

2012

A validation framework for automated essay scoring systems

Lucy Fang Lu

University of Wollongong

Recommended Citation

Lu, Lucy Fang, A validation framework for automated essay scoring systems, Doctor of Philosophy thesis, Faculty of Education, University of Wollongong, 2012. <http://ro.uow.edu.au/theses/3727>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

UNIVERSITY OF WOLLONGONG

COPYRIGHT WARNING

You may print or download ONE copy of this document for the purpose of your own research or study. The University does not authorise you to copy, communicate or otherwise make available electronically to any other person any copyright material contained on this site. You are reminded of the following:

Copyright owners are entitled to take legal action against persons who infringe their copyright. A reproduction of material that is protected by copyright may be a copyright infringement. A court may impose penalties and award damages in relation to offences and infringements relating to copyright material. Higher penalties may apply, and higher damages may be awarded, for offences and infringements involving the conversion of material into digital or electronic form.

**A Validation Framework for
Automated Essay Scoring Systems**

A thesis submitted in fulfilment of the requirements for the award of

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

by

Lucy Fang Lu

B.Comp. Science, M.Com. in Inf. Sys. & Tech.

Faculty of Education

2012

Thesis Certification

CERTIFICATION

I, Lucy Fang Lu, declare that this thesis, submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Education Faculty, University of Wollongong, Australia, is wholly my own work unless otherwise referenced or acknowledged. The document has not been submitted for qualifications at any other academic institution.

Lucy Fang Lu

20 July 2012

ABSTRACT

Writing tests are increasingly being included in large-scale assessment programs and high-stakes decisions. However, Automated Essay Scoring (AES) systems developed to overcome issues of marker inconsistency, volume, speed, cost and so on, also raise issues of score validity. In order to fill a crucial gap identified in the current approaches used to evaluate AES systems, this study develops and applies a framework that draws upon the current theory of validation, for assessing the validity of scores produced from Automated Essay Scoring systems (AES) in a systematic and comprehensive manner.

This thesis provides rationales for, and details of, the five essential components of the proposed AES validation framework. These five components are: 1) the writing traits scored by an AES system, how well they are assessed, and how they relate to the ability being assessed; 2) the validity implications of the type of scoring procedure used by an AES system to derive an overall score; 3) the internal structure of the assessment scores produced by an AES system; 4) the measurement qualities of the scores produced by an AES system; and 5) the consequential aspect of validity evidence. In order to make a convincing argument for AES score validity, evidence must be collected for each component, and the bodies of evidence collected must be evaluated together, in terms of their combined effects on the meaning of the score and the implications of score use.

In order to demonstrate how this framework may be applied, it is used to investigate the validity of scores produced by a particular Automated Essay Scoring system – the Intelligent

Essay Assessor (IEA) for the writing tasks from the Pearson Test of English (PTE) Academic. Five experienced human markers are employed to provide credible alternate measures, as a means to facilitate the examination of IEA scores.

This study demonstrates that the proposed framework is both effective in directing validation efforts, as well as in ensuring a methodical approach to AES validation. Through the application of this framework, the study has collected a wide range of empirical evidence and theoretical rationale, which enables a validity argument to be made for the IEA. Based on evidence collected, a number of recommendations are made with a view to further strengthen the validity of scores produced by the IEA. In addition, the study has illustrated in detail how various theories, including those associated with writing domain and measurement, can be used in conjunction with statistical methods, to collect and investigate evidence that is pertinent to different components of the AES validation framework.

The AES framework proposed in this study can be adapted and applied to the validation of all types of scoring systems. Furthermore, the validation processes undertaken in this study (i.e., first articulating an interpretative argument and then evaluating this argument in a particular test context), are generalisable to validations of all direct performance assessments.

Findings from this study support the position that the validation of AES systems needs to focus on direct evidence linking the scoring method to the intended interpretation and use of the scores. Evidence from this study also calls for careful rethinking of the role of human judgements of essay quality in the evaluation and further development of AES systems.

ACKNOWLEDGEMENTS

I would like to thank the many people who provided encouragement and support during the development of this research project and the writing of the thesis.

I owe my deepest gratitude to my supervisors, Professor Jim Tognolini, Professor Lori Lockyer and Dr Juho Looveer, who have supported me throughout my thesis with their wisdom, patience, encouragement and knowledge. Without them, this study would not have been possible.

I am also indebted to many of my colleagues who encouraged and supported me during this work. In particular, Dr Jenny Donovan, Dr Ruth Habgood, Dr Kelly Stephens and Susan Wright are thanked for their time, editorial comments and advice. I would also like to thank Dr Guenter Plum who provided professional editorial advice for the preparation of the thesis for submission.

Special thanks also go to Margaret Turnbull and the ESL (English as a Second Language) consultants/teachers from the NSW Department of Education and Communities who participated in this research study. Margaret is particularly thanked for her encouragement, assistance and professional discourse.

Finally I would like to thank my husband and son for their continued love, support and patience throughout the study.

TABLE OF CONTENTS

ABSTRACT	i
TABLE OF CONTENTS	iv
List of Tables.....	viii
List of Figures	x
List of Appendices	xii
Chapter 1 Introduction.....	1
1.1 Recurrent Issues in Large-Scale Performance-Based Writing Assessments	3
1.2 Automated Essay Scoring Systems.....	7
1.3 Background to the Study.....	8
1.4 Purpose and Organisation of the Study.....	12
1.5 Limitations	16
Chapter 2 Automated Essay Scoring (AES) Systems	18
2.1 Introduction.....	18
2.2 Project Essay Grade (PEG).....	19
2.3 IntelliMetric	20
2.4 e-rater	24
2.5 Intelligent Essay Assessor (IEA)	27
2.6 Recent Trends in the Development of New AES Models	42
2.7 Chapter Summary	50
Chapter 3 Review of the AES Evaluation Studies	51
3.1 Studies Focusing on the Relationship among Scores Generated by Different Scorers	51
3.2 Studies Focusing on the Relationship Between AES Scores and External Measures	57
3.3 Studies Focusing on the Scoring Process	58

3.4	Other Evaluative Approaches.....	60
3.5	Chapter Summary.....	65
Chapter 4	A Validation Framework for the AES Systems	66
4.1	Concept of Validity	66
4.2	Concept of Validation	68
4.3	A Practical Validation Framework for Automated Essay Scoring (AES) Systems ...	74
4.4	Chapter Summary.....	99
Chapter 5	Pearson Test of English (PTE) Academic Writing Tests and Data Collections Procedures	101
5.1	PTE Academic and Testing Context	101
5.2	Data Collection Procedures	103
5.3	Chapter Summary.....	115
Chapter 6	The Domain of the Writing Ability Construct.....	116
6.1	The Product Approach to Writing.....	116
6.2	The Process Approach to Writing	119
6.3	Chapter Summary.....	123
Chapter 7	Writing Traits Scored by Intelligent Essay Assessor (IEA) and the IEA Scoring Procedure.....	125
7.1	Analysing the IEA Construct Coverage – Relevance and Representativeness	125
7.2	Writing Traits Assessed by IEA for the PTE Academic	128
7.3	The Scoring Procedure Used by IEA	134
7.4	Chapter Summary.....	144
Chapter 8	Correspondence Rates between Human and the IEA (Intelligent Essay Assessor) Overall Scores.....	146

8.1	Reliability of Human Scores Used in the Agreement Analysis.....	147
8.2	Reliability of the Human Scores Acquired from the Pearson Field Tests	152
8.3	Correspondence Rates Between Total Human Scores and the IEA Scores.....	162
8.4	Chapter Summary	175

Chapter 9 Measurement Properties of the Intelligent Essay Assessor (IEA) and Human Scores 177

9.1	Introduction.....	177
9.2.	The Rasch Model Used in this Study.....	179
9.3	Fit Indicators Incorporated in Rasch Analysis.....	183
9.4	Rasch Analysis Performed in this Study.....	185
9.5	Results of Rasch Analysis.....	188
9.6	Chapter Summary	224

Chapter 10 Structural Properties of the IEA (Intelligent Essay Assessor) and Human Scores 227

10.1	Introduction.....	227
10.2	Dimensional Structure in Analytic Writing Scores	228
10.3	Structural Patterns in the Analytic Scores as a Result of Gender Effect	242
10.4	Analysing the IEA and Human Trait Scores in One Two-Dimensional Space	248
10.5	Chapter Summary	255

Chapter 11 Examination of the Individual IEA (Intelligent Essay Assessor) Traits 257

11.1	Introduction.....	257
11.2	The IEA Scoring of the <i>Spelling</i> Trait	259
11.3	IEA Scoring of the <i>Formal Requirement</i> Trait.....	269
11.4	The IEA Scoring of the <i>Content</i> Trait	278
11.5	The IEA Scoring of the Other Four Traits.....	293

11.6	Inter-relations amongst Human and IEA Traits	302
11.7	Chapter Summary	306
Chapter 12	Discussions and Conclusions.....	309
12.1	The AES Validation Framework	309
12.2	A Validity Argument for IEA	311
12.3	Future Work to Strengthen the Validity of Scores Produced by IEA	317
12.4	Implication of this Study for Future AES Research and Development	321
12.5	Concluding Remarks	329
References	331

List of Tables

Table 4.1	Interpretative Argument for a Writing Test (c.f. Kane, 2006, p. 24).....	70
Table 5.1	Directive Given to Test Takers at the Field Tests 2007–2008.....	104
Table 5.2	Descriptions of the Two Sample Prompts	104
Table 7.1	Links Between Traits on the <i>Profile</i> Scale and the IEA Traits	131
Table 8.1	Score Reliability for Analytic and Holistic Scoring and Analytic Traits in a Double-Marking Scenario.....	150
Table 8.2	Dependability Index for Analytic and Holistic Scoring and Analytic Traits in a Double-Marking Scenario – Pearson Field Tests	153
Table 8.3	Mean and Standard Deviation of the Ratings by Two Groups of Markers – Voting	158
Table 8.4	Mean and Standard Deviation of the Ratings by Two groups of Markers – Tobacco.....	158
Table 8.5	Mean and Standard Deviation of the Final Human and IEA Scores	165
Table 8.6	Spearman’s Rank Order Correlation Coefficients (r_s) Between Human Analytic Scores and the IEA Scores	168
Table 8.7	Spearman’s Rank Order Correlation Coefficients (r_s) Between Human Holistic Scores and the IEA Scores	168
Table 8.8	Correlations Between Human (Analytic and Holistic) Final Scores and the IEA Scores – Calculated Without the Two Minimum Requirements	173
Table 9.1	Distribution of Standardised Residuals.....	188
Table 9.2	Fit Statistics for the Voting Prompt	190
Table 9.3	Fit Statistics for the Tobacco Prompt.....	190
Table 9.4	Trait Fit Statistics for the Voting Prompt When the <i>Spelling</i> and <i>Formal Requirement</i> traits Are Removed.....	192
Table 9.5	Trait Fit Statistics for the Tobacco Prompt When the <i>Spelling</i> and <i>Formal Requirement</i> Traits Are Removed.....	193

Table 9.6	Trait Fit Statistics for the Voting prompt – Human Scores.....	195
Table 9.7	Trait Fit Statistics for the Tobacco Prompt – Human Scores.....	196
Table 9.8	Category Frequencies, Average Ability Measures, OUTFIT Statistics and Rasch-Andrich Threshold Measures	207
Table 10.1	The First and Second Factors in the Human and the IEA Analytic Trait Scores	232
Table 10.2	Gender Difference Statistics Across Both Prompts.....	244
Table 11.1	Top Three Traits with the Greatest Number of Unexpected IEA Trait Scores Across the Two Prompts	258
Table 11.2	Descriptions of the IEA <i>Spelling</i> Rating Scale Used for the PTE Academic.	260
Table 11.3	Description of the <i>Formal Requirement</i> Scale	270
Table 11.4	Means and Standard Deviations of Human and IEA <i>Content</i> Scores	280
Table 11.5	Agreement Rates Between the IEA <i>Content</i> Scores and Human <i>Content</i> Scores	282
Table 11.6	Matrix of the Frequency of Occurrence of Human and the IEA <i>Content</i> Scores Across the Two Prompts	285
Table 11.7	Descriptive Statistics of the Trait Scores by the IEA and by Human Markers	294
Table 11.8	Pearson Correlations Between Word Count and the IEA Scores and Between Word Count and the Human Scores	298
Table 11.9	Agreement Rates and Correlations Statistics Across the Traits	300
Table 11.10	Discrepancy Rates at the Score Point Level.....	301
Table 11.11	Inter-correlations Amongst Traits As Assessed by Human Markers and the IEA	304
Table 12.1	Evidence Collected from Using the Proposed AES Framework.....	313

List of Figures

Figure 2.1	IntelliMetric Feature Model (Elliot, 2003, p. 73)	23
Figure 2.2	The <i>e-rater</i> Scoring Model (Quinlan et al., 2009, p. 9)	25
Figure 2.3	IEA Architecture (Landauer et al., 2003, p. 90)	41
Figure 4.1	The Proposed AES Validation Framework and Its Components.....	97
Figure 7.1	Mapping of the Writing Performances Assessed by the IEA to the Construct of Interest.....	133
Figure 7.2	A Schematic Representation of the PTE Academic Writing Score	137
Figure 8.1	Dependability Indices (Based on a Double-Marking Scenario) – Voting	154
Figure 8.2	Dependability Indices (Based on a Double-Marking Scenario) – Tobacco... ..	154
Figure 8.3	Exact Agreement Rates for Australian and Chinese Markers – Voting	160
Figure 8.4	Exact agreement Rates for Australian and Chinese Markers – Tobacco.....	161
Figure 9.1	IEA <i>General Linguistic Range</i> Scale (Adapted from Pearson, 2011b, p. 60)	181
Figure 9.2	<i>General Linguistic Range</i> Rating Scale with Ordered Categories.....	181
Figure 9.3	ICC Graph Obtained for the <i>Development, Structure and Coherence</i> Trait for the Tobacco Prompt	201
Figure 9.4	ICC for <i>Spelling</i> – Voting	203
Figure 9.5	ICC for <i>Spelling</i> – Tobacco	203
Figure 9.6	ICC Graph for Voting	205
Figure 9.7	ICC Graph for Tobacco.....	204
Figure 9.8	ICC Graph for the <i>Formal Requirement</i> Trait for Voting	205
Figure 9.9	Observed Average Measures for Score Categories – Voting	210
Figure 9.10	Observed Average Measures for Score Categories – Tobacco.....	210

Figure 9.11	Category Probability Curve for the <i>Grammar Usage and Mechanics</i> Trait – Voting	212
Figure 9.12	Category Probability Curve for <i>Spelling</i> – Voting	213
Figure 9.13	Variable Map for Voting Prompt	216
Figure 9.14	Variable Map for Tobacco Prompt	217
Figure 9.15	Scatter Plot of Person Measures – Voting	220
Figure 9.16	Scatter Plot of Person Measures – Tobacco Prompt	220
Figure 9.17	Scatter Plot of Person Measures (Five Traits) – Voting Prompt	223
Figure 9.18	Scatter Plot of Person Measures (Five Traits) – Tobacco Prompt	224
Figure 10.1	Plot of the Traits with Contrasting Loadings on the Second Factor in the Human Data – Voting Prompt	234
Figure 10.2	Plot of the Traits with Contrasting Loadings on the Second Factor in the Human Data – Tobacco Prompt	235
Figure 10.3	Plot of Loadings on the Second Factor – Voting (Five IEA Traits)	239
Figure 10.4	Plot of Loadings on the Second Factor – Tobacco (Five IEA Traits)	240
Figure 10.5	Representation of Five Human Traits and Five IEA Traits in a Two-Dimensional Space –Voting	250
Figure 10.6	Representation of Five Human Traits and Five IEA Traits in a Two-Dimensional Space –Tobacco	251
Figure 11.1	Average Trait Marks by Word Count	273
Figure 11.2	Examples of Potential Anomalies Arising from the <i>Formal Requirement</i> Trait Scoring	274

List of Appendices

Appendix A	PTE Academic Writing Scoring Rubrics	368
Appendix B	Scoring Rubrics for the Common European Framework (CEF) Scale	369
Appendix C	Scoring Rubrics for the ESL Composition Profile (Original).....	370
Appendix D	Scoring Rubrics for the Modified ESL Composition Profile.....	372
Appendix E	Scoring Rubrics for TOEFL Independent Writing Tasks	374
Appendix F	Scoring Scheme Used in this Study	375
Appendix G	Background Questionnaire of the Markers	376
Appendix H	G-Study Results – Estimated Variance Components (σ^2) in Holistic and Analytic Human Ratings – Voting.....	377
Appendix I	G-Study Results – Estimated Variance Components (σ^2) in Holistic and Analytic Human Ratings – Tobacco	378
Appendix J	Pearson Correlation (r) Between Human Scores and IEA Scores	379
Appendix K	Category Statistics for Rating Scales Used by Human Markers.....	380
Appendix L	Two Examples of the <i>Spelling</i> Score Anomalies.....	381
Appendix M	An Example of <i>Spelling</i> Scoring Anomalies	383
Appendix N	An Example of Potential Anomaly Arising from the <i>Formal Requirement</i> Trait	384
Appendix O	Two Additional Examples of Potential Anomalies Arising from the <i>Formal</i> <i>Requirement</i> Trait	385
Appendix P	An Example of <i>Content</i> Scoring Anomaly	387
Appendix Q	Matrix of Frequency of Occurrences of the IEA Scores and Human Trait Scores Across the Four Traits	388
Appendix R	Ethics Approval from the Human Research Ethics Committee.....	389
Appendix S	Participant Information Sheet	390

Chapter 1 Introduction

The ability to write well in a variety of formats is a vital skill in today's world. Whether individuals are seeking participation or advancement in social, educational or occupational settings, writing skills are essential to success.

Writing also plays an important role in student learning and its function evolves as students move from compulsory education to higher education. At higher levels of education, the primary function of writing is no longer simply the conveyance of information, but the expansion of knowledge through reflection (Bereiter & Scardamalia, 1987; Purves, Soter, Takala & Vahapassi, 1984; Weigle, 2002). This is because, at that stage of learning, the process of writing tends to become "a two-way interaction between continuously developing knowledge and continuously developing text" (Bereiter & Scardamalia, 1987, p. 12), leading to the strengthening of existing domain knowledge as well as the creation of new domain knowledge (Alamargot & Andriessen, 2002). Writing is also perceived as being closely linked to higher order skills, and expertise in writing is viewed as an indication that students possess the necessary cognitive skills, such as critical thinking and reasoning, required for higher levels of education (Weigle, 2002).

Despite the widely acknowledged importance of writing, employers and university professors have noted and expressed concern regarding the poor writing abilities of both school and university students and graduates. A survey in 1992 of 402 American corporate companies reported by the Associated Press (as cited in Hansen & Hansen, 1997) noted that executives

identified writing as the most valued skill, but added that 80% of their employees at all levels needed to improve. The number of workers needing improvement in writing skills according to the 1992 survey, was 20% greater than in the same survey in 1991. A 2002 survey conducted by Public Agenda, an American non-profit organisation, also reported that more than 70% of employers and college professors rated public high school graduates in America “fair” or “poor” for writing skills (DuPont, 2002). In Australia, 15% of Year 9 students across the country in 2011 were assessed as below the national minimum standard in writing, with the rate as high as 43% in the Northern Territory (Australian Curriculum, Assessment and Reporting Authority, 2011). Additionally, some Australian academics have been reported as attributing the high first-year university dropout rates to poor essay writing skills of those students coming straight from school. These universities have had to offer special remedial classes targeting students with poor essay writing skills to lift retention rates (Bissett & McDougall, 2008).

The importance of writing skills to student success at school and beyond, as well as the issue of poor writing skills amongst students in all disciplines and at all academic levels, have led to the introduction of a performance-based writing test component in an increasing number of standardised testing programs. Internationally, large-scale high-stakes language proficiency assessment programs which now have a performance-based writing component include: the Graduate Management Admission Test (GMAT); the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE). The National Assessment Program –

Literacy and Numeracy (NAPLAN), which was introduced in Australia in 2008, includes a performance-based writing component for each of the Year 3, 5, 7 and 9 testing programs.

As more writing tests are being included in large-scale assessment programs, and more high-stakes decisions (such as school funding allocations, graduation, and admission to higher education programs) are increasingly being informed by writing test results, greater attention is being paid to the assessment of writing. Those who are stakeholders in such tests and the wider public have a right to expect that test results accurately reflect the underlying writing ability of students and that the results are generated with accuracy and reproducibility.

1.1 Recurrent Issues in Large-Scale Performance-Based Writing Assessments

There continues to be a plethora of issues associated with measuring large-scale performance-based writing assessments in spite of the increase in the importance and the use of these measures. A significant body of research exists identifying different factors that may contribute to measurement errors in direct assessment of writing ability (e.g., Carr, 2000; Huot, 1990b; Leckie & Baird, 2011; Lim, 2009; Schoonen, 2005; Weigle, 2002). The main point of consensus is that it is difficult to consistently produce valid measures when assessing writing because of the multiple sources of errors introduced by the complex and multi-faceted nature of performance-based testing (Schoonen, 2005). One constant source of measurement error which influences the reproducibility of scores in writing assessment is the inconsistency among markers. Past studies have demonstrated that markers can give widely different scores

to the same essay or the same marker can give different scores to the same essay at different times (Cooper, 1984; Diederich, 1974; Noyes, 1963). An extreme illustration of this problem was demonstrated in Diederich, French and Carlton's (1961) study, where 300 essays were read by 53 markers on a nine-point scale. The researchers found that 94% of the essays received at least seven different scores (as cited in Huot, 1990b).

Although the reliability of human marking¹ has improved since early studies through various measures such as the provision of more structured training (e.g., Elder, Barkhuizen, Knoch & von Randow, 2007; Knoch, Read & von Randow, 2007), it is doubtful that bias and error in human judgements can ever be completely eliminated. This is because many factors that (consciously or subconsciously) influence human judgements seem to be associated with characteristics inherent in human nature (Linn & Gronlund, 2000; Markham 1976; Myford & Wolfe, 2003; Thorndike & Hagen, 1977).

Factors such as fatigue, loss of concentration arising from boredom, "halo" effects (e.g., judgements affected by previous essays read), appearance (e.g., neatness of handwriting) and markers' tendency to avoid extreme categories of a rating scale, have all been found to contribute to inconsistency and inaccuracy in the scores assigned by human markers (Cooper,

¹ The terms "marking" and "scoring" have the same meaning. Scoring is the more popular term in the Automated Essay Scoring literature which largely originated in America, while marking is the term used more frequently by Australian assessment professionals to describe human evaluation activities. In keeping with this usage, this thesis will use the term marking for general discussions of assessment and for human evaluation activities, and scoring when referring to automated methods of essay evaluation.

1984; Huot, 1990b; Myford & Wolfe, 2003). An individual marker's severity or leniency in marking can also change, sometimes significantly, within one marking period or over time (e.g., Congdon & McQueen, 2000; Myford & Wolfe, 2009). Other studies have demonstrated that variables in the markers' backgrounds, such as the level of professional experience, prior knowledge in marking, teaching foci, linguistic and cultural backgrounds, perceptions of language proficiency and assumptions of language acquisition can all influence markers' marking behaviour and their judgements of essay quality (e.g., Eckes, 2008; Erdosy, 2004; Knoch, Read & von Randow, 2007; Leckie & Baird, 2011). Despite the significant body of research conducted to date, the nature, the causes and the corollaries of "rater effect" are not yet fully understood, making it difficult to fully identify and minimise marker bias in scores (Myford & Wolfe, 2003; Lim, 2009).

In addition to problems related to the consistency and the appropriateness of scores assigned by human markers, the high cost of marking is another limiting factor for large-scale writing assessment (Hardy, 1995; Wainer & Thissen, 1993). In the quest to improve the reliability of human marking, assessment authorities routinely adopt a number of quality control procedures, which invariably have significant cost implications. For example, in the State of New South Wales (NSW), Australia, when marking questions requiring extended responses, such as marking short essays, for the Higher School Certificate (HSC) matriculation examinations, the Board of Studies uses two markers to make independent judgements of each student's response. When the two scores assigned differ by more than the maximum acceptable difference set by the Board, a third or possibly fourth marking of the student's

response is undertaken (Masters, 2002). Although this is common industry practice for high-stakes writing assessment, it is expensive to implement. The cost associated with double-marking is additional to other costs incurred for other quality assurance procedures relating to human marking, such as the selection and training of markers, random checking of assigned marks, and marking of common control scripts (i.e., responses). In some cases, the difficulty of recruiting a large number of professionally trained markers limits the possibility of having a direct performance-based writing test in a large-scale language test program.

A third problem associated with large-scale writing assessments is the time it takes to complete the marking process. This leads to significant delays in communicating scores and other feedback to schools, students and teachers. The routing of essays to marking centres, multiple readings by human markers and adjudication of the scores where significant discrepancies appear, are all time consuming. An important function of educational tests is to assist teachers and students in the conduct of classroom learning. When results are returned to schools promptly, teachers have the option of using the results to plan better their classes and to tailor lessons to specific needs of students. The students can use the results to gain a sharper understanding of their weaknesses and strengths in different subjects and adjust their study pattern accordingly. When results are delayed because of the time it takes to mark the responses, the instructional value of using the test results to improve learning is reduced.

A fourth problem with large-scale writing assessments is related to the generalisability of the measurements (Brown, Hilgers & Marsella, 1991; Moss, Cole & Khampalikit, 1982; Stevens & Clauser, 1996; Swartz, Patience & Whitney, 1985). This is largely due to the sampling

error associated with using a relatively small number of tasks (Kane, Crooks & Cohen, 1999). To obtain a more generalisable measure, it is necessary to use a large number of writing tasks (Shavelson, Baxter & Gao, 1993). However, due to prohibitive cost and time constraints associated with human marking, it is generally not feasible to have a large number of writing tasks included in a language test.

