrasies could disagree with one another on many levels of both qualitative and quantitative decision making, thus resulting in discrepant scores.

To make matters worse, raters may interact with many facets imbedded in an assessment context (e.g., Eckes, 2011). In fact, performance assessment could be thought of as a system comprising a diverse array of facets which could interact with raters, and affect raters' decision making (e.g., Eckes, 2011; McNamara, 1996; Upshur & Turner, 1999). Significantly biased interaction between raters and assessment facets could thus occur, contributing to systematic errors, and ultimately masking and altering the meaning of raw scores. Consequently, the validity of score-based inferences and actions could be effectively jeopardized.

In Interpreting Studies, researchers have realized the pernicious effects of rater variability, calling for robust evaluation of psychometric properties of interpreter performance testing instruments (Angelelli, 2009; Clifford, 2005). But there still lacks sufficient discussion on rater reliability, as Campbell and Hale (2003, p. 217) state that one of the knowledge gaps in translation and interpreting assessment is "a fundamental omission" of a discussion on reliability. Given the pervasiveness and the high-stakes of ICPTs, it is high time that interpreter certification organizations encourage empirical research to shed insight to raters' behaviors in operational ICPTs. To this end, using MFRM represents one of the viable and useful approaches.

### 7.8.3 MFRM: Implications for ICPTs

#### 7.8.3.1 Practical measurement designs

To apply MFRM, one condition must be satisfied: data connectedness or connectivity (Eckes, 2011; Schumacker, 1999). Table 7.12 (a), (b) and (c) present three measurement designs with connected dataset, while the design in Table 7.12 (d) does not meet data connectivity requirement. In this study, the nine raters assessed all the interpretations, representing the optimum design (see Table 7.12 (a)).

However, in large-scale ICPTs such as the China Accreditation Test for Translators and Interpreters (CATTI), for which a total of 50,000 individuals registered in 2012,<sup>43</sup> practical constraints narrow the choice of a rating design. Such constraints typically include the amount of time required for operational rating, rater workload and available financial resources. The fully-crossed/complete design may thus not be operationally feasible, as it requires all raters to assess all interpretations, which entails a large number of raters and heavy workload for each rater. Incomplete designs with data connected represent viable alternatives in practice (see Table 7.12 (b) & (c)).

Table 7.12 (a) Fully crossed/complete design – Connected

Candidate	SI Task	Rater			
		R1	R2	R3	Rn
C01	T01	♦	♦	♦	♦
	T02	♦	♦	♦	♦
C02	T01	♦	♦	♦	♦
	T02	♦	♦	♦	♦
C03	T01	♦	♦	♦	♦
	T02	♦	♦	♦	♦
Cn	T01	♦	♦	♦	♦
	T02	♦	♦	♦	♦

<sup>43</sup> [http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget\\_edu\\_jsp](http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget_edu_jsp)

Table 7.12 (b) Incomplete design – Connected

Candidate	SI Task	Rater			
		R1	R2	R3	Rn
C01	T01	♦	♦		
	T02	♦	♦		
C02	T01		♦	♦	
	T02		♦	♦	
C03	T01			♦	♦
	T02			♦	♦
Cn	T01				♦
	T02				♦

Table 7.12 (c) Incomplete design – Connected

Candidate	SI Task	Rater			
		R1	R2	R3	Rn
C01	T01	♦			♦
	T02	♦			♦
C02	T01		♦		♦
	T02		♦		♦
C03	T01			♦	♦
	T02			♦	♦
Cn	T01				♦
	T02				♦

Table 7.12 (d) displays a disconnected design, thus not suitable for MFRM.

Table 7.12 (d) Incomplete design – Disconnected

Candidate	SI Task	Rater			
		R1	R2	R3	Rn
C01	T01	♦			
	T02	♦			
C02	T01		♦		
	T02		♦		
C03	T01			♦	
	T02			♦	
Cn	T01				♦
	T02				♦

### 7.8.3.2 Rater training and monitoring

MFRM can be used to estimate rater severity/leniency, identify problematic raters (e.g., R08) and examine the direction and the degree of rater variability. An individual rating report, based on Rasch analysis, can be provided to each rater to help them gain an in-depth understanding of their own rating behaviour, and inform where to take corrective action in further training to make raters internally consistent (Eckes, 2011; Knoch, 2011; Weigle, 1998; Wigglesworth, 1993). In such an individualized report, Eckes (2011) recommends including 1) the rater's severity/leniency measure, 2) self-internal consistency indicated by rater infit/outfit indices, 3) frequency of usage of rating scale categories and 4) bias charts that portray rater biased interaction with candidates, tasks and criteria. After a moderation session, behaviors of problematic raters can be compared between the first and the second training sessions, based on the individual rating reports (Knoch, 2011; Weigle, 1998). For example, in this study, misfitting raters such as R08 were identified. She or he can be interviewed and analyzed to determine potential causes of her/his inconsistency. Based on the individual rater report, further training can be provided. A comparison of rating behavior between the first and the second training would help determine the effectiveness of training. In some special circumstances, for example, in which significant inter- and intra-rater inconsistency survives rater training, ratings from misfitting raters could be dropped, if deemed appropriate (Schaefer, 2008).

### 7.8.3.3 Utility of Rasch measurement

MFRM can be applied as a useful complement to traditional correlation-based models to examine rater variability in high-stakes ICPTs. The advantages of MFRM include the ability to: 1) incorporate multiple assessment facets in a model; 2) run analysis for different measurement designs, despite missing data; 3) generate diagnostic statistics on individual raters; 4) examine biased interaction between raters and other facets; and 5) provide individualized feedback to raters for further training and moderation. For ICPTs, MFRM can be also useful in assisting rater training and diagnosing problematic raters. For example, in response to the recommendations proposed to improve NAATI tests (Hale et al., 2012), the NAATI authority agrees in principle to conduct compulsory rater training. Given the strengths discussed, MFRM has much to offer to facilitate NAATI's endeavor.

## 7.9 Conclusion

This paper reports on an application of MFRM to explore rater severity/leniency displayed in an assessment of SI performance. An in-depth understanding of rater behaviour in the assessment context is obtained, which helps decision making as to how to deal with the rater-generated scores for subsequent statistical analysis. It is also hoped that the study has shown to interpreting testers and researchers new possibilities and information proffered by MFRM. Given the strengths of MFRM, it is recommended that it be applied by ICPT developers to detect problematic rating behavior and improve rater training.

## 7.10 References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573.
- Angelelli, C. (2009). Using a rubric to assess translation ability: defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp. 13-47). Amsterdam: John Benjamins.
- Arjona-Tseng, E. (1993). A psychometric approach to the selection of translation and interpreting students in Taiwan. *Perspectives*, 1(1), 91-104.

- Arocha, I. S., & Joyce, L. (2013). Patient safety, professionalization, and reimbursement as primary drivers for National Medical Interpreter Certification in the United States. *The International Journal for Translation & Interpreting Research*, 5(1), 127-142. DOI: ti.105201.2013.a07
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). London: Lawrence Erlbaum.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1), 89-110.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1-15.
- Cai, X. (2009). 全国翻译专业资格(水平)考试分析及其对翻译队伍建设的启示. [Analysis of CATTI tests and implications for translators and interpreters]. *中国翻译*, 1, 60-62.
- Cai, X. H. (2007). 口译评估. [Interpretation and Evaluation]. Beijing: China Translation & Publishing Corporation.
- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Certification Commission for Healthcare Interpreters (2010). *Job Task Analysis Study and Results*. Retrieved from <http://www.cchicertification.org/images/webinars/cchi%20jta%20report-public.pdf>
- Certification Commission for Healthcare Interpreters (2011). *Technical Report on the Development and Pilot Testing of the CCHI Examinations*. Retrieved from <http://www.cchicertification.org/images/pdfs/cchi%20technical%20report%20-%20public%20final.pdf>
- Certification Commission for Healthcare Interpreters (2012). Technical Report on the Development and Pilot Testing of the Certified Healthcare Interpreter™ (CHI™) Examination for Arabic and Mandarin. Retrieved from <http://www.cchicertification.org/images/pdfs/cchi%20arabic%20and%20mandarin%20technical%20report-final.pdf>
- Certification Commission for Healthcare Interpreters (2014). *Candidate's Examination Handbook*. Retrieved from <http://www.cchicertification.org/images/pdfs/candidatehandbook.pdf>

- Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting*, 7(1), 97-13.
- 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.
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185.
- Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37-64.
- Feng, J. Z. (2005). 论口译测试的规范化. [Towards the standardization of interpretation testing]. *外语研究*, 89, 54-58.
- Feuerle, L. (2013). Testing interpreters: Developing, administering, and scoring court interpreter certification exams. *The International Journal for Translation & Interpreting Research*, 5(1), 80-93. DOI: 10.12807/ti.105201.2013.a04
- Fox, C., & Jones, J. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, 45(1), 30-45.
- Green, R. (2013). *Statistical analysis for language testers*. Houndmills: Palgrave Macmillan.
- Hale, S., Garcia, I., Hlavac, J., Kim, M., Lai, M., Turner, B., & Slatyer, H. (2012). *Development of a conceptual overview for a new model for NAATI standards, testing and assessment*. Sydney, Australia. Retrieved from: <http://www.naati.com.au/PDF/INT/INTFinalReport.pdf>
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9(1), 1-11.
- Hlavac, J. (2013). A cross-national overview of translator and interpreter certification procedures. *The International Journal for Translation & Interpreting Research*, 5, 32-65. DOI: 10.12807/ti.105201.2013.a02
- IoL Educational Trust. (2010). *Diploma in Public Service Interpreting: Handbook for candidates*. London, UK. Retrieved from: <http://www.iol.org.uk/qualifications/DPSI/Handbook/DPSIHB11.pdf>

- Jacobs, E. A., Lauderdale, D. S., Meltzer, D., Shorey, J. M., Levinson, W. & Thisted, R. A. (2001). Impact of interpreter services on delivery of health care to limited-English-proficient patients. *The Journal of General Internal Medicine*, 16(7), 468-474.
- Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing*, 28(2), 179-200.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J. M. (2013). *A user's guide to FACETS: Program manual 3.71.2*. Retrieve from <http://www.winsteps.com/a/facets-manual.pdf>
- Lu, M., Liu, C., & Gong, X. F. (2007). 全国翻译专业资格(水平)考试英语口译试题命制一致性研究报告. [How to maintain consistency in CATTI's interpretation tests: A research report]. *中国翻译*, 5, 57-61.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, 12(1), 54-71.
- Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13(4), 425-444.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
- McNamara, T. F., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, 29(4) 555-576.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan.
- Mortensen, D. (1998). *Establishing a scheme for interpreter certification: The Norwegian experience*. Retrieved from <http://folk.uio.no/dianem/report1998.pdf>

- Mortensen, D. (2001). *Measuring quality in interpreting: A report on the Norwegian Interpreter Certification Examination (NICE)*. Oslo, Norway. Retrieved from: <http://folk.uio.no/dianem/IntQuality-Internet.pdf>
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.
- National Accreditation Authority for Translators and Interpreters. (2013). *INT Project discussion paper*. Retrieved from <http://www.naati.com.au/PDF/INT/INT%20Project%20Discussion%20Paper%20-%20November%202013.pdf>
- National Association of the Deaf (2014). *NAD and RID releases the NAD-RID National Interpreter Certification (NIC) Credential Validity, Reliability, & Candidate Performance Report*. <http://nad.org/news/2014/2/nad-and-rid-release-performance-report>
- National Board of Certification for Medical Interpreters (2014). *The National Board of Certification for Medical Interpreters: Certified Medical Interpreter Candidate Handbook*. Retrieved from <http://www.certifiedmedicalinterpreters.org//sites/default/files/national-board-candidate-handbook.pdf>
- National Center for States Courts (2013). *Federal Court Interpreter Certification Examination for Spanish/English: Examinee handbook*. Retrieved from [http://www.ncsc.org/sitecore/content/microsites/fcice/home/About-the-program/~/\\_media/Microsites/Files/FCICE/Final%20Examinee%20Handbook%201-23-2013%20for%20online.ashx](http://www.ncsc.org/sitecore/content/microsites/fcice/home/About-the-program/~/_media/Microsites/Files/FCICE/Final%20Examinee%20Handbook%201-23-2013%20for%20online.ashx)
- Office of China Accreditation Tests for Translators and Interpreters. (2005). 二级口译英语同声传译类考试大纲. 外文出版社. [Syllabus of CATTI Level-two Simultaneous Interpreting Test]. Beijing: Foreign Languages Press.
- PSI Services LLC (2010). *Development and Validation of Oral and Written Examinations for Medical Interpreter Certification: Technical Report*. Burbank, California, USA. Retrieved from: <http://www.certifiedmedicalinterpreters.org//sites/default/files/oral-and-written-medical-i-nterpreter-technical-report-final.pdf>.

- PSI Services LLC (2013). *Development and validation of oral examinations for Medical Interpreter Certification: Mandarin, Russian, Cantonese, Korean, and Vietnamese forms*. Retrieved from <http://www.certifiedmedicalinterpreters.org/sites/default/files/tech-report-development-validation-language-forms.pdf>
- Roat, C. E. (2006). *Certification of Health Care Interpreters in the United States: A Primer, a Status Report and Considerations for National Certification*. Los Angeles, USA, Retrieved from: [http://www.calendow.org/uploadedFiles/certification\\_of\\_health\\_care\\_interpreters.pdf](http://www.calendow.org/uploadedFiles/certification_of_health_care_interpreters.pdf)
- Russell, D., & Malcolm, K. (2009). Assessing ASL–English interpreters: The Canadian model of national certification. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 331-376). Amsterdam: John Benjamins.
- Schaefer, E. (2008). Rater bias pattern in an EFL writing assessment. *Language Testing*, 25(4), 465-493.
- Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested and mixed designs. *Journal of Outcome Measurement*, 3(4), 323-338.
- South African Translators' Institute (2007a). *Guidelines: SASL Interpreter Accreditation Testing*. Retrieved from [http://translators.org.za/sati\\_cms/downloads/dynamic/sati\\_accreditation\\_for\\_sasl\\_interpreting\\_english.pdf](http://translators.org.za/sati_cms/downloads/dynamic/sati_accreditation_for_sasl_interpreting_english.pdf)
- South African Translators' Institute (2007b). *Guidelines: Simultaneous Interpreter Accreditation Testing*. Retrieved from [http://translators.org.za/sati\\_cms/downloads/dynamic/sati\\_accreditation\\_for\\_sim\\_interpreting\\_english.pdf](http://translators.org.za/sati_cms/downloads/dynamic/sati_accreditation_for_sim_interpreting_english.pdf)
- Stansfield, C.W., & Hewitt, W. (2005). Examining the predictive validity of cut scores on a screening test for court interpreters. *Language Testing*, 22(2), 1-25.
- Sudweeks, R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, 9, 239-261.
- Turner, B., Lai, M., & Huang, N. (2010). Error deduction and descriptors – A comparison of two methods of translation test assessment. *The International Journal for Translation*

*and Interpreting Research*, 2(1), 11-23. Retrieved from  
<http://trans-int.org/index.php/transint/article/view/42/66>

- Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing*, 16(1), 82-111.
- Vermeiren, H., Gucht, J. V., & De Bontridder, L. (2009). Standards as critical success factors in assessments: Certifying social interpreters in Flanders, Belgium. In C. V. Angelelli & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 291-330). Amsterdam: John Benjamins.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-319.
- Wu, S. C. (2010). *Assessing Simultaneous Interpreting: A study on Test reliability and Examiners' assessment behavior*. (PhD thesis, Newcastle University upon Tyne, UK). Retrieved from <https://theses.ncl.ac.uk/dspace/bitstream/10443/1122/1/Wu%2011.pdf>
- Youdelman, M. (2013). The development of certification for healthcare interpreters in the United States. *The International Journal for Translation and Interpreting Research*, 5(1), 114-126. DOI: ti.105201.2013.a06
- Yu, D. R. (2005). *T&I Labor Market in China*. Sydney, Australia. Retrieved from: [http://www.ling.mq.edu.au/translation/lmtip\\_china.htm](http://www.ling.mq.edu.au/translation/lmtip_china.htm)

## **An introductory note to Chapter 8**

In Chapter 7, the rater severity/leniency displayed in the assessment of interpreting performance in this research has been examined. Particularly, R08 is found to be most problematic, giving significantly biased scores to a large proportion of the interpreters. However, the generalizability of the rater-generated scores is still not investigated (i.e., the remaining part of **RQ 3**). Chapter 8 therefore seeks to shed insight to score generalizability by employing a modern measurement theory, called generalizability (G) theory, to examine the rater-generated scores in the experiment.

