USING THE RASCH MEASUREMENT MODEL FOR STANDARD SETTING OF THE ENGLISH LANGUAGE PLACEMENT TEST AT THE IIUM

NOOR LIDE BINTI ABU KASSIM

UNIVERSITI SAINS MALAYSIA

2007

USING THE RASCH MEASUREMENT MODEL FOR STANDARD SETTING OF THE ENGLISH LANGUAGE PLACEMENT TEST AT THE IIUM

by

NOOR LIDE BINTI ABU KASSIM

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

March 2007

ACKNOWLEDGEMENTS

Before I thank the individuals and organizations that have contributed to the writing and completion of this thesis, I would like to begin by saying that Allah Most Gracious Most Merciful by His Grace and Will alone has made this journey possible.

First and foremost, I would like to express my sincere and heartfelt gratitude to my supervisor, Associate Professor Dr Abdul Rashid Mohamed, for believing in me and for guiding me throughout this arduous journey. I couldn't have made it without his guidance, support and encouragement. To Dr Wan Mohd Rani Abdullah, my heartfelt gratitude for taking me on as his supervisee, and for helping me through difficult times while I was under his supervision.

My deepest and most heartfelt gratitude to Professor Dr Trevor G. Bond, who is not only my co-supervisor but also mentor and friend, for instigating me to work on standard setting, for developing my understanding of Rasch measurement, for going out of his way to help me, for believing that I can make a difference, and for making my family and I, a part of his, during our stay in Townsville, Australia. I am forever indebted and honoured to have been given a chance to work with such an exceptional individual. To his lovely wife, Marie-Louise Bond, children and grandchildren, my utmost gratitude for their kindness and friendship.

To Professor Dr John "Mike" Linacre, my deepest gratitude and heartfelt appreciation. I can never thank him enough for the endless support and encouragement that he has given me, and for his patience and commitment in helping me with my data analysis and every little question that I had put to him. I am privileged to have been given the opportunity to know this brilliant and kind individual.

My sincere and heartfelt gratitude is also extended to Dr Isarji Bin Hj Sarudin, Dean of the Centre for Languages and Pre-University Academic Development

(CELPAD), and the International Islamic University Malaysia for believing in me and for giving me the study leave and scholarship in my pursuit of academic excellence.

To the lecturers and staff at the School of Educational Studies, USM, my deepest appreciation for their guidance and support. To the School of Educational Studies and the Institute of Postgraduate Studies, USM, my gratitude and appreciation for making it possible for Professor Dr Trevor G. Bond to be appointed as my co-supervisor.

I would also like to express my sincere appreciation to the Australian government, in particular, the Department of Education, Science and Training (DEST) for giving me the opportunity to work with Professor Trevor G. Bond, a leading figure in Rasch measurement, under the Australian Endeavour Postgraduate and Postdoctoral Research Fellowship Award Programme 2004. My sincere appreciation also goes to those who were involved in the programme, particularly, Cynthia Grant and Amy Noone.

To the School of Education of James Cook University, Australia and its dedicated staff, my deepest gratitude for making all the necessary arrangements and making me feel at home. To Professor Annette Patterson, the support and kindness that she has shown me are deeply appreciated and will always be remembered.

Not forgetting my friends and colleagues, my appreciation for their support and encouragement. My special thanks and sincere appreciation to Dr Ainol Madziah Zubairi who has given me tremendous support and helped me get through some of the most difficult moments. Her kindness will never be forgotten.

Last but not least, to my family, my deepest gratitude for their unwavering support, patience and encouragement. And to my beloved parents, husband and daughter, Nana, who have shown me great love and devotion, I dedicate this thesis.

