

University of Tabriz

Faculty of Persian Literature and Foreign Languages

English Language Department

**Dissertation submitted to the Faculty of Persian Literature and Foreign Languages in
partial fulfillment of the requirements for the degree of**

Doctor of Philosophy

In

English Language Teaching

Title

Rater effects in self-assessment, peer-assessment, and teacher assessment:

A multi-faceted Rasch measurement approach

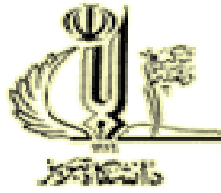
Supervisor: Farahman Farrokhi (PhD)

Advisor: Parviz Azhide (PhD)

Researcher: Rajab Esfandiari

February 2012

In the name of God



University of Tabriz

Faculty of Persian Literature and Foreign Languages

English Language Department

Dissertation submitted to the Faculty of Persian Literature and Foreign Languages in

fulfillment of the requirements for the degree of

Doctor of Philosophy

In

English Language Teaching

Title

Rater effects in self-assessment, peer-assessment, and teacher assessment:

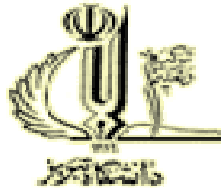
A multi-faceted Rasch measurement approach

Supervisor: Farahman Farrokhi (PhD)

Advisor: Parviz Azhide (PhD)

Researcher: Rajab Esfandiari

February 2012



University of Tabriz

Faculty of Persian Literature and Foreign Languages

English Language Department

We hereby recommend that the dissertation by Rajab Esfandiari

entitled

Rater effects in self-assessment, peer-assessment, and teacher assessment:

A multi-faceted Rasch measurement approach

be accepted in partial fulfillment of the requirements for the degree of

PhD in ELT

Supervisor: Farahman Farrokhi (PhD)

Advisor: Parviz Azhide (PhD)

Examiners:

Internal examiner:

Ali Akbar Ansarin (PhD)

External examiners:

Javad Gholami (PhD)

Biok Behnam (PhD)

To my parents with love

Acknowledgements

When I started a PhD in English Language Teaching at the University of Tabriz some four years ago, I would never have fancied ending up doing language testing. By a strange quirk of fate, many-facet Rasch measurement differs markedly from planned focus on form which I chose to work on for my MA thesis. The printed pages of this dissertation, therefore, hold far more than the culmination of years of study. I would like to take this opportunity to thank those people who, one way or the other, helped me hatch up the idea of Rasch measurement and without whose help this dissertation would have been a complete fiasco.

First and foremost, I would like to offer my sincerest gratitude and express my deepest appreciation to my respectable supervising professor, Dr Farrokhi, whose highly perceptive comments, insightful observations, resourceful guidelines, illuminating insights, enduring inspirations and constructive suggestions are indelibly imprinted in the individual pages of this dissertation. He was so enthusiastic about research issues and, continually and convincingly, conveyed a spirit of adventure in regard to research. Despite his very hectic schedule and immense managerial responsibility, he was very approachable and always accepted me into his office very willingly and kindly, never failing to offer me a hot, freshly brewed cup of tea. His invaluable pieces of advice, persistent help and unfailing support, words of encouragement and life-long academic experience are highly valued and fully appreciated.

I would also like to extend my thanks to my advisor, Dr Azhideh, who made available his support in a number of ways and spared me his precious time to discuss and share with me his immensely hands-on statistical experience which improved the quality of statistical arguments made in this dissertation. His timely provision of comments, valid recommendations and helpful suggestions, and constant assistance enriched the present dissertation to varying degrees. I had

the privilege of being his PhD student in Language Testing course during which period he introduced me with new areas of language testing and methodological innovations employed to unpack issues in language testing including Rasch measurement.

Special acknowledgement must be made of the internal examiner, Dr Ansarin, and external examiners, Dr Gholami and Dr Behnam, whose meticulous reading and fair comments contributed substantially to this dissertation.