1.2 Automated Essay Scoring Systems

Automated Essay Scoring (AES) systems using computers to evaluate and score essays have been developed in response to the issues outlined above (cost, time, reliability and generalisability), for marking essay format writing assessment tasks. Development of AES technology over the last three decades has largely been made possible due to significant advancements in the disciplines of applied linguistics, artificial intelligence, and natural language processing.

Potential benefits of AES systems include consistent and reliable scoring of essays from one scoring scenario to another; increased objectivity and efficiency in the scoring operation; and the reduction of time and financial costs associated with scoring.² Since AES systems make it possible to score a greater number of writing samples per student across different writing

² There are two conditions, under which the claim for improved cost efficiency associated with AES use can be most reasonably made: 1) that “the process required for preparing the automated system to score responses to new test items is rapid and inexpensive to implement”; 2) that the writing tests implemented are standardised tests that have a large number of examinees (Bennett, 2004, p. 1).

tasks, they also have the potential to improve the generalisability of the writing achievement measures.

In addition, AES systems have the capacity to deliver immediate diagnostic feedback to students and teachers. Depending on the quality of the feedback, this could assist teachers to enhance their instructional practices. Such feedback also has the potential to motivate students to write more and to revise more drafts, which is helpful for the further development of their writing expertise.

1.3 Background to the Study

Many evaluative studies have reported relatively high levels of correspondence between the scores produced by AES systems and those produced by human markers (Attali, 2004; Landauer, Laham & Foltz, 2003; Nichols, 2004; Page, 2003; Vantage Learning, 2003a, 2003b). However, despite these positive results and the potential benefits of AES technology, AES systems are yet to be widely accepted by professional educators (Ericsson, 2006; Jones, 2006; McGee, 2006; Rothermel, 2006).

Part of the reason for this lack of broad acceptance is the lack of transparency of these systems (Enright & Quinlan, 2010). In order to accept, trust and effectively use a new technology, consumers need, as a minimum, to be clear about its capabilities in relation to their needs and to have confidence in the results produced.

Currently AES systems are far from transparent (Kelly, 2006). For the majority of the AES systems, details of the textual features being assessed by the automated scoring process, and how they are assessed, are treated as proprietary information by the AES vendors, and are not made available. In the interests of making the AES systems more transparent to educators, clear explanations of the technology including details of the scoring process are needed.

A second issue concerning the acceptance of AES systems relates to consumers' confidence in the validity of AES. Validity here refers to the appropriateness and fairness of the scores assigned by an AES system for a particular purpose. Some critics believe AES systems treat texts as "isolated artifacts" which are "divorced from the broader historical, political, and social contexts and practices" with which writing is necessarily associated (Ericsson, 2006, p. 31). As a result, these critics believe AES systems cannot discriminate exceptional essays from mediocre writing, nor can they understand or appreciate the writer's message in an equivalent manner to human beings (Ericsson, 2006; Jones, 2006; Rothermel, 2006). Many doubt AES systems' ability to assess higher-level skills, such as "identifying evidence of abstract concepts" and "detecting irony or extended use of metaphor or allusion" (comments provided by professional educators in NSW in response to the trial of an AES system for the National Assessment Program for writing). Other researchers have criticised AES for its "over-reliance on surface features of responses, the insensitivity to the content of responses, and to creativity, and the vulnerability to new types of cheating and test-taking strategies" (Yang, Buckendahl, Juskiewicz & Bhola, 2002, p. 393). Some researchers also perceive AES to score indirect features of writing that happen to correlate well with characteristics of

quality writing, and therefore question its ability to score essays that have atypical profiles, such as those that are “well-organised, but with poor mechanics or strong vocabulary but with lots of misspelling” (Calfee, 2000, p. 35, as cited in Wang & Brown, 2007).

It is noted that some of the criticisms of current AES scoring methods may be partly based on “reactions to earlier, outdated AES procedures that tended to rely heavily on the evaluation of surface features, such as the number of words in an essay” (Powers, Burstein, Chodorow, Fowles & Kukich, 2001, p. 2). In the last 10 years, AES systems have evolved from the early, primitive models to more sophisticated methods of scoring. For instance, vendors claim that the number of words in an essay is no longer included in most of the scoring models as a directly-assessed feature (e.g., Burstein, 2003; Elliott, 2003; Landauer et al., 2003).

Nevertheless, there still exists a need to adequately address other substantial concerns regarding validity, if AES is to be more widely accepted. These concerns include the ability of an AES system to assess textual features that are directly linked to qualities of writing, its capacity to measure higher order skills that facilitate the use of language, and its susceptibility to external features that are irrelevant to the writing construct of interest. A “construct” in this thesis refers to “the concept or characteristic that a test is designed to measure” (American Educational Research Association [AERA] et al., 1999, p. 5). In the context of this study, the writing construct of interest is the test taker’s academic writing ability.³ Other validity

³ The nature and components of this ability construct and the characteristics of the test that is used in this study to measure it, will be expanded in subsequent chapters.

concerns include whether scores produced by AES systems can be treated as valid measurements for comparisons of a single ability, and whether the internal structure of the AES scores reflects the theoretical distinctions pertinent to the construct of interest. These are fundamental issues of validity which provide justifications for the interpretation of scores and which need to be addressed by AES validation studies.

Unfortunately, as noted by Attali (2007), there are few AES studies that have addressed these key validity issues for AES systems. Most of the evaluative studies conducted to date have only focused on demonstrating the correspondence between the AES- and human-generated scores as evidence of validity. This type of evidence, though a necessary face-valid metric, does not directly support the intended interpretation and proposed use of test scores (Bennett & Bejar, 1998; Williamson, Bejar & Hone, 1999). A related issue is that there appears to be a lack of a systematic approach towards the collection and examination of validity evidence. Very few studies have attempted to use a systematic approach to examining a wide range of evidence associated with different aspects of validity in order to support the arguments for the validity of the writing scores produced by AES systems.

There are then at least two key issues concerning the acceptance and the adoption of AES systems. First, as potential users, educators need more comprehensive and more detailed information about the scoring processes that are used by AES systems, and also about the key technologies and theoretical frameworks that drive automated scoring. The second issue concerns the need for more robust and comprehensive AES evaluative studies, which use a systematic approach to collect a wide range of direct validity evidence for AES systems.

These two issues are not mutually exclusive; rather they are interrelated. Transparency of a test, including its scoring process, is a key test validity criterion (Frederiksen & Collins, 1989). Researchers (e.g., Baron, 1991; Frederiksen & Collins, 1989; Wiggins, 1993) believe that the scoring criteria (i.e., how performance is scored) and standards of successful performance should be transparent to students, so that they can be more readily internalised by students as self-directive goals (Baron, 1991; Wiggins 1993). Improving transparency and understanding of the scoring system can therefore encourage positive consequences of testing, resulting in strengthened validity of the test (Messick, 1996). In order for AES systems to be more widely accepted and adopted by educators, these two inter-related issues of transparency and adequately proven validity need to be addressed. To progress this, this thesis develops and tests a systematic method for comprehensively examining questions of validity associated with AES systems. It also provides detailed information about AES model building processes and AES technology to assist the development of a deeper understanding of these systems by educators.

1.4 Purpose and Organisation of the Study

The primary purpose of this study is to develop a validation framework, thereby filling a crucial gap identified in the current approaches used to evaluate AES. This thesis argues that the development and implementation of more robust validation processes is one vital step in building consumer confidence and realising the potential of AES systems. The framework developed in this study draws upon the current practices of validation as reflected in the most recent standards for educational testing (American Educational Research Association, 1999).

It aims to provide practical guidance to the systematic collection and examination of validity evidence for AES systems, to ensure that key validity questions (regarding AES systems) are addressed in a comprehensive and robust manner.

The thesis is organised in three parts. The first part (Chapters Two and Three) establishes the need for a structured and coherent approach to exploring validity issues for AES systems. It focuses on existing AES technologies and how they are currently being evaluated. Chapter Two presents a review of the main technologies that drive automated scoring, drawing out their strengths and weaknesses, and discusses validity implications of recent developments in the AES field. This is followed in Chapter Three by a critical review of AES evaluative studies, which identifies key issues and weaknesses in the current approaches used to assess validity of scores produced by AES systems.

The discussions in these two chapters not only help substantiate the necessity for an AES specific validation framework, but also help make AES systems more transparent and comprehensible to educators and policy makers. It is hoped that, by making clear the strength and limitations of the AES technologies, the study will also contribute to the wider acceptance and utilisation of these technologies and hence to the wider realisation of their benefits.

The second part of the thesis (Chapter Four) develops an AES validation framework to facilitate the making of more coherent and robust validity arguments for these systems. It first examines the current understandings of validity and validation as a basis for the development

of the proposed AES framework. The framework, its components, and its application in a validation process, are then discussed in detail.

The utility of this framework is subsequently tested in the third part of the thesis (Chapters Five to Eleven) by applying it to collect and examine validity evidence (both empirical evidence and theoretical rationales) for a key AES system – the Intelligent Essay Assessor (IEA). The validity of the IEA is assessed in the context of the Pearson Test of English Academic (referred to in this thesis as “PTE Academic”).⁴ PTE Academic is a new international academic English proficiency test developed by Pearson Technologies. The test includes a direct performance-based writing component which is scored by the IEA. It is the fairness and appropriateness of the scores assigned by the IEA for this component that will be the focus of the third part of the thesis. The AES framework developed will be partly assessed by its ability to address the following validity questions concerning IEA:

- 1) How do the aspects of writing performance, as measured through writing traits by the IEA, relate to the writing ability being assessed?
- 2) How well does the IEA assess these writing traits?

⁴ As PTE Academic aims to measure the ability of the test taker “to use English in academic settings” (PTE Academic, p. 42), it is similar in nature to other common English language proficiency tests currently operating in the market such as IELTS (the International English Language Testing System) and TOEFL (Test of English as a Foreign Language). See Chapter Five, Section 5.1 for more details about the test – PTE Academic, and the definition of writing ability measured by the test.

- 3) Empirically, do the writing traits scored by the IEA behave as expected?
- 4) What is the scoring procedure used by IEA to derive the total score as an indicator of overall writing ability? What are the implications of this procedure for the meaning of the score?

Questions 1 and 4 each address key aspects of the IEA scoring process (i.e., what writing traits are scored, how well they are scored and how the overall scores are derived), while questions 2 and 3 focus on the properties of the scores generated from the IEA scoring process. Answers to these questions are essential in justifying score interpretation and use. Furthermore, they are critical to the issue of score defensibility, as the credibility of any scoring system depends upon its capacity to rationally explain how scores are determined.

The last chapter (Chapter Twelve) summarises the evidence collected from using the proposed framework and makes a validity argument for the IEA. It also appraises the usefulness of this framework in guiding an AES validation process, and considers how the AES framework helps to systematically address the validity questions set out above. Future work to strengthen the validity of the IEA scoring based on the evidence collected is also identified, as well as the implications of this study for future AES research work.

1.5 Limitations

While the validation process undertaken in this study has much to offer in relation to providing a strong case for the value and utility of the framework proposed, the small sample size (in terms of both the number of writing prompts⁵ and number of essays per prompt used in this study) means that the test findings from this study concerning IEA may not be as generalisable as would have been desirable. Larger studies with bigger sample sizes should be conducted to confirm the generalisability of the results from this study.

Another limitation concerns the source data used in this study. Due to practical constraints, data used in this study is part of the field test data used by Pearson Technologies (hereunder referred to as ‘Pearson’) to train and validate the IEA scoring model. As some of the essays included in this study may have been used to train IEA, the implication is that the level of correspondence reported between the human-generated and the IEA-generated scores might be over-estimated.

Notwithstanding these limitations, this study takes a significant step towards demonstrating a systematic method of evaluating AES construct validity which can then be used across

⁵ A writing prompt (or prompt) provides a rhetorical context to which a student needs to respond during a writing test. It nominates a writing topic which may or may not include additional stimulus material. See Section 5.2 for samples of prompts used in PTE Academic.

different AES systems, as well as towards making AES systems more transparent and meaningful to both test stakeholders and to the wider public.

Chapter 2 Automated Essay Scoring (AES) Systems

This chapter presents a literature review of the main Automated Essay Scoring (AES) systems available. It focuses on the technologies and theoretical frameworks that drive automated scoring and various approaches used to build the AES systems. Although the intention is to describe the nature of these systems as clearly and as completely as possible, the outcome is constrained by the availability of system information in the public domain due to the commercial proprietary nature of these products.

Before proceeding, this researcher first makes the distinction between an AES system and an AES model, to make clear the intended meaning of these two terms in the following discussions. In this thesis, an AES system is defined as a collection of operational scoring models that share the same philosophical foundations and technological frameworks that underpin the methods of scoring. An AES model refers to a scoring model that is developed for operational use for a specific prompt or for a specific assessment program.

2.1 Introduction

Automated Essay Scoring (AES) is a relatively new field with only a 40 year history. The main theoretical frameworks on which various AES systems are based include those from the fields of natural language processing, artificial intelligence, cognitive science and computational linguistics.

Currently there are four main commercial AES systems available: Project Essay Grade (PEG), *e-rater*, IntelliMetric and the Intelligent Essay Assessor (IEA). Despite their differences in the technologies used to score the textual features, the common approach adopted to build the operational scoring models invariably involves first identifying a set of measureable features that are approximations of the construct of interest, then modelling these features to maximise the correspondence of the AES-generated scores with some external criteria. This is then followed by model testing and validation using separate data sets to check the generalisability of model performance and fine-tuning the model. Once the model demonstrates an acceptable level of consistency and accuracy in scoring across different real data sets, it can be put into operational use (Yang et al., 2002, p. 394).

The following sections describe the main types of technologies underlying the four AES systems. There is an emphasis on the IEA because it is the system that will be the focus of this study.

2.2 Project Essay Grade (PEG)

Project Essay Grade (PEG) was the first AES system developed. It was developed by Ellis Page and his colleagues in 1966. Earlier versions of this system used 30 computer quantifiable predictive features to approximate the intrinsic features valued by human markers. Most of these features were surface variables such as the number of paragraphs, average sentence length, length of essay in words, and counts of other textual units (Page, 1966, 1968).

In order to overcome the reliance on surface structures, which threatened construct validity and made the system vulnerable to cheating, a revised version was released in the 1990s. The new version included some natural language processing tools such as grammar checkers and part-of-speech taggers (Page, 1994, 2003; Page & Petersen, 1995). As a result, the new version is said to attend to richer and more complex text features that are more closely linked to the underlying competencies that are required to be measured by the writing tests (Ben-Simon & Bennett, 2007).

Project Essay Grade has been reported as being able to provide scores for separate dimensions of writing such as content, organisation, style, mechanics (i.e., mechanical accuracy, such as spelling, punctuation and capitalisation) and creativity, as well as providing an overall score (Keith, 2003). However, the exact set of textual features underlying each dimension as well as details concerning the derivation of the overall score are not publicly disclosed (Ben-Simon & Bennett, 2007; Page, 2003; Shermis, Koch, Page, Keith & Harrington, 2002).

2.3 IntelliMetric

IntelliMetric is generally regarded as the first essay scoring system that extensively uses the technology of artificial intelligence (Elliot, 2003). It was developed by Vantage Learning and was first released for commercial use in 1998 (Elliot, 2003). At the present time, it is used in conjunction with human markers to assess the quality of essays written for the Analytical Writing Assessment (AWA). AWA is a part of the Graduate Management Admission Test (GMAT) (Talento-Miller, Siegert & Taliaferro, 2011).

Although the key technologies underpinning the IntelliMetric modelling process remain protected by various patents, some insight into the key features of this process can be gleaned from the high-level descriptions provided by Mikulas and Kern (2006) of Vantage Learning. The first feature of the process is the so-called “neuro-synthetic” approach, which is used to build scoring models. The process extensively uses artificial intelligence technologies such as neural nets to imitate the mental process used by human experts to “acquire, store, access and use information” (Vantage Learning, 2003c, p. 5).

A second feature of the modelling process relates to the *iterative* nature of the process to acquire its knowledge of the scoring rubrics⁶ from the training data. Using patent technologies such as CogniSearchTM and Quantum ReasoningTM and a built-in error reduction function, the modelling process infers the scoring rubrics by identifying the characteristics that are valued by human markers and associating them with each score point through many iterations.

A third feature of the model learning process is its claimed capacity to ignore the noises/anomalies that come with the training data set and to focus itself on “the overall pattern of information and the preponderance of the evidence” (Mikulas & Kern, 2006, p. 2).

⁶ For consistency, a common set of terms describing key components of a marking operation are used throughout the thesis. A scoring rubric refers to a set of criteria and standards typically linked to learning objectives. It includes one or more traits (i.e., characteristic of writing) for which performance is measured, definitions and examples that illustrate the trait(s) being measured, and a rating scale for each trait. Standards and rules by which traits are assessed are referred to as scoring criteria. Definitions are referred to as descriptors. A rating scale refers to a measurement instrument used to record the results of markers’ observations. It consists of a number of score points that define a measurement continuum.

A fourth feature is the use of multiple mathematical models to make independent judgements of the quality of an essay via different evaluation methods. This technique is unique amongst AES systems. Vantage Learning views this technique as akin to the employment of multiple human markers in a high-stakes marking scenario making independent evaluations of the same essay (Elliot & Mikulas, 2004; Mikulas & Kern, 2006). Relationships between the outputs from these mathematical models and the final score are modelled by a patented optimisation technique. Although the end results of such model-building techniques are the documented superior correspondence rates between scores produced by IntelliMetric and by human markers (e.g., Dikli, 2006; Rudner, Garcia & Welch, 2006), the limitation is that the exact nature of the complex relationships within IntelliMetric models is difficult to describe, explain or to evaluate.

Overall, IntelliMetric analyses more than 400 semantic, syntactic and discourse level features for any given essay, by using patented technologies in morphological analysis, spelling recognition, collocation and word boundary detection (Rudner et al., 2006; Vantage Learning, n.d.).

In operational scoring, IntelliMetric can provide overall scores as well as scores on each of five broad dimensions of writing: Focus/Coherence, Organisation, Elaboration/Development, Sentence Structure and Mechanics/Conventions (Elliot, 2003). Figure 2.1 shows the

descriptions of each dimension and the relationships between feature classes and the five dimensions.⁷

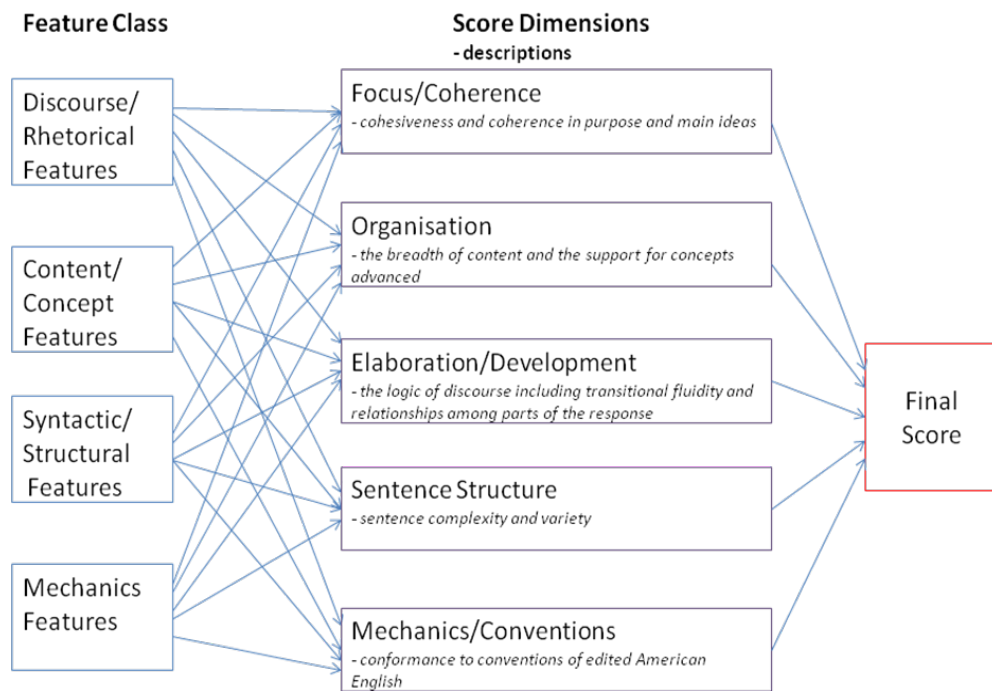


Figure 2.1 IntelliMetric Feature Model (Elliot, 2003, p. 73)

It is evident from Figure 2.1 that the five score dimensions in IntelliMetric are built from all feature classes, regardless of whether categories of features are conceptually related to a dimension (e.g., mechanics features can be contributing to the coherence dimension). While

⁷ According to the Vantage Learning website (Vantage Learning, n.d.), IntelliMetric currently “analyzes more than 400 semantic-, syntactic-, and discourse-level features to form a sense of meaning”. It “provides a holistic score as well as scores within five major domains: Focus and Meaning, Organisation, Content and Development, Language Use and Style and Mechanics and Conventions”. Though the names for the five domains are different from Elliot (2003), there is no evidence to suggest that the main technology frameworks used for scoring (as described in this section), or the way in which overall scores are derived (as depicted on Figure 2.1), have changed.

this type of model-building process may boost the performance of the system, it raises questions about the interpretability and the meaning of the scores produced.

2.4 *e-rater*

The *e-rater* was developed by the Educational Testing Service (ETS) in America during late 1990s. It uses an approach to model building and feature analysis that is based on analysing patterns across a large number of real world texts, such as field-collected first-draft student essays of a particular genre. The *e-rater* is used operationally in conjunction with a human marker for the Graduate Record Examination (GRE) Issue and Argument tasks since 2008 (Bridgeman, Trapani & Attali, 2009) and for the TOEFL Independent tasks since 2009 (Attali, 2009).

The current *e-rater* model structure includes a set of eight features: grammar, usage, mechanics, style, organisation, development, lexical complexity, and content (Attali & Burstein, 2006). At a surface level, the eight features reflect the generally accepted dimensions in essay writing that human markers emphasise in their marking processes. Most of the features are measured through underlying microfeatures. Figure 2.2 displays the internal structure of the current *e-rater* scoring model, with the full set of features and underlying microfeatures (Quinlan, Higgins & Wolff, 2009, p. 9).

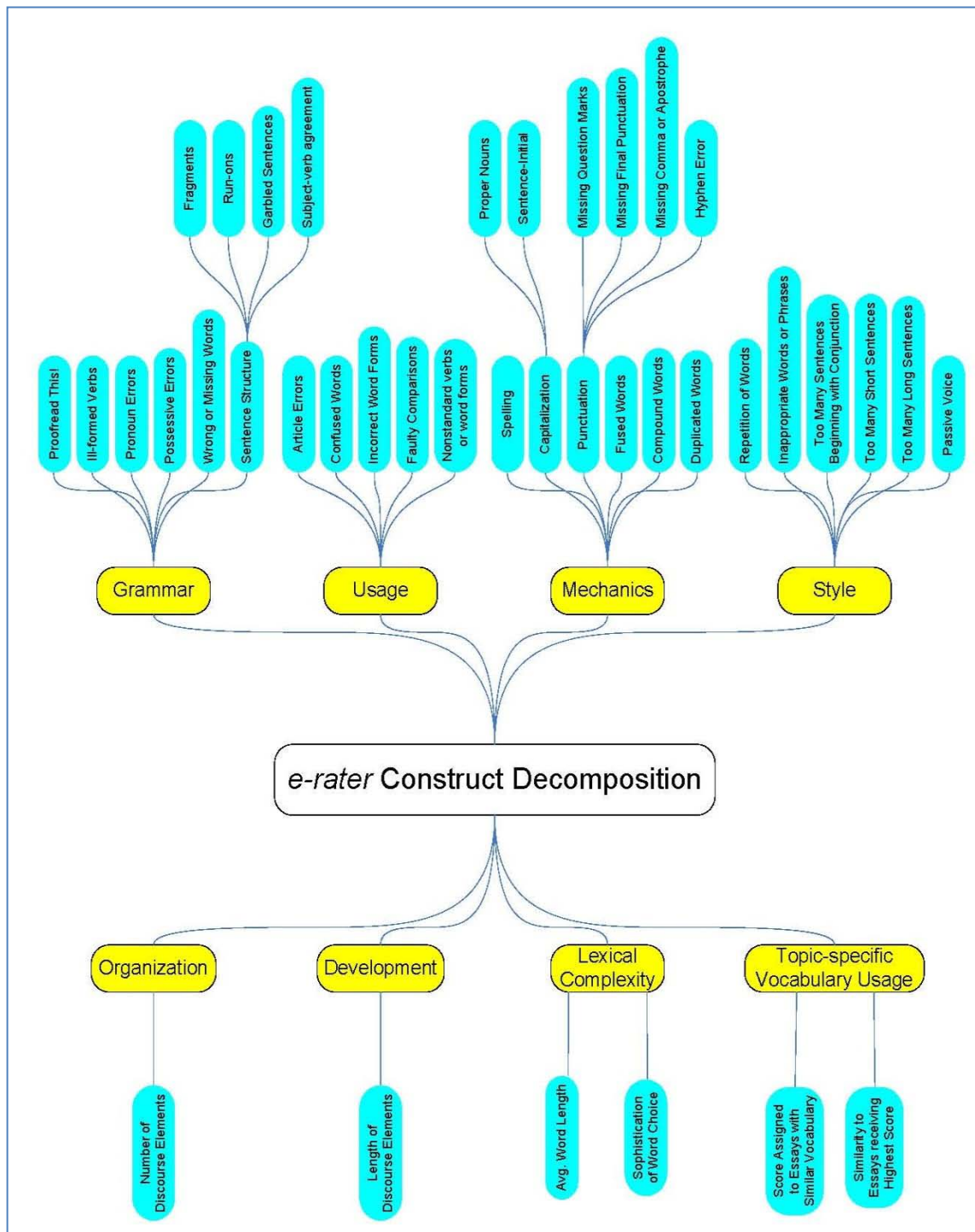


Figure 2.2 The *e-rater* Scoring Model (Quinlan et al., 2009, p. 9)

As shown in Figure 2.2, for example, the style feature score is derived from scores on six microfeatures, which relate to various aspects of writing style, such as too many short or long sentences, or too many sentences beginning with conjunction words. In operational settings, in order to calculate an overall score for an essay, the *e-rater* uses regression analysis to determine the weighting of the feature scores and uses this weighting to combine feature scores to an overall score (Enright & Quinlan, 2010).