Noor Lide Abu Kassim

March, 2007

TABLE OF CONTENTS

PAGE

Acknowledgements..... ii

Table of Contents..... iv

List of Tables..... xvii

List of Figures..... xxv

Abstrak..... xxx

Abstract..... xxxii

CHAPTER 1: INTRODUCTION

1.0 Introduction…… 1

1.1 Context of study…… 5

1.1.1 The CELPAD Placement System (1995 -2000)…….. 6

1.1.1.1 Instructional Component of the Placement System (1995–2000) ……. 6

1.1.1.2 Assessment Component of the Placement System (1995–2000)……. 7

1.1.2 The CELPAD Placement System (2001 - 2004)…….. 11

1.1.2.1 Instructional Component of the Placement System (2001–2004) ……. 12

1.1.2.2 Assessment Component of the Placement System (2001–2004)……. 13

1.1.3 Shortcomings of the Assessment Component of the Placement System……. 16

1.2 Problem Statement…… 18

1.3 Purpose of Study…… 22

1.4 Objectives of Study…… 22

PAGE

1.5 Research Questions…….. 23

1.5.1 Adequacy of the EPT…….. 23

1.5.2 Efficacy of the OSS..…….. 25

1.6 Rationale of Study…… 26

1.7 Significance of Study…… 27

1.8 Delimitation of Study…… 28

1.9 Definition of Terms…… 29

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction ……..... 33

2.1 Standards within the Educational Framework..... 33

2.2 Standard Setting: Terms and Definitions 38

2.2.1 Content Standards..…… 39

2.2.2 Performance Standards..…… 40

2.2.3 Cutscores..…….....…….... 44

2.2.4 Standard Setting.....…….....…… 46

2.3 Classification of Standard Setting Methods…… 48

2.3.1 Meskauskas' Classification (1976).....…… 49

2.3.2 Hambleton and Eignor's Classification (1980).....…… 49

2.3.3 Berk's Trilevel Classification Scheme..... 50

2.3.4 Jaeger's Test-centred and Examinee-centred Classification Scheme..... 51

2.3.5 Burton's Classification (1977)..... 52

2.3.6 A Priori and A Posteriori Classification..... 54

2.4 Standard Setting Methods for Criterion-referenced Tests 54

2.4.1 Non-objective Test-centred Methods..... 55

2.4.1.1 The Nedelsky Method..... 55

2.4.1.2 The Angoff Method..... 56

2.4.1.3 The Ebel Method..... 57

PAGE

2.4.2 Limitations of Non-objective Methods..... 58

2.4.3 The Objective Standard setting method: An Alternative..... 63

2.5 Validation of Performance Standards: Frameworks..... 65

2.5.1 Kane's Validation Framework (1992, 1994, 2001)..... 67

2.5.1.1 Procedural Evidence of Validity..... 70

2.5.1.2 Validity Checks Based on Internal Criteria..... 71

2.5.1.3 External Validity Checks..... 72

2.5.2 Norcini and Shea (1997)..... 76

2.5.3 Hambleton's (2001) Criteria for Evaluating a Standard Setting Study..... 78

2.6 Measurement and Standard Setting..... 80

2.6.1 The Measurement Process..... 80

2.6.2 Limitations of the Classical Test Theory..... 83

2.6.3 The Rasch Measurement Model..... 85

2.6.3.1 Basic Principle of the Rasch Measurement Model..... 85

2.6.3.2 Requirements for Useful Measurement..... 87

2.6.3.3 Requirements of the Rasch Measurement Model..... 88

2.6.3.4 The Many-facet Rasch Model (MFRM) for Measurement 89

2.6.3.5 Capabilities of the Rasch Measurement Model..... 91

2.7 Assessment and Standard Setting..... 93

2.8 Standard Setting Validation Studies..... 99

2.9 Conclusion..... 105

CHAPTER 3: CONCEPTUAL FRAMEWORK

3.0 Introduction ……..... 106

3.1 Standard Setting Methodology..... 107

3.1.1 Criteria for Selection of Standard Setting Method..... 107

3.1.1.1 Berk (1986)..... 107

3.1.1.2 Loomis and Bourque (2001)..... 110

PAGE

3.1.1.3 Summary of Proposed Criteria..... 110

3.1.2 The OSS as a Defensible Standard Setting Method..... 111

3.1.3 Efficacy of the OSS..... 117

3.1.4 Validity Evidence on the Efficacy of the OSS..... 119

3.2 Adequacy of the EPT..... 122

3.2.1 Validity Evidence on Adequacy of Test Quality..... 123

3.3 Measurement Theory..... 125

3.3.1 The Rasch Measurement Model..... 125

3.3.1.1 Utility of the Rasch Measurement Model in Resolving Measurement Issues..... 125

3.3.1.2 Utility of the Rasch Measurement Model in Resolving Standard Setting Issues..... 127

3.4 Conclusion..... 129

CHAPTER 4: METHODOLOGY

4.0 Introduction..... 131

4.1 Research Framework..... 131

4.2 Procedure of Study..... 132

4.3 Adequacy of Test Quality..... 135

4.3.1 Data and Data Collection Procedures..... 135

4.3.2 Description of the EPT Subtests..... 135

4.3.2.1 Grammar Subtest..... 136

4.3.2.1.1 Content and Description..... 136

4.3.2.1.2 Skills Tested and Targeted Level of Difficulty 137

4.3.2.1.3 Method of Scoring..... 140

4.3.2.2 Reading Subtest..... 140

4.3.2.2.1 Content and Description..... 140

4.3.2.2.2 Skills Tested and Targeted Level of Difficulty 142

4.3.2.2.3 Method of Scoring..... 146

4.3.2.3 Writing Subtest..... 146

PAGE

4.3.2.3.1 Content and Description..... 146

4.3.2.3.2 Method of Scoring..... 146

4.3.3 Subjects..... 148

4.3.4 Data Analyses..... 151

4.4 The Standard Setting Study..... 153

4.4.1 Definition of Goals for Standard Setting Study..... 153

4.4.2 Selection and Training of Standard Setting Judges..... 155

4.4.3 Definition of Performance Level Descriptions..... 158

4.4.4 Estimation of Cutscores: Grammar and Reading Subtests..... 161

4.4.5 Estimation of Cutscores: Writing Subtest..... 165

4.4.6 Estimation of Cutscores: Compensatory Approach..... 167

4.5 Efficacy of the OSS..... 168

4.5.1 Procedural Validity..... 168

4.5.1.1 Data and Data Collection Procedures..... 168

4.5.1.2 Data Analyses..... 168

4.5.2 Internal Validity..... 169

4.5.2.1 Data and Data Collection Procedures..... 169

4.5.2.2 Data Analyses..... 169

4.5.3 External Validity..... 172

4.5.3.1 Data and Data Collection Procedures..... 173

4.5.3.2 Subjects..... 174

4.5.3.2 Data Analyses..... 174

4.6 Conclusion..... 174

CHAPTER 5: RESULTS OF DATA ANALYSIS 175

5.1 Adequacy of the English Language Placement Test..... 179

5.1.1 Validity of MCQ Test Items: Item Polarity, Fit, and Unidimensionality..... 179

5.1.2 Construct Definition..... 181

PAGE

5.1.3 Capacity of Items to Yield Results Consistent with Purpose of Measurement..... 181

5.1.4 Validity of Examinee Responses..... 182

5.1.5 Reliability and Validity of Essay Ratings..... 183

5.2 The Grammar Subtest..... 183

5.2.1 Summary of Item Difficulty and Examinee Ability Distributions..... 183

5.2.2 Item Polarity..... 185

5.2.3 Item Fit..... 186

5.2.4 Unidimensionality..... 187

5.2.5 Construct Definition..... 188

5.2.5.1 Construct Definition: Continuum of Increasing Intensity.... 188

5.2.5.2 Construct Definition: Empirical Scaling vs. Expert Judgment..... 193

5.2.6 Capacity of Items to Yield Results Consistent with Purpose of Measurement..... 198

5.2.6.1 Reliability and Separation..... 198

5.2.6.2 Precision of Estimates..... 199

5.2.6.3 Test Targeting..... 200

5.2.7 Examinee Fit..... 202

5.3 The Reading Subtest..... 203

5.3.1 Summary of Item Difficulty and Examinee Ability Distributions..... 203

5.3.2 Item Polarity..... 205

5.3.3 Item Fit..... 206

5.3.4 Unidimensionality..... 206

5.3.5 Construct Definition..... 207

5.3.5.1 Construct Definition: Continuum of Increasing Intensity.... 207

5.3.5.2 Construct Definition: Empirical Scaling vs. Expert Judgment..... 210

5.3.6 Capacity of Items to Yield Results Consistent with Purpose of Measurement..... 215

PAGE

5.3.6.1 Reliability and Separation..... 215

5.3.6.2 Precision of Estimates..... 216

5.3.6.3 Test Targeting..... 216

5.3.7 Examinee Fit..... 218

5.4 The Multiple-choice Subtest..... 219

5.4.1 Summary of Item Difficulty and Examinee Ability Distributions..... 219

5.4.2 Item Polarity..... 220

5.4.3 Item Fit..... 222

5.4.4 Unidimensionality..... 222

5.4.5 Construct Definition..... 223

5.4.5.1 Construct Definition: Continuum of Increasing Intensity.... 223

5.4.5.2 Construct Definition: Empirical Scaling vs. Expert Judgment..... 226

5.4.6 Capacity of Items to Yield Results Consistent with Purpose of Measurement..... 228

5.4.6.1 Reliability and Separation..... 228

5.4.6.2 Precision of Estimates..... 229

5.4.6.3 Test Targeting..... 230

5.4.7 Examinee Fit..... 231

5.5 The Writing Subtest..... 232

5.5.1 Summary of Item Difficulty Locations, Examinee Ability Distribution and Rater Severity..... 235

5.5.2 Summary Statistics, Examinee Reliability and Separation Indexes.. 237

5.5.3 Item Measurement Report..... 238

5.5.4 Category functioning..... 239

5.5.4.1 Content..... 240

5.5.4.2 Organization..... 241

5.5.4.3 Vocabulary..... 243

5.5.4.4 Language Use..... 244

PAGE

5.5.4.5 Mechanics..... 245

5.5.5 Rater Measurement Report..... 246

5.5.5.1 Rater Severity and Interrater Agreement..... 247

5.5.5.2 Intrarater Consistency..... 249

5.5.5.3 Halo..... 249

5.5.5.4 Central Tendency and Restriction of Range..... 250

5.5.6 Correspondence between Raw Ratings and Rasch Measures..... 256

5.5.6.1 Overall Analysis..... 256

5.5.6.2 Individual Raters..... 259

5.5.7 Examinee Fit..... 262

5.6 The EPT Battery: A Compensatory Approach..... 263

5.6.1 Summary of Item Difficulty and Examinee Ability Distributions..... 263

5.6.2 Item Fit and Unidimensionality..... 265

5.6.3 Capacity of Items to Yield Results Consistent with Purpose of Measurement..... 267

5.6.3.1 Reliability and Separation..... 267

5.6.3.2 Precision of Estimates..... 267

5.6.3.3 Test Targeting..... 267

5.6.4 Examinee Fit..... 268

5.6.5 Rater Measurement Report..... 269

5.6.5.1 Rater Severity..... 269

5.6.5.2 Intrarater Consistency..... 271

5.6.5.3 Halo..... 271

5.6.6 Single vs. Double Rating..... 272

5.7 Results of the Standard Setting Study..... 274

5.7.1 Cutscores: Grammar Subtest..... 274

5.7.1.1 The Criterion Points and Final Cutscores..... 274

5.7.1.2 Categorization of Examinees..... 276

PAGE

5.7.2 Cutscores: Grammar Subtest..... 280

5.7.2.1 The Criterion Points and Final Cutscores..... 280

5.7.2.2 Categorization of Examinees..... 283

5.7.3 Cutscores: Writing Subtest..... 284

5.7.3.1 The Criterion Points and Final Cutscores..... 284

5.7.3.2 Categorization of Examinees..... 288

5.7.4 Cutscores: Compensatory Approach..... 289

5.7.4.1 The Criterion Points and Final Cutscores..... 289

5.7.4.2 Categorization of Examinees..... 293

5.8 Efficacy of the OSS..... 294

5.8.1 Procedural Validity..... 294

5.8.1.1 Implementation of the Standard Setting Study..... 295

5.8.1.1.1 Judge Selection..... 295

5.8.1.1.2 Judge Training..... 295

5.8.1.1.3 Procedures for Data Collection: Time Allocation..... 296

5.8.1.1.4 Procedures for Data Collection: Adequacy of Performance Level Descriptions:..... 296

5.8.1.2 Appropriateness of Standard Setting Procedure..... 299

5.8.1.2.1 Judge Expertise..... 300

5.8.1.2.2 Identification of Essential Items..... 301

5.8.1.2.3 Confidence in the Selection of Essential Items..... 301

5.8.1.2.4 Confidence in the Classification of Examinees..... 303

5.8.1.2.5 Confidence in the Standard Setting Method 304

5.8.1.3 Other Issues..... 304

5.9 Internal Validity..... 306

5.9.1 Grammar Subtest..... 306

5.9.1.1 Distribution of Judges' Ratings of Essential Items..... 306

PAGE

5.9.1.2 Descriptive Statistics and Judge Variability..... 310

5.9.1.2.1 Distribution of Judges' Mean Estimates of Essential Items..... 310

5.9.1.2.2 Judge Variability..... 312

5.9.1.3 Facets Analysis..... 318

5.9.1.3.1 Interrater Agreement..... 320

5.9.1.3.2 Intrajudge Consistency..... 321

5.9.1.3.3 Item Displacement..... 322

5.9.1.4 Correspondence between Cutscores and Performance Level Descriptions..... 323

5.9.2 Reading Subtest..... 330

5.9.2.1 Distribution of Judges' Ratings of Essential Items..... 330

5.9.2.2 Descriptive Statistics and Judge Variability..... 333

5.9.2.2.1 Distribution of Judges' Mean Estimates of Essential Items..... 333

5.9.2.2.2 Judge Variability..... 335

5.9.2.3 Facets Analysis..... 341

5.9.2.3.1 Interrater Agreement..... 343

5.9.2.3.2 Intrajudge Consistency..... 344

5.9.2.3.3 Item Displacement..... 344

5.9.2.4 Correspondence between Cutscores and Performance Level Descriptions..... 345

5.9.3 Writing Subtest..... 350

5.9.3.1 Distribution of Judges' Ratings of Essential Items..... 350

5.9.3.1.1 Distribution of Judges' Mean Estimates of Essential Items..... 353

5.9.3.1.2 Judge Variability..... 354

5.9.3.2 Facets Analysis..... 360

5.9.3.3.1 Interrater Agreement..... 361

5.9.3.3.2 Intrajudge Consistency..... 362

PAGE

5.9.3.3.3 Item Displacement..... 363

5.9.3.3 Correspondence between Cutscores and Performance Level Descriptions..... 363

5.9.4 Compensatory Approach..... 367

5.9.4.1 Descriptive Statistics and Judge Variability..... 368

5.9.4.1.1 Distribution of Judges' Mean Estimates of Essential Items..... 368

5.9.4.1.2 Judge Variability..... 368

5.9.4.2 Facets Analysis..... 373

5.9.4.3.1 Interrater Agreement..... 374

5.9.4.3.2 Intrajudge Consistency..... 374

5.9.4.3.3 Item Displacement..... 375

5.10 External Validity..... 377

5.10.1 Grammar Subtest..... 377

5.10.1.1 Differentiation of Cutscores..... 377

5.10.1.2 Appropriacy of Classification Information..... 378

5.10.1.3 Comparisons with SPM English 1119..... 379

5.10.1.4 Comparisons of Pass Rates of Two Cohorts on Parallel Forms of the EPT..... 380

5.10.2 Reading Subtest..... 382

5.10.2.1 Differentiation of Cutscores..... 382

5.10.2.2 Appropriacy of Classification Information..... 382

5.10.2.3 Comparisons with SPM English 1119..... 383

5.10.2.4 Comparisons of Pass Rates of Two Cohorts on Parallel Forms of the EPT..... 384

5.10.3 Writing Subtest..... 385

5.10.3.1 Differentiation of Cutscores..... 385

5.10.3.2 Appropriacy of Classification Information..... 385

5.10.3.3 Comparisons with SPM English 1119..... 386

PAGE

5.10.3.4 Comparisons of Pass Rates of Two Cohorts on Parallel Forms of the EPT..... 387

5.10.4 Compensatory Approach..... 388

5.10.4.1 Differentiation of Cutscores..... 388

5.10.4.2 Appropriacy of Classification Information..... 388

5.10.4.3 Comparisons with SPM English 1119..... 389

5.10.4.4 Comparisons of Pass Rates of Two Cohorts on Parallel Forms of the EPT..... 390

CHAPTER 6: CONCLUSION 391

6.1 Introduction..... 391

6.2 Summary of Findings..... 393

6.2.1 Adequacy of the EPT..... 393

6.2.1.1 Validity of Items..... 393

6.2.1.2 Construct Definition..... 394

6.2.1.3 Capacity to Yield Results Consistent with Purpose of Measurement..... 396

6.2.1.4 Validity of Examinee Responses..... 397

6.2.1.5 Rater Effects..... 398

6.2.1.6 Rating Scale Functioning..... 399

6.2.2 Efficacy of the OSS..... 399

6.2.2.1 Procedural Validity..... 400

6.2.2.2 Internal Validity..... 403

6.2.2.3 External Validity..... 405

6.3 Discussion..... 407

6.3.1 Adequacy of the EPT..... 407

6.3.2 Efficacy of the OSS in Producing Valid and Defensible Cutscores 412

6.3.3 Generality of the OSS..... 416

6.3.3.1 Multiple Cutscores..... 416

6.3.3.2 Diverse Item Types..... 418

PAGE

6.3.4 Utility of the Rasch Measurement Model..... 419

6.3.4.1 Resolving Measurement Issues..... 419

6.3.4.2 Resolving Standard Setting Issues..... 420

6.4 Implications..... 422

6.4.1 Test Development and Improvement..... 422

6.4.2 Logistic and Administrative Constraints..... 424

6.4.3 Rating Scale Development..... 425

6.4.4 Construct Definition and Validity Inquiry..... 426

6.4.5 Construct Validation in Language Testing..... 427

6.4.6 Efficiency of the CELPAD Placement System and Other Similar Systems..... 428

6.4.7 Standard Setting Theory and Practice..... 429

6.5 Limitations of Study..... 432

6.6 Conclusion..... 434

6.7 Recommendations for Further Research..... 435

6.7.1 Utility of MFRM for the Quantification of Cutscores..... 435

6.7.2 Developmental Sequence of Grammatical Ability..... 436

REFERENCES 437

APPENDICES

Appendix 1 Description of the EPT and Sample Questions.....

Appendix 2 Subtests used in the 2004 EPT Administration.....

Appendix 3 Item Judgment Forms – Grammar & Reading.....

Appendix 4 Quantification of Judges' Mean Estimate of Essential Items.....

Appendix 5 Minutes of the IIUM Senate Meeting No. 279.02.....

Appendix 6 Quantification of Judges' Mean Estimate of Essential Items for Compensatory Approach.....

Appendix 7 Self-Report Evaluation Form for Standard Setting Judges.....