A deep sense of gratitude and special thanks are due to professor Myford, of the University of Illinois at Chicago, who assisted me with the many-facet Rasch measurement (MFRM) from inception to completion of this dissertation. She always encouraged me to move on, felt duty bound to keep me working and moving forward, and was so curious to be well-informed of my work progress. She was so kind, generous, supportive, patient, forgiving and caring, always being there to lend her toddler a hand, who had just heard MFRM and could barely articulate it. I do appreciate her care and attention, and I will always remember her in my life.

I am deeply indebted to Dr Schaefer, of the Ochanomizu University, who was my supervising professor when I was a visiting scholar at Ochanomizu University in Tokyo, Japan for a period of six months. He was very kind to let me share his experience in bias analysis, lend me his files of MFRM courses he had registered for on line, and encourage me to ask him for many other books and papers which he provided me with very generously.

I would be remiss if I failed to extend my heartiest appreciations to Dr Ansarin, the current head of the English Language Department, who taught me courses in ESP and Psycholinguistics in MA and PhD programmes. Tremendous complements should also go to Dr Sabouri, the former head of the English Language Department. Thanks are also due to

Dr Andrade, of the University at Albany; Dr Linacre, the Research Director of Winsteps.com; and Dr Baghaei, of the Islamic Azad University, Mashhad Branch.

Special thanks should be given to two of my best friends who have the attitude and substance of geniuses and who have always been readily accessible to help me with their computer knowledge: Dr Mahmoodi, of the University of Tabriz; Dr Vaez Dalili, of the University of Isfahan; and Mr Imam Ali Mirzaei. They have been my constant companions over the years, and I have always enjoyed their company.

Last but not least, I feel obliged to acknowledge my well-behaved students at Imam Khomeini International University in Qazvin, although words alone cannot express the thanks I owe to them.

Abstract

Surname: Esfandiari	First name: Rajab
Title of Dissertation: Rater effects in self-assessment, peer-assessment, and teacher assessment: A Multi-faceted-Rasch Measurement approach	
Supervisor: Farahman Farrokhi (PhD)	
Advisor: Parviz Azhide (PhD)	
Degree: Doctor of Philosophy (PhD)	
Major: English	
Field: English Language Teaching	
University: University of Tabriz	
Faculty: Faculty of Persian literature and foreign languages	
Dissertation pages: 179	
Graduation Date: 17/11/1390	
Key words: Many-facet Rasch measurement; self-assessment; peer-assessment; analytic rating scale; rater effects; rater-mediated second language assessment	
Abstract	
<p>In performance assessment in general and in second language performance assessment in particular, raters are expected to assign ratees ratings, using a rating scale. In the process of rating, raters may commit some errors. These unwanted, rater-dependent sources of variability, which are unrelated to the students' abilities and manifested in various ways, could endanger the fairness and validity of decisions made based on the assigned ratings. Rater errors contribute to construct-irrelevant variance, and, if they are not detected and treated appropriately, they may result in obscuring an examinee's score and threaten the validity and fairness of second language performance assessment. As a result, they deserve further new research and investigation in second language performance assessment. Rater effects—severity/leniency effect, bias effect,</p>	

central tendency effect and halo effect—have been more or less researched either in L1 or in L2, but rarely has any single study striven to address these effects in self-assessment, peer-assessment, and teacher assessment. The present study is an endeavor to employ a multi-faceted Rasch model (MFRM), a relatively newly developed measurement model, to detect these errors in three types of assessment: self-assessment, peer-assessment, and teacher assessment. To that end, 194 assessors—188 self-assessors and peer-assessors and six teacher assessors—were employed to assess 188 essays written by Iranian English majors at two-state run universities in Iran, using a 6-point analytic rating scale. The data were collected and analyzed, using Facets 3.68.1. to answer the research questions. The results of the MFRM analysis showed that of the three assessor types, teacher assessors were the most severe while self-assessors were the most lenient, although there was a great deal of variability in the levels of severity that assessors within each type exercised. MFRM also revealed differing patterns of severity and leniency among the three assessment types. For example, self-assessors and teacher assessors showed the opposite pattern of severity/leniency as peer-assessors toward the highest and lowest ability students. The results of further Facets analysis showed that the three types of assessor did not exhibit any sign of centrality either at group level or at individual level. Finally, Facets analysis showed that, at group level, the assessors did not exhibit any sign of halo effect, but, at individual level, all assessor types displayed considerable halo effect. Further analysis revealed that assessor types were unanimous about halo effect on four items, and that self-assessor showed more of a halo effect compared to the other two assessor types. The present study has possible implications for rater training, concurrent validity of peer ratings, and construction of rating scales. Most importantly, though cautiously, peer-assessors could be employed as an alternative to teachers for rating purposes.