G theory has been widely used in second language testing research, but has not found its way to ICPTs. In ICPT, classical test theory (CTT) approach, operationalized by various correlation coefficients (e.g., Pearson's *r*) and Cronbach's *alpha*, has been traditionally used by ICPT developers. Although the CTT approach is conceptually more accessible, it can only investigate a single source of measurement error at a time. G theory represents a powerful extension of CTT, in that it is capable of taking into consideration main and interaction effects of multiple assessment facets in one go. It is hoped that applying G theory to the rater-generated scores would strengthen the experimental results, and ultimately contribute methodologically to enhancing the generalization inference in the ICPT validity argument.

## Chapter 8 Investigating score reliability in English/Chinese interpreter certification performance testing: A generalizability theory approach<sup>44</sup>

**Abstract.** *As a property of test scores, reliability constitutes an important psychometric consideration and underpins validity of measurement results. A review of interpreter certification performance tests (ICPTs) reveals that 1) although reliability check has been recognized as an important concern, its theoretical importance overshadows the operational efforts to measure score reliability, and 2) while multiple sources of measurement error could contribute variability to total score variance, rater effects have been regarded as the only source of error, and modeled through classical test theory (CTT) in the form of inter-rater reliability coefficients. The present study was therefore initiated to investigate score reliability for a rater-mediated assessment of English-to-Chinese simultaneous interpreting (SI), using generalizability (G) theory. Results show that 1) the information completeness (InfoCom) ratings were more dependable than those of fluency of delivery (FluDel) and target language quality (TLQual), 2) adding tasks was more effective in raising dependability for InfoCom than using extra raters, but the effect was reversed for FluDel and TLQual, and 3) three different weighting schemes produced small variations in the composite score dependability, but the InfoCom rating scale accounted for the largest proportions of the composite universe-score variances. These results were discussed in terms of English/Chinese ICPTs.*

### 8.1 Introduction

As a sub-field of language testing, interpreter performance testing and assessment is lesser-known to many language testers (Chen, 2011). Despite the limited publicity, interpreter performance assessment has been widely conducted for various purposes. One of the important purposes is to certify interpreters for different professional settings (e.g., international conferences, court, medical, public service setting). Although interpreter certification

---

<sup>44</sup> Part of the findings in the chapter was presented at the Biennial Conference of Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), at the University of Queensland in Brisbane, Australia, 27-29 November 2014, and also at the Monterey Forum, at the Middlebury Institute of International Studies at Monterey, California, USA, 28-29 March 2015. A revised version of the chapter is under the 4<sup>th</sup> round of peer review in the journal of *Language Assessment Quarterly* as: Han, C., (under review). Investigating score reliability in English/Chinese interpreter certification performance testing: A generalizability theory approach..

performance testing (ICPT) has a relatively short history, it has witnessed a rapid development across the world (Hlavac, 2013). New ICPTs are developed in different countries (e.g., see Angelelli, 2007; Liu, 2013), and previous ICPTs grow from strength to strength. For instance, according to the China News Service,<sup>45</sup> the China Accreditation Test for Translators and Interpreters (CATTI) has grown rapidly in terms of the number of test candidates. 2012 alone witnessed a total of 50,000 individuals registering for CATTI's tests, making it one of China's largest testing programs. Like IELTS or TOEFL, ICPT is high-stakes, because such testing could produce consequential impacts on individual test takers (Yu, 2005), certification organizations (Clifford, 2005) and recipients of interpreting services (Jacobs, et al., 2001).

Despite their fast development and possible consequences, ICPTs have rarely been subjected to a rigorous validation process (Clifford, 2005; Hale et al., 2012; Sawyer, 2004, pp. 96-102, also see Chapter 2), where multiple strands of validity evidence are generated to support intended score-based inferences and actions. One important strand of validity evidence concerns the reliability or generalizability of test scores (Messick, 1989). Although interpreting educators and testers have called for a robust psychometric evaluation of ICPTs (Angelelli, 2009; Clifford, 2005), it seems that little effort has been made so far. Only a few researchers have empirically examined the score reliability of interpreter performance assessment (e.g., Wu, 2010). Some interpreter certifying bodies also provide "reliability coefficients" in their testing manuals (Roat, 2006). Overall, in the interpreter performance testing literature the approach to score reliability represents what psychometricians call "classical test theory" (CTT) approach. But attention has been increasingly focused on generalizability (G) theory that promises stronger capability of estimating true score and error variances by partitioning the total score variance into various main and interaction variances (e.g., Brennan, 2001a; Shavelson & Webb, 1991).

Against the backdrop, the present study was initiated to apply G theory to investigate issues related to score reliability in an experimental rater-mediated assessment of English-to-Chinese simultaneous interpreting (SI).<sup>46</sup>

---

<sup>45</sup> [http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget\\_edu\\_jsp](http://www.chinanews.com/edu/2013/01-09/4474762.shtml?flashget_edu_jsp)

<sup>46</sup> There are usually four different forms of interpreting including simultaneous interpreting (SI), consecutive interpreting (CI), sight translation (SiT) and dialogue interpreting (DI). SI is performed when an interpreter listens to a source-language speech while interpreting simultaneously in target language. During CI, an interpreter usually listens to a speaker's speech for a few minutes while taking notes, and then interprets what the speaker has said when the s/he stops. SiT involves an interpreter reading of a text from a source language into a target language simultaneously. During DI, an interpreter usually interprets dialogue-like interactions rather than speeches.

## 8.2 Literature review

In this section, an overview of ICPT practice in major countries is provided, with a special focus on English/Chinese ICPTs in China. Then, an in-depth review of the issues related to score reliability for ICPT is conducted, compared to second language testing literature that applies G theory.

### 8.2.1 An overview of the ICPT practice

Across different countries, the common goal of ICPT is to ensure that interpreters have the minimum level of knowledge and abilities required to practice interpreting in a given target domain. ICPTs are therefore criterion-referenced tests, with test scores being interpreted against a set of pre-determined standards. There are also some differences between ICPTs. Different ICPTs are, for example, designed to accommodate different modalities of interpreting (i.e., signed & spoken language interpreting). The Association of Visual Language Interpreters of Canada (AVLIC) is a case in point, which develops a national certification testing system for American Sign Language (ASL) interpreters in Canada. The biggest difference, perhaps, is that different ICPTs are developed to certify different types of interpreters working in different settings or target domains (see Hale et al., 2012; Hlavac, 2013; Roat, 2006). For example, some ICPTs are tailored for interpreters working in legal settings (e.g., the US Federal Court Interpreter Certification Examination/FCICE), some for interpreters in medical settings (e.g., the US National Board of Certification for Medical Interpreters/NBCMI), and still others for public services settings (e.g., the UK Diploma in Public Service Interpreting/DPSI test, Australia's National Accreditation Authority for Translators and Interpreters/NAATI tests). As a result of the different practice domains, divergent types of interpreting tasks and varying task topics are included in ICPTs to ensure content relevance and representativeness. An ICPT that is designed for public service settings could be dominated by SI, CI and SiT tasks (e.g., the DPSI test), while an ICPT that targets high-level international conference settings may only sample SI tasks (e.g., the CATTI level IV test).

In China, competent interpreters are most needed to facilitate the country's increasing economic, social and cultural exchanges with the other parts of the world. As a result, the focus of ICPT is to certify English/Chinese interpreters primarily working in conference settings. A

number of ICPTs have been developed, and are currently administered at both national and local levels. At the national level, two tests are of interest: the CATTI and National Accreditation Examinations for Translators and Interpreters (NAETI). At the local level, two tests are also widely recognized: the Shanghai Interpretation Accreditation test (SIA), and the English Interpreting Certificate (EIC) developed by Xiamen University.

In the ICPTs mentioned above, test scores constitute one of the most important evidence to help make certification decisions. Score reliability therefore should be one of the top priorities on the agenda of the interpreter certifying bodies. Specifically, ICPT developers and publishers have the fundamental responsibilities to obtain and report reliability-related evidence and errors of measurement (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). However, in practice score reliability has not been treated with due rigor in ICPT. Some certifying bodies seem not to have considered score reliability (e.g., NAETI, EIC), as relevant information could not be found in any published materials. This is probably why Campbell and Hale (2003) warn about the knowledge gap of test reliability in translation and interpreting assessment. Other certifying organizations have proposed models of reliability estimation, but concrete reliability estimates could not be found in testing manuals (e.g., CATTI, NAATI). Still other certifiers focus only on rater effects (e.g., FCICE), although multiple sources of measurement error could contribute to total score variance. The rater effects, particularly inter-rater reliability, are primarily modeled through the CTT approach, which permits estimation of a single source of measurement error at a time. Reliability analysis therefore seems to be an under-emphasized and under-researched area for ICPT. Considering that extensive research has been done on score reliability in the field of second language testing, much could be learnt from the mature discipline.

### 8.2.2 G theory in second language testing and its relevance to ICPT

In second language testing literature, to investigate score reliability both univariate and multivariate G-theory analyses have been conducted. In general, two phases are involved in the G theory: a G study and a D study. In the univariate G study, variance components (VCs) associated with each source of measurement error are estimated. In the follow-up D studies, the VC estimates are used to find an optimal measurement design for achieving a desirable level of score reliability. Two types of reliability-like coefficients can be calculated: an index of

generalizability ( $p^2$ ) for norm-referenced interpretations, and an index of dependability ( $\Phi$ ) for criterion-referenced interpretations, the latter of which is particularly relevant to tests such as ICPTs. Moreover, apart from the VC estimates, the multivariate G-theory analysis is capable of estimating covariance components (CCs), which provides some new information about how persons' universe scores and errors co-vary. Specifically, the multivariate G theory has two important applications: 1) to produce correlations among the universe scores that reveal the true relationships among scores on multiple rating dimensions (Xi, 2007), and 2) to analyze the dependability of composite scores based on different weighting schemes (Brennan, 2001a; Xi, 2007).

Applying the G theory, language testers have gained in-depth understandings of the effect of the number of tasks and/or raters on score reliability for different types of discrete language test (e.g., Brown, 1984, 1999; Zhang, 2006), and especially for performance assessments (e.g., Bachman, Lynch & Mason, 1995; Gebril, 2009; Lee, 2006; Lee & Kantor, 2007; Lynch & McNamara, 1998; Xi, 2007). Previous studies show that while increasing the number of tasks or raters generally contributed to higher score reliability, the relative efficiency was different. For example, adding tasks was found to be more effective in raising reliability coefficients than recruiting more raters in writing assessments (Gebril, 2009; Lee & Kantor, 2007) and in speaking assessments (Lee, 2006; Xi, 2007). In contrast with the abundant research in the second language testing, very little empirical evidence is currently available about the impact of tasks and/or raters on score reliability for ICPTs. For ICPTs, the number of interpreting tasks range from two to six, with some tests sampling different types of interpreting task (i.e., DPSI, FCICE, NATTI). For instance, the NATTI professional-level test uses two DI tasks, two SiT tasks and two CI tasks. The number of raters used in ICPTs ranges from two to three. In Canadian AVLIC test, for example, three independent raters are involved (Russell & Malcolm, 2009). Particularly, most relevant to the present study is the CATTI level IV English/Chinese SI test that samples four SI tasks, and uses a double-rating procedure ( $n_t = 4$ ,  $n_r = 2$ ). Given the lack of research for the CATTI test, the operationalized measurement design needs to be examined to see how the number of tasks and/or raters would affect score reliability.

Another strand of research in the second language testing that used the G theory pertains to the reliability of composite scores. Several studies examined how different combinations of divergent types of tasks would affect the composite score reliability for writing (Gebril, 2010; Lee & Kantor, 2007), and for speaking (Lee, 2006). The most relevant research to the present

study was a multivariate G-theory analysis of the dependability of the composite score based on equal weighting ( $n = 0.2$ ) of five different rating scales (Sawaki, 2007). In Sawaki's study, nominal weights that often represent test developer's desired weights are differentiated from effective weights, that is, the degree to which individual rating scales empirically or statistically contribute information to a composite score (Wang & Stanley, 1970). The study shows that although the Grammar rating scale was assigned a weight of 0.2, it empirically accounted for more than 30% of the composite universe-score variance, while the other scales explained only about 15% to 18%. These results suggest that the Grammar rating scale empirically contributed more information to the composite universe-score variance than the other scales. However, such insight has not been developed in ICPT. Although three general rating dimensions (i.e., information completeness, fluency of delivery and target language use) have been consistently used in ICPTs, different types of weighting schemes are available: 1) conventional unit weighting scheme (Scheme I) (see The Institute of Linguists Educational Trust [IoLET], 2010), 2) empirical schemes (Scheme II) based on survey findings (see Kurz, 1989), and 3) subjective schemes (Scheme III) often derived from expert judgment (see Yang, 2005). It would thus be interesting to investigate what effects different weighting schemes have on the composite score dependability for interpreting, and what empirical contributions the weighting schemes make to the composite score variance.

### 8.3 Research questions

Given that the ICPT research is still in its infancy (Clifford, 2005), particularly given the lack of robust treatment of score reliability for ICPTs, the present study was initiated to gain initial insight to the following three questions in relation to rater-mediated assessment of English-to-Chinese simultaneous interpreting (SI).

- 1) What would be the impact of increasing the number of SI tasks and/or raters on the dependability of information completeness, fluency of delivery, and target language quality ratings?
- 2) What would be the impact of different weighting schemes proposed *a priori* on the composite score dependability?
- 3) What would be the empirical contributions of differentially weighted rating scales to the composite score variance?

## 8.4 Method

### 8.4.1 Participants: Interpreters and raters

Using snowball sampling, 32 Beijing-based interpreters were recruited to participate in the simulated assessment. The interpreters had Mandarin Chinese as their L1 and English as their L2. Averaged at 31 years old, the group consisted of 13 male and 19 female interpreters, 81.3% of whom had a master's degree in translation and interpreting, or language-related majors. See more demographic information of the interpreters in Chapter 5.

A total of nine postgraduate English/Chinese interpreting students were recruited as raters in the study. They all had experience of assessing interpreting performance for a regional certification test in China. They also received rater training provided by the researcher, as is described in the section of 8.4.3.

### 8.4.2 Materials: SI tasks and rating scales

Four SI tasks were carefully designed for an experiment to assess interpreting performance. Specifically, the tasks were designed to vary in speech rate and speakers' accent so as to reflect real-life practices, but were comparable in other aspects (e.g., word count, topics, register, lexical complexity). For a detailed and complete description of task development and task characteristics, please refer to Chapter 5. In addition, the task type and the task content were similar to those used in the CATTI level IV SI test.

A descriptor-based rating scale was used to assess SI performance. The scale consisted of three 8-point subscales, focusing on information completeness (InfoCom), fluency of delivery (FluDel) and target language quality (TLQual). Each sub-scale was further divided into four 2-point bands with descriptors (see Appendix K). The scales were trialed in a small-scale study and revised prior to the operational use. Preliminary evidence based on Rasch-Andrich thresholds and fit statistics from Rasch analysis suggested that the subscales functioned properly overall (for more details, please see Chapter 7).

### 8.4.3 Rater training

Four days before the rater training, all source-language texts (in English) were sent by the researcher to the raters. They were required to familiarize themselves with the content of the source texts.