PUBLICATION LIST

LIST OF TABLES

		PAGE
Table 2.1	*Standards* and Standard Setting.....	67
Table 2.2	Hambleton's (2001) Criteria for Evaluating a Standard Setting Study.....	79
Table 4.1	Grammar Elements Included in the Grammar Test Specifications...	136
Table 4.2	Ordering of Grammar Test Items Based on Expert Judgment According to Targeted Difficulty Level.....	138
Table 4.3	Reading Micro-skills Included in the Test Specifications.....	141
Table 4.4	Ordering of Reading Test Items Based on Expert Judgment According to Targeted Difficulty Level.....	143
Table 4.5	Breakdown of Subjects by Programme Used in the Data Analyses	149
Table 4.6	Breakdown of Essays Scored by Raters.....	150
Table 4.7	Breakdown of Judges According to Designation.....	157
Table 4.8(a)	Performance Level Descriptions for Grammar.....	159
Table 4.8(b)	Performance Level Descriptions for Reading.....	160
Table 4.8(c)	Performance Level Descriptions for Writing.....	160
Table 4.9	Summary of Research Objectives, Corresponding Research Questions, Data Sources and Data Analysis Procedures.....	175
Table 5.1	Item Polarity (Grammar Subtest).....	185
Table 5.2	Item Statistics According to Measure Order (Grammar Subtest).....	187
Table 5.3	Table of Standardized Residual Variance (Grammar Subtest).....	188
Table 5.4	Descriptive Statistics of Item Measures by Targeted Proficiency Level (Grammar Subtest).....	195
Table 5.5	Reliability of the Grammar Item Difficulty Estimates.....	198
Table 5.6	Reliability of Examinee Measures as Determined by the Grammar Subtest.....	199

PAGE

Table 5.7(a)	Percentage of Examinees with Infit Mean-squares below 0.8, between 0.8 and 1.2, and above 1.2 (Grammar Subtest).....	203
Table 5.7(b)	Percentage of Examinees with Outfit Mean-squares below 0.8, between 0.8 and 1.2, and above 1.2 (Grammar Subtest).....	203
Table 5.8	Item Statistics in Measure Order (Reading Subtest).....	205
Table 5.9	Table of Standardized Residual Variance (Reading Subtest).....	206
Table 5.10	Descriptive Statistics of Item Measures by Targeted Proficiency Level (Reading Subtest).....	211
Table 5.11	Reliability of the Reading Item Difficulty Estimates.....	215
Table 5.12	Reliability of Examinee Measures as Determined by the Reading Subtest.....	216
Table 5.13(a)	Percentage of Examinees with Infit Mean-squares below 0.8, between 0.8 and 1.2, and above 1.2 (Reading Subtest).....	218
Table 5.13(b)	Percentage of Examinees with Outfit Mean-squares below 0.8, between 0.8 and 1.2, and above 1.2 (Reading Subtest).....	218
Table 5.14	Item Statistics According to Measure Order (Multiple-choice Subtest)	221
Table 5.15	Standardized Residual Variance (Multiple-choice Test).....	223
Table 5.16	Descriptive Statistics of Item Measures by Targeted Proficiency Level (Multiple-Choice Subtest).....	227
Table 5.17	Reliability of the Multiple-choice Item Difficulty Estimates.....	229
Table 5.18	Reliability of Examinee Measures as Determined by the Multiple-choice Subtest.....	229
Table 5.19(a)	Percentage of Examinees with Infit Mean-squares below 0.8, between 0.8 and 1.2 and above 1.2 (Multiple-choice Subtest).....	232
Table 5.19(b)	Percentage of Examinees with Outfit Mean-squares below 0.8, between 0.8 and 1.2, and above 1.2 (Multiple-choice Subtest).....	232
Table 5.20	Number of Examinees in Each Subset.....	233
Table 5.21	Standard Deviations for the Estimation of Average Test Discrimination.....	234
Table 5.22	Examinee Reliability Index.....	238
Table 5.23	Item Measurement Report.....	238

PAGE

Table 5.24	Category Use and Step Difficulty (Content).....	241
Table 5.2	Category Use and Step Difficulty (Organization).....	242
Table 5.26	Category Use and Step Difficulty (Vocabulary).....	243
Table 5.27	Category Use and Step Difficulty (Language Use).....	245
Table 5.28	Category Use and Step Difficulty (Mechanics).....	246
Table 5.29	Rater Measurement Report.....	248
Table 5.30(a)	Raters with Infit Mean-square of above 1.5.....	249
Table 5.30(b)	Raters with Outfit Mean-square of above 1.5.....	249
Table 5.31	Raters with Infit and/or Outfit Mean-square of 0.5 and below.....	250
Table 5.32	Rating Distribution for All Raters by Criteria.....	251
Table 5.33(a)	Rating Pattern A (Ratings in All 4 Categories).....	252
Table 5.33(b)	Rating Pattern B (Concentration of Ratings in Category 2).....	253
Table 5.33(c)	Rating Pattern C (Concentration of Ratings in Category 1).....	253
Table 5.33(d)	Rating Pattern D (Concentration of Ratings in Category 3).....	253
Table 5.33(e)	Rating Pattern E (Ratings in 3 Upper Categories with Concentration in Category 3).....	254
Table 5.33(f)	Rating Pattern F (Ratings in 3 Upper Categories with Concentration in Category 2).....	254
Table 5.33(g)	Rating Pattern G (Ratings in Middle Categories with Concentration in Category 3).....	255
Table 5.33(h)	Rating Pattern H (Ratings in Middle Categories with Concentration in Category 2).....	255
Table 5.34	Summary of Examinee Fit Statistics (Writing Subtest).....	262
Table 5.35(a)	Percentage of Examinees with Infit Mean-squares below 0.5, between 0.5 and 1.5 and above 1.5 (Writing Subtest).....	262
Table 5.35(b)	Percentage of Examinees with Outfit Mean-squares below 0.5, between 0.5 and 1.5, and above 1.5 (Writing Subtest).....	262
Table 5.36	Item Measurement Report (Compensatory Approach).....	266
Table 5.37	Examinee Reliability and Separation Indexes (Compensatory Approach).....	267

PAGE

Table 5.38(a)	Percentage of Examinees with Infit Mean-squares below 0.5, between 0.5 and 1.5 and above 1.5 (Compensatory Approach).....	269
Table 5.38(b)	Percentage of Examinees with Outfit Mean-squares below 0.5, between 0.5 and 1.5, and above 1.5 (Compensatory Approach).....	269
Table 5.39	Rater Measurement Report (Compensatory Approach).....	270
Table 5.40(a)	Raters with Infit Mean-square of above 1.5 (Compensatory Approach)..	271
Table 5.40(b)	Raters with Outfit Mean-square of above 1.5 (Compensatory Approach).....	271
Table 5.41	Raters with Infit and/or Outfit Mean-square of Below 0.5 (Compensatory Approach).....	271
Table 5.42	Grammar Subtest Criterion Points Estimated With ± 1.6 Standard Errors.....	274
Table 5.43	Grammar Subtest Final Cutscores.....	276
Table 5.44	Frequency and Percentage of Examinees by Proficiency Level (Grammar Subtest).....	279
Table 5.45	Mean Ability and Distribution Statistics by Proficiency Level (Grammar Subtest).....	279
Table 5.46	Reading Subtest Criterion Points Estimated With ± 1.6 Standard Errors.....	280
Table 5.47	Reading Subtest Final Cutscores.....	282
Table 5.48	Frequency and Percentage of Examinees by Proficiency Level (Reading Subtest).....	283
Table 5.49	Mean ability and Distribution Statistics by Proficiency Level (Reading Subtest).....	283
Table 5.50	Writing Subtest Criterion Points with ± 1.6 Standard Errors.....	284
Table 5.51	Writing Subtest Final Cutscores.....	286
Table 5.52	Frequency and Percentage of Examinees by Proficiency Level (Writing Subtest).....	288
Table 5.53	Mean Ability and Distribution Statistics by Proficiency Level (Writing Subtest).....	288
Table 5.54	Criterion Points Estimated With ± 1.6 Standard Errors (Compensatory Approach).....	289
Table 5.55	Compensatory Approach Final Cutscores.....	291

PAGE

Table 5.56	Frequency and Percentage of Examinees by Proficiency Level (Compensatory Approach).....	293
Table 5.57	Mean Ability and Distribution Statistics by Proficiency Level (Compensatory Approach).....	293
Table 5.58	Success of Judge Training.....	295
Table 5.59	Adequacy of Time Allocation.....	296
Table 5.60	Adequacy of Performance Level Descriptions.....	297
Table 5.61	Adequacy of Performance Level Descriptions (Grammar).....	297
Table 5.62	Adequacy of Performance Level Descriptions (Reading).....	298
Table 5.63	Adequacy of Performance Level Descriptions (Writing).....	299
Table 5.64	Identification of Essential Items.....	301
Table 5.6	Confidence in Deciding the Essentiality of Items.....	302
Table 5.66	Confidence in Deciding Essentiality of Items by Subtest.....	302
Table 5.67	Confidence in Classification of Examinees.....	303
Table 5.68	Efficacy of the Standard Setting Method.....	304
Table 5.69	Distribution of Grammar Items across Criterion Point by Individual Judges.....	307
Table 5.70	Frequency of Judges' Selection of Grammar Items by Criterion Point (Grammar Subtest).....	309
Table 5.71	Descriptive Statistics of Judges' Mean Estimates of Essential Items (Grammar Subtest).....	312
Table 5.72	Distribution of Items across Criterion Point for Most Severe and Most Lenient Judges (Grammar Subtest).....	320
Table 5.73	Judge Measurement Report (Grammar Subtest).....	321
Table 5.74	Item Measurement Report (Grammar Subtest).....	323
Table 5.75	Cutscores and Hierarchical Ordering of Test Items (Grammar Subtest).....	325
Table 5.76	Item Description and Performance Level Description for Cutscore 1 (Grammar Subtest).....	326
Table 5.77	Item Description and Performance Level Description for Cutscore 2 (Grammar Subtest).....	327

PAGE

Table 5.78	Item Description and Performance Level Description for Cutscore 3 (Grammar Subtest).....	328
Table 5.79	Item Description and Performance Level Description for Cutscore 4 (Grammar Subtest).....	329
Table 5.80	Distribution of Reading Items across Criterion Point by Individual Judges.....	331
Table 5.81	Frequency of Judge Selection of Reading Items by Criterion Point.....	332
Table 5.82	Descriptive Statistics of Judges' Mean Estimates of Essential Items (Reading Subtest).....	335
Table 5.83	Judge Measurement Report (Reading Subtest).....	343
Table 5.84	Item Measurement Report (Reading Subtest).....	345
Table 5.85	Cutscores and Hierarchical Ordering of Test Items (Reading Subtest)	346
Table 5.86	Item Description and Performance Level Description for Cutscore 1 (Reading Subtest).....	347
Table 5.87	Item Description and Performance Level Description for Cutscore 2 (Reading Subtest).....	347
Table 5.88	Item Description and Performance Level Description for Cutscore 3 (Reading Subtest).....	348
Table 5.89	Item Description and Performance Level Description for Cutscore 4 (Reading Subtest).....	349
Table 5.90(a)	Distribution of Individual Judges' Ratings for Content across Criterion Point (Writing Subtest).....	350
Table 5.90(b)	Distribution of Individual Judges' Ratings for Organization across Criterion Point (Writing Subtest).....	351
Table 5.90(c)	Distribution of Individual Judges' Ratings for Vocabulary across Criterion Point (Writing Subtest).....	352
Table 5.90(d)	Distribution of Individual Judges' Ratings for Language Use across Criterion Point (Writing Subtest).....	352
Table 5.90(e)	Distribution of Individual Judges' Ratings for Mechanics across Criterion Point (Writing Subtest).....	353
Table 5.91	Descriptive Statistics of Judges' Mean Estimates (Writing Subtest)....	354
Table 5.92	Judge Measurement Report (Writing Subtest).....	362
Table 5.93	Item Measurement Report (Writing Subtest).....	363