The *e-rater* relies heavily on Natural Language Processing (NLP) and Information Retrieval (IR) technologies to extract text features. NLP is not an overarching theory but a collection of tools that “apply computational methods to analyse characteristics of electronic fields of text or speech” (Burstein, 2003, p. 115). As an example, the following paragraph describes how *e-rater* assesses the syntactic features by capitalising on the research in the NLP field (as summarised from Burstein, 2003; Burstein, Kukich, Wolff, Lu & Chodorow, 2001; Burstein, Marcu & Knight, 2003).

How *e-rater* Assesses Syntactic Features

In order to assess syntactic features of an essay, *e-rater* first uses a part-of-speech tagger (Ratnaparkhi, 1996) to assign labels to all words in an essay (e.g., noun, verb, and preposition). A syntactic structure analyser (Abney, 1996) is then used to identify and assemble phrases into trees based on sub-categorisation information for verbs (Grishman, MacLeod & Meyers, 1994). A computer program is subsequently used to identify the number of different types of clauses (such as complement clauses, subordinate clauses, infinitive

clauses, relative clauses) and occurrences of the subjunctive modal auxiliary verbs for each sentence in an essay. A possible measure of syntactic variety can then be created by calculating the ratios of syntactic types per essay and per sentence.

Other techniques from the NLP and IR fields utilised by the *e-rater* models include:

- applying Standardised (Word) Frequency Index (as devised by Breland, Jones, and Jenkins, 1994) across all words in an essay to measure lexical complexity
- using content vector analysis, based on a vector-space model, to evaluate topical content of an essay (Salton, Wong & Yang, 1975) and
- using surface cue words, non-lexical syntactic structure cues and terms to denote discourse elements in essays according to the discourse classification schema (e.g., Cohen, 1984; Litman 1996; Quirk, Greenbaum, Leech & Svartik, 1985).

2.5 Intelligent Essay Assessor (IEA)

The system used in this validity study was the Intelligent Essay Assessor (IEA). The IEA was developed in 1998 by Knowledge Analysis Technologies (KAT), which was later acquired by Pearson in 2004.

The Intelligent Essay Assessor (IEA) is currently used to score written responses for the Pearson Test of English (PTE) Academic. Of the four AES systems currently available, the IEA is the only one which claims that it can “measure factual knowledge based on semantic

content” and “is the only essay evaluation system in which meaning is dominant”

(Knowledge Analysis Technologies, 2001, as cited in McGee, 2006, p. 80).

Since the IEA is the only essay scoring system in the market with a claimed strong emphasis on the evaluation of content, it is appropriate to begin with a consideration of the importance and feasibility of content assessment in automated scoring.

2.5.1 Importance and Feasibility of Content Assessment in Automated Scoring

There has been considerable debate amongst AES developers about what aspects of writing an AES system should be designed to measure. One key issue in this regard concerns the evaluation of essay content. Some developers question the need to develop sophisticated tools to assess content since machines (i.e., the AES systems) can never read nor understand the meaning of a text as do human beings. Furthermore, they believe that “writing teachers”, in general, focus on the writing skills (such as the rhetorical aspects of the communicative process and language skills), rather than the correctness of content, in the assessment of writing (Shermis et al., 2002). Hence they regard it as neither necessary nor practical to develop tools to understand the meaning of an essay in large-scale assessments. Other researchers contend that content is the most important component in a scoring model (Landauer, Laham & Foltz, 2001, as cited in Attali, 2007) and that it is possible to develop a semantic model to capture the essence of the content in an essay.

It needs to be clarified that there are two different questions in this debate. The first is how important content is to the overall quality of an essay. The second is whether or not a machine can be designed to grasp the meaning of content within an essay. The answer to the first question is not clear-cut; it depends largely on the purposes of the essay assignments. If an essay assignment is for students to demonstrate their understanding of the subject matter or how they apply the factual knowledge learned in the study through critical thinking, the focus of the essay assessment would necessarily include the breadth and depth of the conceptual content contained in a response. This evaluative emphasis on content quality also applies to writing assessments performed in the context of content-based academic writing instruction, where writing is linked to concurrent study of subject matter in one or more academic disciplines (Shih, 1986).

On the other hand, if the writing assessment is about measuring students' basic writing skills (such as a writing test for Year 3 students), the evaluation of the quality of a written product would necessarily focus on basic sentence construction, spelling, language convention and word choice. Quality or quantity of ideas is of a lesser concern to the markers and may be a relatively insignificant aspect in the marking rubrics.

It can be argued therefore that the relative emphasis human markers place on the content to the overall quality of writing increases as the writing moves along a developmental continuum. On this continuum, writing moves from basic writing, such as knowledge telling, to more complex writing, such as the "analysis" and "argument" types of writing, as defined by Kiniry & Strenski (1985). At the high end of this continuum, when writing is for adapting and

transforming individual's domain knowledge, evaluation of essay content would be an essential focus of essay assessment. This type of writing; that is, writing for knowledge transforming, is regarded the most critical in academic writing (Weigle, 2002).

Some researchers (e.g., Attali, 2007; Attali & Burstein 2006) have presented evidence that AES scoring models, which exclude content assessment and focus solely on the language aspects and structure of an essay, can reach high agreement rates with scores from human markers for certain types of expository writing. However, agreement rates do not substitute for construct validity, a point that will be expanded further. Excluding content evaluation from an AES scoring model, when content is regarded as an essential part of the construct being measured, significantly weakens the validity of the scoring model, with scores generated being contaminated by construct under-representation.

The second question in the debate of content assessment by an AES system is “can a machine be trained to derive the underlying meaning of a text?” This question is most often answered on philosophical grounds. The traditional wisdom or belief has always been that the machine cannot appreciate the meaning of a text as well as a human being (Kemp, 1992). However, a group of researchers (Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, and others) have presented a different view. They showed that it was possible to develop a model to simulate human understanding of the meanings in a response. In 1990 they published a seminal paper which examined the use of a technique called “Latent Semantic

Analysis” in the context of information retrieval (Deerwester, Dumais, Landauer, Furnas & Harshman, 1990)⁸.

Latent Semantic Analysis (LSA) provides a way to infer a semantic structure in a corpus through the condensation of local co-occurrences. One measure of local co-occurrences is the number of times a word appears in a context. The term “context” here and in the subsequent sections, means a small section of text which has coherent meanings, such as a paragraph. Although the model is strictly a mathematical one, it was conceptualised by these researchers as a theory of knowledge representation and acquisition. LSA was first used in the scoring of essays in 1995. In 1998 the model became a significant component of a new automated essay scoring system – Intelligent Essay Assessor (IEA), which is the system of interest for this study.

Since the primary focus of IEA is the evaluation of content and LSA is the fundamental theory used to assess content, it is therefore necessary to describe LSA in detail in order to illustrate how IEA evaluates essay quality.

⁸ The first paper on Latent Semantic Analysis was presented at a Conference on Human Factors in Computing in 1988. See Dumais, Furnas, Landauer and Deerwester, (1988), “Using latent semantic analysis to improve information retrieval”, in Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281–285.

2.5.2 Latent Semantic Analysis (LSA)

2.5.2.1 What Is Latent Semantic Analysis? – An Overview

Latent Semantic Analysis (LSA) is a machine learning method that infers the meaning relations among words and passages through mathematical computations applied to a large corpus of text (Landauer et al., 2003). There are three key assumptions underpinning the LSA approach to deriving meaning from text (Landauer et al., 2003).

First, LSA assumes that the ways in which words and sets of words relate to each other in a semantic space are largely constrained by the aggregates of all the word contexts in which a given word does or does not appear (Landauer et al., 2003). For example, suppose two words “doctor” and “patient” have a relatively higher frequency of co-appearing in the same context, than other pairs of words, aggregated over many contexts. The higher likelihood of co-appearance is assumed to mean that these two words have relatively greater similarity in meaning, hence should be closer in distance in a semantic space, than other pairs of words. Under this assumption, LSA uses measures of local co-occurrences, for example, the number of times a word appears in a context, as a basis to determine the similarities of the meanings amongst words, paragraphs and documents.

The second assumption is that, in order to better simulate the semantic relations amongst text units (such as words, paragraphs, documents), it is necessary to reduce the dimensions of the semantic space in which these text units originally appear. One rationale underpinning this assumption is that, there is a lot of natural noise in the raw co-occurrences data. By

condensing the original dimensions to a smaller (but still large) number of optimal dimensions, it might eliminate some of the natural noise in the raw data. Another reason for condensing the original semantic space is that it might provide a way to induce indirect meaning relations amongst words that may never appear in a joint context. It is common that many related words do not appear in one text. For example, words “liver” and “heart”, which are associated with each other through a common object “body”, might not co-occur in one text if the text is presenting a view on only one part of the object. If semantic relations are built based on actual occurrences of co-appearance, no meaning relation can be derived for these two words. However, by condensing the semantic space through dimension reduction, indirect, higher-order association between these two words might be established through their respective associations with the common object “body”. In essence, the original space can be seen as indicating direct relationships between meanings of words based on their *actual* co-appearances in joint contexts from an input corpus. The reduced space can be seen as indicating the similarity of every word to every other word whether or not they have ever occurred in a common text window. Dimension reduction may therefore provide a means to deal with some of the complex phenomena in English, such as synonyms and compounds (Foltz, Kintsch & Landauer, 1998).

The third assumption is that, LSA assumes that the meaning of a text is simply the sum of the meaning of each of the words it contains. For example, LSA would derive the same meaning from two different sentences such as “Chicken lay eggs” and “Eggs lay chicken” because the two sentences contain the same words. How words are ordered within a sentence or how

sentences are arranged within a paragraph or a text is irrelevant in LSA's determination of the semantics of the text.

2.5.2.2 Mathematical Description of Latent Semantic Analysis

To construct a k -dimension semantic space, LSA needs to be trained on a large representative corpus, such as an encyclopaedia, text books or source material on a particular topic. The first step involves LSA dividing the corpus into small chunks of texts with coherent meanings, such as paragraphs. These small text bodies, referred to as "contexts", are then analysed and transformed into a matrix of local co-occurrences, consisting of rows of unique word types by columns of contexts in which word types appear. Each cell represents the frequency of the word type appearing in the particular context.

Next, each cell value is transformed by a log entropy function similar to the following:

$$\frac{\ln(1 + \text{cell frequency})}{\text{entropy of the word over all contexts}}$$

Entropy is a well-known measure that is frequently used in probabilistic information retrieval models. Calculated using a formula, $\sum p_i \ln p_i$, it represents the aggregated probability of a word appearing in all contexts belonging to one domain represented by an input corpus.

Hence entropy is a measure of a word's information value in this general domain. The larger the entropy of a word, the less contextual-specific meaning it carries and the less information it conveys. The combined log entropy function is a well-documented method in the field of

information retrieval which has been shown to improve the associative relationship derived from untransformed co-occurrence data (Landauer & Dumais, 1997).

The final step in LSA involves a dimension reduction process using the Singular Value Decomposition (SVD) technique. SVD (Golub & Kahan, 1965) is a long-known matrix factorisation technique that exposes the underlying structure of a matrix. Deerwester et al. (1990) patented the use of SVD in the context of information retrieval in 1988.

In essence, for the purpose of semantic analysis, LSA uses SVD to construct a condensed “concept space” which retains only the most important dimensions in the original matrix space expressed in the input corpus. In this concept space, the meaning of each word is expressed as the aggregate of its position to each of these “concepts” or dimensions. It is through this “mapping” of a word to common concepts that the meaning of this word can be compared to that of another, even though the two words have not co-appeared in a joint context.

The optimal number of dimensions (k) retained in this concept space is empirically determined. Landauer & Dumais (2008, p. 3) report that, the useful range of k is between 50 and 1000, with 300 ± 50 most often being the best.

Experiments conducted by LSA founders (e.g., Foltz, 1996; Landauer, Foltz & Laham, 1998; Landauer, Laham, Rehder & Schreiner, 1997) have demonstrated that LSA can derive deeper

semantic relations than mere co-occurrences and the dimension reduction technique significantly improves the structural relations in the derived concept space.

Once an optimal semantic space is constructed, all text units, including words, paragraphs and documents, can then be represented as vectors in this space. The similarity in the meanings of any two text units can be computed as the cosine measure between vectors.

In summary, an LSA model first transforms a large representative corpus to a matrix consisting of co-occurrence-based measures. This is followed by a dimension reduction process to produce a semantic space at an optimal dimensionality, in which any one text unit is mapped to another through their respective relevance to dimensions or concepts extracted from the original corpus.

2.5.2.3 Limitations of Latent Semantic Analysis

There are two limitations of LSA which affect its ability to accurately derive meanings from a text. The first limitation is that LSA ignores the order of words or arrangement of sentences in its analysis of the meaning of a text. A text is simply treated as a “bag of words” – an unordered collection of words. As such, the meaning of a text as derived by LSA is not the same as that which could be understood by human beings from grammatical, syntactic relations, logic, or morphological analysis. As noted by Landauer and Dumais (2008), LSA is a mathematical model which describes semantic relatedness of two text units through distances in a condensed space. Hence the nature of the semantic relations as derived by LSA

is *spatial*, instead of strictly linguistic. Similarly, dimensions resulting from the SVD technique may be meaningful on the mathematical level, but may not be interpretable in natural language processing. The effect of this limitation on the IEA's measurement capability when assessing content is probably best demonstrated in McGee's (2006) study. This study took well-written and meaningful essays and turned them into nonsensical ones by changing the sequence of sentences, replacing key words with antonyms thus reversing the true value of the propositions and varying the order of the words within sentences. IEA assigned virtually unchanged high scores to all the revised essays. It may be argued that, if students are clever enough to creatively construct responses to fool the machine, they could probably generate a good essay as well (Shermis et al., 2002). However a more serious concern to educators is that IEA may not reliably score essays in which students quote phrases from text books but do not use them in a meaningful way, due to a lack of understanding of the material (Wohlpert, Lindsey & Rademacher, 2008).

The second limitation is that LSA does not deal with polysemy (i.e., one word with multiple meanings). This is because each word is represented in the semantic space as a single point and its meaning is the *average* of all its different meanings in the corpus (Landauer & Dumais, 2008).

A final word of caution around LSA is the difficulty in finding large enough representative corpora to train a robust model. Landauer & Dumais (2008) indicated that for most language simulations, "corpora supplying less than 20K word types in less than 20K passages are likely

to yield faulty results” (p. 3). Inferior training data increases the chance of anomalies in the proximity of words in the semantic space.

Although the developers of LSA concede that LSA is not a complete linguistic model (Landauer & Dumais, 2008), they argue that this should not limit the use of LSA because the utility of a semantic model in an AES system depends on the sufficiency with which this model simulates human judgements on content. They cite results from various experiments which demonstrate that LSA’s judgements on content were close to those of the human experts, as evidence supporting their argument (Landauer et al., 2003). The accuracy of the IEA scoring of content and the associated validity issues are matters that will be pursued in Chapter Eleven.

2.5.2.4. The Use of Latent Semantic Analysis in Intelligent Essay Assessor

Latent Semantic Analysis is integrated into the IEA as an analytic tool for assessing two aspects of content – the quality and quantity of content, and the conceptual flow of an essay.

Assessing Essay Content

In order to assess the quality of the content, the IEA compares the content in the target essay (i.e., the essay that needs to be assessed) to that in an essay with known quality (e.g., a pre-scored essay or an expert essay or source material). This is achieved by calculating the cosine measure of any two vectors representing the target essay and a training essay in an LSA space derived from the training corpora. In order to assess the quantity of content, the IEA uses the

length of the vector which represents the target essay in the derived semantic space as a proxy measure for how many underlying abstract concepts the essay has covered. The overall content score for the target essay is computed as a weighted sum of the quality and quantity scores, after normalisation and regression analysis (Landauer et al., 2003; Landauer et al., 1998).

Assessing Coherence

In order to assess the conceptual flow of an essay, the IEA computes an average cosine measure of all pairs of vectors of all adjoining sentences in this essay. This measure is intended to gauge how much common content two adjacent sentences share, a reflection of the degree of conceptual flow from one sentence to the next. Similarly, the semantic relatedness of one sentence to the whole of the paragraph or to the whole of the essay can also be computed to signal the extent to which discussions in an essay have stayed on track (Landauer et al., 2003). Studies have demonstrated that LSA's effectiveness in capturing the continuity of lexical semantics is superior to simple measures of literal word overlap, and that an LSA-derived coherence measure is highly correlated with text comprehension and readability measures (Foltz et al., 1998).

Detecting Plagiarism

Latent Semantic Analysis is also used in the IEA to flag essays which might be copies of others. As part of a standard scoring process, the IEA uses LSA to compare the meaning of each essay to all others in a set of essays. If a very high cosine measure between two essay vectors is detected, the two essays are flagged for further investigation. In detecting plagiarism, LSA is not affected by the reordering of sentences or paraphrasing (due to its “bag of words” method of analysing text units) or the use of synonyms, all of which may be hard for human markers to detect, particularly after markers have been subjected to long hours of marking (Landauer et al., 2003).

2.5.3 Intelligent Essay Assessor Architecture

In addition to marking essay content and coherence, a general IEA scoring model also marks for style and mechanics, which includes assessment of writing features such as grammar, spelling and punctuation. Figure 2.3 is the IEA architecture (Landauer et al., 2003, p. 90) which shows the components of a general IEA scoring model.

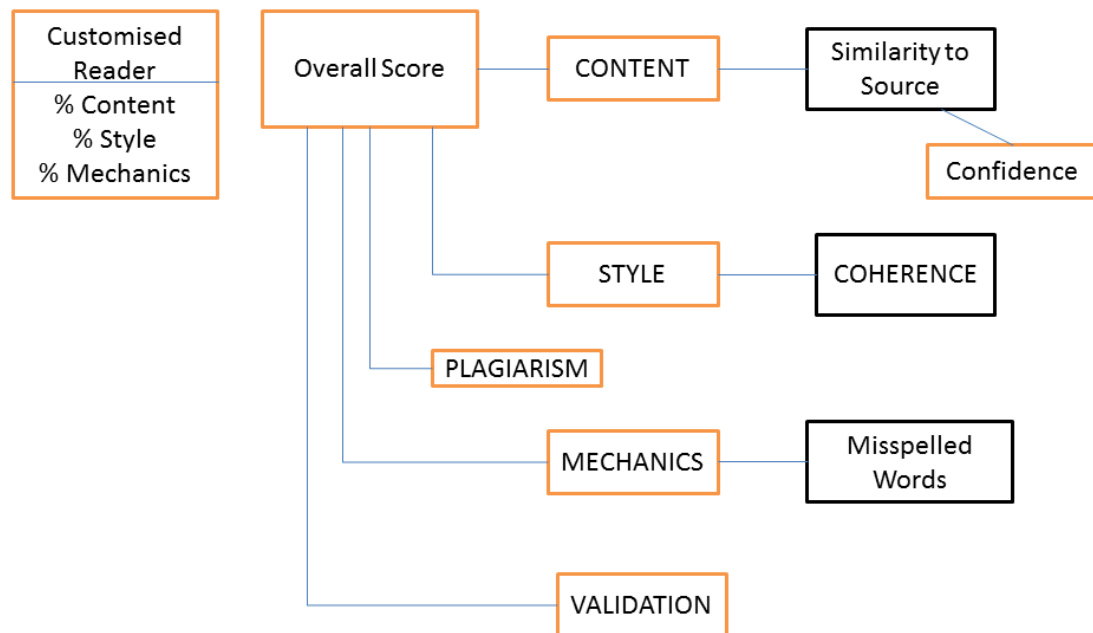


Figure 2.3 IEA Architecture (Landauer et al., 2003, p. 90)

In 2009, the IEA was selected as the automated scoring system to evaluate the quality of responses written to the writing tasks (i.e., prompts) in the PTE Academic. For this test, the IEA assesses the following seven aspects of writing performance: Content, Formal Requirements (a length requirement), Grammar Usage and Mechanics, Vocabulary Range, General Linguistic Range, Spelling, Development, Structure and Coherence. More details of the PTE scoring criteria will be discussed in later chapters. Apart from well-articulated theories of assessing content and coherence, there is little publicly revealed information regarding how other textual features are measured in the IEA.

The preceding sections described various AES systems as well as some obvious issues arising from the model-building processes that may have a deleterious effect on the appropriateness of the scores generated from these systems. In recent times, there have been some concerted efforts from vendors to improve these systems in an attempt to address the validity concerns from educators.

2.6 Recent Trends in the Development of New AES Models

Overall, it can be said that the direction of enhancements that are being made to AES models in recent times is to strengthen the validity and substantiveness of the scores generated by these models. Three emerging trends in the development of new AES models are discussed.

2.6.1 Moving from Prompt-specific Models to More Generic Models

Automatic Essay Scoring models have always been built using a prompt-specific approach; that is, the model is trained on a set of pre-scored essays written to a specific prompt before it is used to score new essays written to the same prompt (Attali & Burstein, 2006; Mikulas & Kern 2006). This ensures high levels of agreement rates between AES- and human-generated scores since the model is specifically trained for each prompt.

A number of issues exist with regard to this model-building approach. First, it results in a standard requirement of building an AES model; that is, a certain number of pre-scored essays are needed to train and calibrate an AES model to score a prompt (Burstein, 2003; Chung & O'Neil, 1997; Elliot, 2003; Landauer et al., 2003; Rudner & Liang, 2002). This requirement

poses some practical problems for a wider application of AES systems. To start with, the need to collect a set of pre-scored essays every time a new prompt is operationalised and then to build a model specific to this prompt using the training data collected, can be both time-consuming and costly to the users. Furthermore, in order to avoid being trained on situational variables or markers' subjective judgements which are likely to result in deficient models, AES models must use a training data set that meets certain quality criteria (e.g., having sufficient coverage across each score point particularly at the tails of the achievement scale). This requirement poses further practical and financial burdens on schools and education systems that want to use AES technology extensively.

A further issue with this prompt-specific approach is that it results in potentially idiosyncratic scoring models that are sensitive to the characteristics of training data sets, such as those related to the markers, essays and examinees that are used to produce the training data. One group of markers whose judgements are used to train one scoring model for one prompt may have different views of writing quality than another group of markers whose judgements are used to train another model for another prompt. Therefore this approach tends to produce different models with different scoring criteria for each prompt, even if these prompts belong to the same assessment program and are meant to be interchangeable⁹ (Attali & Burstein

⁹ An example of an assessment program can be a general English proficiency test such as Test of English as a Foreign Language (TOEFL). For this assessment program, writing prompts which are part of the item bank for the same independent writing component of this test are meant to be interchangeable.

2006). This issue therefore raises the concerning possibility of incomparability of AES scores across different but parallel writing prompts.

In light of the issues raised above, some AES developers (e.g., Attali & Burstein, 2006; Mikulas & Kern, 2006) have started investigating the efficacy of a different approach – building generic scoring models which are prompt-independent. This approach involves calibrating the models using a combination of training data sets from multiple prompts that belong to the same program. Once the models are demonstrated to be invariably effective across different prompts, they are used to score all prompts that belong to the same assessment program. This model-building approach not only allows the automated scoring of a great variety of prompts without the need for collecting pre-scored training essays for all of the prompts, but also ensures that consistent scoring criteria are applied across prompts within the same assessment program.

One possible drawback of this new approach is that the performances of the resultant generic models might be compromised by the fact that the models are not built on prompt-specific training sets. However, a limited number of available studies demonstrate that such models (*e-rater* and IntelliMetric) can be built without significant loss in their predictive performance; that is, without significant loss in their abilities to reproduce human scores (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Mikulas & Kern, 2006).

It is worthwhile noting that a potentially significant issue associated with prompt-independent models is that they do not, and cannot, assess the content of the essays since they are not

trained on prompt-specific essays. As stated above, the omission of content analysis in a scoring model may have serious deleterious effects on AES construct validity, if the quality and quantity of content are part of the definition of good writing and therefore are meant to be part of the scoring rubrics. The latter is particularly true for scoring those essay assignments that require detailed analysis of a specific topic. Although the available studies (e.g., Ben-Simon & Bennett, 2007; Mikulas & Kern, 2006) show that generic models do not suffer significant loss in predictive performance when compared to prompt-specific models, it is not clear to what extent the observed results may be due to the ineffectiveness of these AES models in capturing the breadth and depth of content in the first instance. In other words, had the prompt-specific models been able to measure content more accurately, the comparative results may well have been very different.

2.6.2 Moving from Empirical Methods to Construct Driven Methods of Model Building

In recent years, researchers (e.g., Attali, 2007; Bennett, 2004; Quinlan et al., 2009) have called for the development of AES models to be driven by expert and/or theoretical understandings of the construct domain, rather than by empirical methods, in order to strengthen the validity of scores produced. One critical issue is the way features selected for inclusion in the scoring model and the weights associated with these features, are determined during the process of building an AES model.

Until now, most AES models have used automated statistical optimisation methods to select predictive text features and determine the weights associated with these features for the

purpose of calculating an overall score for each essay. Although this data-driven approach maximises AES agreement with human markers, it nevertheless has potentially harmful effects on the construct validity. For example, features selected and weights determined in this way may not be meaningful or valid and may be difficult to describe and explain to the users of the test results. As pointed out by Attali & Burstein (2006), “difficulty in communicating the inner structure of the scoring model is a threat to the face validity of AES” (p. 13). Furthermore, the use of the data-driven approach to select features and determine the weighting schemes aggravates the idiosyncrasies of the scoring criteria already existent in prompt-specific models and makes the AES-generated scores even more inconsistent and incomparable across prompts.