Following the preparation, the raters participated in a 5-hour training. The training had five stages: 1) an introduction to the SI performance assessment; 2) a familiarization session in which the assessment criteria and a rating sheet were introduced and explained in detail; 3) a practice session during which the raters used the rating sheet to assess two random samples; 4) a norming session in which the raters assessed five pre-anchored performances and discussed results; 5) a pilot session in which the raters assessed four performances in a row.

#### 8.4.4 Procedures

In the simulated assessment, interpreters were asked to perform SI in the four tasks. When they completed an SI task, they took a short break. Performances in all tasks were audio-recorded with consent. The interpretation recordings were then randomly distributed to the trained raters, and were assessed independently. Overall, a fully-crossed measurement design was used in which each rater assessed each interpreter's performance on each SI task, using the three subscales (i.e., 32 interpreters  $\times$  4 tasks  $\times$  9 raters  $\times$  3 criteria). The raters were also provided with the source-language texts to help them check and compare the original information against the renditions. The raters assessed a batch of four recordings before having a short break.

#### 8.4.5 Data analysis

Regarding the G-theory design, the interpreter/person facet (denoted as  $p$ ) was treated as the object of measurement. SI tasks ( $t$ ) and raters ( $r$ ) were defined as the random facets, because both tasks and raters could be regarded as random samples selected from their respective universe of interest. The three assessment criteria ( $c$ ) were modeled as the fixed facet, because these criteria represented different dimensions of SI quality, and were thus not exchangeable with others.

Table 8.1 describes the decomposition of total score variance (take the InfoCom ratings for example) into seven variance components for the univariate G study. The first source of variability, attributable to the object of measurement, arose from systematic individual differences among the interpreters in terms of their performance, which is also known as universe-score variability. The remaining six sources of variability were associated with the

two measurement facets (i.e., tasks and raters), and they introduced errors to sample-to-universe generalization. For example, overall inter-rater inconsistency would increase uncertainty when generalizing from scores given by the particular group of raters to scores provided by a universe of admissible raters. In addition, interaction among interpreters, raters and tasks could also pose potential threats to generalization. For instance, a rater would provide consistently lower-than-warranted scores to a particular interpreter, while another rater would give consistently lenient ratings. Furthermore, in the residual, given only one observation per cell of interpreter-by-task-by-rater matrix, the three-way interaction was confounded with unidentified and/or random errors.

Table 8.1 Decomposition of total variance into variance components for the InfoCom ratings

Source of variability	Description of variance component	Notation
1. Interpreters ( $p$ )	• Universe-score variance (object of measurement)	$\sigma_p^2$
2. SI tasks ( $t$ )	• Main effect for all interpreters due to their performance inconsistency from one task to another	$\sigma_t^2$
3. Raters ( $r$ )	• Main effect for all interpreters due to rater severity	$\sigma_r^2$
4. Interpreter * Task ( $pt$ )	• Interaction effect: inconsistent from one task to another in a particular interpreter's performance	$\sigma_{pt}^2$
5. Interpreter * Rater ( $pr$ )	• Interaction effect: inconsistent rater severity towards a particular interpreter	$\sigma_{pr}^2$
6. Task * Rater ( $tr$ )	• Interaction effect: inconsistent rater severity from one task to another for all interpreters	$\sigma_{tr}^2$
7. Interpreter * Task * Rater ( $ptr$ ), <i>error</i>	• Residual: unique three-way $ptr$ interaction, unidentified and unmeasured facets that potentially affect the measurement, and/or random errors.	$\sigma_{ptr,e}^2$

To address research question 1, a univariate G-theory analysis was conducted, in which a generalizability (G) study with a  $p \times t \times r$  design was first carried out to estimate the variation contributed by the object of measurement, the facets, and their combinations, to the total amount of variation in the observed scores of SI quality ratings, for a situation where only one task and one rater are used for assessment on each rating dimension. Then, decision (D) studies characterized by a  $p \times T \times R$  design were conducted, in which different combinations of tasks

and raters (including the design of four tasks and two raters in the CATTI test) were examined to find an optimal measurement design for achieving a desirable level of score reliability for each rating dimension. For the univariate analysis, EduG 6.1e was used (Cardinet, Johnson, & Pini, 2010), because it greatly simplifies data entry procedures.

To address research questions 2 and 3, a multivariate G-theory analysis was performed. That is, the multivariate G and D studies, denoted as  $p^{\bullet} \times t^{\bullet} \times r^{\bullet}$  and  $p^{\bullet} \times T^{\bullet} \times R^{\bullet}$ , were conducted. Apart from the information that resulted from the univariate analysis, the multivariate G study yields covariance estimates between the rating scales. Since the multivariate D studies are capable of calculating reliability indices for different composite scores based on different weighting schemes, it is therefore possible to examine how score reliability changes as a function of different sets of nominal weights (i.e., research question 2). Moreover, the multivariate approach to the composite score analysis has an additional advantage of producing the effective weights of rating scales for composite universe-score and relative/absolute-error variances. Especially, the effective weights of rating scales for a composite true-score variance provides information on how much empirical or statistical contribution they make to differentiate among examinees based on their true differences in a given ability summarized by a composite (Sawaki, 2007). This type of information makes it possible to examine research question 3. For the multivariate G-theory analysis, mGENOVA 2.1 was used (Brennan, 2001b).

## 8.5 Results

### 8.5.1 Univariate G-theory analysis

#### 8.5.1.1 Univariate G study

Table 8.2 presents the VC estimates from the G study for the baseline situation where only one task and one rater is involved. The VC estimates was then used as baseline measures in subsequent D studies.

Table 8.2 Univariate G study for one task and one rater for each rating dimension

Source of variation	VC	VC estimate, (% of total variance)		
		$c_1 = \text{InfoCom}$	$c_2 = \text{FluDel}$	$c_3 = \text{TLQual}$
Person ( $p$ )	$\sigma_p^2$	1.52, (45.6%)	0.71, (32.4%)	0.68, (33.8%)
Task ( $t$ )	$\sigma_t^2$	0.27, (8.0%)	0.19, (8.9%)	0.03, (1.3%)
Rater ( $r$ )	$\sigma_r^2$	0.16, (4.7%)	0.25, (11.6%)	0.18, (9.0%)
$pt$	$\sigma_{pt}^2$	0.32, (9.5%)	0.18, (8.4%)	0.14, (6.9%)
$pr$	$\sigma_{pr}^2$	0.25, (7.4%)	0.21, (9.4%)	0.22, (10.8%)
$tr$	$\sigma_{tr}^2$	0.05, (1.4%)	0.02, (0.7%)	0.02, (0.9%)
$ptr, e$	$\sigma_{ptr,e}^2$	0.78, (23.4%)	0.62, (28.5%)	0.75, (37.3%)

Note: VC = variance component;  $c$  = criteria

#### 8.5.1.2 Univariate D studies

Based on the VC estimates in Table 8.2, D studies were conducted. One of the designs that are particularly relevant to the study include four SI tasks and two raters, a scenario similar to that of the CATTI level IV SI test. Consequently, the D study for four tasks and two raters was carried out. Table 8.3 shows the VC estimates, indices of dependability and standard error of measurement (SEM) for each rating dimension. As can be seen in Table 8.3, the VC estimates for the Person facet were largest across the rating dimensions, with over 66% of the proportions of the score variances attributable to the ability differences among the interpreters. In contrast, the proportions of variance accounted for by the Task facet, the Person-by-Task and the Task-by-Rater interactions were relatively small. For instance, the VC estimates for the Task-by-Rater interaction were virtually zero across the dimensions. It is also worth noting that the two rater-related sources of variation (i.e., the Rater facet and the Person-by-Rater interaction) had relatively large proportions of variance across the rating dimensions. Specifically, the proportions of score variance attributable to the Rater facet became increasingly larger, from 4.0% for InfoCom to 8.8% for TLQual, and to 11.2% for FluDel. This pattern indicates that averaged over the persons and the tasks, the raters differed in terms of severity/leniency from one rating dimension to another. In addition, the relatively large VC estimates for the Person-by-Rater interaction suggest that there were differences regarding the rank-ordering of the interpreters across the assessment criteria. Lastly, the residual VCs show

that relatively large proportions of variance were attributable to the triple-order interaction and/or unmeasured error that was not captured in the D study design.

As can also be seen in Table 8.3, while the dependability index for InfoCom ( $\Phi = 0.77$ ) was approaching the minimally accepted value of 0.80 (e.g., Cardinet et al., 2010; Shavelson & Webb, 1991), the  $\Phi$  values for FluDel and TLQual (i.e., 0.64 and 0.67) were well below 0.80. Given these results, alternative measurement designs were explored to compare SEMs and dependability indices so as to find an optimal design with a desirable level of score reliability.

Table 8.3 D study results based on four SI tasks and two raters

Source of variation	VC estimate (% of total variance)		
	InfoCom	FluDel	TLQual
Person ( <i>p</i> )	1.52 (76.9%)	0.71 (63.7%)	0.68 (66.7%)
Task ( <i>t</i> )	0.068 (3.4%)	0.048 (4.3%)	0.008 (0.8%)
Rater ( <i>r</i> )	0.08 (4.0%)	0.125 (11.2%)	0.09 (8.8%)
<i>pt</i>	0.08 (4.0%)	0.045 (4.0%)	0.035 (3.4%)
<i>pr</i>	0.125 (6.3%)	0.105 (9.4%)	0.11 (10.8%)
<i>tr</i>	0.006 (0.3%)	0.003 (0.3%)	0.003 (0.3%)
<i>ptr,e</i>	0.098 (5.0%)	0.078 (7.0%)	0.094 (9.2%)
<b>Dependability (<math>\Phi</math>)</b>	0.77	0.64	0.67
<b>Absolute SEM</b>	0.67	0.64	0.58

Note: VC = variance component; SEM = standard error of measurement

### 8.5.1.3 Alternative measurement designs

Alternative combinations of one to six raters and one to eight tasks were explored in order to gain insight to how the number of tasks and/or raters used would affect SEM and score dependability. Figure 8.1 shows the magnitude of SEM as a function of the number of SI tasks and/or raters for absolute (Abs) decision making. As shown, for InfoCom, with the increasing number of tasks, and/or raters,  $SEM_{Abs}$  would be reduced, indicating improved measurement precision. Particularly, for InfoCom, the average contribution made by one additional task to reduce  $SEM_{Abs}$  (i.e.,  $\Delta Task/6 = 0.55/6 = 0.092$ ) was slightly larger than that of one extra rater (i.e.,  $\Delta Rater/6 = 0.46/6 = 0.077$ ). However, this contribution was reversed for FluDel and TLQual.

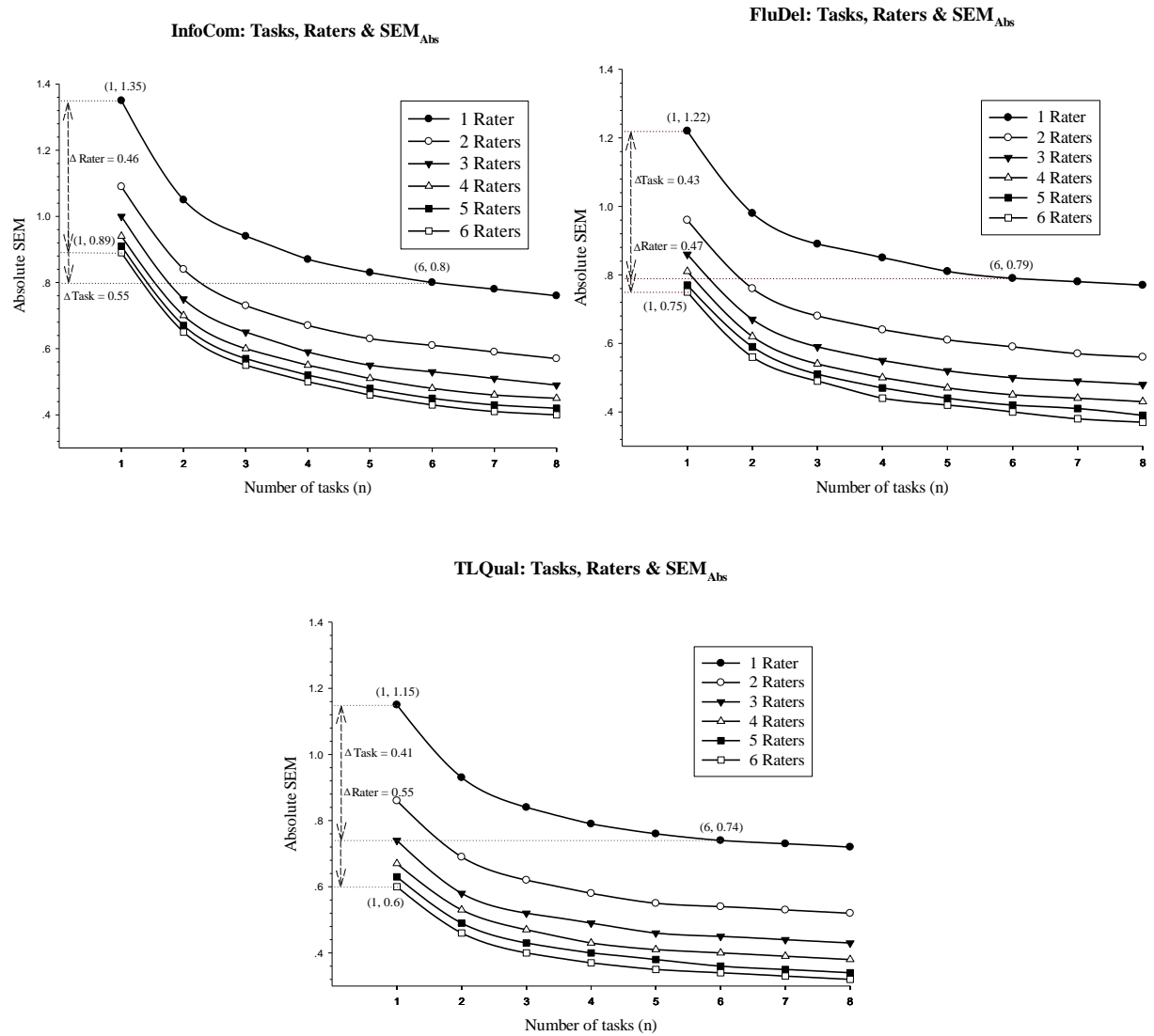


Figure 8.1 SEM<sub>Abs</sub> as a function of No. of tasks and raters

Figure 8.2 displays the magnitude of  $\Phi$  as a function of the number of SI tasks and/or raters. As can be seen in the figure, across the rating dimensions, sampling more tasks and/or using more raters would generally improve score dependability. But the marginal effect became increasingly diminished with more tasks and/or raters, and the  $\Phi$  values seemed to level off beyond a certain point. In addition, for InfoCom, sampling more tasks would do a better job of raising  $\Phi$ ; but for FluDel and TLQual, recruiting additional raters would be more effective. Consequently, a measurement design that helps achieve a desirable level of dependability for one rating dimension may not prove equally effective for the other dimensions. For example, as shown in Figure 8.2, the design of four tasks and three raters helps achieve adequate reliability for InfoCom (i.e.,  $\Phi = 0.81$ ), but would fail to do so for both FluDel and TLQual. For another example, with four tasks, as many as five raters would be

required to achieve a  $\Phi$  value greater than 0.80 for TLQual, but still fall short for FluDel. These results indicate that given the same design the FluDel ratings were least dependable, while the InfoCom ratings were most dependable.