PAGE

Table 5.94	Item Description and Performance Level Description for Cutscore 1 (Writing Subtest).....	365
Table 5.95	Item Description and Performance Level Description for Cutscore 2 (Writing Subtest).....	366
Table 5.96	Item Description and Performance Level Description for Cutscore 3 (Writing Subtest).....	366
Table 5.97	Item Description and Performance Level Description for Cutscore 4 (Writing Subtest).....	367
Table 5.98	Descriptive Statistics of Judges' Mean Estimates of Essential Items (Compensatory Approach).....	368
Table 5.99	Judge Measurement Report (Compensatory Approach).....	374
Table 5.100	Item Measurement Report (Compensatory Approach).....	376
Table 5.101	Cutscores (Grammar Subtest).....	377
Table 5.102	Distance between Cutscores (Grammar Subtest).....	378
Table 5.103	Descriptive Statistics (Grammar Subtest).....	378
Table 5.104	Descriptive Statistics: SPM English 1119 (Grammar Subtest).....	380
Table 5.105	Comparability of Pass Rates on Two Tests of English Proficiency (Grammar Subtest).....	380
Table 5.106	Descriptive Statistics of Cohorts 1 and 2 by Proficiency Level (Grammar Subtest).....	381
Table 5.107	Cutscores (Reading Subtest).....	382
Table 5.108	Distance between Cutscores (Reading Subtest).....	382
Table 5.109	Descriptive Statistics (Reading Subtest).....	383
Table 5.110	Descriptive Statistics: SPM English 1119 (Reading Subtest).....	383
Table 5.111	Comparability of Pass Rates on Two Tests of English Proficiency (Reading Subtest).....	384
Table 5.112	Descriptive Statistics of Cohorts 1 and 2 by Proficiency Level (Reading Subtest).....	384
Table 5.113	Cutscores (Writing Subtest).....	385
Table 5.114	Distance between Cutscores (Writing Subtest).....	385
Table 5.115	Descriptive Statistics (Writing Subtest).....	386

PAGE

Table 5.116	Descriptive Statistics: SPM English 1119 (Writing Subtest).....	386
Table 5.117	Comparability of Pass Rates on Two Tests of English Proficiency (Writing Subtest).....	387
Table 5.118	Descriptive Statistics of Cohorts 1 and 2 by Proficiency Level (Writing Subtest).....	387
Table 5.119	Cutscores (Compensatory Approach).....	388
Table 5.120	Distance between Cutscores (Compensatory Approach).....	388
Table 5.121	Descriptive Statistics (Compensatory Approach).....	389
Table 5.122	Descriptive Statistics: SPM English 1119 (Compensatory Approach)..	389
Table 5.123	Comparability of Pass Rates on Two Tests of English Proficiency (Compensatory Approach).....	390
Table 5.124	Descriptive Statistics of Cohorts 1 and 2 by Proficiency Level (Compensatory Approach).....	390

LIST OF FIGURES

		PAGE
Figure 1.1	The English Language Curriculum Structure of the International Islamic University Malaysia.....	1
Figure 1.2	The English Language Placement Test Structure.....	10
Figure 1.3	The Present English Language Placement System (2001-2004).....	12
Figure 2.1	Performance Standard.....	43
Figure 2.2	Berk's Trilevel Classification Scheme.....	51
Figure 2.3	Kane's Framework for Examining Validity of Performance Standards and Cut Scores.....	75
Figure 2.4	Characteristics of Performance Assessment.....	90
Figure 2.5	Messick's Facets of Test Validity.....	97
Figure 3.1	Translation of Performance Level Description into Corresponding Cutscore for SR and CR Items.....	119
Figure 3.2	Visual Representation of Conceptual Framework of Study.....	130
Figure 4.1	Procedure of Study.....	134
Figure 4.2	Essay Marking Profile.....	147
Figure 4.3	Linking Procedure using Common Person Equating.....	152
Figure 4.4	Cutscores and Corresponding Benchmarks for the English Language Support Courses.....	155
Figure 4.5(a)	Comparing Student Measure and Criterion Point on the Measured Scale.....	162
Figure 4.5(b)	Estimating Student and Criterion Locations within Error Band.....	163
Figure 4.5(c)	Marking out the Confidence Interval Using the Normal Distribution	163
Figure 4.5(d)	Marking out Criterion Region within ± 1.6 Standard Error of Measurement.....	154
Figure 4.5(e)	Marking out Student Region within ± 1.6 Standard Error of Measurement.....	165

PAGE

Figure 4.5(f)	Adjusting Student Measure to Lower Boundary (-1.6 S.E.) to Guarantee Quality.....	165
Figure 4.6	Translation of Performance Level Description into Corresponding Cutscore for SR and CR Items.....	166
Figure 4.7	Data Matrix for Facets Analysis.....	172
Figure 5.1	Wright Map for the Grammar Subtest.....	184
Figure 5.2	Spread, Overlaps and Gaps of Grammar Test Items.....	191
Figure 5.3	Stacks and Gaps (Grammar Subtest).....	192
Figure 5.4	Scaling of Test Items Based on Expert Judgment and Empirical Calibration (Grammar Subtest).....	194
Figure 5.5	Means of Targeted Items by Proficiency Levels Estimated with ± 2.0 Standard Errors(Grammar Subtest).....	196
Figure 5.6	Item Ordering by Grammar Elements and Targeted Proficiency Level.....	197
Figure 5.7(a)	Means of Item Calibrations and Examinee Ability with ± 2.0 Standard Errors (Grammar Subtest).....	201
Figure 5.7(b)	Targeting of the Grammar Subtest (SD).....	202
Figure 5.8	Wright Map for the Reading Subtest.....	204
Figure 5.9	Spread, Overlaps and Gaps of Reading Test Items.....	208
Figure 5.10	Stacks and Gaps (Reading Subtest).....	209
Figure 5.11	Scaling of Test Items Based on Expert Judgment and Empirical Calibration (Reading Subtest).....	211
Figure 5.12	Means of Targeted Items by Proficiency Levels Estimated with ± 2.0 Standard Errors (Reading Subtest).....	212
Figure 5.13(a)	Item Ordering by Skill Area and Targeted Proficiency Level (Reading Subtest).....	214
Figure 5.13(b)	Item Ordering by Skill Area and Targeted Proficiency Level (Reading Subtest).....	214
Figure 5.14(a)	Means of Item Calibrations and Examinee Ability with ± 2.0 Standard Errors (Reading Subtest).....	217
Figure 5.14(b)	Targeting of the Reading Subtest (SD).....	217
Figure 5.15	Wright Map of 75 Multiple-choice Items.....	220
Figure 5.16	Spread, Overlaps and Gaps of Multiple-Choice Test Items.....	224

PAGE

Figure 5.17	Stacks and Gaps (Multiple-choice Subtest).....	225
Figure 5.18	Scaling of Multiple-choice Test Items Based on Expert Judgment and Empirical Calibration (Multiple-choice Subtest).....	226
Figure 5.19	Means of Targeted Items by Proficiency Levels Estimated with ± 2.0 Standard Errors (Multiple-choice Subtest).....	228
Figure 5.20(a)	Means of Item Calibrations and Examinee Ability Estimated with ± 2.0 Standard Errors (Multiple-choice Subtest).....	230
Figure 5.20(b)	Targeting of the Multiple-choice Subtest (SD).....	231
Figure 5.21	Examinee Ability, Rater Severity and Item (Criteria) Difficulty.....	236
Figure 5.22	Probability Curves (Content).....	241
Figure 5.23	Probability Curves (Organization).....	242
Figure 5.24	Probability Curves (Vocabulary).....	244
Figure 5.25	Probability Curves (Language Use).....	245
Figure 5.26	Probability Curves (Mechanics).....	246
Figure 5.27	Scatterplot of Raw Scores and Essay Measures.....	257
Figure 5.28	Histogram Showing Distribution of Examinee Raw Scores or Ratings.....	258
Figure 5.29	Histogram Showing Distribution of Examinee Essay Measures.....	259
Figure 5.30	Scatterplot of Raw Scores and Essay Measures of Selected Raters Displaying Rating Inconsistencies.....	260
Figure 5.31	Scatterplot of Raw Scores and Essay Measures of Raters with High Rating Consistency.....	261
Figure 5.32	Student Ability, Rater Severity and Item Difficulty Distributions (Compensatory Approach).....	264
Figure 5.33	Means of Item Calibrations and Examinee Ability within 2.0 Standard Errors (Compensatory Approach).....	268
Figure 5.34	Plot of Examinee Measures Based on Two Different Essay Ratings by Different Examiners and on Two Different Occasions....	273
Figure 5.35	Grammar Subtest Criterion Points marked with ± 1.6 Standard Errors.....	275
Figure 5.36	Grammar Subtest Final Cutscores Applied to Examinee and Item Distributions (Wright Map).....	277

PAGE

Figure 5.37	Adjusting Examinees' Calibrated Measures by -1.6 Standard Errors	278
Figure 5.38	Reading Subtest Criterion Points Marked with ± 1.6 Standard Errors	281
Figure 5.39	Reading Subtest Final Cutscores Applied to Examinee and Item Distributions (Wright Map)	282
Figure 5.40	Writing Subtest Criterion Points Marked With ± 1.6 Standard Errors	285
Figure 5.41	Writing Subtest Final Cutscores Applied to Examinee and Item Distributions	287
Figure 5.42	Compensatory Criterion Points Marked with ± 1.6 Standard Errors	290
Figure 5.43	Final Cutscores Applied to Examinee and Item Distributions (Wright Map)	292
Figure 5.44	Distribution of Judges' Mean Estimates for the Four Criterion Points (Grammar Subtest)	311
Figure 5.45	Boxplots of Judges' Mean Estimates for the Four Criterion Points (Grammar Subtest)	313
Figure 5.46	Judges' Estimation of the Four Criterion Points (Grammar Subtest)	314
Figure 5.47	Individual Judges' Estimation of the Four Criterion Points (Grammar Subtest)	315
Figure 5.48	Calibrations of Judge Severity, Criterion Points and Test Items (Grammar Subtest)	319
Figure 5.49	Distribution of Judges' Mean Estimates for the Four Criterion Points (Reading Subtest)	334
Figure 5.50	Boxplots of Judges' Mean Estimates for the Four Criterion Points (Reading Subtest)	336
Figure 5.51	Judges' Estimation of the Four Criterion Points (Reading Subtest)	337
Figure 5.52	Individual Judges' Estimation of the Four Criterion Points (Reading Subtest)	338
Figure 5.53	Calibrations of Judge Severity, Criterion Points and Test Items (Reading Subtest)	342
Figure 5.54	Boxplots of Judges' Mean Estimates for the Four Criterion Points (Writing Subtest)	355
Figure 5.55	Judges' Estimation of the Four Criterion Points (Writing Subtest)	356

PAGE

Figure 5.56	Individual Judges' Estimation of the Four Criterion Points (Writing Subtest).....	357
Figure 5.57	Calibrations of Judge Severity, Criterion Points, Test Items and Rating Categories (Writing Subtest).....	361
Figure 5.58	Cutscores, Hierarchical Ordering of Items and Rating Categories (Writing Subtest).....	364
Figure 5.59	Boxplots of Judges' Mean Estimates for the Four Criterion Points (Compensatory Approach).....	369
Figure 5.60	Judges' Estimation of the Four Criterion Points (Compensatory Approach).....	370
Figure 5.61	Individual Judges' Estimation of the Four Criterion Points (Compensatory Approach).....	371
Figure 5.62	Calibrations of Judge Severity, Criterion Points, Test Items and Rating Categories (Compensatory Approach).....	373