The use of automated techniques to maximise the agreement rates may also result in undesirable statistical side effects, such as less variability in AES scores than in the human scores (Attali & Burstein, 2006). The tendency of the AES models to produce more compact scores than the human markers is observed in a number of international and local studies (e.g., Ben-Simon & Bennett, 2007; Davies & Gralton, 2009; Rudner et al., 2006; Wang & Brown, 2007). The fact that AES models may provide less discriminating scores across the whole achievement scale could be of significant concern to test administrators, depending on the purposes and requirements of the assessments. One remedial procedure adopted by *e-rater* is to have an additional scaling process which forces the machine scores to have the same distributional properties as human scores (i.e., the same mean and standard deviation) (Attali & Burstein, 2006).

In order to create more defensible and meaningful AES models, some vendors have started moving away from data-driven approaches to construct-driven methods of developing models. For example, ETS (the *e-rater* developer) has experimented with offering domain experts greater judgemental control over the construct representation when theoretical or other considerations are present (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007). Experts, through panel discussions, determine which dimensions of writing quality need to be scored in a particular testing context and the relative importance of each dimension to the judgement of the overall quality. ETS research studies (e.g., Attali 2007; Attali & Burstein, 2006) have demonstrated that expert-determined weights are no less efficient than the optimal weights found through statistical analysis.

This method of model building highlights a significant benefit of AES; that is, AES technologies can be used to achieve much more refined construct control (Bennett, 2004). While human markers may have difficulty in dealing simultaneously with multiple features and weighting them appropriately and consistently, AES systems can re-tune the models easily based on a new set of scoring criteria and apply them consistently.

2.6.3 Moving from Summative to Formative Assessment

Domain theory (Hayes & Flower, 1980) suggests that writing is a recursive process and that good writing practice involves multiple drafts and revisions. Therefore it is desirable that writing assessment tools allow for students to submit a draft, receive formative writing feedback and make subsequent revisions based on feedback instructions. However, the

development of AES systems as assessment tools has been largely focused on the product rather than the process of writing (Bennett, 2004; Dikli, 2006; Quinlan et al., 2009). As pointed out by Dikli (2006), the product approach views writing assessment as a summative practice with AES development primarily focusing on instant essay scoring in a high-stakes testing context. The process approach, on the other hand, views writing assessment as a formative practice, with AES vendors developing analytic frameworks using AES technologies to provide real-time feedback to students and teachers to assist the writing process. Since formative assessment has been demonstrated to produce specific educational gains (e.g., Brewer, 2004; Henly, 2003; Justham & Timmons, 2005; Peat & Franklin, 2002), and its purpose is *for* learning rather than *of* learning, this type of assessment is considered to be important and may even be “at the heart of effective teaching” (Black & William, 1998, p. 140).

In the current market, MY Access![®] (developed by Vantage Learning), CriterionSM (developed by ETS) and WriteToLearn[®] (developed by Pearson) are examples of online applications that utilise AES technologies to support formative assessment. These applications allow students to save their first and subsequent drafts in the computer and revise these drafts based on the feedback from the computer and the teachers (Dikli, 2006; *WriteToLearn*, 2011). WriteToLearn[®] is also a prime example of using AES technologies to provide immediate content-based instructions. This type of diagnostic feedback is important in learning because it helps reinforce important concepts in a subject and assists students to think more constructively about the missing ideas or correct information before writing

revisions. All three instruction-based AES applications deliver feedback and essay scores to students immediately and as a result, they are likely to encourage more effective practice of writing by students. Although practice has long been regarded by writing experts as an essential factor in developing writing expertise (Alamargot & Chanquoy, 2001; Becker, 2006; Rijlaarsdam & van den Bergh, 1996), in reality, students' writing practice has been limited by teachers' capacity to respond thoughtfully to students' writing in a timely manner (Elbow, 1981; Grimes & Warschauer, 2010). The advent of these instruction-based systems helps reduce the burden on teachers to respond while encouraging more practice by students, therefore potentially generating a greater positive effect on students' writing skills.

In order to further improve students' positive experience with AES in support of their learning, researchers have recommended future enhancements to the AES instruction-based systems (e.g., Burstein, Marcu, Andreyev & Chodorow 2001; Scharber, Dexter & Riedel, 2008). These include the need to align the development of these instruction-based systems closely with research work on writing revision and the purpose and function of responses to students' writing (Scharber et al., 2008), and the need to adapt feedback to individual student's writing skill and writing task (Burstein et al., 2001).

There is no doubt that, as AES becomes more grounded in credible theories of writing proficiency and as the use of AES moves more towards formative assessment, the credibility of AES and the educational benefit it brings will greatly increase.

2.7 Chapter Summary

This overview of theories and technologies underpinning the major AES systems has identified that the scoring methods used to assess many writing features by AES systems remain opaque, and that the complex mathematical relationships derived from sophisticated models are often difficult to understand and difficult to relate to theories of writing ability. These issues affect score interpretation and ultimately validity, and need to be resolved before significant educational benefits of these systems can be realised. The more construct-driven methods of building AES models now being experimented with by AES vendors are designed to support improvements in both rigour and transparency.

As a basis for developing a new framework for assessing score validity to support these new (and productive) directions, the next chapter delves deeper into the issue of the validity of AES systems, focusing on how it is currently being assessed and the gaps to be addressed.

Chapter 3 Review of the AES Evaluation Studies

Since the inception of the first Automated Essay Scoring (AES) system in the 1960s, a large number of evaluation studies have been conducted on the efficacy and validity of the AES systems (e.g., Attali & Burstein, 2006; Landauer et al., 2003; Nichols, 2004; Page, 2003; Vantage Learning, 2003a, 2003b). The majority of these studies have been conducted by developers of these systems.

The main evaluation approaches used in the AES studies can be classified into three categories focusing on: 1) the relationship among scores generated by different scorers; 2) the relationship between essay scores and external measures; and 3) the scoring processes used by the AES systems (Yang et al., 2002).

This chapter considers the main findings of the evaluation studies conducted so far, organised by the approaches used in these studies as classified by Yang et al. (2002). Other approaches that are significant and not covered by the abovementioned classifications are also discussed. In addition, potential problems associated with some of the approaches used are examined in order to gain a sound understanding of what is needed for future research studies in this area.

3.1 Studies Focusing on the Relationship among Scores Generated by Different Scorers

As validity evidence for AES systems, most of the evaluation studies have focused on the agreement rates that occur between scores given by AES and by human markers on the same

essay (Attali, 2007; Burstein, 2003; Elliot, 2003; Landauer et al., 2003; Page, 2003; Yang et al., 2002).

The most commonly used measures of agreement fall into the following categories:

- Exact agreement rates – the proportion of essays that human markers and an AES model agree on the exact score point level;
- Exact + adjacent agreement rates – the proportion of essays that human markers and an AES model agree within 1 score point difference;
- Correlation rates – the correlations between human and AES scores;
- Cohen's Kappa Index – the agreement rates adjusted by chance agreement. This recognises that human markers and an AES model can agree by chance alone.

Other measures used include comparisons of distributional properties (e.g., mean and standard deviation of human and automated scores) and standardised mean score difference between human and automated scores. Regardless of which measures are used, generally speaking, reasonable levels of agreement between the human and AES scores have been observed for all four AES systems (Attali, 2007; Ben-Simon & Bennett, 2007). Furthermore, many of these studies have also reported that, when agreement rates between two human markers are compared to the agreement rates between an AES system and an individual marker, AES systems are no less consistent than human markers (Burstein, 2003; Elliot, 2003; Landauer et al., 2003; Page, 2003).

Although agreement rates are relatively straightforward measures, there are some common issues which complicate the interpretation of these measures. First, when studies use the scores obtained from human markers as the external criterion measures, it is frequently the case that in these studies, the validity of scores from human markers is assumed rather than rigorously evaluated. The literature presents a strong argument that human markers are fallible and frequently have difficulty in assigning scores that are consistent and valid (e.g., Bennett & Bejar, 1998; Cooper, 1984; Diederich, 1974; Noyes, 1963). This raises the question of using human judgment as the gold standard. As cautioned by Messick (1989), where potentially contaminated external criterion measures (such as human judgments) are used in the construct validation process, the validity of the external criterion measures must first be examined and any findings must be interpreted within the context of how valid these criterion measures are. These considerations underpin the analysis in Chapter Eight of the quality of the scores produced by the human markers used in this study.

To the extent that a human marker's judgment maybe impaired by bias and subjectivity, it is not desirable for an AES model to achieve perfect agreement rates with one human marker. However, if there is a measure of "true score" for essay quality, it is desirable for the AES model to achieve a high level of agreement with such a measure (Yang et al., 2002). In this regard, some AES studies used average ratings of a large number of markers as an estimate of true score (e.g., Elliot, 2003; Shermis et al., 2002). Although this method does improve the reliability of the external criterion measures, it is still arguable that true scores constructed in this way can be treated as the unequivocal standard, particularly when markers might still be

biased in a uniform way. For example, McColly (1970) noted that human markers tended to agree in “their reaction to appearance” when reading handwritten essays. Markers within the same group might also have the same tendency for leniency or severity. This type of “correlated errors of measurement” can make up a significant portion of inter-rater agreement rates which give misleading indications of the quality of human scores (Werts, Breland, Grandy & Rock, 1980). When human markers share a common bias or error, averaging their scores does not remove the bias or errors associated with human scores. These issues mean that any inferences from the results of the human-AES score comparisons are both complicated and hazardous.

The second issue in interpreting agreement rates between scores from human markers and AES systems relates to how these rates are reported in the studies. Most of the studies report overall correspondence rates at the prompt level, without revealing disaggregated rates at each of the score points. An issue associated with this reporting method is that agreement rates between human markers and AES models may well be different for essays at different writing proficiency levels, and rates averaged across score points often mask the extent of discrepancies at a micro level.

For example, a study conducted by Burstein, Kukich, Wolff, Lu and Chodorow (1998) on the accuracy of *e-rater* scoring showed that, while *e-rater* had comparable correlation rates to those between the two human markers at the prompt level, large differences emerged at the two highest score points. While the rate of discrepancy between two human markers was 7% at the highest score point, the corresponding rate of discrepancy between *e-rater* and an

individual marker was much higher (31% to 34%). The study was based on the *e-rater* scoring of 500 Graduate Management Admissions Test essays and 200 Test of Written English essays. On the other hand, Wang & Brown (2007) demonstrated that the IntelliMetric system could not score essays at the lower end of the achievement scale well, based on its scoring of a sample of essays produced by 107 developmental writing students from a Hispanic serving institution in South Texas. The study found that IntelliMetric assigned a much lower failure rate (i.e., 2.8%) than did the human markers for the same students in the sample (i.e., 27.1%). These results suggest that while the machine (i.e., the AES models) and the human markers may converge on the scoring of mediocre essays, the differences could be quite significant at the high or low proficiency levels. One possible reason why the AES models cannot handle essays at the two tails of the achievement scale very well is that these essays are often diverse in the content, style and language used and thus have quite unique characteristics that may be difficult to be exemplified in a set of training essays. In the case of exceptionally high quality essays, it might also be a sign of the AES suffering from construct under-representation. In such a case, an AES model has either placed little emphasis or could not score well on personal voice, originality, creativity or graceful style. The possibility that an AES model might not be able to handle essays at both ends of the performance continuum would be of concern to those stakeholders who want a scoring tool that can assess essay quality equally well across the whole achievement scale. For these stakeholders, it would therefore be necessary to have agreement rates investigated at a micro-level. This type of analysis forms part of the investigations carried out in Chapter Eleven.

The third issue in interpreting the agreement rates is that measures of proportionate agreement rates are very sensitive to the number of score points, the number of essays used and the marginal distribution of scores (Yang et al., 2002). This issue needs to be taken into account when results from different studies are compared. When accessible, Kappa rates (Cohen, 1960) which adjust agreement rates for chance agreements should be used to form judgements, in conjunction with other types of measures such as simple agreement rates. This thesis reports Kappa rates as well as various other types of correspondence rates between human and IEA scores, when investigating the accuracy of the IEA scoring at the writing feature level.

Perhaps the most serious problem with the use of the agreement rates in AES studies is mistaking these rates as direct construct validity evidence. Although the agreement rates are a requisite criterion for evaluating the usefulness of an AES system, they do not provide answers to fundamental validity questions such as whether the writing features scored by an AES system are relevant to and representative of the construct of interest. It has been demonstrated that a scoring system can achieve a reasonable level of agreement with the human markers by evaluating only the surface level features such as number of words and number of paragraphs in an essay (Chodorow & Burstein, 2004; Kaplan et al., 1998; Page, 1966). Therefore to address the validity questions more convincingly and substantively, other types of empirical and theoretical investigations are required.

3.2 Studies Focusing on the Relationship Between AES Scores and External Measures

A large number of validity studies used correspondence rates as evidence. Considerably fewer AES studies, on the other hand, used the relationship between AES scores and external measures (i.e., measures of the same or similar construct) as evidence of the validity for AES systems. The studies that have been carried out have generally aimed to gather convergent evidence that served to strengthen construct validity; or identify such sources of invalidity as AES scoring essay features that were irrelevant to the writing ability being measured (Messick, 1996).

The types of external measures used in the available studies include: students' achievements in subjects dependent on writing; students' performance in multiple choice writing tests or other similar writing tests; teachers' assessments of students' writing ability; self-evaluations of writing skills; and, self-reported accomplishments in writing. Overall these studies show inconsistent and incomplete results, partly due to the limited number of external measures used in any one study (e.g., Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Elliot, 2003; Landauer et al., 2001; Petersen, 1997; Powers, Burstein, Chodorow, Fowles & Kukich, 2002; Weigle, 2010).

3.3 Studies Focusing on the Scoring Process

A very small number of AES studies have examined the AES scoring process, through both quantitative and descriptive evaluation methods, in order to accumulate direct evidence of validity for scores generated by these systems.

Some key validity questions that may be addressed through the scientific inquiries into the AES scoring process include: 1) What features are assessed by AES in its scoring process and how do they relate to the generally accepted dimensions of essay quality? 2) What are the precision and depth of AES measurement capabilities with regard to the assessment of these features? 3) Is the method used by an AES model to select and weight features theoretically sound? Answers to these questions not only help test administrators make informed decisions regarding the use of AES, but also help AES developers prioritise future work in order to further strengthen AES score validity.

The limited number of studies which focused on the AES scoring process to explore these questions have used the following main approaches:

- 1) statistical methods (e.g., Confirmatory Factor Analysis) to examine the internal structure of *e-rater* feature scores in order to ascertain the nature of the scoring model and *e-rater*'s construct coverage (Attali, 2007; Attali & Powers, 2008);

- 2) conceptual investigation comparing the IEA scoring process to the framework of markers' cognitive rating processes (Nichols, 2004);
- 3) linking *e-rater* measurement capabilities to the construct of interest as defined by the cognitive writing process as well as defined by the quality of written products to evaluate construct relevance (Quinlan et al., 2009);
- 4) judging the *e-rater* features against the standard of human annotation (Burstein, Chodorow & Higgins, 2007);
- 5) examining statistical anomalies in how *e-rater* features/micro-features perform on a large corpus of student essays (Quinlan et al., 2009);
- 6) examining the dimensional structure derived from *e-rater* feature scores and analysing uneven profiles of performances across different aspects of writing for second language learners (Lee, Gentile & Kantor, 2008).

The findings from these studies are mixed, with most presenting empirical evidence showing partial construct coverage by AES systems, evidence of AES measuring essay features irrelevant to the construct, and issues relating to the accuracy of the AES scoring at the writing feature level.

3.4 Other Evaluative Approaches

A significant evaluative approach that is not covered by the classifications of Yang et al. (2002) is one that focuses on the extent to which AES scores can be influenced by external factors, such as test-taking strategies, cheating and typing skills. Examining the AES's sensitivity to extraneous factors that are not directly linked to the construct of interest provides discriminant evidence that helps address some of the unique and persistent concerns of AES.

One of these concerns is AES's perceived vulnerability to new types of cheating, test-taking strategies and bad-faith writing. This concern is not entirely unfounded, considering some earlier AES products were heavily reliant on surface and non-linguistic features that were highly coachable (Kaplan et al., 1998; Page 1966, 1968). Evidence that alleviates this concern is considered essential to the use of AES in high-stakes tests. Perhaps one of the most comprehensive studies conducted in this area is Powers et al. (2001). The researchers invited writing experts to compose essays with the sole intention of tricking the machine (*e-rater*) to award scores that were either higher or lower than deserved. In order to help the writers formulate certain test strategies or writing approaches to beat the machine, details of the machine scoring method were explained to the experts beforehand. The study found that it was relatively more difficult to trick the machine to award a lower than deserved score by writing good but unusual papers, than to dupe the machine into assigning a higher score by stressing the linguistic and structural features attended to by the machine. The results

suggested that *e-rater* (v1) was not ready to be used as a sole scorer and it should be paired with human markers in high-stakes assessments (Powers et al., 2001).

Studies of a similar nature include those conducted by McGee (2006) and Jones (2006) whose attempts to fool IEA and IntelliMetric respectively demonstrated that neither of these two AES systems can effectively score order and coherence aspects of writing. Both AES systems did not react appropriately to the doctored samples of writing where sentences were deliberately re-ordered within each essay to reduce the rhetorical effectiveness of the writing. Jones (2006) also found that IntelliMetric could not discriminate between concise and superfluous writing. He tampered succinctly written essays by carefully adding redundancy and superfluity; yet these revised essays received higher than original scores from IntelliMetric.

Other studies (e.g., Higgins, Burstein & Attali, 2006; Rudner et al., 2006) investigated whether AES could successfully flag different types of anomalous writing, such as off-topic essays written to other prompts, essays that were simple repetitions of prompts, essays that comprised multiple repeated texts and essays that were made up of some genuine responses and some repetitions of the prompts. Results of these studies show that various AES systems can effectively identify most, if not all, of these types of bad-faith essays.

Another critical issue concerning the validity of AES relates to the extent that AES scores are influenced by typing skills, since AES is more likely to be implemented in computer-based tests¹⁰ (due to practical and economic considerations) which require students to type essays. It is necessary to establish that the introduction of a new scoring technology does not disadvantage a sub-group of students, in this case, students who cannot type well or quickly. One available study (i.e., Vantage Learning, 2001) demonstrated that only a small portion of variance in students' writing scores as assessed by IntelliMetric, is attributable to variance in students' typing abilities. However, it needs to be noted that results from these types of studies are highly sensitive to the student populations used and situational variables in the studies.

A further factor that is not directly related to the writing ability but could potentially influence the scores generated by the AES systems is essay length. Jones (2006) conducted two experiments to show how length seemed to be a disproportionately large factor in IntelliMetric scoring. In the first experiment, he combined two essays which had each been given a score of 7 by IntelliMetric (on a scale of 0–10) but each had a position that contradicted the other. The combined essay received a score of 10. In the second experiment, he combined two essays which had each received a score of 7 but were written on two completely different topics. The resultant essay received a score of 9. These experiments

¹⁰ Computer-based tests are those tests that are programmed and administered to students on computer. For these tests, students also need to submit their responses on computer.

indicate that essay length, even though not a direct scoring criterion used by IntelliMetric, seems to play a more significant role than the scoring criterion of “focus” in the machine’s judgement of writing quality. The results also indicate that in a real testing situation, IntelliMetric might have difficulty in assessing appropriately when students “have contradicted themselves and when they have gone off the topic” (Jones, 2006, p. 101).

Acknowledging the influence essay length might have on the AES models, some *e-rater* studies (e.g., Attali, 2007; Chodorow & Burstein, 2004) compared differences in the sensitivity of various *e-rater* models to the essay length as evidence of higher validity for some models. For example, Chodorow and Burstein (2004) presented evidence that a later model, e-rater01, had a stronger relationship with scores from human markers than an earlier model (e-rater99), after the influence of essay length was removed from both sets of relationships.

Although a large number of AES studies have been conducted, the quality of these studies, in particular the depth and breadth of the validity questions addressed and the manner in which these questions are addressed, has raised further issues.

Feuer, Towne & Shavelson (2002, as cited in Bennett, 2004) identified a recent emphasis on scientifically based research as a prerequisite for the purchase and use of educational programs and products. Bennett (2004, p. 4), in his review of studies concerning automated scoring, further emphasised the importance of rigour in these studies by stressing “Rigor is particularly important for automated scoring because without scientific credibility, the

chances of general use [of automated scoring] in operational testing programs are significantly diminished”.

Bennett (2004, p. 4), in the same review, went on to note a number of methodological weaknesses which were observed in some studies, ranging from breaching the very basic rudiments of scientific investigations such as “failing to describe the examinee populations”, to more subtle flaws such as “using only a single prompt, which offers little opportunity for generalisation to any universe of tasks”, and to perhaps the “most subtle but pernicious flaw” which is “mistaking machine-human agreement for validation”. He also noted the lack of rigorous scientific investigations into the validity of automatic scores from a measurement perspective emphasising that “automated scoring is first and last about providing valid and credible measurement” (Bennett, 2004, p. 7).

Bennett concluded that there was a need to conduct more rigorous scientific research that could help build a strong argument for the validity of automatic scores, and that the “validity argument must rest on an integrated base of logic and data” (Bennett, 2004, p. 7).

It would seem that the quality of most AES research work conducted to date is compromised by a lack of a structured and systematic approach towards the collection and examination of the validity evidence for AES systems. Most of the available studies have relied on few types of evidence in each study to support their overall validity arguments. This fragmented approach prevents construct validity questions being addressed comprehensively, which in

turn precludes a more convincing argument for the connection between the scoring method and the intended score interpretations.

The necessity for a more systematic and a more robust approach to scientific research of AES validity, and the need for more direct forms of evidence that go beyond the agreement rates between human and machine generated scores, is echoed by other researchers (e.g., Attali, 2007). It is these demands for more comprehensive forms of validity evidence, and for this evidence to be collected and evaluated using a more integrated manner, which are the focus of this study.

3.5 Chapter Summary

The issues raised above demonstrate a number of weaknesses in current approaches being used to assess the validity of scores produced by AES systems. As a consequence, a key problem has been that, with few exceptions, validity in the AES field has generally not been pursued in a robust and rigorous manner.

A significant factor contributing to the above-mentioned problem is the current lack of a comprehensive validation framework that could be used to guide the collection and evaluation of validity evidence relevant to AES systems in a systematic and integrated manner. The next chapter (Chapter Four) attempts to fill this void in the AES research field by proposing a validation framework specific to the AES systems.

Chapter 4 A Validation Framework for the AES Systems

This chapter proposes a practical framework that can be used to guide the collection and examination of validity evidence for scores produced from Automated Essay Scoring (AES) systems. The utility of this framework will be demonstrated in the remaining chapters through applying it to assessing the validity of scores assigned by a particular AES system – the Intelligent Essay Assessor (IEA) for writing tests of the Pearson Test of English (PTE) Academic.

First the concept of validity is considered as it applies to the construct validation processes. This is followed by a discussion of validation which establishes not only the need for an AES specific validation framework but also the shape it must take.

4.1 Concept of Validity

Traditionally, validity has been seen as the degree to which a test measures what it purports to measure (Cureton, 1951; Lado, 1961). Most recent theories view validity as the degree to which empirical evidence and theoretical rationales support the intended interpretations and use of test scores (AERA et al., 1999; Kane, 2006; Messick, 1989, 1995). As a consequence, validity is not a property of the test, but rather of the inferences drawn from, and actions based on, test results. According to Messick (1989), there are two distinct aspects to this concept of validity, both of which have been widely acknowledged by researchers as important to the justifications of score interpretation and its use.

The first aspect is the evidential basis of validity; that is, the reasoning and the empirical evidence that support the interpretations as well as uses of scores given a particular context. A central consideration in score validation is thus the collection of empirical evidence to establish (or challenge) the link between the score and the intended interpretation of the score, as well as evidence to identify and evaluate possible counter interpretations (Clauser, 2000; Cronbach, 1971; Kane et al., 1999; McNamara & Roever, 2006; Messick, 1989). A proposed interpretation has little or no credibility if there are equally plausible rival interpretations (Kane et al., 1999; Messick, 1989).

Whether the appropriateness of score interpretation holds across different testing contexts, or across different test population groups, is a persistent and perennial empirical question. Validity thus is an evolving property and construct validation an ongoing process (Messick, 1996). An implication of this concept for AES research is that, as AES systems continue to evolve at a fast pace, there exists both a great need as well as an immense challenge for researchers to conduct independent and comprehensive studies to address validity issues as they arise from any new developments in the AES technologies.

The second aspect to validity is the consequential basis of validity (Messick, 1989). This aspect relates to the value implications of the score meaning and the intended and unintended consequences resulting from test interpretation and use in both the short and the long term (Messick, 1996). What elements should be included in the consequential aspect of validity, however, and how they may be evaluated, continues to be debated (Brennan, 2006; Kane,

2006; Xi, 2007). The issue relating to what to include and why, in the context of validating AES scores, is taken up in more detail in Section 4.3 below.

Though the broad concept of validity is useful, it does not, in itself, necessarily provide clear guidance for the validation of test score interpretation or use (e.g., where to begin or where to focus validation efforts) (Kane, 2006). For that purpose, a validation framework is needed.

4.2 Concept of Validation

Modern validation frameworks, which all draw on the current interpretations of validity, include those developed by Kane (2002, 2004, 2006) in educational measurement, versions of which are extended in language testing by Bachman (2005) and by Chapelle, Enright and Jamieson (2008). These frameworks are developed to assist in the validation of a whole test, of which scoring is an integral part.

A key notion underlying all of these frameworks is that there is a sequence of inferences involved in the interpretations of performance assessments, leading from observed performances to the conclusions and to the decisions based on performances. Furthermore, the validity of a proposed interpretation of score and its use depends on the plausibility of all the inferences involved in this sequence (Kane, 2006). There must be supporting evidence for the credibility of each one of these inferences in order for the whole sequence to be credible.