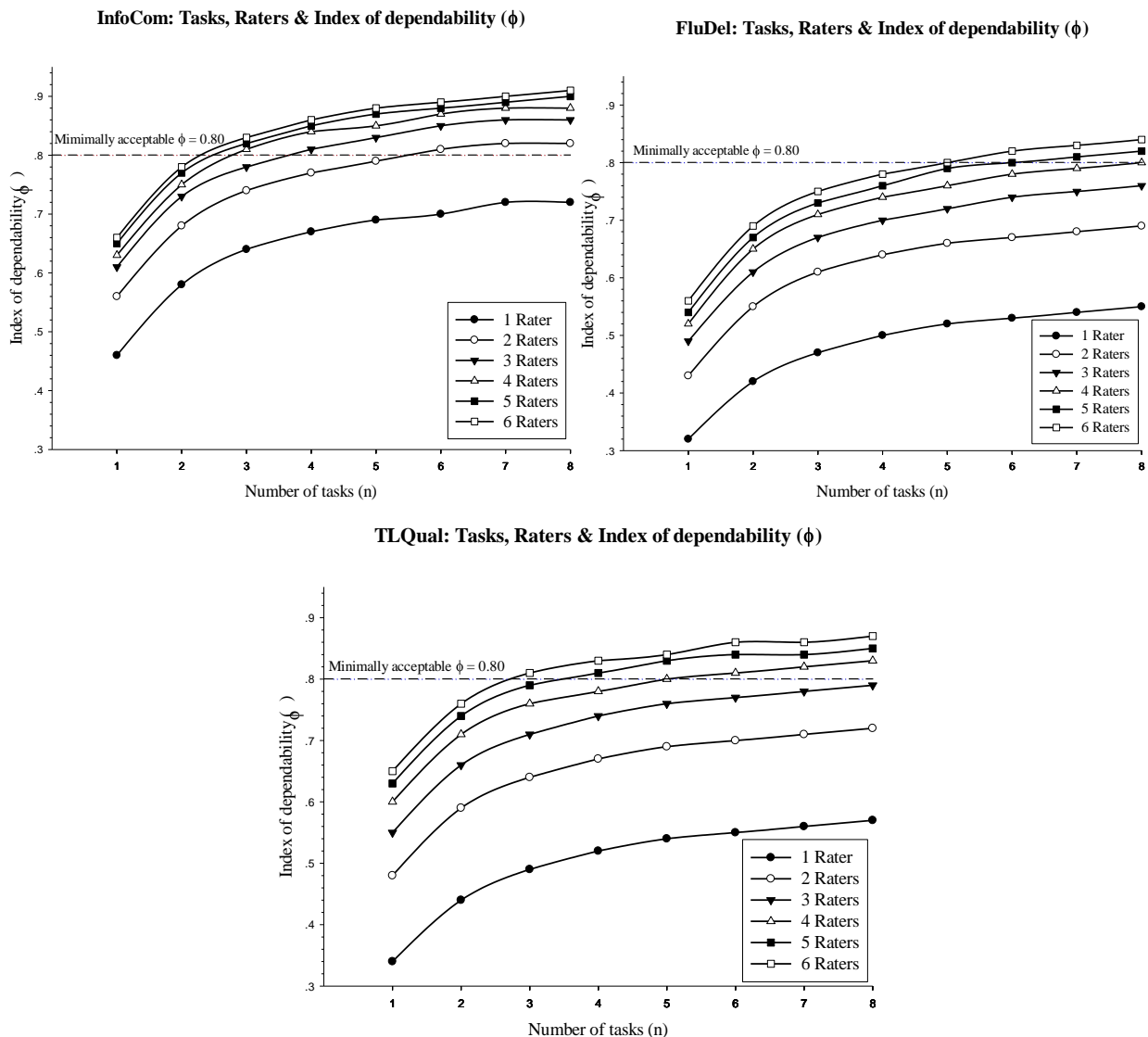


Figure 8.2 Index of dependability as a function of No. of tasks and raters

### 8.5.2 Multivariate G-theory analysis

A multivariate G study was first conducted for one task and one rater. Additional information yielded was covariance component (CC) estimates (see Table 8.4). As can be seen in the table, the CC for persons ( $p$ ) show how persons' universe scores on InfoCom, FluDel and TLQual co-varied with one another. The high CC estimates between InfoCom and FluDel (0.90), as well as between InfoCom and TLQual (0.85) indicate that the persons who performed well on InfoCom also tended to do well on both FluDel and TLQual. The relatively lower CC estimate

between FluDel and TLQual (0.65) suggests a weaker inter-relationship, compared to the previous two.

Table 8.4 Estimated variance-covariance components (VCC) from the multivariate G study for one task and one rater

Source of variation		VCC estimates (% of variance & covariance)		
		InfoCom ( $c_1$ )	FluDel ( $c_2$ )	TLQual ( $c_3$ )
Person ( $p$ )	$c_1$	<b>1.50 (45.5)</b>		
	$c_2$	0.90 (54.5)	<b>0.70 (32.1)</b>	
	$c_3$	0.85 (52.8)	0.65 (41.1)	<b>0.68 (33.7)</b>
Task ( $t$ )	$c_1$	<b>0.26 (7.9)</b>		
	$c_2$	0.13 (7.9)	<b>0.19 (8.7)</b>	
	$c_3$	0.07 (4.3)	0.08 (5.1)	<b>0.03 (1.5)</b>
Rater ( $r$ )	$c_1$	<b>0.15 (4.5)</b>		
	$c_2$	0.00 (0.0)	<b>0.25 (11.5)</b>	
	$c_3$	0.04 (2.5)	0.20 (12.7)	<b>0.18 (8.9)</b>
$pt$	$c_1$	<b>0.32 (9.7)</b>		
	$c_2$	0.20 (12.1)	<b>0.18 (8.3)</b>	
	$c_3$	0.19 (11.8)	0.14 (8.9)	<b>0.14 (6.9)</b>
$pr$	$c_1$	<b>0.24 (7.3)</b>		
	$c_2$	0.13 (7.9)	<b>0.21 (9.6)</b>	
	$c_3$	0.17 (10.6)	0.16 (10.1)	<b>0.22 (10.9)</b>
$tr$	$c_1$	<b>0.05 (1.5)</b>		
	$c_2$	0.00 <sup>a</sup> (0)	<b>0.02 (0.9)</b>	
	$c_3$	0.00 (0)	0.01 (0.6)	<b>0.02 (1.0)</b>
Residual: $ptr, e$	$c_1$	<b>0.78 (23.6)</b>		
	$c_2$	0.29 (17.6)	<b>0.63 (28.9)</b>	
	$c_3$	0.29 (18.0)	0.34 (21.5)	<b>0.75 (37.1)</b>

Note: <sup>a</sup> The negative value of -0.01 was set to 0.

Based on the multivariate G study results, D studies were conducted to investigate research question 2 by comparing the dependability of the composite scores based on the three

weighting schemes (i.e., IoLET, 2010; Kurz, 1989; Yang, 2005) in six measurement designs. The six designs include combinations of four to six tasks and two to three raters. These designs were chosen because in operational ICPTs the number of tasks and raters range from four to six, and two to three, respectively.

Table 8.5 compares the dependability index ( $\Phi$ ) of the composite scores based on the three weighting schemes for the six measurement designs.

Table 8.5 Dependability of composite scores based on the three weighting schemes

<b>Weight</b>	<b>Nominal weights</b>			<b>Dependability (<math>\Phi</math>)</b>					
	InfoCom	FluDel	TLQual	$n_{\text{task}} = 4$	4	5	5	6	6
<b>schemes<sup>a</sup></b>	$(w_1)$	$(w_2)$	$(w_3)$	$n_{\text{rater}} = 2$	3	2	3	2	3
IoLET (2010)	0.33	0.33	0.33	0.75	0.80	0.77	0.82	0.78	0.83
Kurz (1989)	0.38	0.32	0.31	0.76	0.80	0.78	0.82	0.79	0.83
Yang (2005)	0.50	0.30	0.20	0.77	0.81	0.79	0.83	0.80	0.84

Note: <sup>a</sup>  $w_1 + w_2 + w_3 \approx 1$  due to rounding.

As can be seen in the table, firstly, for the design of four tasks and two raters (the same to that of the CATTI test), the composite score dependability was below 0.80 across the weighting schemes. Secondly, compared to the design of four tasks and two raters, using an extra rater (i.e., the design of four tasks and three raters) was more effective in raising the  $\Phi$  value than adding one more task (i.e., the design of five tasks and two raters). More importantly, the design of four tasks and three raters resulted in an acceptable level of composite score reliability across the weighting schemes, while using five tasks and two raters failed to do so. Thirdly, holding the measurement design constant, the  $\Phi$  values did not vary drastically, despite the different sets of nominal weights. In other words, different weighting schemes did not lead to considerable variation in composite score dependability, given the same design.

To address research question 3, the multivariate D studies calculated effective weights or empirical contributions of the respective rating scales to the composite universe-score and absolute-error variances. Table 8.6 summarized the composite score analysis results for the design of particular interest: four tasks and two raters, as it resembles that of the CATTI test.

Table 8.6 Composite score analysis results: Effective weights for the design of four tasks and two raters

Effective weights contributing to	Weight schemes: Nominal weights <sup>a</sup>		
	InfoCom ( $w_1$ )	FluDel ( $w_2$ )	TLQual ( $w_3$ )
	0.33	0.33	0.33
Universe score variance (%)	42.33	29.35	28.32
Absolute error variance (%)	33.30	34.28	32.42
	0.38	0.32	0.31
Universe score variance (%)	47.27	27.28	25.45
Absolute error variance (%)	38.90	31.81	29.29
	0.50	0.30	0.20
Universe score variance (%)	60.51	24.14	15.35
Absolute error variance (%)	54.92	27.66	17.41

Note: <sup>a</sup>  $w_1 + w_2 + w_3 \approx 1$  due to rounding.

As can be seen in Table 8.6, the results show that the InfoCom rating scale accounted for the largest proportions of the composite universe-score variances across the weighting schemes, ranging from 42.33% to 60.51%, and that the empirical contributions of the InfoCom rating scale to the composite absolute-error variance were also largest (38.90%, 54.92%), except for the unit weighting scheme. Taken together, these results indicate that the InfoCom rating scale contributed relatively more information to both the composite universe-score and absolute-error variances, compared to the other scales.

## 8.6 Discussion and implications

The results from the univariate and the multivariate G-theory analyses are worthy of further discussion, especially in light of ICPT.

Firstly, for the measurement design of particular interest (i.e., four tasks and two raters), the relatively large VC estimates associated with the raters as shown by the Rater facet and the Person-by-Rater interaction warrants more discussion. For the Rater facet, the smallest VC estimate was with InfoCom, and the largest with FluDel. This indicates that the raters

showed less variation using the InfoCom rating scale than the FluDel scale. This is probably because in evaluating the InfoCom dimension the raters checked the interpretations against the original source texts line by line. The source texts then provided a consistent reference for rater judgment. When evaluating the FluDel dimension, the raters largely depended on their individually internalized standards of fluency to arrive at a conclusion. As a result, more subjectivity and instability was involved in rater judgment, leading to relatively larger VCs for FluDel than InfoCom. The relatively large Rater VC estimate for FluDel could also be attributed to lack of sufficient practice and norming in the rater training. Enhanced rater training should have been provided so that the raters could use the FluDel rating scale more consistently. Furthermore, the large Rater VC estimate for FluDel emerged probably because the rating scale and its associated rubrics had not been rigorously developed and validated.

Secondly, the D study for four tasks and two raters did not lead to a relatively large Person-by-Task interaction effect, an effect that has been observed in different types of performance assessment (e.g., Brennan, 2000; Mehrens, 1992). The non-existence of the effect is probably because in the present study only one type of interpreting task (i.e., SI) was used, and the four tasks also shared similar topics and other characteristics. In addition, the participants were all Beijing-based interpreters, which could be regarded as a homogenous group. However, the finding does not rule out the possibility that the interaction effect would occur in the real-life ICPTs that sample a small number but divergent types of interpreting tasks that focus on different topics (e.g., FCICE, DPSI, NAATI), and that target potentially heterogeneous groups of test candidates. The ICPTs usually involve a small number of tasks probably due to financial and logistical constraints; they also sample different types of interpreting task and select different topics in order to adequately represent the practice domain of interest. For example, the DPSI test consists of two SI tasks, two CI tasks and two SiT tasks, and each task may have a different topical focus (e.g., mediating in medical encounters, legal hearings, police questioning). Although these types of interpreting task essentially involve oral translation from one language to another, they differ in terms of the interpreting mode and subject matter knowledge. As a result, a large Person-by-Task interaction could emerge in such ICPTs as NAATI and DPSI, which would ultimately affect score-based inferences. Empirical studies therefore need to be initiated to examine the Person-by-Task interaction and its effects on score dependability, based on authentic test scores from the operational ICPTs.

Thirdly, regarding the design of four tasks and two raters, the dependability coefficients for the scores of each rating dimension and for the composite scores were lower than the minimum threshold of 0.80, as can be seen in Table 8.3 and Table 8.5, respectively. The results help raise a concern over score dependability of the CATTI SI test, partly because the test also employs the same measurement design (i.e.,  $n_t = 4$  and  $n_r = 2$ ). It is also because there have been no rigorous empirical investigations into score reliability for the test.

The D-study results would also be of help to the CATTI authority if adequate dependability for scores of each rating dimension needs to be achieved, especially for the less dependable scores of the FluDel and TLQual ratings. Figure 8.2 show that for  $\Phi = 0.80$  to be achieved for the FluDel scores, more than six raters are needed if the number of tasks is held at four (as is currently operationalized in the CATTI SI test); or more than eight tasks are required if the number of raters remains as a constant of two; or a design of five tasks and six raters is used. Each way, it would be infeasible for the CATTI test administration, given usually limited amount of resources. The situation is similar in the case of the TLQual ratings. As a result, it seems that simply adding more raters and/or tasks is not particularly practical for FluDel and TLQual.

To improve the dependability of FluDel and TLQual ratings, some fundamental strategies may instead need to be prioritized. For one thing, the quality of rating scales can be further improved, especially considering the fact that there have been no empirically driven and rigorously validated rating scales (with rubrics) for the purpose of interpreter performance assessment, despite several preliminary studies (e.g., Lee, 2008; Tiselius, 2009). Efforts should therefore be initiated to develop and validate descriptor-based rating scales, particularly for the FluDel and the TLQual criteria. For another thing, rigorous and constant trainings that are generally lacking for ICPT can be institutionalized and provided to raters, as is also suggested by interpreting researchers (e.g., Hale et al., 2012; Roat, 2006; Sawyer, 2004). In the training, raters should be well-informed with rating scales, provided with ample practices, and monitored across different sessions.

Fourthly, given the same measurement designs (see Table 8.5), the  $\Phi$  values based on the different weighting schemes in the multivariate D studies did not differ considerably (i.e., within a range of 0.02). It therefore seems that the different sets of nominal weights could be used interchangeably. However, the higher covariance estimates between InfoCom and FluDel, as well as between InfoCom and TLQual (see Table 8.4) could be interpreted from a

substantive perspective that that InfoCom is the central quality dimension, as the universe scores of other rating dimensions highly co-varied with those of InfoCom. After all, the fundamental purpose of interpreting is to communicate message and information (i.e., InfoCom) between different languages. From a statistical perspective, the composite score dependability was indeed slightly larger when more nominal weights were assigned to InfoCom. This is largely because the InfoCom ratings were more dependable than both those of FluDel and TLQual, as can be seen in Figure 8.2. More importantly, based on the effective weights in Table 8.6, the InfoCom rating scale empirically contributed the largest amount of information to both composite universe-score and absolute-error variances. Consequently, from a substantive point of view, given the centrality of InfoCom in interpretation quality assessment, the rating dimension should be given more weight. From a statistical perspective, it is also desirable to assign relatively more weight to InfoCom, if a composite score needs to be reported.

## **8.7 Limitations and conclusion**

Overall, the study has three limitations. Firstly, the study would gain more insight and produce direct impacts on ICPTs, if the dataset was derived from an operational certification performance test. Secondly, the study only investigates one interpreting task type (i.e., SI tasks), which limits its implications for those tests that sample divergent types of interpreting tasks. Thirdly, given the unavailability of fully validated rating scales for interpreting performance assessment, the scales used in the present study may have introduced scale-related variances, especially in the case of FluDel and TLQual.