MENGGUNAKAN MODEL PENGUKURAN RASCH BAGI PENETAPAN STANDAD UNTUK UJIAN PENEMPATAN BAHASA INGGERIS DI IIUM

ABSTRAK

Dengan penggunaan skor sempadan dan standad untuk membuat keputusan-keputusan pendidikan yang berciri “high-stakes”, pelbagai usaha seharusnya dibuat untuk mencari kaedah-kaedah penetapan standad yang tidak dipertikaikan. Kajian ini adalah merupakan salah satu usaha kearah matlamat tersebut. Tujuan utama kajian ini adalah untuk menyelidik keberkesanan Kaedah Penetapan Standad Objektif, yang dilandaskan kepada Model Pengukuran Rasch, di dalam pembinaan skor sempadan berganda yang sah dan boleh dipertahankan bagi ujian-ujian yang menggunakan jenis item yang pelbagai. Kaedah Penetapan Standad Objektif yang diperkenalkan oleh Stone (1996) untuk penetapan satu skor sempadan bagi ujian yang menggunakan item-item berbentuk item pilih telah dibuktikan boleh menghasilkan keputusan yang sah. Walaubagaimanapun keberkesananya bagi ujian yang menggunakan jenis item yang pelbagai dan keberkesanannya untuk membina skor sempadan berganda masih belum ditentukan secara empirikal. Oleh kerana kualiti ujian yang digunakan di dalam sesuatu kajian penetapan standad boleh menjejas kesahan skor-skor sempadan yang dihasilkan dan kesahan klasifikasi pelajar, isu-isu berkaitan dengan penilaian juga perlu diambil kira. Begitu juga dengan model pengukuran yang digunakan. Ia harus berkemampuan untuk menghubungkan pencapaian pelajar (di dalam ujian) dan kedudukan mereka (berdasarkan standad yang ditetapkan) dengan konstruk yang diukur secara terus. Ia juga harus berkemampuan untuk menukar bilangan betul kepada ukuran linear jeda yang tidak bergantung kepada sampel atau ujian yang digunakan. Selain itu, teori pengukuran yang digunakan juga harus berkeupayaan untuk menyelesaikan isu-isu penting di dalam pengukuran dan penetapan standad. Di dalam kajian ini,

keberkesanan Kaedah Penetapan Standad Objektif telah diuji di dalam konteks Ujian Penempatan Bahasa Inggeris yang ditadbirkan di IIUM. Didapati bahawa dengan penggunaan Kaedah Penetapan Standad Objektif penetapan skor sempadan berganda bagi ujian yang menggunakan jenis item yang pelbagai boleh dilakukan dengan mudah tanpa menjejas kesahan skor sempadan atau standad yang dihasilkan. Selain dari itu, penggunaan Kaedah Penetapan Standad Objektif juga membolehkan tahap pencapaian yang diingini diterjemahkan secara terus kepada konstruk yang diukur. Ini memberikan makna sebenar kepada standad yang dihasilkan dan bukan sekadar nisbah jawapan betul. Model pengukuran Rasch juga telah dibuktikan berguna di dalam menyelesaikan isu-isu asas di dalam pengukuran dan penetapan standad. Namun begitu, harus diingat bahawa sebaik mana sekali pun sesuatu kaedah penetapan standad yang digunakan, hasil sesuatu kajian penetapan standad tetap dipengaruhi oleh kualiti ujian, kebolehan pakar, diskripsi tahap pencapaian yang diinginkan dan lain-lain variabel di dalam proses penetapan standad. Perkara ini jelas ditunjukkan di dalam kajian ini. Oleh yang demikian, langkah-langkah sesuai harus diambil bagi menangani isu-isu di atas agar kesahan skor sempadan yang diperolehi tidak terkompromi.

USING THE RASCH MEASUREMENT MODEL FOR STANDARD SETTING OF THE ENGLISH LANGUAGE PLACEMENT TEST AT THE IIUM

ABSTRACT

With the use of cutscores and standards for making high-stakes educational decisions, efforts should be made to search for more defensible standard setting methods. This study represents an effort to this end. The main intent of this study is to investigate the efficacy of the Objective Standard Setting Method (OSS), which is based on the Rasch Measurement Model, in constructing multiple cutscores that are valid and defensible on tests utilizing diverse item types. The OSS, which was developed by Stone (1996) to set a single cutscore on tests utilizing selected-response items, has been demonstrated to yield valid results. However, its efficacy in handling other item types and the construction of multiple cutscores has yet to be empirically established. As the quality of the tests used in the standard setting endeavour influences the validity of derived cutscores as well as the validity of examinee classification, assessment-related issues are also of major concern. Measurement theory is one other aspect that requires serious consideration. The need for a measurement model that transforms counts correct into interval linear measures that are neither sample-bound nor test-bound, and at the same time references an examinee's performance (on the test) and status (based on the standards set) directly to the measured construct cannot be underrated. The same applies to the capacity to resolve important measurement and standard setting issues. In this study the efficacy of the OSS was examined in the context of the English Language Placement Test conducted at the IIUM. It was found that with the use of the OSS, multiple cutscores on diverse item types can be easily set without compromising the validity of the derived cutscores or standards. Additionally, with the

use of the OSS, the desired level of attainment can be directly translated onto the measured construct and, thus, allowing the standards set to have real meaning and not just proportions of correct answers. The Rasch measurement model has also proved to be useful in resolving fundamental issues in measurement and standard setting. However, one cautionary word is necessary. Regardless of how sound a standard setting method is, the results of a standard setting study are bound to be impacted by test quality, judge competency, performance level descriptions and other variables in the standard setting process. This has been demonstrated very clearly in this study. Steps must, therefore, be taken to address all these issues to ensure that the reliability and validity of derived cutscores and standards are not compromised.

CHAPTER 1

INTRODUCTION

1.0 INTRODUCTION

The concept of a performance standard, which deals with the question of "how good is good enough" with respect to the attainment of educational standards, has been the subject of considerable attention, and is considered to be one of the most controversial issues in educational measurement (Linn & National Centre for Research on Evaluation, Standards, and Student Testing, 2003; Zieky, 2001; Cizek, 2001). This is hardly surprising as there are a number of well-founded reasons for the controversy surrounding the use of performance standards.

The first pertains to the accuracy or appropriateness of the standard set. Setting too high or too low a performance standard wields lasting consequences on many stakeholders (Popham, 2000). If excessively high standards are set on a high-stakes competency test, failure to attain the set standards could result in unwarranted sanctions for schools (Linn et al., 2003) as well as inequitable penalties on students (Popham, 2000). Conversely, if excessively low standards are set, detrimental consequences on the value of education will result (Popham, 2000).

The second relates to the negative consequences that result from their use, or rather misuse, for purposes of educational accountability. In discussing the utility of performance standards for different uses of assessments, Linn et al. (2003) states with consternation, "performance standards have been mandated or become the preferred method of reporting assessment results where the standards are not essential to the use" (p. 1). This, he asserts, is of no real consequence in situations "when there are no requirements of achieving them, but it is another matter altogether when there are serious sanctions for falling short" (p. 3).

The next central cause of dissent on the use of performance standards concerns issues related to assessment, particularly with the introduction of large-scale high-stakes standardized testing. Contentions against the use of high-stakes standardized tests are not without legitimate reasons. The impact of large scale standardized testing on educational policy decisions is considerable (Airasian & Madaus, 1983). So is the negative impact on instruction and learning resulting from narrowing of the curricula due to focused teaching (e.g., Zieky, 2001; Linn, 2000).

The common practice of using scores from a single standardized test for different and possibly conflicting decision-making purposes is another valid reason for the controversy over the use of performance standards. The danger of over reliance on a single measure of student performance is argued by the Pennsylvania State Education Association (2003), and caution against it is explicated in *The Standards* (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999, p. 146):

> In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decisions.

Another assessment-related issue raised concerns test quality and testing procedures. Bolon (2000) in discussing school-based standardized testing in the United States, points out that errors in standardized test scores are 'enormous'. Errors in the scoring of Vermont's mathematics and language arts tests in 1998 and 1999 (Bowman as cited in Zieky, 2001) are cases in point. Poor test quality has also been reported in relation to standardized tests in New York. Hursh (2005, p. 612) states that "...almost every recent standardized examination in New York has been criticized for having poorly constructed, misleading, or erroneous questions".

Misclassifications of students due to measurement error (Hambleton, 1978; Crocker & Algina, 1986) and the indiscriminate application of cutscores set on one

form on an alternate form (e.g., Lee, Lewis, Hanson & Egan, 2002) are yet other legitimate causes for concern (Jaeger, 1993). The issue of protecting "innocence" and guaranteeing "quality" (Wright & Grosse, 1993) is a nontrivial matter, and one that requires serious and careful deliberation.

However, the most important reason for the controversial use of performance standards has to do with the judgmental nature of the standard setting process in which cutscores that correspond to pre-specified performance levels are established. The lack of objectivity due to use of human judgment in constructing cutscores instead of "a straightforward process of parameter estimation" (Kane, 2001, p. 81), to some experts, renders performance standards arbitrary, and thus invalid at worst or imprudent at best (e.g., Glass, 1978; Burton, 1977). Glass (1978) in his highly controversial paper argues,

> To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. Arbitrariness is no bogeyman, and one ought not to shrink from necessary decisions because they may be arbitrary. However, arbitrary decisions often entail substantial risks of disruption and dislocation. Less arbitrary is safer (p. 258).

The concern regarding the arbitrariness of the process in which cutscores are established is also expressed by Kane (2001).

> …one source of arbitrariness arises from the fact that the score scale is generally a continuous function of the level of achievement. As a result, there is no simple and obvious way to choose a particular point on the score scale as the cutscore, and there is no compelling reason why it could not be set a little higher or a little lower (p. 81).

Despite the controversy that shrouds the use of performance standards, there are legitimate grounds for their use in educational decision-making (Hambleton, 1978: Burton, 1977; Popham, 2000; Cizek, 2001; Linn et al., 2003). In contexts where assessments are used for certification or licensure, performance standards are deemed essential (Linn et al., 2003). What is considered a minimal level of competency needs to be clearly ascertained to "protect the public from incompetent

practitioners" (Linn et al., 2003, p. 2). Though the problems of misclassifications cannot be avoided (Ebel, as cited in Hambleton, 1978), standards still need to be set "[as] there are legitimate practical reasons that require that a decision be made" (Linn et al., 2003, p. 2).

Performance standards are also essential to provide invaluable feedback for continued curricular and instructional improvement (Burton, 1977; Linn, 2000). They allow for "tracking progress of achievement for schools, states or the nation" (Linn et al., 2003, p. 3) and more importantly, for the monitoring and improvement of student learning. The standard-based educational reform in the US and the literacy movement in Australia are cases in point. In the classroom context, performance standards provide educators with a diagnosis of what is lacking and corrective measures that need to be taken as a result of acceptable or unacceptable performance (Burton, 1977).

The setting of performance standards inevitably involves human judgment and, therefore, is not infallible. However, this does not mean that the setting of educational standards should be avoided as standards are crucial in the educational decision-making process. What needs to be borne in mind is that there must be clear and valid reasons for the use of performance standards in order to avoid undesirable consequences. Standard setting is a highly technical (Marzano & Kendall, 1997) and judgmental process (Messick, 1975; Hambleton, 1978; Glass, 1978; Jaeger, 1993; Kane, 1994; Linn et al., 2003). Therefore, it has to be handled with great prudence and a consciousness of what it entails and the stakes involved, as appropriately argued by Popham (2000),

> …when human beings apply their judgmental powers to the solution of problems, mistakes will be made. However, the fact that judgmental errors are possible should not send educators scurrying from such tasks as the setting of standards. Judges and juries are capable of error, yet they do the best job they can. Similarly, educators are now faced with the necessity of establishing performance standards, and they, too, must do the best job they can. That educational performance standards need to be set in order for instructional

> decisions to be made is indisputable. That those standards will, in the final analysis, be set judgmentally is equally indisputable. However, that all judgmental standards must be arbitrary is decidedly disputable. Approaching the standard-setting task seriously, taking full cognizance of available data and the preferences of concerned constituencies, need not result in an enterprise that is arbitrary and capricious. On the contrary, the act of standard-setting can reflect the very finest form of judgmental decision-making (p. 372).

With greater demands for better quality education and greater improvements in student learning and achievement, the role of performance standards has come to the forefront and needs to be dealt with openly (Popham, 2000). However, great care needs to be exercised to ensure that whatever standards are set are not only theoretically and psychometrically defensible but also take into consideration the educational context they serve and the people whose lives they affect.