The validation process thus starts with the making of an interpretative argument that lays out all the inferences, and the supporting assumptions on which these inferences depend. These inferences and assumptions serve to identify areas where validation efforts should be deployed. As Kane (2006) explicitly noted, “[t]he kinds of validity evidence that are most relevant”, and that need to be collected and evaluated “are those that support the main inferences and assumptions in the interpretative argument, particularly those that are most problematic” (p. 23). Consequently, an interpretative argument that is clearly specified in sufficient detail for a particular test provides a basic framework for test validation.

Table 4.1 presents an example of an interpretative argument that has been specified for a writing test, used for making university-level academic program admission decisions. Such a test is similar in nature to the writing component of the Pearson Test of English (PTE) Academic, details of which will be expanded in the next chapter.

Table 4.1

Interpretative Argument for a Writing Test (c.f. Kane, 2006, p. 24)

I1	Scoring	from an observed performance (a sample of writing performance) to an observed score
	A1:	the scoring rubrics are appropriate
	A2:	the scoring rubrics are applied accurately and consistently
I2	Generalisation	from observed score to universe score (i.e., expected score on the universe of generalisation)
	A1:	the sample of writing performance is representative of the universe of generalisation over writing tasks, occasions and test conditions
	A2:	sampling errors associated with replications of the measurement procedure are small
	A3:	the sample of writing performance is produced under conditions consistent with the measurement procedure
I3	Extrapolation	from universe score to the level of actual writing skill in real-world educational settings
	A1:	the sample of writing performance is related to writing skill in higher education settings
	A2:	there are no skill irrelevant sources of variability that would seriously bias the interpretation of scores as measures of level of actual writing skill
I4	Decision	from conclusion about level of writing skill to decision of admission to a university program
	A1:	overall, positive consequences associated with the decisions outweigh the negative consequences
	A2:	the performance standard (and the cutscore) set for admission decisions reflect the minimum level of writing skill required to study in a university program
	A3:	candidates whose scores fall below the cutscore are unlikely to succeed in the university program

Note: I1, I2, I3, I4 denote inferences; A1, A2, A3 denote assumptions.

Table 4.1 outlines the sequence of the four major inferences involved in the interpretation and use of results from such a writing test. These are: 1) the scoring inference that links a sample of writing performance to an observed score; 2) the generalisation inference that links the observed score to the expected score over relevant parallel versions of the tasks and raters; 3)

the extrapolation inference that links the expected score to writing performance in real-world academic contexts; and 4) the decision inference that links the conclusion about the level of writing skill to a decision made about the test taker based on the test score. Table 4.1 also specifies some of the key assumptions that support each individual inference. For example, with regard to the credibility of the scoring inference drawn from the scoring of a writing sample, two general assumptions have been specified. These are that the scoring rubrics are reasonable and appropriate (A1), and, that they are applied correctly and consistently (A2).

The second step in a validation process involves constructing a validity argument, by evaluating the interpretative argument in a particular context. This entails both the collection and appraisal of empirical evidence and reasoning for each inference and its associated assumptions, as well as the integration of multiple pieces of evidence into a coherent argument that either supports or challenges the proposed interpretation or use of the score (Kane, 2006).

The concept of a validation process as described above not only provides direction as to how a validation process should be conducted, but it also reinforces the notion that in order for a convincing validity argument to be made, all the main inferences and assumptions involved in the interpretation and use of test results must be identified so that they can be adequately examined.

In the context of collecting validity evidence with a view to constructing a convincing validity argument for writing scores produced from AES systems, it thus becomes clear that it is

imperative to have a framework that articulates critical assumptions that are pertinent to AES. These assumptions should not be limited to just those supporting the scoring inference, but should also include those supporting the generalisability, extrapolation and decision inferences, because of the “ripple effect” AES can have on the sequence of inferences that extends beyond the scoring inference (Clauser, Kane & Swanson, 2002, p. 420). Such a framework can provide practical guidance regarding the focus of AES validation efforts, because it identifies key assumptions for which evidence must be collected and examined.

A review of relevant literature indicates that such an AES validation framework does not currently exist. Using Kane’s (2006) argument-based approach to validation, Clauser et al. (2002) presented preliminary discussions of validity issues in a broad context of automated scoring including but not limited to AES. Drawing upon the work of Clauser et al. (2002), Xi, Higgins, Zechner & Williamson (2008) and Enright & Quinlan (2010) applied the argument-based approach to the validation of a speaking practice test and to the evaluation of one AES system respectively. Although the work of these researchers was useful in terms of demonstrating a method of presenting validity arguments for automated scoring, it did not extend to the development of an AES specific framework that listed critical validity assumptions that needed to be examined by AES studies.

The AES validation framework proposed in the next section fills this void. It is developed with the explicit goal of identifying key AES-related assumptions, and the inferences they support, to ensure that all important validity issues underlying the appropriateness of AES scores are addressed in a structured and comprehensive manner.

Within the framework, the key assumptions are phrased as validity questions for investigations, rather than as general statements. This is in contrast to the manner of their presentation in other more general frameworks (e.g., Kane, 2006); and emphasises the key concept that the plausibility of key assumptions supporting the interpretation and use of scores must not be taken for granted; they must be evaluated.

As it is difficult to outline all possible validity questions (i.e., assumptions) due to the multi-faceted nature of a measurement process, this framework represents an initial attempt to make explicit the most critical and important assumptions that need to be addressed to build a case for validity, with the intention that the framework can be further expanded by other researchers. A consequence of this attempt has been the inclusion in the framework of some key validity assumptions (such as those related to measurement and structural aspects of validity) that are critical to the validation of the scores produced by AES systems. These assumptions either have not been made explicit in other relevant studies (e.g., Enright & Quinlan, 2010; Xi et al., 2008), or have been stated but not elaborated upon in reference to AES in the more general test validation frameworks (e.g., Kane, 2006).

An attempt has also been made to specify the validity questions within the framework in such a manner that, while general, they are sufficiently clear that they can be adapted for different tests that use different AES systems. The remainder of this thesis demonstrates how this might be achieved by applying this framework to assess the appropriateness of scores assigned by the IEA for the PTE Academic tests.

In addition to the concepts of validity and validation which form a significant theoretical basis for the development of the proposed AES framework, the general validity criteria proposed by Messick (1996) for the interpretation and use of performance assessments have also been considered in developing the framework. These criteria (i.e., content, substantive, structural, external, generalisability and consequential aspects of validity) have been adapted and integrated, where appropriate, into the AES validation framework. The precise meanings of these aspects of validity will be made clear when the framework is elaborated in the following section.

4.3 A Practical Validation Framework for Automated Essay Scoring (AES) Systems

The proposed framework aims to provide practical guidance for the collection and evaluation of validity evidence to support or challenge the link between the AES scoring method and the intended interpretation and use of the resulting scores. This framework proposes that, in order to build a convincing validity argument for an AES system, validation efforts should focus on at least the following five areas:

- 1) writing traits¹¹ scored by the AES operational model;

¹¹ From this point onwards, the term “writing trait” will be used to describe the dimension of the writing that an AES model analyses in its scoring process. All AES models score a variety of textual features which are then aggregated to broad categories for reporting. These broad categories represent readily recognisable dimensions of writing, such as organisation, mechanics, vocabulary range, and so on. The choice of the term “writing trait”

- 2) scoring procedure used by the AES operational model;
- 3) measurement aspect of score validity;
- 4) structural aspect of score validity; and
- 5) consequential aspect of score validity.

Each of the five areas (referred to hereunder as “components” of the framework) and the relevant validity questions for which evidence must be accumulated and evaluated, will be discussed in more detail below.

Component One – Writing Traits Scored by the AES Operational Model

Writing ability is a latent construct; that is, it is neither directly observable nor directly measurable. Rather it is inferred from the scores given by markers considering the manifestation of student performances produced by the underlying writing ability. The traits of the writing performance that are taken into account by an AES model have a profound impact on the score interpretations. Consequently, evaluating the strength of the link between the writing ability being measured and aspects of performances as scored by the AES model is

is to be consistent with the term used by Pearson to describe dimensions of writing scored by IEA, which will be the AES system of interest for the remaining chapters. Other vendors use different terms.

a critical part of an AES validation process. There are three central issues that need to be considered.

The first is the extent to which the writing traits scored by an AES model are relevant to the writing ability being assessed in a writing test. A major source of invalidity in this regard is construct-irrelevant variance (Messick, 1989) – that is, an automated method may be measuring extraneous traits of students' writing and consequently awarding scores in ways that are irrelevant to the ability intended to be assessed. An example where this threat to validity may occur is when an AES model cannot differentiate typographical errors from genuine spelling errors. In such a situation, scores awarded for the spelling aspect of writing are unduly influenced by an external ability (e.g., ability to type accurately on a computer) which bears no relevance to the spelling competency being assessed.

It is stressed that, in determining the relevance of the writing traits assessed by an AES model to the ability being assessed, the specificity of the testing context must be taken into account. As skills and knowledge that are of interest to test consumers are potentially different for tests with different purposes, the definition of the writing ability that is intended to be measured varies from one testing context to another. Accordingly, the first step in validating the scores produced by an AES system is to clearly define the ability that needs to be measured by the writing test. This thesis takes this approach to validating the IEA-generated scores. In the following two chapters, the PTE Academic testing context is described, and then the characteristics of the writing ability intended to be captured by IEA are examined.

A second issue in the analysis of writing traits scored by an AES model is whether these traits are actually representative of the writing ability domain that needs to be captured by a writing test. There is a need to investigate whether all important parts of the target ability domain are covered by the AES model. In this regard, a major source of invalidity is construct under-representation (Messick, 1989) – that is, automated methods may fail to recognise traits that are relevant to the writing ability of interest. According to Kane (2006), evidence of such invalidity, as well as of the above-mentioned construct-irrelevant source of invalidity, renders less valid the extrapolation inference. This inference links the universe of generalisation (aspects of writing assessed) to the target domain (aspects of writing that are of interest to test users). This is a critical part of the overall validity argument for AES models, as these models are built on the premise that they identify and select traits that are accessible for quantification in the scoring model. It is therefore important to ensure that the AES models are not leaving out any traits that are difficult to measure (or cannot be measured well), but which are salient to good writing, as defined for a particular context.

The two issues above are both content validity criteria and are often referred to in the literature as construct relevance and construct representativeness (Messick, 1996). One method to investigate these criteria is to map the aspects of writing performances scored by an AES model to the ability of interest defined for a specific testing context, and vice versa. How this may be achieved, and the rationale underpinning the techniques used are illustrated and discussed in Chapters Six and Seven.

The final issue in the analysis of writing traits scored by an AES model concerns how accurately and reliably an AES model assesses each of these traits. Four types of evidence can be accumulated. The first is the criterion-related validity evidence. This can be obtained by comparing scores assigned by an automated method with those assigned by expert human markers when measuring the same aspect of performance. The second type of evidence is the external form of validity evidence, which can be acquired by analysing relationships between AES scores and independent measures (or non-assessment behaviour) on the same or similar traits of performance. The third type of evidence is the predictive form of validity evidence, which can be accumulated through analysis of the relationship between AES scores and students' grades/progression in future university writing or academic programs. Within Kane's (2006) test validation framework, the criterion-related evidence tests the strength of the scoring inference. The external and predictive forms of evidence support or challenge the plausibility of the extrapolation inference which links the AES scores to the test-takers' writing performance in a real-world academic environment.

The fourth type of evidence needed for the accuracy and reliability of AES scoring of writing traits relates to the generalisability of the AES (trait) scores. In the context of writing assessments, the focus is usually the extent to which scores can generalise over different facets of a measurement process (e.g., raters, tasks and testing occasions). An obvious advantage of automated scoring is that it removes random errors associated with human marking (e.g., errors arising from the rater effect, or from the interactions a rater has with rating occasions, with tasks, and/or with test takers). This is because an AES system can

consistently apply the same scoring criteria from one rating occasion to another. However, as Clauser et al. (2002) point out, there are unique issues associated with automated scoring that may impact on the generalisability of AES scores. In order to claim the validity of AES scores, there needs to be supporting evidence for the following:

- 1) Generalisability of AES models over somewhat different but parallel model development procedures. For operational use, an AES model is typically trained on scores produced by a sample of human markers on a sample of occasions using a sample of essays. Such a model may not generalise to another model that is developed from using a similar (but not identical) procedure, one which might have used another sample of equally qualified human markers or used another sample of essays to train the model.
- 2) Generalisability of AES scores across tasks. AES models are typically trained to predict human scores over one task or over a number of tasks. By this design, AES may be disproportionately capturing aspects of the human score variance that do not generalise across tasks (Clauser et al., 2002, p. 422).
- 3) Generalisability of AES scores over time and over different test populations. In operational settings, AES models may be continuously reviewed and updated in response to either advancements in the automated scoring technology or to the changes in the characteristics of test populations, as models gain more acceptance and roll out to more areas. These model updates impact on the generalisability and comparability of scores produced by AES models over time. It is noted that, while some efforts have been made to understand the generalisability of AES models over alternate test forms (e.g., Attali, 2007; Bridgeman, Trapani & Williamson, 2011), there has been very little, if any, research done concerning the

generalisability of AES models over different but parallel model development processes, or over time as a result of model updates. These issues should become research foci for future studies because evidence from investigating these issues can reveal to what extent AES models are capturing model-specific, candidate-specific, or task-specific variance that is not relevant to the construct.

In addition to evidence related to the accuracy and reliability of AES scoring of writing traits, when there is a requirement for an AES model to produce an overall score as an indication of the overall quality of an essay, the above-mentioned four types of evidence should also be obtained for the overall scores.

To summarise, the key validity questions that need to be addressed for the writing traits component of the AES validation framework, include:

1. Are the writing traits scored by an AES model representative of the writing construct intended to be assessed?
2. Are the writing traits scored by an AES model relevant to the writing construct intended to be assessed?
3. Can these writing traits be accurately assessed by the AES model?
4. When there is a requirement to produce overall scores, how well do the overall scores align with those assigned by experienced human markers and with independent measures?

5. Can AES (trait and overall) scores obtained under one condition generalise to scores that would be expected to be obtained under different but parallel conditions?

Component Two – Scoring Procedure Used by the Operational AES Model

In operational settings, AES models are typically required to produce a total score as an indication of the overall quality of an essay. The type of scoring procedure used by an AES model to evaluate the overall quality of an essay has impact on the meaning of a score, because different scoring procedures assume different theories of writing quality, a point that will be elaborated upon later. An AES model can use one of the following two types of procedures to derive an overall score: 1) analytic scoring, where an AES model evaluates each essay on various writing traits and then combines the trait scores to an overall score; or 2) holistic scoring where an AES model awards one overall score to an essay (e.g., by comparing it to a set of pre-scored essays) without appraising individual traits. Review of the AES systems in Chapter Two suggests that the main AES models currently available invariably employ an analytic scoring procedure to derive an overall score. A critical question for these models is: How does an AES model select and combine writing traits to produce an overall score in real rating situations?

There are currently a number of variations in the methods used by different AES models to select and combine traits. Some select traits using statistical analyses that aim to best predict human scores (e.g., Page, 1966, 1994). This results in different combinations of traits and different emphases being placed on each of the traits from one scoring occasion to another.

Others have a more fixed set of traits in the scoring model but these traits may or may not vary in their relative importance to the overall score, depending on the testing contexts (Attali & Burstein, 2006).

Some insight into how different types of scoring procedures might impact on the validity of the scores generated can be gained from a review of the research concerning the impact of different rating scales on the validity of scores produced by human markers. This is because most of the discussions about rating scales arise from the tendency of different scales to give markers different levels of flexibility to include certain writing traits that markers feel important, as well as to offer markers different levels of flexibility to adjust the emphasis that certain traits receive in the actual scoring.

It is therefore useful to present a short literature review of rating scales to illuminate the validity and reliability issues that might be applicable to AES scoring procedures. Reliability issues are included in these discussions because reliability is inextricably linked to validity, and is considered a pre-requisite of validity (Weigle, 2002). If a scoring model cannot score the same essay consistently from one occasion to another, test stakeholders and the wider public cannot be expected to have any confidence in the appropriateness and fairness of the decisions and inferences that are made on the scores resulting from the tests.

Another reason for including a discussion of rating scales at this time is that some AES models (such as the IEA model used for the PTE Academic writing prompts) already use a particular rating scale to generate writing scores. The following discussions, relating to the

impact of rating scales on the validity and reliability of scores, are then explicitly relevant to the validation of scores produced by these AES models.

This thesis discusses two main types of rating scales used in the human marking process – analytic rating scales and holistic rating scales. A holistic scale is defined as one which “uses a single global numerical marking to rate a composition, while an analytic rating scale uses several subscales, which may or may not be summed or averaged to form a composite total, to rate characteristics of a composition separately” (Carr, 2000, p. 209). Under these definitions, the scoring procedure adopted in a human marking process involving the use of a holistic scale corresponds to an AES holistic scoring procedure, whereas the scoring procedure involving the use of an analytic scale mirrors that of an AES analytic scoring procedure.

A major point in the discussion about rating scales is the notion that a rating scale represents, explicitly or implicitly, the theoretical basis and assumptions upon which the test is founded (Goulden, 1992; McNamara, 1996). For example, as Goulden (1992, as cited in Barkaoui, 2007) explains, analytic scales which sum the sub-scale (trait) scores to derive an overall score assume that the sum of the parts is “exactly equal to a valid score for the whole and, by evaluating the parts, the marker has evaluated the whole” (Goulden, 1992, p. 265). On the other hand, a holistic scale assumes that “the whole is not equal to the sum of the parts”, rather, “the whole is equal to the parts *and* their relationships”, and hence performance should be assessed as a whole entity holistically (Goulden, 1992, p. 265). These assumptions, which also underpin analytic and holistic scoring procedures used in AES scoring processes, imply different theories about writing quality.

Scales consisting of different components and different weightings for the components also represent different theories about writing ability. For example, the ESL Composition Profile analytic rating scale developed by Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey (1981) (Appendix C) represents the scale developers' view of what constitutes effective written communication. This view is predicated upon there being five important dimensions of written prose (as measured through five traits) that are salient to the effectiveness of written communication, namely content, organisation, language use, vocabulary and mechanics. These dimensions have different levels of impact on the overall communicative effectiveness, with content being the most significant factor, and mechanics the least. Therefore, validating the appropriateness of a rating scale against the specific testing context includes developing logical arguments that support or challenge these theories and assumptions that are embodied in the scale. Once the appropriateness of the scale is validated, empirical evidence should be accrued to demonstrate the degree to which the scale can be consistently and accurately applied across different testing populations and contexts. This set of procedures should form a part of the overall process for investigating the validity of scores produced by those AES models that already use a particular type of a rating scale in the scoring processes.

A number of studies have considered issues concerning the fairness, appropriateness and consistency of human-generated scores when different types of scales are used (Carr, 2000; Goulden, 1994; Hamp-Lyons, 1991, 1995; Hamp-Lyons & Kroll, 1997; Perkins, 1983). The main validity and reliability issues identified in the literature stem from the degree of freedom a rating scale allows for the inclusion, exclusion, or emphasis of certain writing traits in the

actual scoring process. For example, analytic scoring limits the traits to just those on the scale and it controls the level of the importance of each trait to the overall performance when an overall score is required. In terms of reliability, this reduces personal choice and the level of subjectiveness in the marking process and leads to greater overall consistency in scoring (e.g., Brown & Bailey, 1984; Goulden, 1994; Hamp-Lyons, 1991; Veal & Hudson, 1983).

However, in terms of validity, an analytic rating scale may force markers to ignore important or relevant qualities that may affect the overall quality of performance (Barkaoui, 2007) or force the markers to isolate textual features from context (Perkins, 1983). Although there are validation studies (Cumming, 1990; Cumming & Mellow, 1996; Hamp-Lyons, 1991, 1995; Weigle & Lynch 1996) that lend support to the argument that more than a single score is needed to adequately describe students' performances, other studies have demonstrated problems associated with markers distinguishing between multiple subscales (e.g., Hamp-Lyon & Henning, 1991). Another important consideration regarding analytic scales is how component scores are aggregated, either with equal or differential weightings, to form a single score. Different weighting schemes can change the writing test results and affect the meaning of the overall scores produced and the types of information a test provides.

In contrast, markers using holistic rating scales can include additional traits they feel are important and use personal judgments to control the level of importance of each trait to the overall score. This leads some proponents (e.g., White, 1984) to claim a holistic rating scale is more valid than an analytic scale as it more closely reflects "the authentic, personal reaction of a reader to a text" (White, 1984, p. 409). However, other researchers have questioned the

validity of scores produced from holistic rating scales on various grounds. Goulden (1994) contends that the use of holistic scales tends to result in “an idiosyncratic set of supplemental traits different from those written in the basic guide” (p. 74). This leads to a persistent question of what it is that a holistic score is measuring and whether such a score represents a single construct (Carr, 2000). Although there are a few validation studies (Homburg, 1984; Huot, 1990a; Vacc, 1989) that present some evidence of holistic scales measuring certain traits of writing, findings are not consistent in terms of the traits identified (Carr, 2000). Markers using holistic scales seem to struggle to agree on the specific writing traits that make one essay superior to another (Hamp-Lyons, 1990), or to agree on the relative importance of each trait’s contribution to the overall score (Breland & Jones, 1982). These problems contribute to the continual problem in holistic scoring – less than desirable inter-marker reliability (e.g., see Cooper, 1984, for a review of writing assessment).

These discussions are relevant to those AES operational models whose scoring procedures have the potential to alter the combination of writing traits selected for scoring and the contribution each trait makes towards the overall score, from one scoring scenario to another. The implications of such scoring procedures for the validity of the scores produced are then similar to those arising from the use of holistic scales—that is, the meaning of the scores produced is potentially different across scoring scenarios, resulting in non-comparable scores from one scoring scenario to another.

It is clear from the above review that the choice of a holistic or an analytic scoring procedure has different implications regarding validity and reliability. The appraisal of the

appropriateness of the type of scoring procedure is usually complicated and needs to consider a variety of factors including construct validity, score generalisability, practicality, score utility, purpose of testing, and authenticity (Clauser, 2000; Weigle, 2002). This is true for procedures used by both AES and human markers.

In order to investigate the validity implications associated with the AES scoring procedures, the following questions should be explored to help clarify the AES writing construct:

1. What is the rationale for the procedure used by an AES model to select and to combine traits to produce a single score? What are the validity and reliability implications of such a procedure?
2. What are the assumptions and theories of the writing construct that are embodied in the rating scales internalised by AES models? Are they appropriate for the testing contexts?

Within Kane's (2006) framework, evidence from this component provides either support for, or rebuttal of, the credibility of the extrapolation inference that links the AES scores to the actual performance of test takers in a real-world academic environment. This is because the validity questions listed above for this component help clarify the connection between the AES scoring processes and the intended interpretation and use of scores produced from these processes.

Component Three – Measurement Aspect of Score Validity

The implicit requirement of using a single score to summarise a student's writing performance is the need for the score to be represented on a single or uni-dimensional scale. Only when this requirement is met, can the scores of two or more students be compared on the measurement scale in a meaningful way. Where an AES model needs to summarise scores of various traits in a single score as a requirement of the scoring, the validation of AES scores must therefore encompass the examination of the extent to which writing traits exhibit the empirical consistency that is expected of them as expressions of a single underlying construct.

Furthermore, where an AES model uses rating scales to score various writing traits, there is also the requirement that these rating scales are functioning as expected. For example, an expectation of a functioning rating scale is that a higher score category on the scale indicates a higher underlying ability, and vice versa. AES scores must meet these essential requirements before they can be considered to be useful measurements for the purposes of comparisons of a single ability.

Although examinations of these requirements are now routinely carried out for human scores through psychometric analyses, they have not become a norm in the AES validation process. In fact, most studies simply assume that AES scores possess the necessary measurement properties rather than rigorously examining whether empirical data supports such an assumption. In those cases, lacking the crucial measurement aspect of evidence significantly weakens the validity argument constructed for the AES systems under investigation. The

concept of these issues and empirical methods of examining whether data meets these measurement requirements will be considered further in Chapter Nine.

For present discussions, evidence should be collected in relation to the following two questions when investigating whether AES scores meet the essential measurement requirements:

1. Is there empirical evidence of writing traits scored by an AES system measuring a single ability construct?
2. Are the rating scales used to score the individual traits functioning as intended?

Component Four – Structural Aspect of Score Validity

The structural aspect of validity is a validity criterion proposed by Messick (1996) for performance-based assessments. As pointed out by Messick (1996), the “theory of construct domain should guide the rational development of construct-based scoring criteria and rubrics”, and in return, “the internal structure of the assessment should be consistent with what is known about the internal structure of the construct domain” (Messick, 1996, p. 10). Loevinger (1957) refers to this property of rational scoring models as structural fidelity.

Validation studies for AES can accumulate evidence by comparing the internal structural patterns derived from the trait scores assigned by the AES model to the expected interrelations among the different traits of writing performance, either deduced directly from

domain theory or observed from empirical investigations into the human scoring processes. This type of evidence should provide an indication as to whether an AES system has been developed in a rational manner and whether the system is measuring the right achievement construct.

In summary, evidence collected from the following questions should provide either support for, or rebuttals of, the claim of validity made for an AES system:

1. Does the internal structural pattern in the AES scores confirm the theoretical distinctions about the construct?
2. Is the internal structural pattern in the AES scores consistent with that in the scores from human experts?
3. Is the internal structure in the AES scores consistent with a theoretical view of writing as a number of inter-correlated yet conceptually distinct dimensions?

Both the structural and measurement components of the framework examine the empirical inter-trait relationships among AES scores against those expected from domain theories. Evidence from these two components therefore provides essential backing for the claim that AES scores capture aspects of writing performance that reflect the underlying writing abilities. As such, of the main inferences that support the intended interpretation and use of test scores, these two components contribute to the credibility of the extrapolation inference, which links observed scores to aspects of test-takers' actual performance on relevant writing tasks in a real world.