Despite the limitations, the study has some main findings. It shows that given the same measurement design, the InfoCom ratings were more dependable than those of FluDel and TLQual. In addition, different weighting schemes do not seem to produce drastically different dependability coefficients for the composite scores. However, more weighting is recommended to be given to InfoCom for both substantive and statistical reasons.

In summary, ICPT is developing with good momentum. But related research on test development and validation seems to be lacking. Given G theory's capability in disentangling multiple sources of error by estimating variance components associated with measurement

facets, and given its informative D study results, it is recommended that G theory be applied for the development of ICPTs to buttress test validity arguments.

## 8.8 References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angelelli, C. (2007). Assessing Medical Interpreters. *The Translator*, 13(1), 63-82.
- Angelelli, C. (2009). Using a rubric to assess translation ability: defining the construct. In Angelelli, C & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies* (pp.13-47). Amsterdam: John Benjamins.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, 12(2), 239-257.
- Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24(4), 339-353.
- Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springer-Verlag.
- Brennan, R. L. (2001b). *Manual for mGENOVA Version 2.1*. Retrieved from: <http://www.education.uiowa.edu/centers/casma/computer-programs#8f748e48-f88c-6551-b2b8-ff00000648cd>
- Brown, J. D. (1984). A norm-referenced engineering reading test. In A.K. Pugh & J.M. Ulijn (Eds.), *Reading for professional purposes: Studies and practices in native and foreign languages* (pp. 213-222). London: Heinemann Educational Books.
- Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, 16(2), 217-238.
- Campbell, S., & Hale, S. (2003). Translation and interpreting assessment in the context of educational measurement. In G. Anderman & M. Rogers (Eds.), *Translation today: trends and perspectives* (pp. 205-224). Clevedon: Multilingual Matters.
- Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Routledge.
- Chen, J. (2011). Language assessment: Its development and future - An Interview with Lyle F.

- Bachman. *Language Assessment Quarterly*, 8(3), 277-290.
- Clifford, A. (2005). Putting the exam to the test: Psychometric validation and interpreter certification. *Interpreting*, 7(1), 97-131.
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit all? *Language Testing*, 26(4), 507-531.
- Gebril, A. (2010). Bringing reading-to-writing and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, 15(2), 100-117.
- Hale, S., Garcia, I., Hlavac, J., Kim, M., Lai, M., Turner, B., & Slatyer, H. (2012). *Development of a conceptual overview for a new model for NAATI standards, testing and assessment*. Retrieved from: <http://www.naati.com.au/PDF/INT/INTFinalReport.pdf>
- Hlavac, J. (2013). A cross-national overview of translator and interpreter certification procedures. *The International Journal for Translation & Interpreting Research*, 5, 32-65. DOI: 10.12807/ti.105201.2013.a02
- Jacobs, E. A., Lauderdale, D. S., Meltzer, D., Shorey, J. M., Levinson, W., & Thisted, R. A. (2001). Impact of interpreter services on delivery of health care to limited-English-proficient patients. *The Journal of General Internal Medicine*, 16(7), 468-474.
- Kurz, I. (1989). Conference interpreting - User expectations. In D. L. Hammond (Ed.), *Coming of Age: Proceedings of the 30th Annual Conference of the American Translators Association* (pp. 143-148). Medford, NJ: Learned Information.
- Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, 2(2), 165-184.
- Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing*, 7(4), 353-385
- Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23(2) 131-166.
- Liu, M. H. (2013). Design and analysis of Taiwan's Interpretation Certification Examination. In Tsagari, D & R. van Deemter (Eds.), *Assessment Issues in Language Translation and Interpreting* (pp. 163-178). Frankfurt am Main: Peter Lang.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.

- Mehrens, W. A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice*, 11(1), 3-9.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Roat, C. E. (2006). *Certification of Health Care Interpreters in the United States: A Primer, a Status Report and Considerations for National Certification*. Retrieved from: [http://www.calendow.org/uploadedFiles/certification\\_of\\_health\\_care\\_interpreters.pdf](http://www.calendow.org/uploadedFiles/certification_of_health_care_interpreters.pdf)
- Russell, D., & Malcolm, K. (2009). Assessing ASL–English interpreters: The Canadian model of national certification. In Angelelli, C. V. & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 331-376). Amsterdam: John Benjamins.
- Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, 24(3), 355-390.
- Sawyer, D. B. (2004). *Fundamental aspects of interpreter education: Curriculum and Assessment*. Amsterdam & Philadelphia: John Benjamins.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park: Sage.
- The Institute of Linguists Educational Trust. (2010). *Diploma in Public Service Interpreting: Handbook for candidates*. Retrieved from: <http://www.iol.org.uk/qualifications/DPSI/Handbook/DPSIHB11.pdf>
- Tiselius, E. (2009). Revisiting Carroll's scales. In Angelelli, C. V. & H. E. Jacobson (Eds.), *Testing and Assessment in Translation and Interpreting Studies: A Call for Dialogue between Research and Practice* (pp. 95-121). Amsterdam: John Benjamins.
- Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 40(5), 663-705.
- Wu, S. C. (2010). *Assessing Simultaneous Interpreting: A study on Test reliability and Examiners' assessment behavior*. (PhD thesis, Newcastle University upon Tyne, UK). Retrieved from <https://theses.ncl.ac.uk/dspace/bitstream/10443/1122/1/Wu%2011.pdf>
- Xi, X. M. (2007). Evaluating analytic scoring for the TOEFL<sup>®</sup> Academic Speaking Test (TAST) for operational use. *Language Testing*, 24(2), 251-286.
- Yang, C. S. (2005). 口译教学研究 – 理论与实践. [Studies on Interpreter Training:

- Theories and Practices]. Beijing: China Translation and Publishing Corporation.
- Yu, D. R. (2005). *T&I Labor Market in China*. Retrieved from:  
[http://www.ling.mq.edu.au/translation/lmtip\\_china.htm](http://www.ling.mq.edu.au/translation/lmtip_china.htm)
- Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing*, 23(3), 351-369.

## Chapter 9 Summary and Conclusions

### 9.1 Introduction

This final chapter recapitulates the key findings of the research, integrates the empirical and methodological findings to the theoretical ICPT validity argument, presents the research contributions, acknowledges the limitations, examines the implications of the key findings, provides recommendations for further research, and concludes the whole thesis.

### 9.2 Summary of the thesis

As sketched out in the Introduction (Chapter 1), the practice of *interpreter certification performance testing* (ICPT) is growing, with different types of *interpreter certification performance tests* (ICPTs) being developed and administered across the world. Notwithstanding the pervasiveness of ICPTs, it seems that there has been very limited academic research conducted to explore and examine foundational aspects of ICPT, such as reliability and validity. This interdisciplinary mixed-methods research was therefore initiated to build a validity foundation for an ICPT. In particular, it attempted to address fundamental questions facing ICPT, including construct definition and validation of ICPTs. Three major research questions (RQs) are posed and answered in the thesis. The three RQs are:

- 1) What are the characteristics of the real-life English/Chinese conference interpreting practice in China?
- 2) What is the possible interplay between characteristics of simultaneous interpreting (SI) tasks, interpreting ability, and SI performance quality?
- 3) How multifaceted Rasch measurement and generalizability theory can be incorporated into interpretation testing and assessment to investigate rater severity/leniency and score variability?

A multi-phase mixed-methods research (MMR) design was implemented to investigate the research questions (RQs) in a sequential manner. For the theoretical part of the research, the argument-based approach to test validation and the interactionalist approach to construct definition were imported from language testing and assessment research to inform theoretical discussions of ICPT. For the methodological part of the research, the multifaceted Rasch

measurement (MFRM) and the generalizability (G) theory were introduced and demonstrated to interpreting researchers and testers to help analyze rater variability and its effects on score reliability in ICPT. For the empirical part of the research, the conference interpreting practice in China, especially the characteristics of SI tasks, was empirically profiled through a diary (n = 11) and a follow-up survey (n = 140). The interactions between the three components in the construct model were also investigated by conducting a factorial experiment (n = 32).

Although the research findings and implications of Chapters 2, 3, 4, 5, 6, 7, and 8 have been summarized in individual chapters, this section integrates and synthesizes the key research findings for the thesis.

Chapter 2 responded to the lack of systematic validation research for ICPTs by building the validity argument. The chapter provides the rationale for validation of ICPTs, tracks the evolution of validity theory, and ultimately elaborates a roadmap designed to help ICPT testers collect validity evidence to support intended test score interpretations and uses. In particular, the chapter identifies two potential weaknesses in the ICPT validation: 1) the lack of a construct theory for ICPTs, which is expected to impute substantive meanings (both trait- and performance-referenced) to test scores, and 2) the lack of up-to-date methodologies to gain in-depth understandings of rater variability and its effects on score reliability for the rater-mediated ICPTs. The two potential weaknesses therefore became the topics of subsequent studies.

Chapter 3 attempted to theorize a construct model for ICPTs, based on an interactionalist approach to construct definition (e.g., Chapelle, 1998), and on literature from Interpreting Studies. Consisting of two main components, namely, characteristics of simultaneous interpreting (SI) tasks and interpreting ability, the construct model hypothesized that interpreting performance was a function of SI tasks, interpreting ability and the interactions between them. The model has the potential to help organize testers' thoughts on design of ICPTs, and to justify the trait- and behaviourist-based approach to test score interpretations (i.e., both trait- and performance-referenced). However, for the model to be useful for the design of the ICPTs in China, answers needed to be provided to at least two questions: 1) What are the characteristics of SI tasks in the real-life interpreting practice in China? and 2) What is the relationship between task characteristics, interpreting ability, and interpreting performance quality?

Chapter 4 therefore sought to profile characteristics of interpreting practice in China, by using a diary and a survey. The main results include: 1) the conference-related materials such as PPT and draft speech texts were received by the interpreters most frequently, 2) although the interpreters were found to perform a greater variety of SI tasks, seven specific task varieties were significantly more frequently performed, for example, SI (DiaIntr), SI with PPT (ShortAbun) and SI with Text (ShortAbun), 3) fast speech rate (FSR) and strong accent (StrA) were found to be among some of the most frequent factors contributing to SI difficulty.

Chapter 5 and Chapter 6 explored the relationship between task characteristics, interpreting ability, and interpreting performance. Specifically, Chapter 6 reported the results from the investigation into the effects of SI task characteristics (represented by FSR and StrA), based on part of the experiment data. Overall, the study showed a pattern of mixed impacts of the speed factor on InfoCom, FluDel and TLQual dimensions of SI performance, and a consistent pattern of detrimental effects of the accent factor across the dimensions.

Chapter 6 reported the results from the preliminary analysis of the effects of task characteristics on strategy use, and of the relationship between strategy use and interpreting performance. In general, it was found that in English-to-Chinese SI the interpreters utilized a variety of interpreting strategies, but employed syntactic transformation and substitution most frequently. They also used strategy clusters to cope with complex source-language segments. In addition, it seems that the source-language speech rates considerably affected how the strategies of syntactic transformation and substitution were used in English-to-Chinese SI, but the accent factor did not produce the same effect. Furthermore, the preliminary results suggest that the more strategies were used, the better performance was. But the positive effect did not hold across the strategies used.

Chapter 7 and Chapter 8 shifted the focus to rater/score reliability in the experiment. Concentrating on the rater variability, Chapter 7 sought to help interpreting testers gain insightful understandings of complexities and nuances of rater behavior in ICPTs, by applying multifaceted Rasch measurement (MFRM) to the quantitative experiment data. It was found that R08 was most problematic, as s/he was internal self-inconsistent and significantly biased toward a large proportion of the interpreters, the SI tasks and the criteria.

Chapter 8 focused on the score reliability, especially the effects of the number of raters and tasks used on score dependability, from the perspective of generalizability (G) theory. The quantitative experiment data was again analyzed. Overall, the results showed that 1) the score

dependability for the operational measurement design (i.e., four tasks and nine raters) was high, and 2) although increasing the number of raters and/or tasks would help improve score reliability, the effect was not equal across the criteria.

### 9.3 Linking the empirical and methodological findings to the theoretical ICPT validity argument

This section links the empirical studies in Chapters 4, 5 and 6 and the methodological explorations in Chapters 7 and 8 to the overarching theoretical validity argument proposed in Chapter 2. The linking is also visually displayed in Figure 9.1.

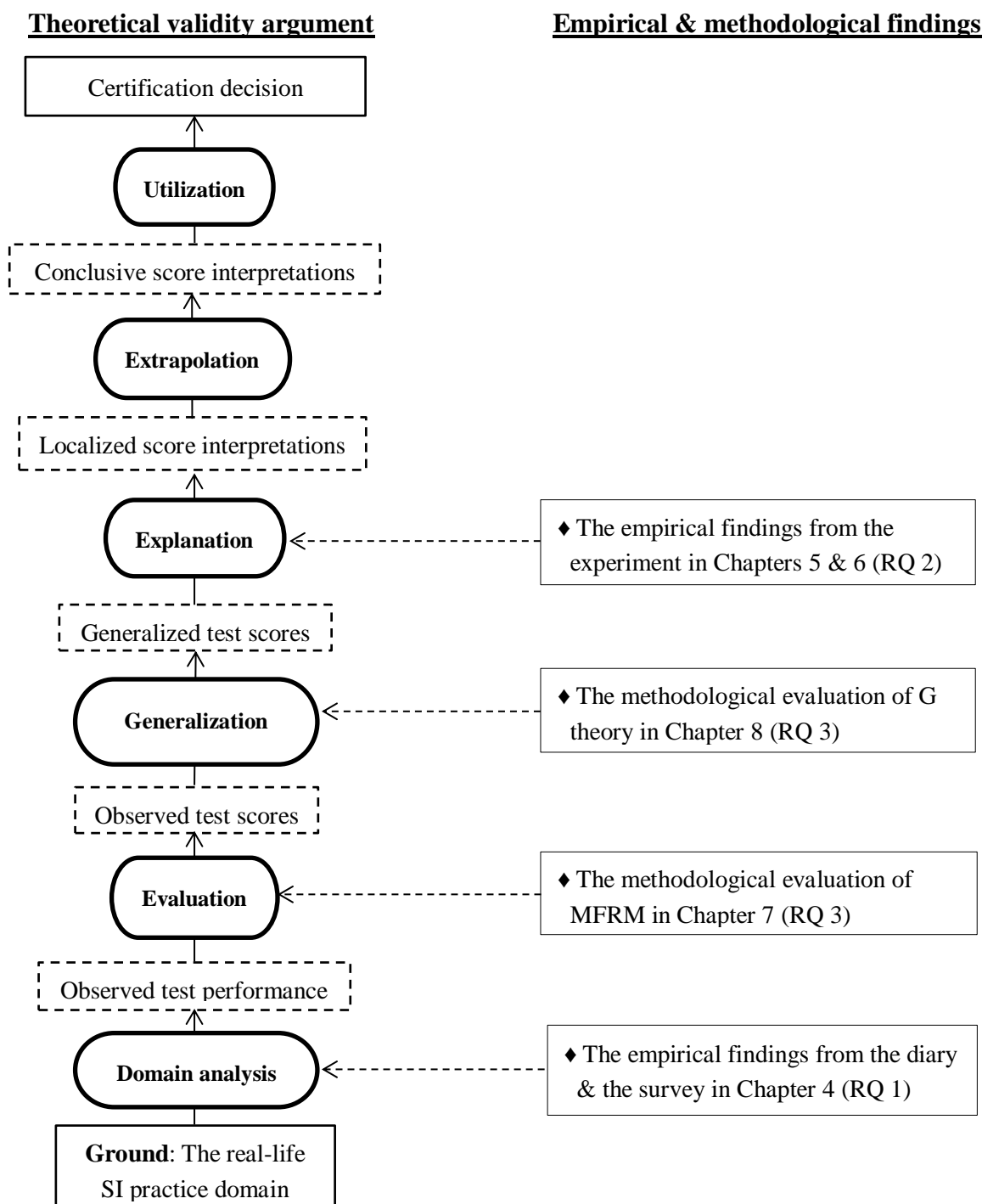


Figure 9.1 A visual display of linking the empirical and methodological findings to the theoretical validity argument

As can be seen in the figure, the domain analysis and modeling inference requires that empirical data describing real-life interpreting practice domains be collected and used to design ICPTs (see Chapter 2). The findings generated from the diary study and the survey reported in Chapter 4 contributed preliminary empirical evidence describing the interpreting practice domain in China.