List of Tables

Table 2. 1. Principled Evaluation of Alternatives to Assessment	37
Table 3. 2. Illustration of Rating Designs for Three-Facet Rater-Mediated Performance Assessments	78
Table 3. 3. Schematic Representation of the Rating Design of the Study	80
Table 4. 4. Category Statistics	94
Table 4. 5. Assessor Type Measurement Report	96
Table 4. 6. Assessor Type-Items Bias/Interaction Analysis	99
Table 4. 7. Frequency of Assessor Type-Items Bias/Interactions	101
Table 4. 8. Assessor Type-Students Bias/Interaction Analysis	102
Table 4. 9. Frequency of Assessor Type-Students Bias/Interactions	105
Table 4. 10. The Group-Level Statistics Indicators	107
Table 4. 11. Assessor Type-by-Item Report from Facets Output Analysis for Halo Effect	109
Table 4. 12. Category Statistics for Assessor Type	112
Table 4. 13. Student Measurement Report for Central Tendency Effect	112
Table 4. 14. Item Measurement Report for Central Tendency Effect	118
Table 4. 15. Assessor Type Measurement Report for Central Tendency at Individual Level ..	119
Table 4. 16. Category Statistics for Self-Assessor	120

List of Figures

Figure 3. 1. Connectedness of the Data for Present Study	81
Figure 4. 2. Variable Map	92
Figure 4. 3. Probability Curves for Students	95
Figure 4. 4. Bias Analysis for Assessor Type (Assessor-Type-Items Interactions)	100

List of Abbreviations

MFRM	Many-facet Rasch measurement
Df	Degree of freedom
L2	Second language
L1	First language

For correspondence with the researcher:

E-mail: raesfandiari@gmail.com

Refereed journal publications based on this dissertation:

1. Farrokhi, F., & Esfandiari, R. (2011). A Many-Facet Rasch Measurement Model to Detect Halo Effect in Three Types of Raters. *Theory and Practice in Language Studies*, 1(11), 1531-1541.
2. Farrokhi, F., Esfandiari, R., & Schaefer, E. (to be published in September 2012). A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Self-assessment, Peer-assessment, and Teacher assessment. *Journal of Applied and Basic Scientific Research*.
3. Farrokhi, F., Esfandiari, R., & Vaez Dalili, M. (2011). Applying the Many-Facet Rasch Model to Detect Centrality in Self-assessment, Peer-assessment and Teacher assessment. *World Applied Science Journal*, 15 (11), 76-83.
4. Farrokhi, F., Esfandiari, R., & Schaefer, E. (to be published in May 2012). A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in three types of assessment. *JALT Journal*.
5. Esfandiari, R., Myford, C., & Farrokhi, F. (under review). Severity Differences among Self-assessors, Peer-assessors, and Teacher Assessors Rating EFL Essays. *Assessing Writing*.

National and international conference presentations based on this dissertation:

1. Esfandiari, R., & Farrokhi, F. (2011). A Many-Facet Rasch Measurement Approach to Rater Effects. Paper Presented at the Ninth Annual International TELLISI Conference, Illam, 2011.
2. Esfandiari, R., Schaefer, E., & Farrokhi, F. (2011). Detecting Rater Effects in Three Types of Assessment. Paper Presented at the Tenth Annual JALT PAN-SIG Conference, Matsumoto, Nagano, 2011.