Component Five – Consequential Aspect of Score Validity

Although consequences have always been an integral part of the validity concept, the range of consequences that need be considered in a test validation procedure has evolved over time (Cronbach & Gleser, 1965; Guion, 1974; Messick, 1975). In contrast to the traditional focus on immediate positive and negative consequences arising from test use (such as direct benefits and costs), the contemporary view of the consequential aspect of validity extends to the consideration of social consequences and adverse impact on individuals and groups in the evaluation of the legitimacy and validity of test use (Cronbach, 1988; Kane, 2006; Messick, 1989, 1995, 1996). Although the types of social consequences that should be included in the validity concept are still subject to debate, the current generally accepted view is that all negative consequences that can be directly traced to a source of invalidity in the measurement procedure (such as construct under-representation or construct irrelevance) count against validity and therefore must be evaluated as a part of the validity argument (current edition of *Standards for Educational and Psychological Testing*, AERA et al., 1999).

This current view of validity dictates that an important part of the validation for AES systems is the evaluation of any negative consequences for individuals and groups directly attributable to bias in the scoring processes used by these systems and/or to the deficiencies in measurement capabilities of these systems. This type of evaluation is especially important when AES models are not yet robust in analysing certain aspects of writing, particularly those reflecting high-level reasoning such as the writer's ability to construct a logical argument in a concise and coherent manner (see Jones, 2006; Matthews, 2004; McGee, 2006). When these

aspects of writing are part of the writing construct being measured, and when the AES models are used in high-stakes tests where decisions (based on scores produced by AES systems) can have a significant impact on the rights and life chances of individuals, the negative consequences arising from AES measurement incapability must therefore be evaluated. The need for this type of validation forms the basis of most of the discussions and investigations around the individual trait scoring by IEA in Chapter Eleven, where issues associated with IEA scoring that might result in particular testing cohorts being disadvantaged, are pursued.

Another implication of the current broader concept of validity, which goes beyond immediate benefits and costs, is the necessity to accumulate evidence associated with the long- and short-term impact of AES uses, particularly in high-stakes tests, on instruction and learning. It is long established that assessment practices can have a profound impact on study behaviour, teaching procedures and indeed the curriculum itself (Crooks, 1988; Fredricksen & Collins, 1989; Kane et al., 1999; Odell, 1981). As automated essay scoring represents a significant deviation from the traditional method of assessing writing, it is critical that evidence concerning the impact on teaching and learning arising from AES use is collected and evaluated, so that the validity arguments concerning these systems can be made more convincingly to key stakeholders.

In determining the specific kinds of impact or consequences requiring validation efforts, priority should be given to those that have already been identified as significant by key stakeholders (Kane, 2006). In this regard, one potentially significant negative consequence identified by the teaching profession is that the use of AES in classrooms and standardised

tests might discount the complexity of written communication, which in turn is likely to impoverish students' understandings of writing (Cheville, 2004). For example, AES might give students an impression that good writing is just correct writing. In addition, the use of AES is also perceived to encourage "formulaic and highly standardised writing" from students (Rothermel, 2006, p. 209). This may result in students being less prepared to respond to different and complex rhetorical situations in real world settings (Broad, 2006; Drechsel, 1999).

These concerns are not unfounded, since some evidence already exists regarding the impact of the use of AES in high-stakes tests on the way students are taught to write. For example, a website tutoring students for the Analytical Writing Assessment (AWA) within the Graduate Management Admission Test (GMAT) recommends to students that they be "conformist" because the AES system used to mark the AWA tasks "is not programmed to appreciate individuality, humour, or poetic inspiration". Students are also taught to follow certain organisational structures in their writings and use certain transitional phases to "help the computer identify concepts between and within the paragraphs" (*How to tackle the Analytic Writing Assessment?* n.d.).

When students are taught to write to computer algorithms as a result of the use of AES, some significant questions need to be addressed as part of the validity arguments for the AES systems. These questions include: 1) What is the effect of AES technologies on the social and cognitive aspects of writing? 2) Would students change their writing process, if they know a machine is evaluating their essays? For example, would they place more emphasis during

writing on the traits they believe the machine values such as mechanical accuracy, and put less effort into aspects of writing that they think the machine cannot reward appropriately such as creativity, logic of argument or unique ideas? 3) Would the use of AES as an instructional tool result in students having an inferior educational experience, such as losing the opportunity to engage in constructive and meaningful dialogues with those teachers who take appropriate stances and read essays critically? Though a limited number of studies (e.g., Herrington & Moran, 2006) have started probing some of these issues, more work is required to adequately evaluate the consequences of the use of AES on students' understandings of writing as a complex meaning-making rhetorical activity, on their learning focus, and on the way students plan, compose and revise.

Another potential consequence considered to be significant by professional educators is that using AES to mark essays might increase the separation of assessment from teaching, which could reduce a positive “wash back” effect from essay evaluation activities to the curriculum and to instruction (Broad, 2006; Herrington & Moran, 2006). These educators argue that having teachers evaluating students' writing is an important part of effective teaching. A close link between teaching and assessment needs to be maintained in order to realise the pedagogical and educational benefits of assessment.

On the other hand, AES systems can have positive effects on teaching and learning and it is equally important to measure these effects. A limited number of studies have examined the classroom use of the instruction-based AES systems. Results are mixed, with only some studies finding recognisable benefits of instruction-based AES systems, such as easier

classroom management for teachers and increased motivation to write and revise for students (e.g., Grimes, 2008; Grimes & Warschauer, 2008, 2010; Warschauer & Grimes, 2008). These benefits are of course in addition to the widely recognised immediate benefits of AES such as consistent implementation and enforcing of scoring standards, reductions in cost and in the time required for large-scale marking operations.

Overall, due to the complex connections between the assessment method and subsequent changes in instructional practices and/or in students' learning behaviour, more studies are needed to examine the potential impact that AES scoring systems can have on curricula, instruction and study focus (Weigle, 2002). Broadly speaking, some key questions that need to be explored for the consequential component of validity include:

1. Would the introduction of AES as a new scoring method disadvantage or advantage certain groups of students?
2. Would the use of automated scoring systems lead students to change their study behaviour (e.g., only focus on improving those text features that are assessed or can be assessed by the machine)?
3. How does the AES shape students' writing processes and products?
4. What is the impact of AES systems on instruction and writing curricula?

Summarising the AES validation framework and issues around its application

This thesis argues that, in order to construct a comprehensive network of validity evidence to either support or challenge the intended interpretation and use of AES scores, validation efforts should focus on five areas (i.e., the five components of the proposed validation framework). These five areas are: 1) traits of writing performances scored by an AES model; 2) the type of scoring procedures used to produce an overall score; 3) the structural properties of scores; 4) the measurement properties of scores; and 5) the consequential aspect of score validity. Evidence collected from these five areas helps build a coherent and comprehensive argument for or against the appropriateness and fairness of the scores produced by AES scoring processes.

The above sections presented rationales for, and details of, these five components of the proposed AES validation framework. Figure 4.1 now assembles these components and the associated key validity questions into a visual overview of the framework.

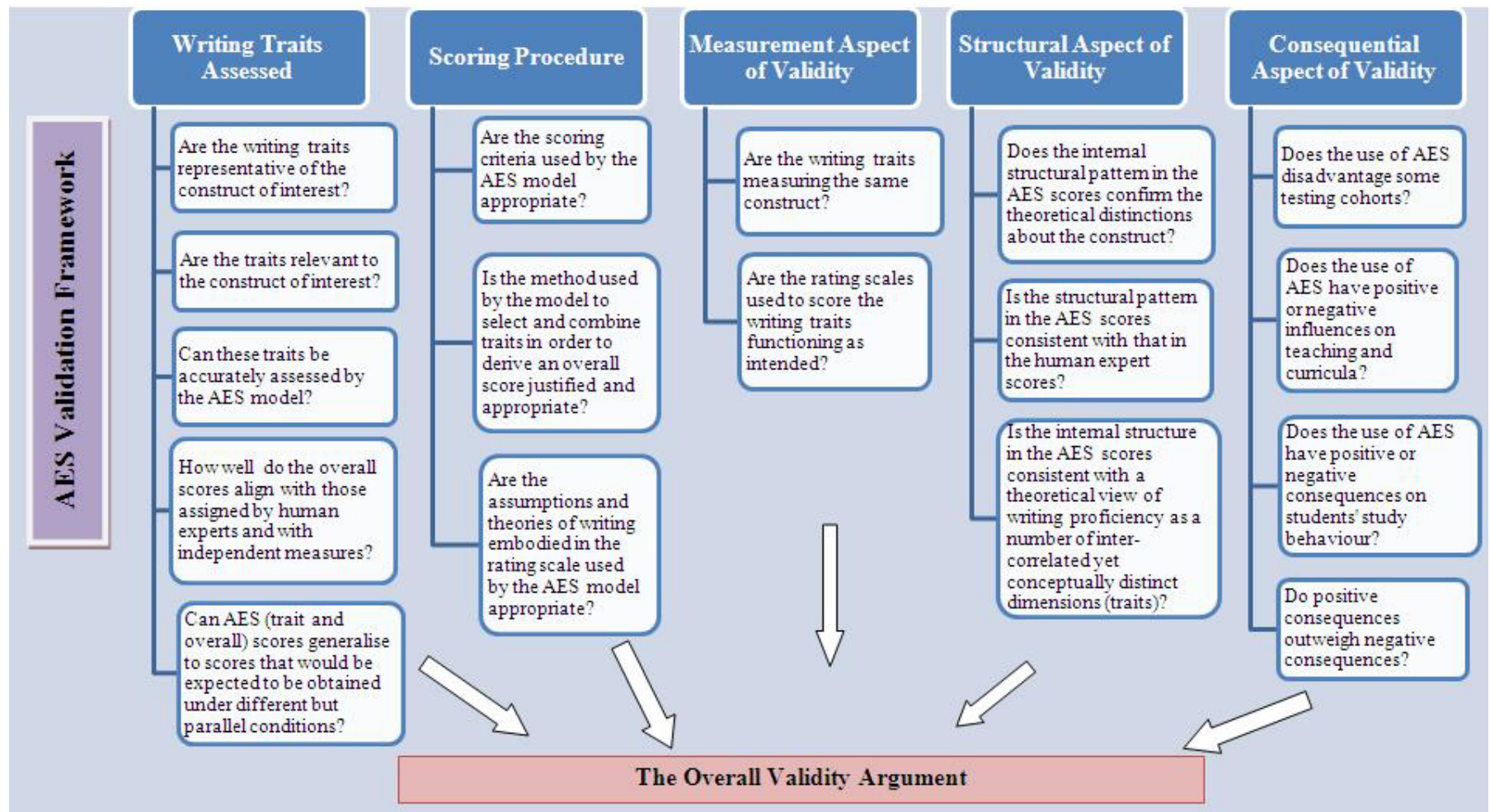


Figure 4.1 The Proposed AES Validation Framework and Its Components

Evidence collected through the application of the proposed AES framework needs to be evaluated as a whole in order to construct a convincing validity argument. This is consistent with the contemporary view of validity as a “unified concept”. Taking this approach, the different forms of evidence pertinent to the five components should be treated as “interdependent and complementary” forms of validity evidence, which must be interpreted together rather than viewed as “substitutable” or “separate” validity types (Messick, 1996, p. 15).

There are a number of possible methods that could be used to evaluate evidence in an integrated fashion. One approach is to assess the combined impact relevant evidence has on the plausibility of the main inferences supporting the intended interpretation and use of test scores. As stated previously, to claim the validity of test results, there must be supporting evidence for the credibility of each of these inferences. A validity argument for AES can therefore be built by first assessing the strength of each inference using evidence collected according to the AES framework, and then evaluating the combined impact the bodies of evidence have on the credibility of the whole sequence of inferences – that is, the combined effects evidence for an AES system has on the meaning of the score and the implications of score use.

In the process of evaluating the strength of the validity argument for an AES system, there are no hard-and-fast rules as to which evidence should be accorded pre-eminence. The emphases one places on different types of evidence will depend on the intended use of the tests results, and on the nature of the tests (e.g., whether or not the tests are high-stakes). For example,

when considering score validity for a low-stakes test where AES scores on certain writing traits will be used by teachers to improve classroom instruction in the short-term, the consequential form of validity evidence might be given relatively less importance than other forms of evidence such as the accuracy of the AES scoring of the designated writing traits. However, if it is a high-stakes test where test results can have a significant effect on individuals or on overall pedagogy, the consequential form of validity evidence should be given relatively more importance in the making of an overall validity argument.

It is noted that, although the framework is developed to guide the collection and examination of validity evidence concerning AES systems, the concepts underpinning the framework are generalisable to validation of any scoring systems, including those that are developed to automatically measure speech, pronunciation and neuropsychological characteristics.

4.4 Chapter Summary

A practical framework for investigating the validity of AES-generated scores has been developed in this chapter, in an effort to promote a more structured, coherent and integrated approach to collecting and evaluating validity evidence for AES systems.

The remainder of this thesis uses this framework as a guide to examine the validity of scores generated by a particular AES system – the Intelligent Essay Assessor (IEA) – in the context of the Pearson Test of English (PTE) Academic. Although it is difficult to investigate all components of the framework comprehensively in one study, the aim of this study is to

demonstrate the merits of the proposed validation framework through applying it to a main AES system. In particular, this study intends to demonstrate how a range of different types of evidence relevant to different aspects of validity can be collected and evaluated.

The next three chapters (Chapters Five, Six and Seven) investigate the first and the second components of the proposed framework for the IEA. Chapter Eight investigates the overall correspondence rates between human markers and the IEA scores, which is related to the first component of the proposed framework. Chapters Nine and Ten inspect the third and fourth components of the framework respectively (i.e., the measurement and structural properties of the IEA-generated scores). Chapter Eleven provides further empirical evidence and theoretical rationales regarding the appropriateness and the accuracy of the IEA scoring at the trait level. The analyses performed in Chapter Eleven relate to the first component of the proposed framework.

Although the consequential component of the framework is not a research focus for this study, implications and consequences arising from the possible deficiencies in IEA are discussed whenever these issues emerge from the analysis.

The next chapter describes the PTE Academic and PTE Academic writing tests, for which the IEA scores are generated. Discussions therein facilitate the examination of the boundaries and the structure of the construct domain that is of interest to the users of the PTE Academic writing test results. This is followed by an account of the data collection and sampling procedures used in this study.

Chapter 5 Pearson Test of English (PTE) Academic

Writing Tests and Data Collections Procedures

5.1 PTE Academic and Testing Context

Pearson Test of English (PTE) Academic is a relatively new international computer-based academic English language test developed by Pearson Technologies (hereunder referred to as “Pearson”). It was launched world-wide in October 2009 and is designed to measure the academic language competency of international students who wish to study academic programs where English is the principal language of instruction (Pearson, 2011b, p. 42). As a result, a common use of the PTE Academic test results is to determine whether a student applying for admission to a university/college program has the requisite academic proficiency in English. At the time of writing, the test is “recognised by over 80% of universities and colleges in the UK” and “by over 150 institutions in Australia”. It has also been accepted by the UK Border Agency (UKBA) and the Australian Department of Immigration and Citizenship (DIAC) for student visa applications (*PTE Academic Australia*, n.d; *PTE Academic UK*, n.d).

Four communicative skills (Listening, Reading, Speaking and Writing) are assessed in PTE Academic. The writing skill is measured through test-takers’ performance on two integrated tasks—one reading-and-writing and one listening-and-writing—and on one independent writing task. The reading-and-writing task presents test takers with a text and asks them to summarise the content of the text in one sentence. The listening-and-writing task asks test

takers to listen to a short lecture and then write a brief summary to summarise the key points in the lecture. The independent writing task presents test takers with a prompt and requires them to write an argumentative essay of 200 to 300 words in response within 20 minutes (Pearson, 2011b). These three types of tasks reflect the kinds of writing that students often encounter in academic contexts.

For each essay written to an independent writing task, PTE Academic produces scores on a number of writing traits, as well as a total score. The total score contributes to the overall PTE Academic test score for the individual test taker (Pearson, 2011b, p. 7). The focus of this thesis is the validity of the total score produced by the Intelligent Essay Assessor (IEA) for an independent writing task. The choice of independent, rather than integrated writing tasks in this study is because these tasks represent the most typical tasks in a writing test – write an essay to an impromptu topic under timed conditions. Compared to integrated writing tasks, independent writing tasks are also more relevant to the writing ability construct as test-takers' performances on these tasks are much less influenced by other abilities, such as reading comprehension. For simplicity, these independent writing tasks are referred to in the remainder of this thesis as PTE Academic writing prompts.

From the above descriptions of PTE Academic, the construct of interest in this study (i.e., the writing ability that is purported to be measured by the IEA for PTE Academic writing prompts), can more precisely be defined as the test-taker's ability to write to achieve a communicative goal in an academic environment. This definition will be further expanded in the next chapter to make clear the skills, knowledge and competencies that are part of the

construct and that are meant to be assessed by IEA. It is also noted that the decisions made on the PTE Academic test results are most likely to be high-stakes decisions, since the test results are typically used for university/college admission determinations. These decisions are usually difficult to reverse and any errors in the decisions made are difficult to correct. Discussions in the remainder of this thesis about the validity of the total scores produced by IEA will be made in these contexts.

The next section describes how the sample data (essays, human scores, writing prompts) were acquired for this study and the characteristics of this data.

5.2 Data Collection Procedures

5.2.1 Data Provided by Pearson

Essays used in this study were sourced from Pearson in 2009. They were part of the data Pearson acquired from its field tests administered between 2007 and 2008. These tests were conducted to test the new PTE Academic instrument, as well as to calibrate items and to train and validate the IEA model. According to Pearson, the test takers recruited for the field tests (more than 10,000 in total) were students who “had a similar level of language proficiency to that of the prospective PTE Academic test takers” (Pearson, 2011a, p. 4). Table 5.1 shows the directive provided to the test takers for the writing prompts at the field tests under simulated testing conditions.

Table 5.1

Directive Given to Test Takers at the Field Tests 2007–2008

DIRECTIVE

You will have 20 minutes to plan, write and revise an essay about the topic below. Your response will be judged on how well you develop a position, organise your ideas, present supporting details, and control the elements of standard written English. You should write 200–300 words.

At the request of the researcher, Pearson randomly selected two writing prompts from the field tests and provided all the essays that were acquired through the administration of the two prompts. Table 5.2 describes the two sample prompts selected – the Voting and the Tobacco prompts.

Table 5.2

Descriptions of the Two Sample Prompts

Prompt ‘Voting’

In some countries around the world, voting is compulsory. Do you agree with the notion of compulsory voting? If voting is compulsory in a democratic society, what are some conclusions we can draw about the nature of democracy?

Prompt ‘Tobacco’

Tobacco, mainly in the form of cigarettes, is one of the most widely-used drugs in the world. Over a billion adults legally smoke tobacco every day. The long term health costs are high – for smokers themselves, and for the wider community in terms of health care costs and lost productivity.

Do governments have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke, or are such decisions up to the individual?

Source: provided by Pearson May 2009.

During the field tests, each essay was scored by the Intelligent Essay Assessor (IEA) on seven analytic traits. The same essay was also double-marked by human markers on five of these seven IEA traits, using the IEA scoring criteria (Appendix A).¹² The only two IEA traits that were not marked by human markers were the *Spelling* trait and the *Formal Requirement* trait, which was mainly a length criterion. This was because the IEA was considered to be able to accurately and objectively score these two traits, therefore required no data from human markers, for training or validation purposes (Pearson, 2009). In addition, each essay was also double-marked by markers using the holistic Common European Framework (CEF) rating scale (Appendix B). For both analytic and holistic scoring, where two scores assigned by human markers were different by more than one score point, a third marking was acquired.

Altogether Pearson provided 216 essays written to the Voting prompt and 223 written to the Tobacco prompt. Data was complete for 391 out of the 439 essays received. For each of the 391 essays, the following data was included:

- scores assigned by the Intelligent Essay Assessor (IEA) on the seven analytic traits using the IEA scoring criteria (Appendix A);
- two human scores (or three if an adjudication was necessary) for each of the IEA traits, except for the *Spelling* and *Formal Requirement* traits, assigned by markers using the same IEA scoring criteria;

¹² Details of the traits assessed by the IEA are provided in Chapter Seven.

- two holistic scores (or three if an adjudication was necessary), assigned by markers using the holistic Common European Framework (CEF) rating scale (Appendix B).

5.2.2 Process of Acquiring Human Scores for this Study – Rating Scales and Markers

In order to examine the IEA-generated total scores and the structural and measurement properties of the IEA trait scores, a set of credible human scores were required for use as criterion measures. To achieve this, this study used a separate human marking process to acquire an independent set of human scores, rather than using the human scores from the field tests as provided by Pearson. There were three considerations for this study design. The first was that there were no human scores for two traits – *Formal Requirement* and *Spelling* – in the data package, as markers from the field tests did not mark these two traits. This made it impossible to use the human trait scores provided to derive total scores, for comparisons with the corresponding total scores produced by the IEA¹³. A second consideration was that the quality of the holistic scores from the Pearson field tests, specifically in terms of score generalisability, was found to be less than desirable.¹⁴ A third consideration was that human

¹³ The IEA derives a total score by summarising the trait scores over the seven traits IEA scored. If the same scoring procedure were used to generate a total score for the human scoring method, scores for all seven traits from human markers would be required. Although one might assume that the IEA scoring of *Spelling* and *Formal Requirement* should be accurate, therefore IEA scores for these two traits could be used as human scores in the calculation of a total score, evidence to the contrary is presented in Chapters Nine and Eleven. Specifically, there are issues around the accuracy of the IEA scoring of spelling.

¹⁴ Evidence of the generalisability of holistic scores from Pearson field tests is presented and discussed in Chapter Eight.

scores provided by Pearson for the five traits assessed were produced by markers using the IEA scoring criteria. As argued in the previous chapter, the appropriateness of the IEA scoring procedure, including that of the scoring criteria used by the IEA to score traits, on its own, is a factor that could impact on the validity of the scores generated. Therefore, to fully appraise the validity implications of using IEA to score essays, including those implications that might arise from the IEA scoring procedure, it was necessary for this study to simulate an alternate human marking process, which would otherwise have taken place, had IEA not been chosen as the scorer. This alternate process needed to occur under the following conditions that were deemed to be appropriate for the PTE Academic test context: 1) using a double-marking process that was typical for high-stakes marking; 2) using a rating scale that had been adequately validated for use in a context similar to that of the PTE Academic; 3) using markers who had suitable assessment and teaching experience to mark the sample essays, with training provided to ensure consistent applications of the rating scale(s) prior to marking. Scores generated from this marking process under these conditions could be regarded as credible alternate human measures which could then be used to validate the IEA scores. The empirical evidence to further support the quality of human measures generated from such a process is collected and examined in Chapters Eight to Ten.

The first step to acquire independent human measures under this study design was to select appropriate rating scale(s) for use in the marking process. Since both types of scales (i.e., analytic and holistic scales) were being used in test contexts similar to that of the PTE Academic, one each that were suitable for use by the PTE Academic writing tests, were first

identified. The use of both types of scales in this study allowed for comparisons of IEA total scores with those scores from a typical human analytic marking process as well as with those from a typical human holistic marking process.

The ESL Composition Profile (Jacobs et al., 1981, hereunder referred to as the *Profile*, Appendix C) was the analytic rating scale chosen. The *Profile* is generally regarded as one of the best known and most widely used analytic scales for second language writing assessment (Lee et al., 2008; Weigle, 2002). It has been adopted by numerous college-level writing programs for testing and placing international students into North American Universities and English Language Institutes. Supporting evidence for the appropriateness of this use of the scale is provided in Jacobs et al. (1981, pp. 74–79). A small modification was made to the scale prior to its use in this study. The original version of the scale required markers to assign a much refined numerical score for each analytic trait. The modified scale retained the four mastery levels within each trait but only required markers to assign a fixed score corresponding to a mastery level, such as a score of 0, 1, 2 or 3. This equated to the use of a 0–3 sub-scale for all traits (Appendix D). This level of discrimination was deemed to be sufficient for the purpose of this study.

The Independent Writing Scale (Appendix E) used to assess the Test of English as a Foreign Language (TOEFL) iBT independent writing tasks was the holistic rating scale chosen for this study. Launched in 2005, the iBT (Internet-based test) is the latest version of TOEFL and is currently used for university/college admissions purposes in more than 130 countries (Educational Testing Service, 2011). The scale was considered appropriate for use in this

study because the writing ability that was intended to be captured by a TOEFL iBT independent writing task was very similar to that intended to be measured through a PTE Academic independent writing task. Both tests aim to measure students' writing ability in an academic setting (Educational Testing Service 2006; Pearson, 2011b). There have been numerous studies providing supporting validity evidence for the appropriateness of the use of this scale in the context of assessing test-takers' academic writing ability (e.g., Chapelle, Enright & Jamieson, 2008; Cumming, Kantor, Baba, Eouanzoui & Erdosy & James, 2006; Educational Testing Service, 2007).

The next step to acquire human measures involved selecting suitable markers. Before markers were selected, ethics approval for this research project was sought from the Human Research Ethics Committee (HREC) of the University of Wollongong (Ethics Number: HE09/130). Approval to conduct the research was granted by HREC on 19 May 2009 (see approval letter from HREC at Appendix R). The Multicultural Programs Unit (MPU) of the New South Wales Department of Education and Communities (NSW DEC) provided assistance to the marker selection process. A senior official of the unit first dispatched an information package concerning this project, including a Participant Information Sheet (Appendix S), to a pool of experienced English as Second Language (ESL) teachers/consultants the unit employed. The official then helped select markers for this study from those who expressed interest. The selection criteria for markers required them to have experience in teaching Stage 6 English

subjects¹⁵ to senior high school students and experience in marking short essay questions in English papers in high-stakes assessment situations such as the Higher School Certificate (HSC) English examinations.