To support the evaluation inference, one piece of validity evidence, among others, must show that rater training is conducted effectively, and rater behavior is analyzed carefully (see Chapter 2). The methodological exploration of multifaceted Rasch measurement (MFRM) in Chapter 7 revealed its huge potential for contributing to effective rater training by providing an in-depth analysis of nuanced rater behavior. The application of MFRM in the operational ICPT is expected to generate robust validity evidence to the evaluation inference.

To legitimate the generalization inference, empirical evidence needs to show that systematic variance of test scores is largely attributable to examinees, not to raters, tasks, and their interaction effects, and test scores are generalizable across an admissible sample of raters and tasks (see Chapter 2). The methodological exploration of G theory in Chapter 8 demonstrated an advanced analytic technique that could be utilized to examine the effects of multiple assessment facets on score generalizability and to produce indices of generalizability and dependability for test scores. The use of G theory in the operational ICPT can generate rigorous statistical validity evidence that enhances the generalization inference.

Finally, the empirical findings derived from the factorial experiment and reported in Chapters 5 and 6 to address RQ 2 contributed to a better understanding of the interactionist construct model that could be used to impute both trait- and performance-referenced meanings to test scores, thus strengthening the explanation inference in the validity argument.

#### **9.4 Strengths, contributions and limitations of the research**

This section discusses strengths, contributions and limitations of the research. While the strengths may inform design of other similar studies and contribute to Interpreting Studies, the

limitations represent the problematic areas future research should be aware of and hopefully address.

#### 9.4.1 Strengths and contributions of the research

Overall, the research has three strengths. The first strength is that the research was interdisciplinary, building on the literature from Interpreting Studies and language testing and assessment. While testing-related literature provided new perspectives, expanded frameworks and sophisticated psychometric models, literature from Interpreting Studies provided concrete materials to flesh out testing and assessment theories. The cross-fertilization has thus contributed to broadening the field of language testing on the one hand, and generating in-depth discussion within the field of Interpreting Studies on the other.

The second strength of the research is the use of various MMR designs to gain a better understanding of the phenomenon of interest. For example, in Chapter 4, to profile conference interpreting practice in China, the sequential-exploratory MMR design was used, in which the diary and the survey generated both qualitative and quantitative descriptions of the interpreting practice. For another example, in Chapter 5, a convergent-parallel MMR design combining the quantitative and qualitative approaches was implemented to gain both emic descriptions and general statistical patterns of the effects of FSR and StrA on interpreting performance. The implementation of rigorous MMR designs has therefore contributed to Interpreting Studies by demonstrating a means to cross-validate results from different methodological paradigms, and providing opportunities for further reflections on research results when inconsistencies occur between quantitative and qualitative findings.

The final strength of the research has to do with the experimental design and the measurement design. In particular, a strong quantitative experimental design was operationalized through a 2×2 factorial repeated-measures experiment. This type of experiment is regarded as one of the most robust designs to establish a cause-and-effect relationship between independent variables and dependent variables (Johnson & Christensen, 2012). In addition, the four SI tasks used in the experiment were carefully designed and calibrated, based on the framework of SI task characteristics. By doing so, the effects of the extraneous variables that may confound the interpretations of the experiment results were minimized as much as the research sources could afford. Furthermore, all the experiment participants were active SI

practitioners, and the majority of them had a postgraduate interpreting degree, which compares well to other similar studies.

Regarding the measurement design, a fully-crossed measurement design was implemented, in which each rater provided three ratings to each performance by each interpreter. A total of 3456 data points were thus generated (i.e.,  $32 \text{ interpreters} \times 4 \text{ tasks} \times 9 \text{ raters} \times 3 \text{ measures}$ ). Compared to previous similar studies, the present design had the potential to generate more accurate and reliable measurement outcomes. The enhanced measurement precision and reliability were reflected by the relatively low estimates of standard error of measurement (SEM), and the high values of dependability indices ( $\rho^2$ ) derived from the G-theory analysis. In summary, the use of the experimental design, the measurement design, and the G-theory analysis represents a methodological contribution to Interpreting Studies. Similar studies in future can draw upon these strengths to better answer research questions.

#### 9.4.2 Limitations of the research

Despite the strengths outlined above, the research has a number of limitations. Specifically, the research findings are limited in the following five aspects. First, the sample sizes constitute one of the major limitations. In the survey, although a relatively large sample ( $n = 140$ ) was obtained, the sample may not adequately represent the target population, due to the non-probability sampling. In the experiment, although a total of 32 interpreters participated, they were all from Beijing and differed in terms of their SI experience. Interpreters working in other large cities in China (e.g., Shanghai) and highly experienced interpreters (e.g., AIIC members) could not be recruited due to the constraint of research sources. The difficulty of obtaining sufficient and properly-controlled data is also lamented by Gile (1998) and Liu (2011) and Pöschhacker (2004).

Second, to seek a better understanding of the construct model, authentic data from the real-life ICPTs would provide more insight. This type of data, however, was not available for the present research. The researcher was also unable to have access to any authentic test data for analysis, despite multiple attempts made by the researcher to search for such data. This is probably because such data is usually kept confidential by certifying organizations, and external access is thus generally restricted.

Third, in generating the empirical evidence for the construct model, although the effects of FSR and StrA on interpreting performance were investigated through the MMR design, the

remaining interactions (i.e., in Chapter 6) were not rigorously examined, particularly the relationship between the interpreters' strategy use and the quality of their interpreting performance. This is partly because only eight higher-performing interpreters' data were analyzed and partly because there has not been a reliable and accurate method to measure interpreters' strategic competence (i.e., strategy use). Nevertheless, there were some emerging results that indicate a general positive relationship, but more robust studies are needed to reveal the underlying or the lack of association.

Fourth, in assessing SI performance, the three rubrics-based rating scales used by the raters were not empirically developed and validated. Although some preliminary Rasch-based statistics supported the proper functioning of the rating scales, the content of the rubrics in the scales needs to be justified based on rigorous empirical evidence. As a result, the scales may have affected the measurement outcomes by introducing rater-scale interactions. In addition, the use of postgraduate students as raters could be controversial, in that they lack sufficient interpreting experience and exposure to SI practice. The use of student interpreters was due to such practical reasons as unavailability of a relatively large number of experienced raters and lack of sufficient funding for rater recruitment. In the research, to offset the potential drawbacks of using students as raters, a five-hour rater training was thus provided (see Chapter 7).

Finally, in the experiment, only two characteristics, namely, speech rate and accent, were investigated for uni-directional SI (i.e., English-to-Chinese). This decision was made for two reasons: firstly, the limited resources for this PhD research, particularly the available funding, constrained the scope of the research; and secondly, by including bi-directional SI (i.e., SI between English and Chinese) in the repeated-measures experiment, there would be as many as eight treatment conditions, which is challenging for both the researcher and potential participants. On the one hand, operationalizing such an experimental design and developing properly-controlled source-language texts in each treatment condition would represent a daunting task for the researcher. On the other hand, asking interpreters to perform in eight SI tasks would introduce potential threats to internal validity such as attrition (e.g., loss of participants due to heavy workload) and altered cognitive state (e.g., tiredness due to repeated measurement).

## 9.5 Implications and recommendations

This section discusses the implications of the research findings for Interpreting Studies, interpreter education, and language testing, particularly ICPT.

First, the research has methodological implications for Interpreting Studies, in that it has identified several elements of good practice for implementing a quantitative experimental design: 1) using a repeated-measures experiment design works well with a small sample size, and it uses participants as their own controls. Given that small sample sizes are typical in interpreter-related experiments (Gile, 1998), the repeated-measures design is recommended for future researchers; 2) it is important to ensure that a complete dataset is collected for analysis, because missing data causes problems to inferential statistical analysis and may further reduce already small sample sizes. It is recommended that precautionary methods be taken to guard against the loss of data; 3) in developing interpreting tasks to be used in experiments, both qualitative and quantitative indicators could be produced to characterize independent variables, and to maintain consistency of EVs across tasks. It is recommended that sufficient efforts be invested to control extraneous variables (EVs) that may produce confounding effects on experiment results (e.g., Dillinger, 1989). 4) in the case of using raters to assess interpreting performance, it is recommended that multiple raters be used to reduce rater-related measurement error. As can be seen in Chapters 8, to some degree, the increase of the number of raters would contribute to a higher value of dependability index. It would also be desirable to calculate G coefficients to evaluate the measurement reliability.

Second, the research findings in Chapter 6 could have implications for interpreter education. As indicated by the findings, overall the eight higher-performing interpreters employed a variety of interpreting strategies in a flexible manner. For example, the interpreters were able to use a sequential combination of strategies to address the potential problems effectively. In addition, in English-to-Chinese SI, due to the left-branching structure in Chinese, syntactic transformation was most frequently used by the interpreters. Interpreter trainers may need to develop proper SI materials, and provide specific training to students to master how to perform syntactic transformation in some special circumstances. Interpreter educators may also need to focus on training students to employ strategy clusters to solve processing difficulties in SI.

Third, the research also has implications for the general field of language testing, particularly ICPT. For interested language testing researchers, the research provides an

overview of the status quo of ICPT. The research also enables language testers to appreciate the difficulty of assessing two languages simultaneously in a single test, and to understand the challenges and opportunities facing ICPT. Hopefully, more language testers would show interest in assessing interpreters, and collaborate with interpreting researchers to advance ICPT.

For interpreting researchers and testers, the research findings are meaningful in the following four specific aspects. To begin with, the proposed assessment use argument (AUA) could be adjusted and adapted for different purposes of interpreter performance testing. Next, the empirical data describing interpreting practice in China could be used to help interpreting testers to revisit and re-analyze the design of the current ICPTs, in terms of the relevance and representativeness of test content. The interactionalist construct model also helps testers to re-think the degree of interactiveness between test takers and ICPTs, and how test tasks could be developed and characterized to elicit and engage desirable knowledge, skills and strategies of test candidates. Finally, for interpreter certification organizations, the research findings indicate a need to think carefully about the number of raters and tasks to be used in high-stakes testing, and provide effective rater training. Particularly, given that certification bodies often have limited financial and logistical resources, and may not have in-house experts to test the generalization inference, a balancing act perhaps is critical. Although it is generally advisable not to sacrifice validity for the sake of reliability (Brennan, 2000), the decision to emphasize on one test quality over the other depends ultimately on specific contexts and purposes. On the one hand, it is generally desirable to sample different types of tasks in a certification test so as to approximate the real-life domain of interest. On the other hand, it is necessary that raters be adequately trained before actual rating. In addition, at least a *third-rater adjudication* procedure should be implemented, as suggested by Angelelli (2009) and also practiced in some ICPTs (see Liu, 2013).

## **9.6 Future research**

This research has identified a number of areas that merit further research to advance ICPT. The potential research pertains to two broad areas of ICPT: test design and development, and test validation.

For better test design and development, more efforts could be focused on construct definition for ICPT. Regarding the interactionalist construct model, it is important to obtain

more knowledge of characteristics of real-life interpreting practice through empirical “domain analysis”, and to gain in-depth understandings of how different interpreting tasks engage relevant knowledge, skills, abilities and strategies in order to inform sound test design.

In addition, interpreting testers could follow established test development procedures in language testing and assessment (e.g., Bachman & Palmer, 1996) to write test tasks and assemble these tasks into a test. When doing this, several questions need to be answered: how can test specifications be developed and standardized, how test content relevance and representativeness could be established, and how difficulty of multiple forms of a test could be made comparable.

Given the lack of an empirically developed and validated rubrics-based rating scale for high-stakes interpreter testing, interpreting researchers may also need to focus more attention on developing reliable and easy-to-use assessment rubrics that capture the test construct of interest. Although there are a number of rating scales available in the interpreting literature (e.g., Lee, 2008; Tiselius, 2009; Wang et al., 2015), it seems that rigorous studies have not been initiated to empirically demonstrate their reliability, validity, and utility. Much work therefore needs to be done.

Furthermore, rater variability in ICPTs needs to be further researched, as it constitutes one of the major sources of measurement error. A series of substantive questions need to be answered in the context of ICPT. For example, is it justifiable to have raters to assess interpretations into their B language(s)? Do raters who have different L1 behave similarly when assessing interpretations into and from their L1? How can rater training be effectively conducted to reduce rater variability?

For the test validation, empirical studies could be initiated by certifying authorities to rigorously investigate the validity of score-based inferences and uses, following the proposed validation roadmap in Chapter 2 (i.e., the AUA). Such validation studies would serve as a self-interrogation tool to examine test quality and credibility. Lack of validity evidence to support a given inference indicates that further research could be done, and a particular aspect of a test could be improved or revamped.


## 9.7 References

- Angelelli, C. (2009). Using a rubric to assess translation ability: Defining the construct. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 13-47). Amsterdam: John Benjamins.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement*, 24(4), 339-353.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70), Cambridge, UK: Cambridge University Press.
- Dillinger, M. L. (1989). *Component Processes in Simultaneous Interpreting* (Doctoral thesis, McGill University, Canada). Retrieved from [http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object\\_id=39215&local\\_base=GEN01-MCG02](http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object_id=39215&local_base=GEN01-MCG02)
- Gile, D. (1998). Observational studies and experimental studies in the investigation of conference interpreting. *Target*, 10(1), 69-93.
- Johnson, B., & Christensen, L. (2012). *Educational Research: Quantitative, qualitative and mixed approaches* (4th ed.). Thousand Oaks: Sage.
- Lee, J. (2008). Rating scales for interpreting performance assessment. *The Interpreter and Translator Trainer*, 2(2), 165-184.
- Liu, M. H. (2011). Methodology in interpreting studies: A methodological review of evidence-based research. In B. Nichoemus & L. Swabey (Eds.), *Advances in Interpreting Research: Inquiry in action* (pp. 85-119). Amsterdam: John Benjamins.
- Liu, M. H. (2013). Design and analysis of Taiwan's interpretation certification examination. In D. Tsagari & R. van Deemter (Eds.), *Assessment issues in language translation and interpreting* (pp. 163-178). Frankfurt: Peter Lang.
- Pöchhacker, F. (2004). *Introducing Interpreting Studies*. Shanghai: Shanghai Foreign Language Education Press.
- Tiselius, E. (2009). Revisiting Carroll's scales. In C. Angelelli & H. E. Jacobson (Eds.), *Testing and assessment in translation and interpreting studies* (pp. 95-121). Amsterdam: John Benjamins.

Wang, J-H., Napier, J., Goswell, D., & Carmichael, A. (2015). The design and application of rubrics to assess signed language interpreting performance, *The Interpreter and Translator Trainer*, 9(1), 83-103.