1.1 CONTEXT OF STUDY

The International Islamic University Malaysia (IIUM), unlike most other government-funded institutions of higher learning in Malaysia, uses English as one of its mediums of instruction for both its matriculation programme, and its postgraduate and undergraduate programmes. As such, it is essential that its students possess an appropriate level of English Language proficiency in order to cope with the rigorous demands of academic study in the English Language. The need to ensure that students have the necessary language skills to succeed in their academic study, and the need to provide those who are lacking in the skills required with remedial instruction are greatly recognized, and are of serious concern to the university.

In the effort to meet this need, the Centre for Languages and Pre-University Academic Development (CELPAD) of the International Islamic University Malaysia has been charged with the task of assessing the English language proficiency of incoming students and providing English Language support courses to those who require them. Hence, the placement system adopted by CELPAD, like many other

placement systems as described by Sawyer (1996), has been designed to constitute an assessment component that estimates students' probability of success in standard first-year courses as well as an instructional component that provides underprepared students with instruction in the language skills and knowledge they need to succeed in the standard first year courses.

1.1.1 The CELPAD Placement System (1995 -2000)

From the years 1995 to 2000, CELPAD adopted a three-tiered curriculum structure in its instructional component at the Matriculation Centre of the IIUM. Its assessment component, on the other hand, was a two-part battery comprising five subtests. The curriculum structure (inevitably the assessment component as well) was revised in 2001 as a response to the reports of the declining standards of English language proficiency among students. The following brief description of the 1995 – 2000 curriculum and assessment system provides the necessary background to the issue at hand.

1.1.1.1 Instructional Component of the Placement System (1995 – 2000)

The first tier of the 1995-2000 curriculum structure comprised two sub-levels: Core Competence Lower (CCL) and Core Competence Upper (CCU). These courses focused on the development of English Language fluency and form in meaningful contexts through the aural/oral direct approach. The English Language grammar system "was taught inductively in given contexts, and discourse was predominantly dealt with at the sentential level". At this level, reading and writing skills were of secondary concern (Centre for Pre-University Academic Development, 1993, p. 47).

The courses conducted in the second tier, on the other hand, focused on the development of academic English language skills. These courses dealt with study skills – which involved the abilities, techniques, strategies which are used when

reading, writing or listening for study purposes as well as the use of general academic English. The courses conducted at this level were the Skills Focused Course (Listening and Speaking) and the Skills Focused Course (Reading and Writing).

The third tier, English for Specific Academic Purposes (ESAP), comprised subject-specific English language courses focusing on the types of discourse specific to the needs of individual Kulliyyahs or faculties (e.g., Economics, Engineering and Architecture) at the Main Campus.

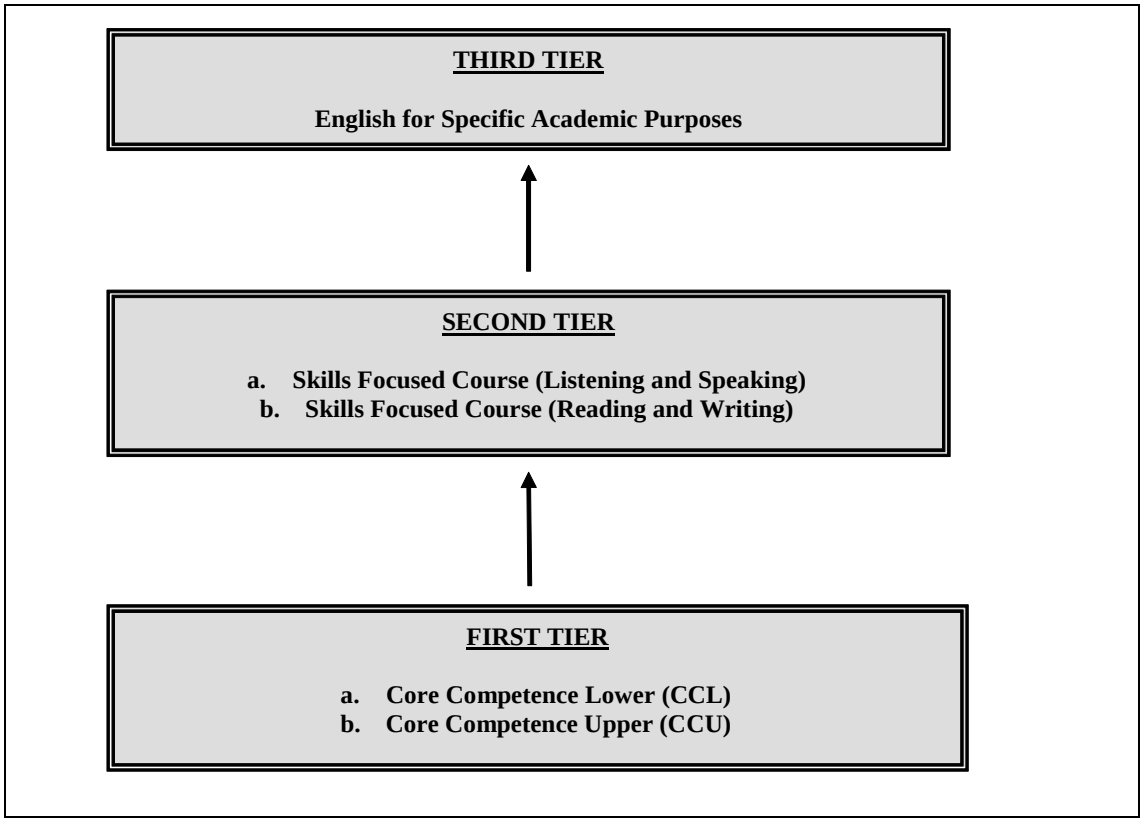

Figure 1.1 The English Language Curriculum Structure of the International Islamic University

1.1.1.2 Assessment Component of the Placement System (1995 – 2000)

The assessment component of this placement system, the English Language Placement Test (EPT), was a two-part placement test battery. It served two main functions. The first was to ascertain whether incoming students met the required minimum level of English language proficiency for purposes of undertaking

undergraduate or postgraduate studies. Its second function was to place those who did not demonstrate the required level of English language proficiency (and therefore did not qualify to undertake content subject courses) into relevant language support courses for remediation. In the context of the Matriculation Centre of the IIUM, the EPT served yet another function: an exit/entry requirement. Matriculation students had to meet the minimum English language requirement in order to graduate from the matriculation programme to gain entry into the Kulliyyah (faculty) at the Main Campus.

The EPT, a criterion-referenced test, was based on the view that language ability is partially divisible rather than unitary following the current view of second language proficiency (see Bachman & Palmer, 1996). Thus, the test was structured according to the kinds of language skills that are seen to define language ability: reading, writing, listening, and speaking. The assessment of grammatical ability was also included in the EPT as it is a widely accepted notion that knowledge in grammar underlies the ability to use language to express meaning (e.g., Bachman, 1990; Canale & Swain, 1980), and it is a common feature of high-stakes language proficiency tests (Hughes, 1989).

The first part of the EPT battery, the EPT Core Test, was a general proficiency test focusing mainly on the assessment of grammatical competence and performance. It consisted of five sections: completion passage (grammar), error identification and error correction (grammar), reading comprehension, standard cloze and paragraph writing. Students who achieved the Minimum Basic Adequate Score (50% above) on this test were allowed to proceed to Part II of the placement test. On the other hand, those who failed to fulfil the minimum requirement were placed in the relevant first tier proficiency courses (CELPAD, 1993, p. 53).

Students scoring 34% and below were placed into the Core Competence Lower course (CCL) while those scoring between 35% to 49% into the Core

Competence Upper course (CCU) (Refer to Figure 1.1). In this part of the EPT battery, two cutscores were set. The first, which was 50%, served to separate those who were eligible to proceed to part II of the EPT and those who would be placed into the first tier of the curriculum (instruction component). The second cutscore, 35%, separated examinees who belonged in the first tier of the curriculum structure into the two groups: CCL and CCU.

The second part of the EPT, on the other hand, consisted of a battery of skills-based tests covering reading, writing, listening and speaking skills. These tests aimed at assessing competency and performance in those four language skills. Students who did not perform adequately in these tests were placed in the respective skills-based courses whereas those who attained a Band 6 on all the skills tests, on the other hand, were considered to have achieved the minimum language requirement and, therefore, were exempted from language support courses at the matriculation centre (Figure 1.2).

After being placed into the relevant language support courses, and undergoing a semester of instruction students were required to re-sit the relevant subtests of the placement battery. Those who were placed in the first tier courses were required to re-sit the Core Test. The same cutscores were applied. Examinees who met the 50% cut point advanced to the skills tests and those who did not were given further remediation in the relevant language support courses.

The same procedure was applied to students in the skills-based courses. If they attained the expected criterion level, which is Band 6 on all the skills tests, they were deemed to have met the minimum English language requirement. This meant that they had achieved the required standard. Upon completion of their matriculation courses and successfully passing end-of-semester examinations, these students would gain entry into the Kulliyyah (i.e., faculty) at the Main Campus.

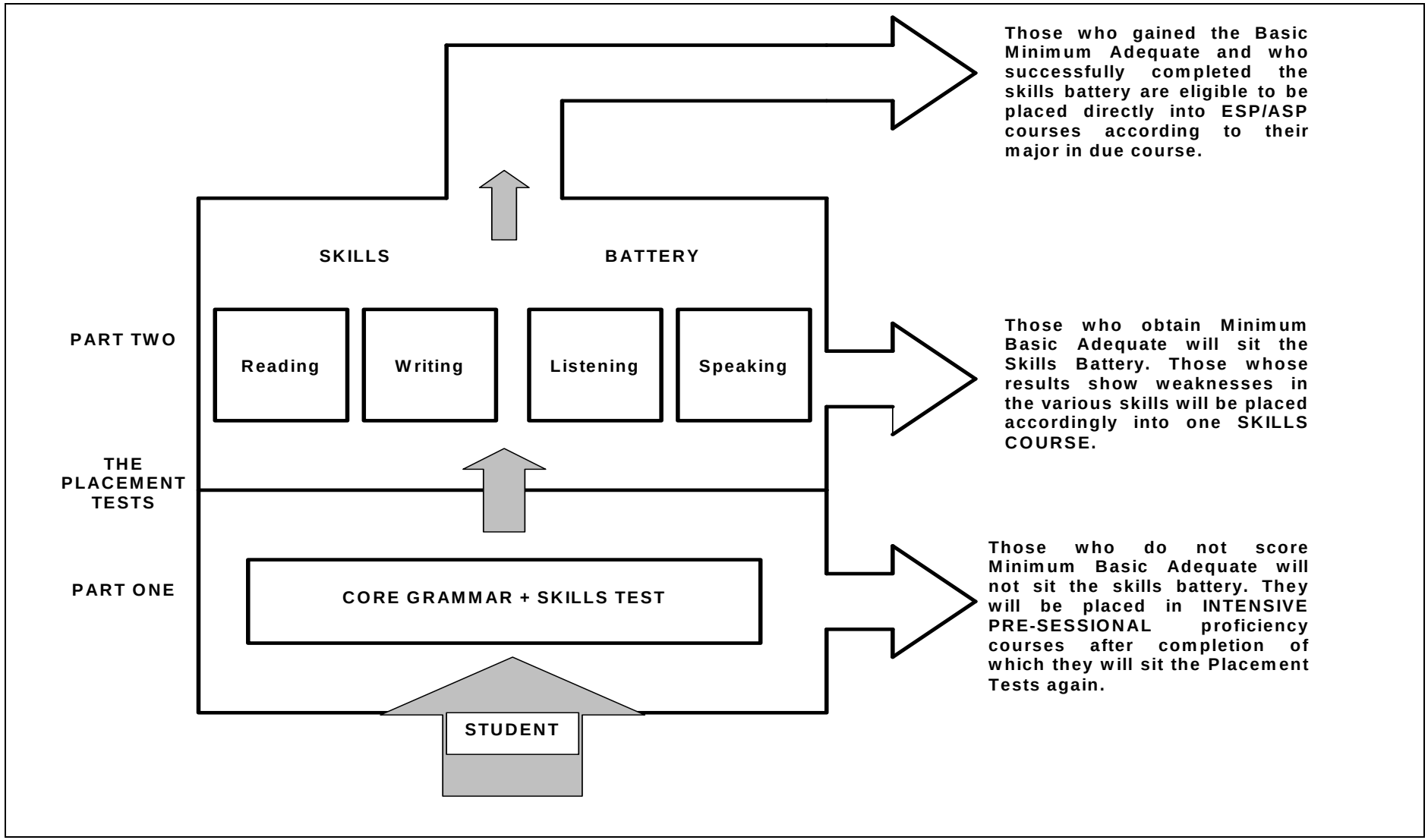

Figure 1.2: The English Language Placement Test Structure

(Source: CELPAD, 1993, p. 54)

Before discussing the kind of standard used in this placement system, it is of relevance, at this juncture, to briefly describe two types of standards generally used to facilitate educational decision-making. The first, 'relative standard', is expressed as a number or percentage of examinees, and is considered most appropriate for examinations where the purpose is to identify a certain number of examinees for admission or placement (Norcini, 2003). For this kind of standard, norm-referenced information is generally used. Cutscores are generally decided based on actual student performance on a test.