Markers who met the criteria for this study were required to have considerable prior knowledge in applying different types of rating scales to score essays of the persuasive genre, the genre of writing tested in the PTE Academic. For instance, as ESL teachers, these markers were often required to use analytical rating scales to mark essays and to provide diagnostic feedback on various writing aspects to students. As HSC markers, on the other hand, they were rigorously trained to mark consistently using a holistic rating scale. This study required human markers to score essays using both types of rating scales. Markers who met the selection criteria were considered to have a higher likelihood, than others, of producing consistent human scores with minimal training.

Ultimately, five experienced markers were selected. They were all female.

Based on the questionnaires (Appendix G) completed by the markers before the marking session, three had Bachelor Degrees in Arts and two had a Masters Degree in Arts. All acquired a Diploma in Education, which is an essential credential for teaching in NSW

¹⁵ Senior high school students in NSW Australia must study and sit the Higher School Certificate examination (i.e., university entrance examination) for at least one of the Stage 6 English subjects in order to gain admission to universities.

schools. On average, each marker had more than 10 years of experience in teaching Stage 6 (matriculation level) English subjects; and four years' experience in marking HSC English examinations. Four markers indicated on the questionnaires that they had not heard about the AES systems before. The one marker who indicated an awareness of these systems had her doubts that these systems would be suitable for scoring essays written by ESL students. She believed "the machine would not be able to accurately mark higher order traits such as content, coherence and organisation". She considered this to be not helpful for those ESL students who were still developing their English language skills, and who had different levels of proficiency across different dimensions of writing. These students would require accurate feedback on the analytic writing traits to improve their writing.

5.2.3 The Sampling Procedures

The next step in the data collection process involved sampling essays from two prompts for the marking. The study design was that a sample of 120 essays per prompt would be double-marked analytically using the modified ESL Composition Profile. Of these essays, half would be double-marked using the holistic scale as well¹⁶. Double-marking was considered

¹⁶ The total sample size (i.e., the total number of essays that would be marked holistically or analytically) was partly determined based on the total number of marking hours that the researcher was granted by the Multicultural Program Unit of NSW Department of Education and Communities, after taking out the time estimated for training and moderation processes. The rationale for having more essays double-marked analytically than holistically was that, in this validity study, analytic scores from human markers were used more extensively than holistic scores. For example, human trait scores were used in the investigation of the measurement and structural properties of the IEA scores, in addition to their use in the examination of correspondence rates between human total scores and IEA total scores. Hence there was a desire to maximise the reliability of the results involving analytic scores, within the total resources available. It was thus decided to have a larger sample size for analytic scoring than for holistic scoring.

necessary for a high-stakes test like PTE Academic. When two human (analytic or holistic) scores were different by more than 1 score point, adjudication would take place. The adjudication process was organised by the researcher on another day after the main marking process was completed. Human scores produced from these double-marking processes would then be used as criterion measures for validating the IEA scores.

In order to sample 120 essays for each prompt for marking, the holistic scores produced by the human markers using the Common European Framework (CEF) scoring rubric (0–4) (Appendix B), provided by Pearson in the data package, were used to first examine the spread of writing ability demonstrated by all essays across the two prompts. Far fewer essays received one of the two highest CEF score points (12 essays for Tobacco and 9 for Voting) than those which received a score of 1 or 2. In order to ensure the samples had enough essays at the highest ability level, it was decided to retain all the essays which achieved the two highest CEF score points in the sample for each prompt. Similarly, all 22 Voting essays which received the lowest CEF score point 0 were retained in the sample for the Voting prompt; while a random selection of 20 out of 32 Tobacco essays, which received a score of 0, were included in the sample for this prompt.

The next step of the sampling procedure involved generating a stratified random sample of essays for the middle two achievement levels (i.e., the score points 1 and 2) on the CEF scale, with the ratio of essays at the score point of 2 to those at the score point of 1 similar to the ratio represented in the original full data set. Once a sample of 120 essays per prompt was generated, each sample was then randomly divided into two sets, with random division taking

place on the score point level to ensure a roughly equal number of essays at each score point for each set within a prompt. In accordance with the study design, one set was to be scored both holistically and analytically, and the other set to be scored only analytically.

A scoring scheme (Appendix F) was then devised so that no marker scored the same essay twice using the two different modes of scoring. This design avoided any bias that may otherwise have arisen from possible halo effects as a result of markers recognising an essay and remembering their first impression when engaged in the second mode of scoring. For the same reason, essays were also placed in random order before scoring.

5.2.4 Characteristics of the Test Takers

Of the total 240 test takers who produced the sample of essays across the two prompts, no one responded to both prompts. The average age of these test takers was around 26 years at the time of the test, with a standard deviation of 7 for each prompt. An important characteristic of these test takers was that the majority of them were from a non-English speaking background. The 240 test takers came from 40 different countries and spoke 37 distinct languages. The majority (70%) were born in Asian countries, and a further 10% were born in Europe. As a result, most test takers (67%, or 160 out of 240) spoke an Asian language at home. The most common language background was Chinese (60), this was followed by English (48); Indonesian (18); and, Gujarati, Hindi and Korean, each spoken by 15 test takers.

5.2.5 Scoring Session

Scoring took place on the last day in Term 3 in 2009 (September in Australia¹⁷). At the beginning of the scoring session, markers completed a questionnaire on their background (Appendix G). The researcher then reiterated the purpose of the research project and gave a brief description about the automated essay scoring systems. The researcher also provided information regarding the PTE Academic testing context including the testing conditions and the descriptions of the two prompts used in this study. The two rating scales used in this study were then introduced and explained to the markers.

Markers agreed that the two rating scales chosen by the researcher were appropriate for use in this study as the contexts in which two rating scales were normally applied were similar to the testing contexts of the PTE Academic. Specifically, markers agreed that the modification made to the ESL Composition Profile by the researcher for use in this study was appropriate for this scoring occasion.

In order to ensure a consistent application of the two scales amongst markers, the researcher then provided, for the analytic and holistic scoring separately, a small set of sample essays from two prompts for scoring, discussion and resolving differences.

¹⁷ Australian schools have four terms in each school year.

During the discussions and deliberations associated with the training on the sample essays, it was observed that, whenever there was a question or a different view emerging on the rating scales, markers as a group frequently went back to the purpose of the test, the definition of the writing ability being tested, the essential levels of the writing required in a university academic environment and the directives given to the test takers at the test to try to seek agreements on the application of the scales. Different views on how to deal with atypical essays such as off-topic and extremely short essays were also discussed and resolved. When consensus on the use of the rating scales was reached, markers were allocated a set of essays for marking according to the scoring scheme (Appendix F).

5.3 Chapter Summary

This chapter described the testing context of PTE Academic, the process of acquiring human measures, including the sampling procedures, used in this study and the characteristics of the markers, test takers, writing prompts and essays included in the sample. The next chapter considers the writing ability construct that needs to be measured by the PTE Academic.

Chapter 6 The Domain of the Writing Ability Construct

In order to examine the first and the fourth components of the proposed framework (i.e., the content and the structural aspects of validity), it is necessary to first determine the nature of the construct that is of interest to the users of the test – Pearson Test of English (PTE) Academic. This entails clarifying the skills, knowledge and competencies that are meant to be assessed by the PTE Academic writing prompts. It also includes illuminating the internal structure of the ability construct to guide the accumulation of the structural aspect of validity evidence, which is explored in a later chapter.

This chapter examines the construct domain through two approaches: product (i.e., through articulating what human markers value in a written product); and process (i.e., through examining the various competencies required in the process of writing). Using these two different approaches to explore the construct domain reflects the two broad perspectives educators hold on how writing should be taught and assessed (Quinlan et al., 2009). It is important that both perspectives be considered so that the domain can be defined as completely and as accurately as possible.

6.1 The Product Approach to Writing

Defining the writing ability through the product approach reflects the way in which writing ability has traditionally been measured – that is, through the quality of a written product, such as an essay. Using this approach, the writing ability construct can be characterised by the key

traits of writing quality that are emphasised by human markers when they evaluate the overall communicative effectiveness of a written product. A number of studies have used different methods to identify the key writing traits in an essay valued by markers. These methods include analysing what markers said that they would value in an essay (e.g., Cumming, Kantor & Powers, 2002; Diederich et al., 1961; Jones, 1978); what markers actually valued in a marking process (e.g., Breland & Jones, 1984; Freedman, 1977; Harris, 1977); and, markers' decision-making behaviours while evaluating essays, using think-aloud techniques (e.g., Cumming, 1990; Cumming, Kantor & Powers, 2001, 2002).

What consistently emerges from these studies is that essay quality, as conceptualised by expert markers, is inherently multi-dimensional (Quinlan et al., 2009). Furthermore, there is a degree of consensus amongst markers about the traits they value in an essay or the traits they use to describe their thinking processes while marking. The commonality in the traits that make up effective writing as identified in these studies corroborates the researchers' views that experienced markers focus on more or less the same few traits in a written product – that is, what to say (the content/ideas); how to organise it (organisation of ideas); and how to say it effectively (word choice, sentence fluency, and conventions) (Jacobs et al., 1981).

This commonality in the key traits identified by expert markers is also reflected in descriptions of holistic and analytic rating scales. From a review of the common scales used to evaluate persuasive essays, the genre of writing assessed by PTE Academic, Quinlan et al. (2009) concluded that these scales focus on a relatively stable set of traits of essay quality. Typically, these consist of one or two high-level traits of writing (such as quality and

organisation of ideas) and a few low-level language traits of writing (such as vocabulary range, sentence construction, and conventions).

In summary, writing ability, when reflected in the quality of an essay as a product, can be defined through the few common traits in an essay, which are emphasised by experienced markers. If AES systems truly measure essays in the same way as expert markers, the automated scoring process should resemble the thinking processes of these markers while they evaluate essays. This type of validity evidence has been asserted by some researchers as the ultimate test of score validity for AES systems (Lee & Kong, 2004; Lee et al., 2008). At a minimum, an AES scoring process should attend to at least the same set of traits valued by experts in their thinking processes.

Since analytic scales readily identify the important traits of essay quality, it is rational to use an appropriate and well-constructed analytic rating scale as a representation of the writing construct in order to facilitate the mapping of the writing traits scored by an automated system to those valued by human markers in the same testing context (Lee et al., 2008; Quinlan et al. 2009). This rationale forms the theoretical basis for the approach used in the next chapter to examine the Intelligent Essay Assessor (IEA) construct coverage. Construct coverage, in this thesis, means the extent to which the aspects of writing performance that are assessed by the IEA are relevant to, and representative of, the aspects of performance that need to be captured through PTE Academic writing tests.

6.2 The Process Approach to Writing

A review of the various writing process models developed in the past thirty years (e.g., Bereiter & Scardamalia, 1987; Hayes, 1996) suggests that writing ability in an academic environment, which is the ability of interest in this study, is influenced by two distinct competencies: *language competence* and *strategic competence*. While the language competence refers to the linguistic resources available to writers to draw upon during the writing process, the strategic competence represents a higher order, non-language-specific ability, which enables an individual to use available language resources in appropriate ways to accomplish a communicative goal.

One influential writing process model exemplifying this view is the Bereiter and Scardamalia (1987) model. This model argues that novice writers and expert writers use different strategies to compose texts. While the former tend to use a knowledge-telling strategy of writing, involving little planning and representing a natural way of writing (i.e., writing down ideas as they occur), expert writers use a knowledge-transforming strategy of writing, which involves significant conscious planning and problem solving activities to reach the communicative goals (Alamargot & Andriessen, 2002). The differences between the novice and expert writers as conceptualised by this model lend support to the view that effective writing in academic environments is not only contingent on the writer's language knowledge for writing down ideas in appropriate linguistic forms, but also on the writer's higher order processing skills for setting goals, planning and organising the content, taking the perspective of readers, and monitoring and evaluating texts against the initial intentions of the writing.

Another significant model – Hayes’s (1996) writing process model – also incorporates a similar notion of strategic competence. His model conceptualises three recursive cognitive processes inherent in writing: text interpretation, reflection and text production. Both the reflection and the interpretation processes involve many mental activities such as reasoning, inference making, problem solving and high level reflective thinking. In Hayes’s (1996) model, an experienced writer is interpreted as having the ability to activate a repertoire of cognitive, meta-cognitive, linguistic and rhetorical strategies in order to construct a coherent and connected piece of writing that meets his/her overall rhetorical goals. According to modern theorists (e.g., Becker, 2006; Galbraith, 2009; Weigle, 2002), it is these writing-specific strategies that separate expert writers from novice writers, given an equivalent level of linguistic knowledge and general strategic competence.

This notion that writing ability in an academic environment is influenced by two distinct competencies is also consistent with contemporary theory in the applied linguistic field concerning communicative language ability (Bachman & Palmer, 1996). Communicative language ability is defined as the ability to use language to accomplish a communicative goal and is seen to be manifested through traditional skills of reading, speaking, listening and writing. Bachman and Palmer (1996) suggest that communicative language ability consists of interactions between two distinct components: aspects of language knowledge and strategic competence. Recognition of a non-linguistic factor enabling language use is considered to be a significant step forward, particularly in facilitating discussions about non-native speakers’ communicative ability (McNamara, 1996; Phakiti, 2008).

Though the two competencies of language knowledge and strategic competence are not directly measured in writing tests, they are manifested in different aspects of the written product. Strategic competence, for example, can be seen as manifested in the content and rhetorical aspects of essays, since it is characterised in terms of such functions as: goal setting, planning (e.g., determining what content is to be retrieved from memory), assessing (e.g., evaluating the adequacy and appropriateness of content to the communicative situation) and control of execution (e.g., organising ideas) (Douglas, 2000). These aspects of writing performance include writing traits such as rhetorical development, argumentation, coherence, use of evidence and organisation of ideas. On the other hand, language knowledge is reflected through the language traits of essays, such as sentence construction, vocabulary range and language conventions.

It can therefore be argued that the main traits of an essay written in an academic context can be conceptualised into two distinct dimensions, since they are manifestations of two distinct competencies. Consequently, it is hypothesised that the structural relations amongst the scores assigned to the various traits in an essay, whether through an automated system or through human markers, exhibit a degree of discriminant evidence, which reflect the conceptual distinction between these two dimensions. In other words, scores awarded to the content and organisation aspects of essays should, to some degree, be independent of those awarded to the language traits of the writing.

In the context of evaluating essays at the college or college entrance test level, a limited number of studies (e.g., Cumming et al., 2002; Lee et al., 2008; Santos, 1988) have presented

empirical evidence that human markers do in fact discriminate between content/rhetoric and language aspects of writing in compositions. An example is Santos' (1988) study which investigated university professors' reactions to the academic writing of non-native speaking students. Santos (1988, p. 84) found that, while professors considered the language errors contained in the sample essays as being linguistically unacceptable, they still gave significantly higher ratings to the content than to the language of these essays. In addition, even though professors rated the overall language of some essays significantly lower than others, they did not rate the content correspondingly lower.

Further to the notion that language knowledge and strategic competence are conceptually and (potentially) measurably distinct, there is also evidence that the two competencies are mutually influential during the process of writing (Becker, 2006; Galbraith, 2009; Kellogg, 1996). Kellogg (1996) argues that the interactive nature of the cognitive processes inherent in writing places extensive demands on the limited working memory capacity. This often leads to a pervasive phenomenon in the process of writing – cognitive overload. Expert writers are believed to have more developed writing skills, rhetorical strategies and more domain and linguistic knowledge to help them to free up working memory space and reduce the overall load on the central executive system (Kellogg, 1996; Becker, 2006, Galbraith, 2009).

However, when writers have very limited linguistic knowledge, they need more cognitive resources to translate ideas into linguistic codes. For example, extensive cognitive resources may be spent on lengthy searches for lexical and syntactic choices. As the capacity of the

memory space is limited, this leaves fewer resources available for higher order processes such as generating content, planning, and organising and developing arguments to support ideas.

The interaction between the two competencies during the writing process suggests that the structural relations among the traits are also likely to be influenced by the intensity of the interactions that may exist between the two competencies during the writing process. The extent to which the higher order traits (such as content and rhetoric aspects of writing) and language traits can be discriminated in practice is therefore likely to vary across different groups of test takers and across different types of essay assignments.

If the IEA scoring model is rationally developed, the internal structure of the assessment (i.e., the interrelations among the scored aspects of performance) should be consistent with the internal structure of the target construct, expected from the domain theory; or consistent with the internal structure of scores given by human experts on the same set of essays. Differences in internal structure of the IEA and human trait scores, including the abilities of the human markers and the IEA to discriminate between the higher order and language traits for the same set of essays will be pursued in Chapter Eleven, as a means to provide evidence to the structural aspect of validity for IEA.

6.3 Chapter Summary

This chapter examined the characteristics of the construct domain of interest, both from the process and the product perspectives. Discussions of the writing process models led to a

hypothesis regarding the structural relations amongst scores assigned to the various writing traits. This hypothesis will be tested in Chapter Eleven. Discussions of the writing ability using a product approach provided the theoretical rationale for using a well-constructed analytic scale as a validation instrument for the purpose of examining the IEA construct coverage. This analysis is the focus of the next chapter.

Chapter 7 Writing Traits Scored by Intelligent Essay Assessor (IEA) and the IEA Scoring Procedure

This chapter demonstrates how evidence pertinent to the first and the second components of the proposed Automatic Essay Scoring (AES) validation framework can be collected and evaluated. It first examines the link between the writing traits scored by the Intelligent Essay Assessor (IEA) for the Pearson Test of English (PTE) Academic writing prompts and the writing ability that is intended to be measured by these prompts. This is then followed by a discussion of the validity implications of the scoring procedures used by the IEA.

7.1 Analysing the IEA Construct Coverage – Relevance and Representativeness

A relevant and well-constructed analytic scale is first chosen as a representation of the target construct domain to evaluate the construct coverage of the IEA. The intention is to address two questions: 1) Are the writing traits assessed by IEA relevant to the writing construct of interest? and, 2) Do these traits represent the relevant construct domain? That is, are all important parts of the writing construct of interest covered by the IEA traits?

For this study, the ESL Composition Profile (the *Profile*) developed by Jacobs et al. (1981), is employed as a validation instrument for the purpose of assessing the content coverage of the IEA writing construct. The *Profile* is one of the best-known analytic scales in second-language assessment and is widely used in testing contexts that are similar to that of the PTE Academic (Lee et al., 2008; Weigle, 2002). Various validation studies (as summarised in

Jacobs et al., 1981, pp. 74–79) have provided evidence of how well this scale assesses non-native speakers' ability to write in an academic environment. Validity evidence includes:

1) concurrent validity – correlations between scores from this *Profile* and writing scores from other tests such as TOEFL; 2) construct validity – through detecting changes in writing ability assessed by the *Profile* before and after instructional writing programs; and 3) some evidence of predictive validity – the extent to which the *Profile* scores predict students' performance in English and other subjects at the university level. As the typical use of the *Profile* scores closely mirrors the use of IEA scores assigned for the PTE Academic writing tests, the writing skills and knowledge that the IEA and the *Profile* scores are purporting to measure can be considered close. Consequently, it is appropriate to use the *Profile* as a validating instrument to analyse the IEA construct coverage.

The *Profile* identifies five essential dimensions of writing (as measured through the five writing traits) that are pertinent to the effectiveness of written communication at the college level: content, organisation, language use, vocabulary and mechanics (see the scoring rubric of the *Profile* at Appendix C). Key criteria for the highest mastery level within each of the traits provide indications of the breadth and depth of the *Profile* assessment coverage for each corresponding dimension. The criteria, adapted from Jacobs et al. (1981, pp. 92–96), are:

<i>Content:</i>	Knowledgeable, substantive, thorough development of thesis, relevant to assigned topic
-----------------	--

<i>Organisation:</i>	Fluent expression, ideas clearly stated and/or supported, succinct, well-organised, logical sequencing and cohesive
<i>Vocabulary:</i>	Sophisticated range, effective word/idiom choice and usage, word form mastery and appropriateness of register
<i>Language Use:</i>	Effective complex constructions, agreement, tense, number, word order and function, articles, pronouns and prepositions
<i>Mechanics:</i>	Spelling, punctuation, capitalisation, paragraphing, and handwriting.

These five writing traits reflect the consensus amongst experienced human markers as to what constitutes effective academic writing from a product point of view (e.g., Cumming, 1990; Cumming et al., 2001, 2002; Freedman, 1977; Harris 1977). Furthermore, they adequately reflect the key knowledge sets and competencies that are embodied in the theoretical writing construct as defined through the writing process. For example, the assessment criteria across the three language traits in the *Profile* scale, such as effective sentence construction, linguistic accuracy and lexical sophistication, measure the level of the language knowledge that is required for writing with clarity, accuracy and fluency, as well as for ensuring essays are written in socially and culturally appropriate forms (Bereiter & Scardamalia, 1987; Hayes, 1996). On the other hand, the assessment criteria across the two higher order traits in the *Profile*, such as the substantiveness of the ideas, appropriateness of the content to the topic, and thorough development of the thesis, measure the underlying strategic competencies which are essential for establishing more conceptual goals to guide the retrieval of content and to construct better structured and more coherent texts for a communicative purpose.

In summary, the *Profile* adequately captures the essential knowledge sets and competencies that need to be measured for decisions concerning admission to university or college programs. The essential concept embodied in this rating scale is that, when the writing ability is measured through the quality of an essay, five essential traits of the writing need to be assessed: content, organisation, language use, mechanics and vocabulary. Each trait provides a slightly different perspective on the communicative effectiveness of an essay; and when interpreted together, they provide a reliable estimate of a test-taker's writing proficiency. This understanding of the writing construct provides a suitable theoretical framework to guide the assessment of the appropriateness and representativeness of the writing traits assessed by IEA.

7.2 Writing Traits Assessed by IEA for the PTE Academic

In order to evaluate the quality of the essays written for the PTE Academic writing prompts, IEA assesses seven writing traits. As a broad indication of the assessment coverage, the following paragraph lists the key criteria for the highest assessed level for each trait that the IEA assessed, as extracted from the PTE Academic writing scoring rubric (see Appendix A). For simplicity, this rubric is also referred to hereunder as the IEA scoring rubric. This rubric¹⁸ remains the most complete descriptions of the traits assessed by the IEA. The vendor of the

¹⁸ The scoring rubric for the PTE Academic writing (Pearson, 2011b) is essentially the same as it was supplied to the author in 2009. The only noticeable difference is in the name of the trait '*Grammar Usage and Mechanics*'. Although the criteria for this trait remain the same, the latest PTE Academic Score Guide (Pearson, 2011b) refers to this trait as "*Grammar*" (p. 60). For the purpose of this thesis, the original trait name "*Grammar Usage and Mechanics*" provided by Pearson in 2009 is used throughout the thesis.

IEA does not publicly disclose the exact properties of the micro-text features included in most of these traits.

<i>Content:</i>	The essay adequately deals with the prompt.
<i>Development, Structure and Coherence:</i>	The essay shows good development and logical structure.
<i>Grammar Usage and Mechanics:</i>	The essay shows consistent grammatical control of complex language; errors are rare and difficult to spot.
<i>General Linguistic Range:</i>	The essay exhibits mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No sign that the test taker is restricted in what they want to communicate.
<i>Vocabulary Range:</i>	The essay shows a good command of a broad lexical repertoire, idiomatic expressions and colloquialisms.
<i>Spelling:</i>	Correct spelling, but there may be one typing error.
<i>Formal Requirement:</i>	A length trait; to receive the highest score, the length must be between 200 and 300 words.

In order to analyse the IEA construct coverage, the characteristics of the writing performances assessed through the traits of the *Profile* are mapped to those assessed through the seven IEA traits, based on the high-level descriptions of the IEA scoring criteria.

When interpreting the results of the mapping, the following two issues need to be borne in mind. First, the mapping is based on the characteristics of writing that are intended to be captured by each of the IEA traits, as indicated by the IEA scoring criteria. It is not based on how well IEA can actually measure what it purports to measure. Evidence for the latter will

be collected through the next few chapters. For example, for the purpose of this mapping exercise, coherence is recognised as being a characteristic of writing that is assessed through the *Development, Structure and Coherence* trait in the IEA scoring model, even though the model may not be able to measure it effectively, as demonstrated by McGee's study (2006). The purpose of this mapping is to ascertain whether there are any major characteristics of writing that should be assessed by IEA, but which are omitted in the IEA scoring rubric; and whether there are any characteristics which are measured by IEA, but which are irrelevant to the writing ability being assessed.

The second issue is that, in the absence of the details of the micro-textual features that are included in the IEA traits, the researcher has necessarily made assumptions based on the researcher's interpretation of the high-level IEA scoring criteria. This occurs when determining the assessment coverage of some IEA traits. For example, the scoring criteria for the IEA *Vocabulary Range* trait refer to a "good command of a broad lexical repertoire, idiomatic expressions and colloquialisms" (Pearson, 2011b). It is assumed that these criteria not only intend to include the assessment of the range of the vocabulary used, but also to encompass an evaluation of the accuracy, effectiveness and appropriateness of the choice of the vocabulary in the context in which it is used. When assumptions are made, they are stated as so in the descriptions of the mappings.

The following table maps the traits assessed by the *Profile* to those assessed by the IEA traits, based on the detailed scoring criteria associated with the *Profile* (Appendix C and Jacobs et al., 1981, pp. 92–96) and the high-level IEA scoring rubric (Appendix A).