## Appendices

### Appendix A: Macquarie University guidelines for thesis by publication

GUIDELINE	
	<b>Higher Degree Research Thesis By Publication Guideline</b>
<b>Purpose</b>	<p>This Guideline provides information to assist Higher Degree Research (HDR) candidates in the preparation of a thesis including published or co-published material prepared during candidature. A thesis prepared in journal article format adds value to the research student experience, encourages timely completion, enhances job prospects and improves the publication outputs and research ranking of the University.</p>
<b>Guideline</b>	<p><b>Eligible Material</b></p> <p>A thesis by publication may include relevant papers, including conference presentations, which have been published, accepted, submitted or prepared for publication for which at least half of the research has been undertaken during enrolment. The papers should form a coherent and integrated body of work, which should be focused on a single thesis project or set of related questions or propositions. These papers are one part of the thesis, rather than a separate component (or appendix).</p> <p><b>Contribution by Co-Authors</b></p> <p>These papers may be single author or co-authored. The candidate must specify his/her specific contribution. The contribution of others to the preparation of the thesis or to individual parts of the thesis should be specified in the thesis Acknowledgments and/or in relevant footnotes/endnotes. Where a paper has multiple authors, the candidate would usually be the principal author and evidence of this should appear in the appropriate manner for the discipline. Examiners can then assess if the quality and extent of the candidate's contribution warrant the award of the degree based on the standard criteria.</p> <p><b>Number and Presentation of Papers</b></p> <p>Each discipline will have a different number of publications that are acceptable as the substantive foundation for a thesis by publication. As a general rule a candidate will need to have enough papers to support the important findings from the research, presented in a logical and coherent way. Most theses by publication have between 2 and 8 papers in combinations of sole and co-authored papers. These papers will normally form thesis chapters and the chronological publication order may be quite different from the way they are sequenced in the thesis.</p> <p>The length of the papers will reflect discipline requirements and journal guidelines. Although it is not necessary to reformat published works in a thesis, it is not enough simply to bind these publications together. The candidate needs to include a critical introduction to the work, sections that link the papers together, and a concluding section that synthesises the material as a</p>

	<p>whole. Above all, candidates must consider the coherence of the thesis as a whole, and the way in which each paper contributes to the overall thesis.</p> <p><b>Preparing for a Thesis by Publication</b></p> <p>Candidates and supervisors should plan a thesis by publication in relation to the timetable of the individual project and the writing conventions and publishing schedules of their discipline in order to make sure that research, writing and journal submission can be undertaken within standard candidature. For instance, in some science disciplines major journals have 10 editions in a year, whereas the major journals in education may publish biannually.</p> <p>Although a thesis by publication may contain some repetition, it is expected that the repetition be minimal so as to facilitate the examination process. Candidates should ensure that any referencing and stylistic inconsistencies between papers are minimised to assist the examiners.</p>
--	--

<b>Contact Officer</b>	Dean, Higher Degree Research
<b>Date Approved</b>	28 November 2013
<b>Approval Authority</b>	Higher Degree Research Committee
<b>Date of Commencement</b>	25 July 2014
<b>Amendment Dates</b>	28 November 2013 – revised guideline approved by Higher Degree Research Committee
<b>Date for Next Review</b>	25 July 2017
<b>Related Documents</b>	<p>Higher Degree Research Thesis Preparation, Submission and Examination  <a href="#">Policy</a> / <a href="#">Procedure</a></p> <p><b>Links</b>  <a href="http://www.hdr.mq.edu.au/">http://www.hdr.mq.edu.au/</a>  <a href="http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_preparation">http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_preparation</a>  <a href="http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_submission">http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_submission</a>  <a href="http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_submission">http://www.hdr.mq.edu.au/information_for/current_candidates/thesis_submission</a>  <a href="http://www.hdr.mq.edu.au/information_for/thesis_examiners">http://www.hdr.mq.edu.au/information_for/thesis_examiners</a></p>
<b>Keywords</b>	Thesis by Publication, thesis with journal articles, thesis with papers, thesis co- authors, thesis co-publication.

## **Appendix B: Final ethics approval**

RE: HS Ethics Final Approval (5201200443) (Condition met)

Fhs Ethics fhs.ethics@mq.edu.au Mon, Aug 13, 2012 at 2:50 PM

To: Ms Helen Marjorie Slatyer <helen.slatyer@mq.edu.au>

Cc: Mr Chao Han <chao.han2@students.mq.edu.au>

Dear Ms Slatyer,

Re: "Building the Validity Foundation for Interpreter Certification Performance Testing"

Thank you for your recent correspondence. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee and you may now commence your research.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

[http://www.nhmrc.gov.au/\\_files\\_nhmrc/publications/attachments/e72.pdf](http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf).

The following personnel are authorised to conduct this research:

Mr Chao Han

Ms Helen Marjorie Slatyer

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).
2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 13th August 2013

Progress Report 2 Due: 13th August 2014

Progress Report 3 Due: 13th August 2015

Progress Report 4 Due: 13th August 2016

Final Report Due: 13th August 2017

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

[http://www.research.mq.edu.au/for/researchers/how\\_to\\_obtain\\_ethics\\_approval/human\\_research\\_ethics/forms](http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms)

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Sub-Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

[http://www.research.mq.edu.au/for/researchers/how\\_to\\_obtain\\_ethics\\_approval/human\\_research\\_ethics/forms](http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/forms)

5. Please notify the Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University.

This information is available at the following websites:

<http://www.mq.edu.au/policy>

[http://www.research.mq.edu.au/for/researchers/how\\_to\\_obtain\\_ethics\\_approval/human\\_research\\_ethics/policy](http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy)

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have final approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of Final Approval to an external organisation as evidence that you have Final Approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of final ethics approval.

Yours sincerely,

Dr Peter Roger

Chair

Faculty of Human Sciences Ethics Review Sub-Committee

Human Research Ethics Committee

\*\*\*\*\*

Faculty of Human Sciences - Ethics

Research Office

Level 3, Research HUB, Building C5C

Macquarie University

NSW 2109

Ph: +61 2 9850 4197

Fax: +61 2 9850 4465

Email: [fhs.ethics@mq.edu.au](mailto:fhs.ethics@mq.edu.au)

<http://www.research.mq.edu.au/>

## Appendix C: Part of a completed diary form

Conference preparation					
No.	Received any conference-related materials or information?	Language (Chinese, English, both)	When received?	Perceived Usefulness	How did you make use of it? (Preparation techniques)
1	●和论坛主办方恰接, 了解论坛的主题及时间和地点。(Contacted the forum organizer to know its theme, time and venue)	中文 (Chinese)	会前3-4天 (3-4 days before the conference)	无法判断 (Not sure)	●根据主办方提供的会议流程和主题, 在网络上搜集相关信息和专业词汇。(Based on the agenda and the forum theme, I searched relevant info. & terminology on the Internet.)
2	●获得论坛第二部分, 即‘专家’论坛部分的一些相关背景资料以及相关内容(Obtained relevant background materials for the 2nd part of the forum: the ‘Expert Forum’)	中文 (Chinese)	会前1-2天 (1-2 days before the conference)	有帮助 (Useful)	●在网络上搜索各个专家的背景信息, 根据其要谈论的内容进行初步准备。(Searched background info. for each speaker & familiarized with possible topics to be discussed.)
3	●获得联合国非物质文化遗产特别代表的发言稿(Obtained the speech script from a UNESCO special representative)	英文 (English)	会议当天 (During the conference)	非常有帮助 (Very useful)	●了解发言人的演讲主旨, 适当的准备语言。(Familiarized with the gist of the speech & did some language preparation.)

## Characteristics of SI tasks

Characteristics	SI task No. 1
<ul style="list-style-type: none"> <li>Before SI, did you receive materials related to the task? (e.g. PPT, conference agenda)</li> </ul>	<p>会前没得到相关材料。(No relevant materials received before the conference)</p> <p>因为无稿件, 所以记不清楚主题内容了。(Because I didn't get any material, I couldn't remember the topics of the speech.)</p>
<ul style="list-style-type: none"> <li>What was the topic of this task?</li> </ul>	论坛开幕式 (Opening of the forum)
<ul style="list-style-type: none"> <li>In which conference component did the task take place? (e.g. Opening, Q&amp;A)</li> </ul>	20 人 (20 people)
<ul style="list-style-type: none"> <li>Estimated No. of interpretation users</li> </ul>	现场发言 (Live presentation)
<ul style="list-style-type: none"> <li>Vehicle of input (i.e. Live, Canned, Both)</li> </ul>	无 (No)
<ul style="list-style-type: none"> <li>Did the speaker(s) speak with an accent?</li> </ul>	发言稿 (Speech script)
<ul style="list-style-type: none"> <li>Did the speaker(s) address audience with the aid of other materials? (e.g. PPT, outline)</li> </ul>	无 (No)
<ul style="list-style-type: none"> <li>During SI, did you use other supplementary materials to help interpreting? (e.g. glossary, draft speech text)</li> </ul>	没有 (No)
<ul style="list-style-type: none"> <li>Any help from booth mate?</li> </ul>	无稿件, 信息密集。(No speech script provided in advance & dense information)
<ul style="list-style-type: none"> <li>What factors contributed to difficulty of the SI task? (e.g. terminology, info. density)</li> </ul>	中进外 (From Chinese to English)
<ul style="list-style-type: none"> <li>Directionality (i.e. Chinese-to-English, English-to-Chinese, Both)</li> </ul>	1 名 (1 speaker)
<ul style="list-style-type: none"> <li>No. of speakers involved in the task?</li> </ul>	10 分钟 (10 minutes)
<ul style="list-style-type: none"> <li>How long did you work in the task?</li> </ul>	论坛结束后 (After the completion of the forum)
<ul style="list-style-type: none"> <li>When did you complete the diary?</li> </ul>	中文发言人应该提供稿件。(The Chinese speaker should provide the speech script to me in advance.)
<ul style="list-style-type: none"> <li>Any comments about the task?</li> </ul>	

## Appendix D: SI task categorization and descriptions

Category	Task variety & Description
SI with Text	<b>SI with text (LongAbun)*</b> Interpreters received a speaker's draft speech text LONG before a task started, and found it to be ABUNDANTLY matched with the speech actually delivered when performing SI.
	<b>SI with text (LongMod)*</b> Interpreters received a speaker's draft speech text LONG before a task started, and found it to be MODERATELY matched with the speech actually delivered when performing SI.
	<b>SI with text (LongBar)*</b> Interpreters received a speaker's draft speech text LONG before a task started, but found it to be BARELY matched with the speech actually delivered when performing SI.
	<b>SI with text (ShortAbun)*</b> Interpreters received a speaker's draft speech text only SHORTLY before a task started, and found it to be ABUNDANTLY matched with the speech actually delivered when performing SI.
	<b>SI with text (ShortMod)</b> Interpreters received a speaker's draft speech text only SHORTLY before a task started, and found it to be MODERATELY matched with the speech actually delivered when performing SI.
	<b>SI with text (ShortBar)</b> Interpreters received a speaker's draft speech text only SHORTLY before a task started, and found it to be BARELY matched with the speech actually delivered when performing SI.

Category	Task variety & Description
SI with PPT	<p><b>SI with PPT (LongAbun)*</b></p> <p>Interpreters received a speaker's PPT LONG before a task started, and found it to be ABUNDANTLY matched with the PPT actually presented when performing SI.</p>
	<p><b>SI with PPT (LongMod)*</b></p> <p>Interpreters received a speaker's PPT LONG before a task started, and found it to be MODERATELY matched with the PPT actually presented when performing SI.</p>
	<p><b>SI with PPT (LongBar)*</b></p> <p>Interpreters received a speaker's PPT LONG before a task started, but found it to be BARELY matched with the PPT actually presented when performing SI.</p>
	<p><b>SI with PPT (ShortAbun)*</b></p> <p>Interpreters received a speaker's PPT only SHORTLY before a task started, and found it to be ABUNDANTLY matched with the PPT actually presented when performing SI.</p>
	<p><b>SI with PPT (ShortMod)</b></p> <p>Interpreters received a speaker's PPT only SHORTLY before a task started, and found it to be MODERATELY matched with the PPT actually presented when performing SI.</p>
	<p><b>SI with PPT (ShortBar)</b></p> <p>Interpreters received a speaker's PPT only SHORTLY before a task started, and found it to be BARELY matched with the PPT actually presented when performing SI.</p>

Category	Task variety & Description
SI without Materials	<b>SI with no materials (NoText)*</b> Interpreters received no speech text in advance, but performed SI for a speaker who relied on a prepared text(s) to deliver his/her speech.
	<b>SI with no materials (NoPPT)*</b> Interpreters received no PPT in advance, but performed SI for a speaker who relied on prepared PPT to make his/her presentation.
	<b>SI with no materials (NoText&amp;PPT):</b> Interpreters received neither text nor PPT in advance, but performed SI for a speaker who relied on both prepared text and PPT to make his/her presentation.
	<b>SI (MonoImprm)*</b> Interpreters might receive background materials in advance (e.g. agenda/program), and performed SI for a monologue in which an individual speaker made an impromptu speech at a conference.
Others	<b>SI (DiaIntr)*</b> Interpreters might receive background materials in advance (e.g. agenda/program), and performed SI for a dialogic interaction component in a conference (e.g. Q&A) which is engaged by more than one interlocutor.
	<b>SI (Audio-/video)*</b> Interpreters might receive background materials in advance (e.g. agenda/program), and performed SI for a pre-recorded audio or video materials played at a conference.

Notes: \*These tasks were empirically identified in the diary study (see Chapter 4); “LONG before” means that draft speech text/PPT was received ONE or MORE days before a conference; “SHORTLY before” means that draft speech text/PPT was received on the same day when a SI task was performed; “ABUNDANTLY matched” means “more than 70%” matched; “MODERATELY matched” means “40% to 60%” matched; “BARELY matched” means “less than 30%” matched.

## **Appendix E: English speech scripts for the four SI tasks**

Please contact the author at [chao.research@gmail.com](mailto:chao.research@gmail.com) to obtain this appendix.

## Appendix F: Indices for characteristics of the four source texts

Quantitative indicators / Text			T <sub>SN</sub>	T <sub>SA</sub>	T <sub>FN</sub>	T <sub>FA</sub>
<b>Lexical characteristics</b>	<b>Length of lexical input</b>	Word count (words)	1264	1275	1243	1250
		Syllables per word (SPW)	1.85	1.84	1.75	1.73
	<b>Lexical density</b>	Lexical density (LD)	0.53	0.53	0.50	0.49
	<b>Lexical sophistication</b>	Lexical sophistication - I (LS-I)	0.32	0.36	0.35	0.34
		Lexical sophistication - II (LS-II)	0.30	0.32	0.30	0.38
	<b>Lexical variation</b>	Type-token ratio (TTR)	0.54	0.51	0.49	0.49
		Lexical variation (LV)	0.71	0.68	0.67	0.66
	Proposition density (ProD)		0.54	0.53	0.55	0.56
<b>Propositional characteristics</b>	No. of Elementary discourse unit (EDU)		145	136	142	141
	Ratio of nucleus to satellite EDU (N/S)		1.34	1.39	1.41	1.35
<b>Syntactic characteristics</b>	<b>Length of source text</b>	No. of sentences (S)	74	72	72	73
		No. of clauses (C)	111	111	111	112
		No. of T-unit (T)	77	75	76	80
		No. of dependent clauses (DC)	37	37	36	38
		No. of coordinate phrases (CP)	37	33	34	35
		Mean length of sentence (words)	17.08	17.71	17.26	17.12
		Mean length of clause (words)	11.39	11.49	11.20	11.16
		Mean length of T-unit (words)	16.42	17.00	16.36	15.63
		Clause per T-unit (C/T)	1.44	1.48	1.46	1.40
		Dependent clause per T-unit (DC/T)	0.48	0.49	0.47	0.48
		Coordinate phrase per clause (CP/C)	0.33	0.30	0.31	0.34
		Coordinate phrase per T-unit (CP/T)	0.48	0.44	0.45	0.44
		T-unit per sentence (T/S)	1.04	1.04	1.06	1.10
<b>Readability index</b>	The Coleman-Liau Index		14.67	13.00	12.67	13.00
	Gunning Fog		15.47	15.33	13.73	13.43

## Appendix G: Interview questions in the experiment

### Interview Questions

- 1) What are the **prominent factor(s)** do you think contribute(s) to the difficulty of this SI task? 你认为这篇讲话哪些突出的因素使得同传变难了?
- 2) How did these factors affect your **SI performance**? (i.e., information completeness, fluency, target language quality) 那你列举出的这些因素如何影响了你的同传表现呢? (信息完整度, 译文流畅度, 语言质量)
- 3) What **strategies** did you employ to deal with these factors? 应对这些因素, 你采取了哪些对策呢, 或者说哪些口译策略呢?
- 4) How did the use of strategies affect your SI performance? (i.e., information completeness, fluency, target language quality) 使用这些口译策略对你的同传结果有什么影响吗? (信息完整度, 流畅度, 语言质量)

## Appendix H: The *post-hoc* questionnaire

### ◆ Demographic information

1. Your gender:

☐ Female

☐ Male

2. Your age: (e.g., 36 years old)

3. Which city do you most frequently work in as a conference interpreter? (Please name only one city)

4. What is the highest level of education you have COMPLETED?

☐ Bachelor's degree

☐ Master's degree

☐ Doctoral degree

Other (please specify)

5. What kind of interpreting training and education have you received? (You may indicate more than one)

☐ Intensive interpreting training course

☐ Interpreting diploma

☐ Postgraduate-level interpreting degree

Other (please specify)

6. You identify yourself as:

☐ A part-time interpreter (e.g., hold a formal job, only interpret part-time)

☐ Freelance interpreter

☐

In-house/staff interpreter

7. Your simultaneous interpreting (SI) experience (in months): (e.g., about 35 months)

8. How many conferences have you provided SI service for? (e.g., about 60 conferences)

9. Please give an estimate of the number of conferences you work for in the past 12 months?

(e.g., about 12 conferences)

♦ **Overall difficulty of source speeches**

10. Please rate the overall difficulty level of the source speeches respectively.

Very easy

Moderate

Very difficult

(1)

(2)

(3)

(4)

(5)

(6)

(7)

Speech 1

Speech 2

Speech 3

Speech 4

Comment

## **Appendix I: Background reading material in the experiment**

Please contact the author at [chao.research@gmail.com](mailto:chao.research@gmail.com) to obtain this appendix.