The second type of standard, 'absolute standard', on the other hand, is expressed as a number or percentage of test questions. This type of standard is used for tests of competence, like final or exit examinations, and tests for certification and licensure (Norcini, 2003). The kind of information used for this type of standard is usually criterion-referenced (or domain-referenced, content-referenced).

In the context of the 1995-2000 placement system, it is clear that absolute standards were utilized. Students were required to achieve a certain percentage of the total score (which represents a given amount or level of language skills) to be considered as having the expected level of English language proficiency. The use of absolute standards was consistent with the nature of the EPT and its the function in determining the threshold level students were expected to achieve in order to gain entry into the Kulliyyah at the Main Campus. It was also consistent with the need to maintain the same standards at each level of the instruction component across the different student intakes.

1.1.2 The CELPAD Placement System (2001 – 2004)

The present English Language curriculum structure at the Matriculation Centre was put into place in 2001. This four-tiered curriculum structure (Figure 1.3) was conceived in an effort to address the declining standards of English proficiency amongst students.

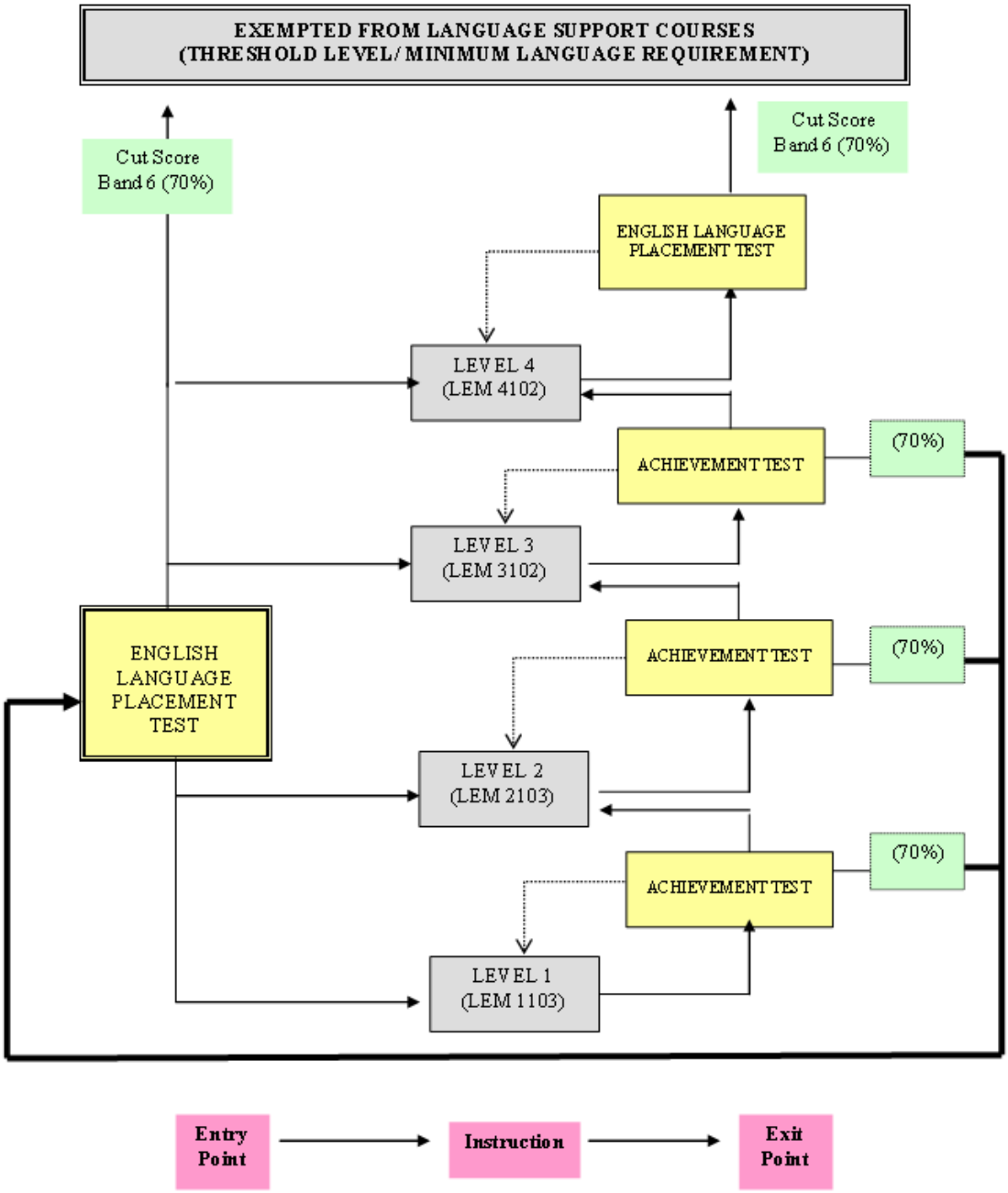

Figure 1.3: The Present English Language Placement System (2001–2004)

1.1.2.1 Instructional Component of the Present Placement System (2001 – 2004)

The first level course serves as a bridging course aiming to familiarise students with “tertiary approaches to language learning” (Centre for Language and Pre-University Academic Development, 2001, p. 4) (Figure 1.3). The primary focus is on reading and writing skills with a strong secondary focus on speaking and

listening skills. A key element in this course is a thorough review of common problem areas of English grammar. A task-based approach is adopted for the teaching and learning of these language skills.

The second level course adopts the same approach and focus though skills taught at this level are those that require greater communicative English language ability. The aim of this course is to further develop students' reading and writing competencies, oral and aural skills as well as to develop "practical application of grammatical structures" (CELPAD, 2001, p. 5).

The third and fourth level courses were designed with a more academic purpose. The primary focus is still on the development of reading and writing skills but with a more academic slant. The secondary focus, as with the lower level courses, is on the development of speaking and listening skills within the academic context. Grammar is integrated with the four language skills with emphasis on "its practical application to extract and produce meaningful sentence level, paragraph level and essay level English" (CELPAD, 2001, p. 5). At all four levels, project work is an important feature. It aims at providing students with the opportunity to apply language skills learnt.

1.1.2.2 Assessment Component of the Present Placement System (2001 – 2004)

With the introduction of a new curriculum structure in 2001, it was inevitable that the assessment component of the placement system was revised to complement the newly implemented instructional component. In the development of the new English Language Placement Test (EPT) two factors were of prime concern. The first was that the revised EPT should complement the newly-introduced instructional component. The second relates to practical considerations which pertain to constraints of time and manpower at the Matriculation Centre. This, therefore, resulted in the development of a much shorter test battery; one that requires only two

test administrations and a total of seven days to process but at the same time attempts to keep the major language aspects assessed intact. Below is a description of the structure of the revised EPT. (A detailed description of the EPT and sample questions are presented in Appendix 1).

Paper 1:

- Section 1 – Completion passage
- Section 2 – Error identification (Sentence and Paragraph Level)
- Section 3 – Reading Comprehension

Paper 2: Essay Writing
Paper 3: Speaking Test

- Section 1 – Short Talk
- Section 2 – Question Time
- Section 3 – Extended conversation

It is essential to mention here that the newly-developed test is based largely on the previous battery it replaces. Decisions as to which subtests to retain and exclude were made based on the results of validation studies of the previous EPT battery (see Noor Lide, 2002; Noor Lide & Isarji Sarudin, 2001, Noor Lide, Ainol Zubairi & Isarji Sarudin, 2001) as well as the constraints faced by the Matriculation Centre. Thus, in the newly-placed EPT battery, the listening skill test has been excluded and the test of writing ability is limited to a single writing task, which is essay writing. The speaking test is administered only to students who meet the minimum requirement (i.e., 70%) on the written tests (Papers 1 & 2).

The major reason for excluding the listening component of the previous EPT battery from the present test is the lack of proper facilities to adequately accommodate the large number of students (about 2,500 examinees per test administration) at the Matriculation Centre. Data interpretation and summary writing, on the other hand, are excluded from the writing test as (1) more time would be needed for the scoring of the subjective section of the EPT than the Matriculation Centre could afford and (2) it has been found that Essay Writing is a sufficient indicator of examinees' general academic writing ability and that it taps skills that are central in general academic writing (Noor Lide & Isarji Sarudin, 2001).

Unlike the previous placement system which utilized absolute standards for placement into the language support courses and entry into the Kulliyyah at the Main Campus, the current system adopts both absolute and relative standards. For exit from the Matriculation Centre and entry into the Kulliyyah, absolute standard is utilized where students are expected to achieve 70% on the written tests and pass the speaking test.

For placement into the language support courses relative standards are used. This move is motivated by the practical needs of the university. As the Kulliyyahs (faculties) require that the majority of their students complete the matriculation programme within one and a half years, it is necessary to place about 80% of incoming students to the Matriculation Centre in the three upper levels of the English language curriculum structure.

With the use of relative standards, no definite cutscores have been set to place students into the language support courses. Instead, norm-referenced information of student performance is presented to the Matriculation Examination Board (which is represented by key members of the respective Kulliyyahs), and the percentage of students to be placed into the respective language support courses is collectively agreed upon by members of the board.

Unlike in the previous placement system, students are required to sit for achievement tests upon completion of the language support courses. Those who attain a score of 50% (a combination of 40% coursework and 60% final exam score) and above are promoted to the next level. Those who fail to do so, on the other hand, are retained. However, two categories of students are re-administered the EPT at the end of the semester. The first group consists of students who have completed Level 4 of the English language support courses. Instead of sitting for an achievement test, they are required to sit for the EPT. If they meet the 70 % cutscore, they are allowed entry into the Kulliyyahs (faculties) at the Main Campus.

Those who fail to do so are required to remain at the Matriculation Centre and undergo further remediation.

The second group consists of those in Levels 1, 2 and 3 who attain 70% on the respective achievement tests. This means that there is a chance for students who have shown considerable improvement in their language proficiency to bypass certain levels of English support courses.

1.1.3 Shortcomings of the Assessment Component of the Placement System

To ensure that the EPT yields reliable and valid interpretations of student performance, a number of validation studies have been carried out. The EPT has been evaluated in terms of its reliability, content validity, concurrent validity and construct validity (e.g., Noor Lide, 2002; Noor Lide, Ainol Zubairi & Isarji Sarudin, 2001; and Noor Lide & Isarji, 2001). Findings of these studies have been used as the basis for further improvements of the EPT. However, there are still some problems inherent in the assessment component of the placement system that have remained unresolved.

The first of these problems pertains to the use of percentage mastery and the raw score scale in estimating and reporting student performance. The assumption that raw scores and percent corrects are “numbers with equal-interval units” (Wright & Linacre, 1996, p. 1) where “one point of score is considered to represent the same amount of ability” (Angoff, 1984, p. 5) is erroneous as raw scores and percent corrects are governed by the ability of the group tested and difficulties of the items on the test. Raw scores and percent corrects, therefore, are arbitrary measures and cannot be treated as equal-interval, linear measures (Wright & Stone, 1979).