Table of contents

Title	Page
Acknowledgements	vi
Abstract in English	ix
List of Tables	xi
List of Figures	xii
List of Abbreviations	xiii
Chapter One	1
1 Introduction	2
1.0. Background to the Study	2
1.1. Statement of the Problem	4
1.2. Significance of the Study	6
1.3. Goals of the Study	7
1.4. Research Questions	8
1.5. Definition of Key Terms	9
1.6. Organization of the Study	10
Chapter Two	11
2 Review of the Related Literature	12
2.0. Introduction	12
2.1. Rater Effects	12
2.2. A Catalogue of Rater Effects	13
2.3. Conceptual Definitions of Leniency/Severity	13
2.3.1. Strategies to Reduce Leniency/Severity	14

2.3.2. Empirical Studies on Leniency/Severity in L1 and L2 Settings	15
2.4. Conceptual Definitions of Halo Effect	17
2.4.1. Approaches to Minimize the Halo Effect	19
2.4.2. Approaches to Detect the Halo Effect	20
2.4.3. Studies Done on Halo Effect Using MFRM	21
2.5. Conceptual Definitions on Central Tendency Effect	23
2.5.1. Studies on Central Tendency Effect in L1	24
2.5.2. Studies on Central Tendency Effect in L2	26
2.6. Bias Effect	29
2.4.1. Previous Bias Analysis Studies	29
2.7. Alternative Assessment	32
2.7.1. Characteristics of Alternative Assessment	35
2.7.2. Purposes and Uses of Alternative Assessment	35
2.7.3. Responsible Test Characteristics of Alternative Assessment	36
2.8. Self-assessment	37
2.9. Peer-assessment	41
2.10. Studies on Self-Assessment, Peer-Assessment, and Teacher Assessment	44
2.11. Studies on Self-Assessment, Peer-Assessment, and Teacher Assessment Employing MFRM	57
2.12. Many-Facet Rasch Measurement Model	62
2.12.1. Defining Many-Facet Rasch Measurement	62
2.12.2. Requirements and Assumptions of the Many-Facet Rasch Measurement	64
2.12.3. Applications of the Many-Facet Rasch Measurement	66

2.13. Chapter Summary	69
Chapter Three	70
3 Methodology	71
3.0. Introduction	71
3.1. Participants	71
3.2. The Rating Scale	72
3.3. Assessor Training and Data Collection Procedures	74
3.4. Rating Design of the Study	77
3.5. Restatement Research Questions	81
Chapter Four	83
4 Data Analysis and Results.....	84
4.0. Introduction	84
4.1. Data Analysis	84
4.2. Results of Preliminary Analysis	90
4.3. Variable Map	91
4.4. Main Results	93
4.4.1. Rating Scale Functioning	93
4.4.2. Severity Differences among Assessor Types	97
4.4.3. Variability in the Levels of Severity within Each Assessor Type	97
4.4.4. Bias Analysis	98
4.4.5. Halo Effect Analysis	106
4.4.6. Central Tendency Effect Analysis	111
Chapter Five	122

5 Discussion, Conclusions, and Pedagogical Implications.....	123
5.0. Introduction	123
5.1. Discussion	123
5. 2. Conclusions	136
5.3. Implications	137
5.4. Avenues for further research	138
5.5. Limitations of the Study	139
References	140
Appendices	155
Appendix A: Essay Rating Sheet	156
Appendix B: ESL Composition Profile	157
Appendix C: Essay Topic	158
Appendix D: Topics	160
Appendix E: The Most Interesting Topic	165
Appendix F: Rated Essay	166
Appendix G: Essay Rating Sheet راهنمای استفاده از	168
Appendix H: Essay	169
Appendix I: A Sample Self-assessed Essay	170
Appendix J: A Sample Peer-assessed Essay	173
Appendix K: A Sample Teacher Assessed Essay	175
Abstract in Farsi	179

CHAPTER One:

INTRODUCTION