## Appendix J: Background information sheet

### Information: Task<sub>SN</sub>

**Setting:** Australia-China Economic and Cooperation Trade Forum (formal setting)

**Conference participants:** High-ranking government officials, business leaders and other dignitaries from both Australia and China

**Speaker:** A senior Australian Government official outlines bilateral relationship between Australia and China.

**Purpose:** To deepen relationship and to promote more exchanges.

**Topics:** Bilateral relationship concerning investment, two-way trade, a shared clean energy future and people-to-people links.

**Task type:** Simultaneous interpreting (SI) for read-aloud text.

**Task material:** Only aural material (audio recordings).

**Directionality:** From English to Chinese.

**Task duration:** Approximately 10 minutes.

**Note:** Interpreters can make reference to Background Materials<sup>47</sup> (with notes) and jot down notes when interpreting.

### Information: Task<sub>SA</sub>

**Conference setting:** Australia–China Business Council, Canberra Networking Day (formal setting).

**Conference participants:** Business leaders and Government officials from both Australia and China.

**Speaker:** A senior Australian Government official highlights Australia-China relations.

**Purpose:** To promote friendship and to deepen cooperation.

**Topics:** Australia Government's approach to the bilateral relations, and implications of China's rise on Australia and the world beyond.

**Task type:** Simultaneous interpreting (SI) for read-aloud text.

**Task material:** Only aural material (audio recordings).

**Directionality:** From English to Chinese.

**Task duration:** Approximately 10 minutes.

---

<sup>47</sup> Background Materials refer to the background reading material that has been provided by the researcher.

**Note:** Interpreters can make reference to Background Materials (with notes) and jot down notes when interpreting.

### **Information: Task<sub>FN</sub>**

**Setting:** Australia-China Forum (formal setting).

**Conference participants:** Government officials, business leaders and other dignitaries from both Australia and China.

**Speaker:** A senior Australian Government official reviews Australia-China relations.

**Purpose:** To promote bilateral relationship and to celebrate the 40<sup>th</sup> anniversary of Australia-China diplomatic ties.

**Topics:** Bilateral relations concerning diplomatic, economic, trade and investment relationship.

**Task type:** Simultaneous interpreting (SI) for read-aloud text.

**Task material:** Only aural material (audio recordings).

**Directionality:** From English to Chinese.

**Task duration:** Approximately 10 minutes.

**Note:** Interpreters can make reference to Background Materials (with notes) and jot down notes when interpreting.

### **Information: Task<sub>FA</sub>**

**Conference setting:** A speech to Australian Studies Centre (formal setting).

**Conference participants:** Government officials and academics from both Australia and China.

**Speaker:** A senior Australian Government official talks about Australia-China bilateral relations.

**Purpose:** To deepen mutual understanding.

**Topics:** Bilateral relations concerning trade, investment, people-to-people links and future opportunities.

**Task type:** Simultaneous interpreting (SI) for read-aloud text.

**Task material:** Only aural material (audio recordings).

**Directionality:** From English to Chinese.

**Task duration:** Approximately 10 minutes.

**Note:** Interpreters can make reference to Background Materials (with notes) and jot down notes when interpreting.

## Appendix K: Interpreting performance assessment sheet

Rater ID:		Recording No.:	Recording ID:
Scoring criteria	Information completeness	Fluency of delivery	Target language quality
<b>Band 4 &amp; Descriptors</b> (Score range 7-8)	> 90% of source text propositional content delivered in target text.	Delivery on the whole fluent, containing only a few disfluencies.	Target language idiomatic and on the whole correct, with only a few instances of unnatural and incorrect usage.
<b>Band 3 &amp; Descriptors</b> (Score range 5-6)	70-80% of source text propositional content delivered in target text.	Delivery on the whole generally fluent, containing a small number of disfluencies.	Target language generally idiomatic and on the whole mostly correct, with several instances of unnatural and incorrect usage.
<b>Band 2 &amp; Descriptors</b> (Score range 3-4)	50-60% of source text propositional content delivered in target text.	Delivery rather fluent. Acceptable, but with regular disfluencies.	Target language is to a certain degree both idiomatic and correct. Acceptable, but contains many instances of unnatural and incorrect usage.
<b>Band 1 &amp; Descriptors</b> (Score range 1-2)	< 40% of source text propositional content delivered in target text.	Delivery lacks fluency. It is frequently hampered by disfluencies to such a degree that they may impede comprehension.	Target language stilted, and lacking in idiomaticity, to such a degree that it may impede comprehension.
<b>Final Band (score)</b>			

## Appendix L: Strategies and definitions

No.	Strategy	Definition
01	Stalling by using neutral material	<ul style="list-style-type: none"> <li>• Producing generic utterings, absent in the ST, which provides no new information, but enable the interpreter to deploy production and to continue listening to the incoming text while avoiding long pauses when faced with comprehension difficulties.</li> </ul>
02	Syntactic transformation	<ul style="list-style-type: none"> <li>• Re-arranging the syntactic order of SL elements at sentence or inter-sentential level to help achieve a natural sounding TT.</li> </ul>
03	Syntactic segmentation	<ul style="list-style-type: none"> <li>• Dividing a long clause into shorter clauses, or one long sentence into shorter sentences, also known as salami technique or chunking.</li> </ul>
04	Changing the order of phrases	<ul style="list-style-type: none"> <li>• Reformulating a cascade of SL phrases, particularly enumerations, into the TT in a different sequence.</li> </ul>
05	Generalization	<ul style="list-style-type: none"> <li>• Replacing a SL segment with a superordinate TL term or a more general speech segment.</li> </ul>
06	Simplification	<ul style="list-style-type: none"> <li>• A lexical or stylistic simplification of the original SL message.</li> </ul>
07	Omission	<ul style="list-style-type: none"> <li>• Reprocessing the SL text through the deletion of superfluous or redundant information by means of a selection of information.</li> </ul>
08	Explanatory addition	<ul style="list-style-type: none"> <li>• A lexical and content expansion aimed at clarifying the message.</li> </ul>
09	Addition to maintain coherence	<ul style="list-style-type: none"> <li>• Explicating TT coherence relations with a view to conferring logical continuity to the text.</li> </ul>
10	Repetition	<ul style="list-style-type: none"> <li>• Repeating previously processed elements as a way of enhancing lexical accuracy by means of synonyms or synonymic phrases.</li> </ul>
11	Paraphrase	<ul style="list-style-type: none"> <li>• Explaining the meaning of a SL term or wording when the interpreter is unable to find the suitable TL correspondent.</li> </ul>

**Appendix L. Continued**

No.	Strategy	Definition
12	Substitution	<ul style="list-style-type: none"><li>• Providing renditions which though different from those originally produced by the speaker, can be plausible in the speech context.</li></ul>
13	Reproduction	<ul style="list-style-type: none"><li>• Leaving a word or phrase (typically an unknown name) as it appears in the ST. The interpreter repeats, as closely as s/he can, what was said in the SL.</li></ul>
14	Repair	<ul style="list-style-type: none"><li>• Self-correcting something that has already been said, which is a mis-representation of the meaning intended by the speaker.</li></ul>
15	Transcoding	<ul style="list-style-type: none"><li>• Translating a SL term or segment into TL word for word.</li></ul>

Notes: ST = source text; TT = target text; SL = source language; TL = target language

## Appendix M: Certification tests reviewed & practice of rater training and calibration

Country	Certification body/test	Literature reviewed	Model of rater reliability estimation	Rater training? & Rater reliability estimate (RRE)
Australia	<ul style="list-style-type: none"> <li>National Accreditation Authority for Translators and Interpreters (NAATI) †††</li> </ul>	<ul style="list-style-type: none"> <li>* Hale &amp; Campbell (2003), Hale et al. (2012), NAATI (2013), Turner et al. (2010)</li> </ul>	♦ Intra- & inter-rater	<ul style="list-style-type: none"> <li>• Yes, but details cannot be accessed (internal report).</li> <li>• RRE: Inaccessible.</li> </ul>
Belgium (Flanders)	<ul style="list-style-type: none"> <li>Social Interpreter Certification Examination (SICE) by Flemish Central Support Cell (COC) ††</li> </ul>	<ul style="list-style-type: none"> <li>* Vermeiren et al. (2009)</li> </ul>	♦ Intra- & inter-rater	<ul style="list-style-type: none"> <li>• Yes, but no detailed descriptions.</li> <li>• RRE: No indices provided.</li> </ul>
Canada	<ul style="list-style-type: none"> <li>Signed language interpreter certification test by the Association of Visual Language Interpreters of Canada (AVLIC) †</li> <li>Conference Interpretation Examination by the Canadian Translators, Terminologists and Interpreters Council (CTTIC) ††</li> </ul>	<ul style="list-style-type: none"> <li>* Russell &amp; Malcolm (2009)</li> <li>* CTTIC website<sup>a</sup></li> </ul>	<ul style="list-style-type: none"> <li>♦ Inter-rater agreement</li> <li>♦ Inaccessible</li> </ul>	<ul style="list-style-type: none"> <li>• Yes, detailed training procedures.</li> <li>• RRE: Above 95% agreement on pass/fail decision.</li> <li>• Inaccessible.</li> </ul>

a. See CTTIC website: <http://www.cttic.org/certification.asp>

**Appendix M. Continued**

<b>Country</b>	<b>Certification body/test</b>	<b>Literature reviewed</b>	<b>Model of rater reliability estimation</b>	<b>Rater training? &amp; Rater reliability estimate (RRE)</b>
China	<ul style="list-style-type: none"> <li>China Accreditation Tests for Translators and Interpreters (CATTI) ††</li> <li>National Accreditation Examinations for Translators and Interpreters (NAETI) ††</li> </ul>	<ul style="list-style-type: none"> <li>* Cai (2007), Cai (2009), Office of CATTI (2005), Lu et al. (2007), CATTI website<sup>a</sup></li> <li>* NAETI website<sup>b</sup></li> </ul>	<ul style="list-style-type: none"> <li>♦ Inaccessible</li> <li>♦ Inaccessible</li> </ul>	<ul style="list-style-type: none"> <li>• Inaccessible.</li> <li>• Inaccessible.</li> </ul>
Norway	<ul style="list-style-type: none"> <li>Norwegian Interpreter Certification Examination (NICE) ††</li> </ul>	<ul style="list-style-type: none"> <li>* Mortensen (1998, 2001)</li> </ul>	<ul style="list-style-type: none"> <li>♦ General concept of reliability</li> </ul>	<ul style="list-style-type: none"> <li>• Yes, but no detailed descriptions.</li> <li>• RRE: Inaccessible.</li> </ul>
South Africa	<ul style="list-style-type: none"> <li>Simultaneous Interpreter Accreditation Testing by South African Translators' Institute (SATI) †††</li> </ul>	<ul style="list-style-type: none"> <li>* SATI website<sup>d</sup>, SATI (2007a, 2007 b)</li> </ul>	<ul style="list-style-type: none"> <li>♦ Inaccessible</li> </ul>	<ul style="list-style-type: none"> <li>• Inaccessible</li> </ul>
UK	<ul style="list-style-type: none"> <li>Diploma in Public Service Interpreting (DPSI) by IoL Educational Trust (IoLET) ††</li> </ul>	<ul style="list-style-type: none"> <li>* IoL Educational Trust (2010)</li> </ul>	<ul style="list-style-type: none"> <li>♦ Inter-rater</li> </ul>	<ul style="list-style-type: none"> <li>• Yes, training &amp; monitoring, no detailed descriptions.</li> <li>• RRE: Inaccessible.</li> </ul>

b. CATTI website: [http://www.catti.net.cn/node\\_74539.htm](http://www.catti.net.cn/node_74539.htm)

c. NAETI website: [http://sk.neea.edu.cn/wyfyzs/xmjs.jsp?class\\_id=26\\_07\\_01\\_01](http://sk.neea.edu.cn/wyfyzs/xmjs.jsp?class_id=26_07_01_01)

d. SATI website: [http://translators.org.za/sati\\_cms/index.php?frontend\\_action=display\\_text\\_content&content\\_id=1783](http://translators.org.za/sati_cms/index.php?frontend_action=display_text_content&content_id=1783)

**Appendix M. Continued**

Country	Certification body/test	Literature reviewed	Model of rater reliability estimation	Rater training? & Rater reliability estimate (RRE)
USA	<ul style="list-style-type: none"> <li>Federal Court Interpreter Certification Examination (FCICE) ††</li> </ul>	<ul style="list-style-type: none"> <li>* National Center for States Courts (2013), Feuerle (2013), Stansfield &amp; Hewitt (2005)</li> </ul>	♦ Inter-rater	<ul style="list-style-type: none"> <li>• Yes, intensive training, no detailed descriptions.</li> <li>• RRE: Inaccessible.</li> </ul>
	<ul style="list-style-type: none"> <li>National Interpreter Certification (NIC) by National Association of the Deaf (NAD) the Registry of Interpreters for the Deaf (RID) †</li> </ul>	<ul style="list-style-type: none"> <li>* NAD (2014), RID website<sup>a</sup>, Roat (2006)</li> </ul>	♦ Inter-rater	<ul style="list-style-type: none"> <li>• Yes, intensive training, detailed descriptions.</li> <li>• RRE: Inaccessible.</li> </ul>
	<ul style="list-style-type: none"> <li>National Board of Certification for Medical Interpreters (NBCMI) ††</li> </ul>	<ul style="list-style-type: none"> <li>* Arocha &amp; Joyce (2013), NBCMI (2014), PSI Services (2010, 2013), Roat (2006),</li> </ul>	♦ Inter-rater	<ul style="list-style-type: none"> <li>• Yes, intensive training &amp; calibration, detailed descriptions.</li> <li>• RRE: <math>r = 0.88 - 0.99</math></li> </ul>
	<ul style="list-style-type: none"> <li>Certified Healthcare Interpreter™ Examination by Certification Commission for Healthcare Interpreters (CCHI) ††</li> </ul>	<ul style="list-style-type: none"> <li>* CCHI (2010, 2011, 2012, 2014), Youdelman (2013)</li> </ul>	♦ Inter-rater, decision consistency	<ul style="list-style-type: none"> <li>• Yes, intensive training &amp; detailed descriptions.</li> <li>• RRE: Intraclass correlation <math>\approx 0.73</math> (for Spanish/English version)</li> </ul>

††† Certifying both spoken & sign language interpreters; †† Certifying spoken language interpreters; † Certifying sign language interpreters.

e. RID website: <http://rid.org/education/testing/index.cfm/AID/86>