The second problem involves the scoring of the essay writing section of the EPT. Rater effects or errors such as rater severity, halo, central tendency and restriction of range that pose serious threats to the quality and accuracy of ratings

(Saal et al., 1980; Engelhard, 1994) had never been properly investigated and adjusted for. The use of interrater reliability as evidence for the reliability of essay writing test scores is also problematic. The notion that interrater reliability – or more accurately, rater agreement – can be considered as a real and sufficient measure of reliability has been questioned by many (e.g., Henning, 1997; Linacre, 1989; Engelhard, 1994) as it fails to give an "accurate approximation of the true ability score". Henning (1997) argues,

> …two raters may agree in their score assignments and both be wrong in their judgments simultaneously in the same direction, whether by overestimating or underestimating true ability. If this happens, then we have a situation in which raters agree, but assessment is not accurate or reliable because the ratings fail to provide an accurate approximation of the true ability score. Similarly, it is possible that two raters may disagree by committing counterbalancing errors in opposite directions; that is where one rater overestimates true ability, and the other rater underestimates true ability. In this latter situation, it may happen that the average of the two raters' scores may be an accurate and reliable reflection of true ability, even though the two raters do not agree in their ratings (pp. 53-54).

The third problem relates to the construct definition of the constructs measured in the EPT. Congruent with common practice, selection of test items has been based on the notion of content representativeness. How far the items selected represent the "continuum of knowledge acquisition" (Glaser, 1994) and define the construct measured in terms of levels of development has been largely ignored. Therefore, interpretations of cutscores and performance standards are at best ambiguous as they are not directly referenced to the construct measured and, therefore, can be interpreted only as the proportion of items correctly answered.

Of all the problems that beset the EPT, the most critical due to its significant impact on students, is the arbitrariness in the way cutscores which determine minimum competency were set. Though the placement system from the years 1995 to 2000 utilized what appears to be 'absolute standards', the determination of cutscores was rather dubious. The determination of 50% on the Core Test as the

'minimum basic adequate score' for allowing students to proceed to the skills-based tests (the second part of the EPT battery) and the 35% cutscore to separate students into the two courses in the first tier of the curriculum structure (Refer to p. 10) had no clear rationale and empirical justifications. Similarly, the establishment of Band 6 (equivalent to 50% of total test score) as the minimum English language requirement for entry to the Main Campus was not empirically determined.

The introduction of relative standards for student placement into language support courses from 2001 onwards was equally, if not more, problematic. The use of norm-referenced information, which is designed to rank and compare students, makes even the semblance of a fixed standard impossible. More importantly, it is inconsistent with the nature of the EPT as a criterion-referenced test. Furthermore, unlike absolute standards, relative standards have the disadvantage of producing standards that are directly dependent on the "existing distribution of scores" (Postlethwaite, 1994, p. 36). Neither do relative standards present concrete evidence as to what students are able to do, as appropriately noted by Glass (1978),

> This approach has more to do with how many students are to be placed in a particular level, and less with what they know and can do. Because criterion / standards were determined normatively and not by direct reference to the behaviours exhibited on the test (p. 243).

1.2 PROBLEM STATEMENT

The arbitrary and inappropriate practice of setting cutscores is not peculiar to the EPT at the IIUM. Stevens and Parkes (2000) in their review of the practices, policies, and procedures used by state-level assessment programmes in the United States for the evaluation of school and school district effectiveness reported that some of the states were found to set cutscores by simply taking quartiles of the distribution of a total test score on a norm-referenced test. Others used scaled scores and percentile ranks which are inappropriate for standard-based reporting.

More troubling is the issue surrounding the achievement levels of the National Assessment of Educational Progress (NAEP), the main assessment programme of the standard-based educational reform in the United States. Several independent evaluators and committees, such as the congressionally mandated evaluation by the National Academy of Sciences (NAS) on the 1996 NAEP results of American students' achievement in key subject areas, have concluded that the Angoff procedure used in the standard setting process for the construction of cutscores to reflect desired performance levels is fundamentally flawed (Pellegrino, Jones & Mitchell, 1999).

The judgment tasks are said to be "difficult and confusing"; raters' judgments of different item types are deemed "internally inconsistent" (Pellegrino et al., 1999, p. 166; and evidence to support the validity of the cutscores and results is lacking (National Centre for Education Statistics [NCES], 2003; Pellegrino et al., 1999). Recommendations have therefore been made against the use of the achievement level-setting results in the NAEP reporting. It is asserted that due to the use of "a methodology that has been repeatedly questioned in terms of its accuracy and validity", the achievement results should be interpreted as suggestive rather than definitive (Pellegrino et al., 1999, p. 167).

Despite these findings, the National Assessment Governing Board (NAGB) has continued the use of the same procedure on the grounds that the Angoff procedure is the "most widely-used formal process for setting of performance standards" in the United States; that it has "over the past 20 years" "withstood many legal challenges"; and that it is backed by "respected expert practitioners in psychometrics and standard setting" (NAGB, 2004, p. 3).

Hambleton, Brennan, Brown and Dodd (2000), in defence of the decision made by NAGB, claim that the recommendations made by the independent evaluators are invalid and "[constitute] a one-sided, incomplete and inaccurate

accounting of the standard-settings conducted" (cited in NAGB, 2004, p. 3). Nonetheless, reports on the legitimacy of the Angoff procedure have impacted confidence in its use.

The 2001 reauthorization law has mandated that the achievement levels derived using the Angoff method "be used on trial basis until the Commissioner of Education Statistics determines that the achievement levels are "reasonable, valid and informative to the public" (NCES, 2003, p. 1). Efforts are now being made to find a more defensible method for the NAEP and this task has been extended to the research community (NCES, 2003).

To date the issue of the "right" standard setting method has remained unresolved as it has been established that different standard setting methods yield different results (e.g., Jaeger, 1993). However, this should not be used as an excuse to justify the continued use of standard setting methods that have been proved to be questionable in terms of their theoretical foundations.

In the last decade, a number of standard setting methods have been developed with promises of greater validity. Given the current state of affairs, it is important to investigate the efficacy of these newly-developed standard setting methods in delivering what they claim. One standard setting method that merits investigation is a Rasch-based method which was pioneered by Wright & Grosse but further developed and refined by Stone into the Objective Standard Setting Method (OSS) (Stone, 2001, 1996).

It is claimed that the main advantage of this Rasch-based method is that it capitalizes on the "two key attributes of a scientific measurement system in the human sciences: the validity of the test being used and the Rasch measurement properties of the resultant scale" (Stone, 1995, p. 452). This is of great significance as the "problem of developing evidence to support an inferential leap from an observed consistency to a construct that accounts for that consistency is a generic

concern of all science" (Messick, 1975, p. 955). Objective Standard Setting provides the inferential leap Messick (1975) was referring to as it captures "that which is most critical to validity: a clear and definable construct" (Stone, n.d.).

Nevertheless, it must be emphasized that standard setting methodology is not the sole issue that needs to be considered to arrive at valid and defensible cutscores and performance standards. A closely related issue which is critical to the validity of the inferences made on the basis of cutscores pertains to the more general issues of assessment. Thomas (1994) elucidates this very clearly. He notes that "applying standards not only defining content and levels [of achievement or performance], but also specifying how learners' achievement will be assessed" (p. 101).

As student achievement or ability is measured by performance on tests, it is imperative that steps are taken to ensure that the quality of the tests used in the measurement of student performance supports the kinds of inferences and decisions made on the basis of the cutscores. Assessment issues that need to be considered include ensuring the validity of the items used, congruence between empirical results and theoretical expectations, validity of responses and consistency of results. These are necessary requirements; in order for tests to "serve as adequate barometers of students' competence" they must satisfy fundamental requirements of sound measurement practice (Jaeger, 1993, p. 487).

Standard setting and assessment inexorably involve measurement. Answers to the questions of reliability and validity of test results, which are core psychometric issues, are derived from mathematical and statistical procedures characterized by the measurement theory employed (Suen, 1990). So are the validity and credibility of derived cutscores. As the key point in educational and psychological measurement is that inferences and interpretations are drawn from scores (Messick, 1981; Suen, 1990; Angoff, 1984) – and by extension cutscores or standards which are set on

tests – it is imperative that the procedures used to derive these interpretations are well-grounded in theoretically sound measurement theory.

1.3 PURPOSE OF STUDY

The primary purpose of this study is to investigate the efficacy of the OSS as a valid and defensible standard setting procedure. However, as issues related to assessment and measurement exert considerable influence on derived cutscores, these are also examined. This study, therefore, involves (1) the investigation of the adequacy of the EPT as a tool for measuring English language proficiency for placement and exemption purposes; (2) the accumulation of empirical evidence on the efficacy of the OSS in yielding multiple cutscores that are valid and defensible on tests utilizing selected-response (SR) items, constructed response (CR) items, and combination of these two item types; and (3) the demonstration of the utility of the Rasch measurement model in resolving measurement and standard setting issues.

1.4 OBJECTIVES OF STUDY

Consistent with the purpose of this study, the main objectives are to:

1.4.1 examine the adequacy of the EPT in defining the construct (i.e., English language proficiency), and measuring examinee performance for placement and exemption purposes ;

1.4.2 examine the efficacy of the OSS in producing multiple cutscores that are valid and defensible – in terms of procedural validity, internal validity and external validity – on tests utilizing SR items, CR items and combination of these two item types.

1.4.3 illustrate the utility of the Rasch measurement model in resolving measurement and standard setting issues.

1.5 RESEARCH QUESTIONS

1.5.1 Adequacy of the EPT

In the context of high-stakes assessment programmes, multiple sources of evidence to support the valid interpretations and use of test results are not only essential but mandatory (Messick, 1989; Jaeger, 1993). Therefore, in this study various types of empirical evidence to illustrate the validity of the EPT are collected and examined. However, as the gathering of validity evidence can be overwhelming, as in most validation research, evidence to support the adequacy of the EPT focuses on several major sources. Hence, the research questions this study seeks to answer with respect to the adequacy of the EPT are limited to the following:

1.5.1.1 Validity of Items:

1.5.1.1.1 To what extent are the items in the EPT subtests working in the same direction to define the measured constructs?

1.5.1.1.2 To what extent are the items in the EPT subtests contributing in the same meaningful and useful way to the construction of the measured constructs?

1.5.1.1.3 To what extent are the items in each of the EPT subtests measuring a single unidimensional construct?

1.5.1.2 Construct Definition:

1.5.1.2.1 To what extent are the items in the EPT subtests separated to define a continuum of increasing intensity?

1.5.1.2.2 To what extent is the empirical scaling of the test items in the EPT subtests consistent with the expectations of CELPAD test constructors?

1.5.1.3 Capacity of the items to lead to results which are consistent with the purpose of measurement:

1.5.1.3.1 To what extent are the EPT subtests able to replicate the ordering of examinees?

1.5.1.3.2 To what extent are the EPT subtests able to separate the measured examinees into five strata of proficiency?

1.5.1.3.3 To what extent are the EPT subtests able to provide a precise measurement of examinee ability?

1.5.1.3.4 To what extent is the sample tested accurately targeted by items in the EPT subtests?

1.5.1.4 Validity of examinee responses:

1.5.1.4.1 To what extent do examinee responses fit the expectations of the Rasch model?

1.5.1.5 Rater Effects:

1.5.1.5.1 To what extent do raters differ in severity?

1.5.1.5.2 To what extent do raters agree with one another in their rating?

1.5.1.5.3 To what extent are raters internally consistent in their rating?

1.5.1.5.4 To what extent are other rater effects present?

1.5.1.6 Rating Scale Functioning

1.5.1.6.1 To what extent is the rating scale used in the assessment of examinees' performance in essay writing functioning usefully?