Table 7.1

Links Between Traits on the Profile Scale and the IEA Traits

Profile traits

<i>Content</i>	The main assessment criteria for this <i>Profile</i> trait include content knowledge, relevance of the ideas to the topic, substantiveness (e.g., whether main aspects of the topic are discussed with sufficient details). These criteria are assessed through the IEA <i>Content</i> trait, as this trait also tries to identify main and/or minor aspects of topic that are discussed in an essay and whether they are supported with adequate details. An exception is that, although the <i>Profile Content</i> trait assesses the originality in the way information is used to support the thesis, it is unclear whether the same is also assessed by the IEA <i>Content</i> trait.
<i>Organisation</i>	This trait can be mapped to the IEA <i>Development, Structure and Coherence</i> trait. Both traits attempt to assess the structure of the text, cohesion and coherence, logical sequencing and development of ideas. An exception is that, while succinctness is part of the scoring criteria for the <i>Profile Organisation</i> trait, it is not explicitly included in the scoring criteria for the IEA <i>Development, Structure and Coherence</i> trait.
<i>Vocabulary</i>	This trait can be mapped to the IEA <i>Vocabulary Range</i> trait. Both traits assess the range of vocabulary used. It is assumed that, as with the <i>Profile Vocabulary</i> trait, the IEA <i>Vocabulary Range</i> trait also assesses whether the choice of vocabulary is accurate and effective for the context in which it is used.
<i>Language Use</i>	A part of the scoring criteria for this trait can be mapped to those covered by the IEA <i>General Linguistic Range</i> trait, as both attempt to measure the writer's ability to use a wide range of language to achieve a communicative goal (e.g., the ability to construct different types of sentences at varying levels of complexities; ability to use language techniques suitable for the context). The rest of the scoring criteria for the <i>Language Use</i> trait focus on the grammatical control of complex language, which are assessed through the IEA <i>Grammar Usage and Mechanics</i> trait.
<i>Mechanics</i>	This trait assesses mechanical accuracy such as spelling, punctuation,

capitalisation and paragraphing. While spelling is assessed through the IEA *Spelling* trait, it may be assumed that punctuation, capitalisation and other types of mechanical errors are assessed through the IEA *Grammar Usage and Mechanics* trait. It is also possible that paragraphing may have been assessed through the IEA *Development, Structure and Coherence* trait as appropriate paragraph breaks contribute to the meaning, coherence and flow of ideas.

The links between the traits of the *Profile* and the IEA traits are schematically represented in Figure 7.1.

The main conclusion from the above analysis is that the IEA does seem to assess all the important parts of the writing construct of interest, as operationalised through the *Profile*. However, it is also clear that the IEA *Formal Requirement* trait does not have a corresponding trait within the *Profile* analytic scale. That is, length is not a criterion that is directly and explicitly assessed by the *Profile*. This raises the concerning possibility that this criterion may introduce construct-irrelevant variance. The rationale to include a length criterion in the IEA scoring model is that it allows an assessment of the writer's ability to summarise relevant and complex information while observing a strict length requirement (John De Jong, Pearson, personal communication, May 17, 2010). However, such an inclusion does have validity implications, which will be discussed in full in Chapter 11.

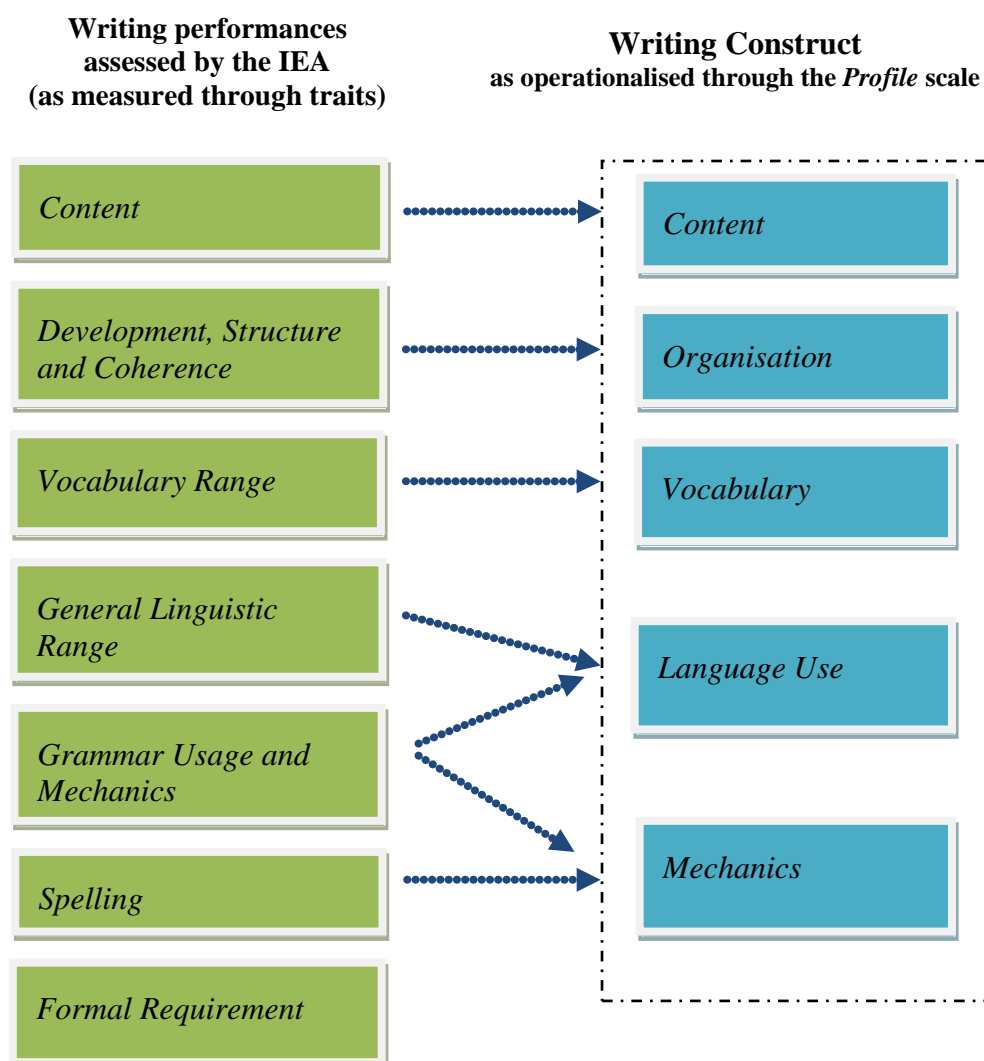


Figure 7.1 Mapping of the Writing Performances Assessed by the IEA to the Construct of Interest

It is acknowledged that this mapping exercise, though essential in order to establish the content coverage of the IEA construct, is limited by its necessary reliance on the very high-level descriptions of the scoring criteria publicly released by Pearson (2011b). It is recommended that such exercise be repeated when there is more detailed information about the micro-textual features that are assessed by the IEA traits.

In the meantime, in order to establish the substantive link between the writing traits assessed by the IEA and the construct domain, there is a need to go beyond face validity evidence and collect more direct evidence. This evidence will emerge from an investigation of the measurement and structural aspects of score validity, as well as of the accuracy of the IEA scoring at the trait level. It will be developed in Chapters Nine, Ten and Eleven.

The next section collects evidence pertinent to the second component of the proposed AES framework. Specifically, it inspects the nature of the IEA scoring procedure and how that might impact on the interpretations and the validity of the scores produced. The appropriateness of the criteria used to score each trait, which is part of the scoring procedure, will be pursued in Chapter Eleven.

7.3 The Scoring Procedure Used by IEA

It has been previously stated that the focus of this part of the study is to investigate the appropriateness of the total writing score assigned by the IEA for an essay written to a PTE Academic writing prompt. According to the PTE Academic Score Guide (Pearson, 2011b, p. 7), the IEA produces a total score by first assessing seven writing traits in accordance with the pre-defined scoring criteria for each trait, and then summarising scores of the traits to derive the total score. From this description, the IEA essentially uses an analytic scoring procedure to score the essays.

There are two key assumptions underlying this scoring procedure. The first assumption is that the quality of a whole piece of writing is exactly the same as the summed quality of the parts, as measured by the analytic traits. The key validity question therefore is: *Is this assumption appropriate for the writing construct intended to be measured?*

Although it is important to accumulate theoretical rationale and/or empirical evidence to verify the appropriateness of this assumption, it is in fact very difficult to do so and is intentionally outside the scope of this study. This is because there is still very little understanding about how different scoring procedures and their respective underlying assumptions are grounded in theory. For example, there does not seem to be any significant effort made to embed holistic scoring in a coherent body of psychological or writing theories (Hunter, Jones & Randhawa, 1996). Even for those who administer it or participate in it, holistic scoring is still largely a black box (Haswell, 2006). The lack of understanding in this area is confounded by empirical observations that markers have difficulty in identifying patterns of relationships among the various writing traits (e.g., Braungart-Bloom, 1986; Marsh & Ireland, 1984; Schoonen, 2005). These factors make it hard to be certain about the complex interactive relationships that may exist among the parts, and between the parts and the whole. This in turn makes it difficult to assess in a robust manner whether the type of procedure the IEA uses to assess the PTE Academic essays is appropriate. This study nonetheless notes that both analytic and holistic scoring procedures are used in similar testing contexts to that of the PTE Academic, and that, although holistic scoring may appear to be more prevalent than the analytic scoring in large-scale language assessments, this may largely be attributable to

practical, cost-effective and face-validity reasons, rather than any conclusive evidence through writing measurement theory.

The second assumption underlying the IEA scoring procedure is that a single score is sufficient to capture the profiles of an individual's performances across different writing traits. This assumption will be tested in Chapter Nine.

An additional observation of the IEA scoring procedure is that it is fixed across all writing prompts for the PTE Academic. This means that there is no option for the IEA scoring model to add or omit any traits, or adjust any weighting for individual traits, when assessing the quality of essays written to different PTE Academic writing prompts. This scoring procedure is different from the prompt-specific procedure, which uses statistical optimisation techniques to determine a potentially idiosyncratic set of traits and weightings for each prompt.

The advantage of using a fixed analytic procedure for all writing prompts is that it ensures consistency in the interpretations of the scores produced for different prompts within the same testing programs. On the other hand, it is critical to ensure that these traits are relevant to, and representative of, the construct domain, and that the weightings associated with the traits are appropriate and justified for the testing context.

7.3.1 The Appropriateness of the Contribution of the IEA Traits towards the Overall Score

Figure 7.2 is a schematic illustration of the IEA scoring model used to derive an overall PTE Academic writing score. The information contained within the brackets indicates the number of score points the rating scale uses to summarise performance on an individual trait.

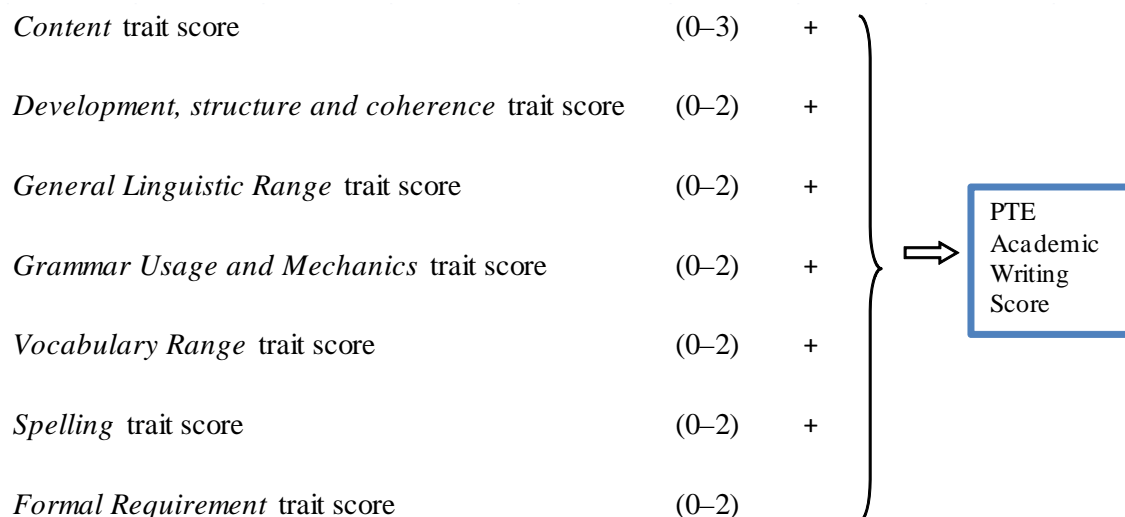


Figure 7.2 A Schematic Representation of the PTE Academic Writing Score

As shown in Figure 7.2, the total score is simply calculated as the sum of the scores across the seven analytic traits. Two observations can be made about the IEA scoring model in Figure 7.2. First, the *Content* trait can contribute more to the overall score than any other trait because it is scored on a 0–3 rating scale while all other traits are scored on a 0–2 scale. The second observation is that spelling is specifically separated from the *Grammar Usage and Mechanics* trait and can contribute the same number of marks towards the overall judgment of

essay quality as other traits such as *Vocabulary Range, Development, Structure and Coherence and Language Use*.

The next two sections examine further whether the contributions of the *Content* and *Spelling* traits make towards the overall score are justified in the PTE Academic testing context.

7.3.2 *Content*

Although it is not universally acknowledged in the literature (Harris, 1977; Raforth & Rubin, 1984), many studies find that markers place more emphasis on content and organisation than on other traits in their qualitative judgments of students' writing at the college level (e.g., Breland & Jones, 1984; Diederich et al., 1961; Freedman, 1979a, 1979b; Huot, 1990a, 1993; Pula & Huot, 1993). For example, to investigate responses from teachers of college-level freshman English to different aspects of writing performance, Freedman (1977, 1979a, 1979b) manipulated college-level essays to be either strong or weak in content, organisation, mechanics and sentence structure, and asked teachers to evaluate those essays. Using analysis of variance techniques, results showed that teachers were most influenced by the content of an essay. Organisation was the second most important influence, and mechanics was the third.

Breland and Jones (1984) reported similar results. Their studies found that the top four traits of writing which most influenced the college English professors when they were evaluating essays, were the same as those the professors perceived to be the most important traits of

writing, albeit in a slightly different order. These traits were overall organisation, use of supporting material, noteworthy ideas and statement of thesis.

A number of other studies which investigated markers' decision-making behaviours also found that content was the primary consideration upon which human markers based their judgments of writing quality (e.g., Connor & Carrell, 1993; Huot, 1988; Vaughan, 1987).

The importance of content relative to all other traits when marking essays is also reflected in the ESL Composition Profile (Appendix C). The *Profile* scale allocates the highest weighting (30% of the total score) to *Content*. This is more than the weighting allocated to *Language Use* (25%) and to *Organisation* and *Vocabulary* traits (each 20%). The developers of the *Profile* scale, Jacobs et al. (1981), believe that when assessing an essay for its overall communicative effectiveness, the primary focus should be on “the semantic content of the communication” rather than “the form” of the essay, because this is what readers in a real world would focus on in a written product (Jacobs et al., 1981, p. 36).

In summary, findings from prior empirical investigations lend support to the IEA allocating more weight to *Content*, relative to all other traits, when combining trait scores to produce a single score indicating the overall quality of an essay written to a PTE Academic prompt.

7.3.3 Spelling

Spelling is separated from the *Grammar Usage and Mechanics* trait in the IEA scoring model, and this trait can contribute the same number of marks to the overall score as all other traits, except the *Content* trait. Implicit in this practice is the belief that, when assessing writing proficiency levels for college/university admission purposes, spelling is one of the determining traits of good writing. Therefore it is given the same weight as the other traits, such as development/organisation, vocabulary, language use, when forming an overall judgement of essay quality.

Although more studies are needed to understand the nature of the complex relationships that exist between spelling and writing ability (Hayes, 2010), results from the limited available studies indicate that the above-mentioned belief does not seem to be consistent with the shared understandings amongst human markers. When assessing essay quality in college academic settings or for general English proficiency tests, the available studies suggest that human markers generally regard spelling (and other types of mechanical errors), as relatively minor aspects of writing proficiency, so long as the errors are localised occurrences and do not impede the overall communicability of the written text (e.g., Breland & Jones, 1984; Lee et al., 2008; Matsuno, 2009; Vann, Meyer & Lorenz, 1984).

Breland & Jones (1984) surveyed 20 college English professors about which characteristics of writing influenced them most in judging brief, impromptu, argumentative type of essays. Of the 20 characteristics provided on the evaluation form (nine relate to discourse; six to syntax;

and five, including spelling, relate to lexical characteristics), spelling was perceived to be the least important influence. Although the same study demonstrated spelling was in fact an important criterion actually used by markers when making qualitative judgements of essay quality, it was found to contribute less to the overall essay scores than each of the nine discourse characteristics (e.g., organisation, rhetorical strategy, and sentence variety), the combined syntactic measure and three other lexical characteristics (i.e., range of vocabulary, level of diction and precision of diction).

Vann, Meyer and Lorenz (1984) provided further evidence that, in assessing academic writing of non-native English speaking students, university teachers showed more tolerance towards spelling errors than they did towards other types of grammatical/linguistic errors such as word order, article omission and subject-verb agreement.

This result also reflects Matsuno's (2009) study which examined self, peer and teacher assessments of Japanese university students' writing. Of all 16 writing traits (such as grammar, introduction and logical sequencing), spelling was found to have been judged the most leniently by all three groups. The researcher (Matsuno, 2009) also suspected that the mechanics traits (i.e., the spelling, punctuation and format traits) might not even be directly related to writing ability, as these traits elicited patterns of responses that were different from the general pattern of response. Matsuno (2009) provided an explanation, noting that "while low ability writers can get mechanical aspects of an essay correct, only good writers can produce high-quality essays in terms of features such as lexical choice, grammatical accuracy and complexity, and logical flow" (p. 87). Therefore mechanics traits, such as spelling, "may

not discriminate well between different levels of writing ability and/or indeed may not be valid measures of writing ability” (p. 87).

The Lee et al. (2008) study of the holistic and analytic scores assigned to 930 TOEFL argumentative essays by human markers, corroborated the Matsuno (2009) study in that it also found that scores for the mechanics trait, which subsumed spelling, were least aligned with those for other analytic traits. In addition, their study found that the mechanics trait contributed least to markers’ overall judgements of essay quality, when compared with all other five traits (i.e., development, organisation, vocabulary, sentence variety/construction, grammar/usage). This led the researchers to conclude that the mechanics trait “plays only a limited role in the holistic rating of essays” (p. 28).

The above studies suggest that there is a degree of consensus amongst human markers that spelling plays a relatively minor role in markers’ judgments of essay quality for the purpose of assessing students’ ability to write in an academic setting. This shared understanding is also reflected in a number of common analytic scales used in these contexts, in which spelling constitutes only one part of a grammar/mechanics trait that also includes other textual features such as grammatical usage and accuracy. For example, the IELTS (International English Language Testing System) writing scale, which is used to assess prospective international students’ academic writing proficiency, does not assess spelling separately but subsumes it into the *Grammatical Range and Accuracy* trait (IELTS, n.d). The ESL Composition Profile (Jacobs et al., 1981), one of the best known analytic scales used to assess ESL students’ writing ability in college settings, conflates spelling into a *Mechanics* trait, which is also

given the least weight in the overall score. The small weighting (i.e., 5% of the total score) reflects the scale developers' view that the mechanics aspect of writing, such as spelling, "though fundamental to communication at even the most rudimentary level, do not always have to be perfect in order that a writer communicate effectively or, for proficiency testing, at least adequately" (Jacobs et al., 1981, p. 35).

The IEA scale specifically separates out spelling from its *Grammar Usage and Mechanics* trait. Further, it treats losing one point on the *Spelling* trait as having the same effect on the overall score of an essay as losing one point on other seemingly more salient traits of writing such as *Development, Structure and Coherence, Language Use* and *Vocabulary Range*. As a consequence, the writing construct being measured by the IEA analytic scale is different from that being measured by other common analytic scales, such as the IELTS writing scale (IELTS, n.d), the ESL Composition Scale (Jacobs et al., 1981) and the 6-trait scoring scale (Spandel & Stiggins, 1990). The particular emphasis that the IEA places on spelling may result from a belief that spelling can be "objectively" and "accurately" measured by the machine scoring engine. However, by placing a seemingly un-substantiated emphasis on spelling, the overall validity of the IEA writing construct may have been compromised. In this regard, Schoonen (2005) stressed that decisions about which writing traits of the texts need to be scored, and which traits should prevail in the scores, "should be determined by the construct(s) one wants to assess and not just by psychometric considerations" (p. 18).

7.4 Chapter Summary

This chapter demonstrated how evidence pertinent to the first component of the AES validation framework could be collected; that is, how an AES writing construct could be examined for its relevance to, and representativeness of, the domain of the target writing construct. The examination of the IEA construct coverage was performed through mapping the writing traits scored by the IEA to the target writing construct as it was operationalised by an existing analytic rating scale appropriately chosen for the context, in this case, the ESL Composition Profile scale. At a high level, while the IEA traits seemed to adequately capture the main knowledge/skill sets required to produce effective written communication in an academic setting, there was a concern that the *Formal Requirement* trait in the IEA model might introduce construct-irrelevant variance. This aspect of validity will be further examined in Chapter Eleven.

This chapter then illustrated how evidence for the second component (i.e., the scoring procedure of the AES framework) can be collected and examined. It found that the IEA used a fixed analytic scoring procedure to assess essays written to PTE Academic prompts, which helped ensure consistency in the interpretations of scores produced across different prompts. The main concern with the procedure the IEA used to combine the trait scores to overall scores lied in the seemingly un-substantiated importance that the IEA rubric placed on the *Spelling* trait. The validity implications of this treatment of the *Spelling* trait will be explored further in Chapter Eleven.

The next chapter focuses on the overall agreement rates between the human and the IEA-generated scores. Although these agreement rates are only considered as face validity evidence, they attest to an important quality of an automated scoring system – the usefulness of such a system.

Chapter 8 Correspondence Rates between Human and the IEA (Intelligent Essay Assessor) Overall Scores

This chapter examines the correspondence rates between the overall writing scores produced by the Intelligent Essay Assessor (IEA) and by the human markers recruited by this study for the same sample essays. An overall writing score refers to the total score given by IEA for an essay as an indication of its overall quality. The question to be addressed here is: how well do the IEA overall scores align with those produced by the human markers, if a typical human scoring process is used to produce scores in the same testing context? A reasonable level of agreement is regarded by many test administrators as a prerequisite for the usefulness of the IEA system, considering the general consensus in the community that scores from human markers represent a fair measure of writing ability. This investigation also addresses the fourth validity question identified for the first component of the AES (Automatic Essay Scoring) framework proposed in Chapter Four. That question relates to the accuracy of the IEA scoring at the overall score level.

However, before human scores are used as “gold standards” to illustrate the utility of the IEA-generated scores, the quality of human scores is first investigated, through analysis of score reliability. The lower the reliability of the human scores, the more hazardous the generalisation from observed human scores to universe scores (i.e., scores that would be expected to be obtained under parallel marking processes). This in turn makes the inferences

drawn from results of comparisons between human observed scores and IEA scores more hazardous.

8.1 Reliability of Human Scores Used in the Agreement Analysis

One way of estimating the reliability of human scores is to decompose the variation across the observed human scores into separate variance components which correspond to different sources of measurement error. This study uses Generalisability theory – G-theory – to estimate multiple sources of error of measurement and to produce reliability coefficients for different rating scenarios (Cronbach, Gleser, Nanda & Rajaratnam, 1972). The reliability coefficients estimated for the double-marking scenario are the focus of this study because the human scores used for the subsequent agreement analysis are the adjudicated scores from the human double-marking process.

The G-theory analysis was conducted using the GENOVA computer program (Crick & Brennan, 1983). Since each person¹⁹ only wrote to one prompt and no markers marked the same essay using two different scoring procedures²⁰, a separate G-theory analysis was performed for overall scores produced from each scoring procedure, and for scores on each analytic trait. These analyses were conducted for each prompt separately. In each analysis, the

¹⁹ From this point onwards, this thesis will use the term ‘person’ to describe the objects of measurement for the PTE Academic writing tests – that is, prospective international students.

²⁰ In this section, a scoring procedure refers to either the analytic scoring procedure involving the use of an analytic scale in the marking process or the holistic scoring procedure involving the use of a holistic scale by human markers in the marking process.

single-facet crossed design ($p \times r'$) was employed in the G-theory analysis to estimate the variance components and reliability coefficients. Persons (p) were the objects of the measurement with ratings (r' – first and second ratings each essay received for a trait or as a whole) as random facets.

Two types of reliability estimates are available from the G-theory analysis – the generalisability coefficient (G-coefficient) and the dependability index (Φ). They relate to different types of decisions. The G-coefficient is used for decisions concerning relative standing of persons (norm-referenced testing), while the dependability index (Φ) is used for decisions concerning the absolute level of performance (criterion-referenced testing) (Shavelson & Webb, 1991). As Pearson Test of English (PTE) Academic scores are normally used for the decisions relating to admission to higher education programs, and usually there are pre-established English proficiency admission standards for these programs (e.g., minimum PTE Academic scores²¹), the users of the PTE Academic tests are more likely to rely on interpretations of absolute rather than relative performance level. Though the users are more inclined to use the dependability indices, both types of reliability indices are reported and interpreted here for the sake of completeness, with the emphasis on the dependability indices.

²¹ For example, “a score of at least 36 is required for UKBA tier 4 student visas for students wishing to study on a course below degree level”, according to the PTE Academic website: <http://pearsonpte.com/TestMe/About/Pages/ukba.aspx>, retrieved February 1, 2012.

In the G-theory analysis performed for overall scores produced from the analytic scoring procedure, the two ratings used were the two composite analytic scores, calculated from combining the first and second ratings each essay received on the five analytic traits respectively, using the original weighting scheme from the ESL Composition Profile (Jacobs et al., 1981). Formula 8.1 lists the calculation detail:

$$\begin{aligned} \text{Composite analytic score} = & 0.3 * \textit{Content} + 0.2 * (\textit{Organisation} + \textit{Vocabulary}) \\ & + 0.25 * \textit{Language Use} + 0.05 * \textit{Mechanics} \end{aligned}$$

Formula 8.1

The two ratings used in the G-theory analysis performed for the overall scores produced from the holistic scoring procedure and for scores on each of the five analytic traits were simply the two holistic ratings, or the two analytic ratings an individual essay received for each trait during the human marking process.

Results

Variance components estimated from the G-theory analysis for the two prompts are attached at Appendices H and I. The following table (Table 8.1) reports the G-coefficients and the dependability indices for overall scores produced from each scoring procedure, and for scores on each analytic trait assessed, based on double-marking scenarios.

Table 8.1***Score Reliability for Analytic and Holistic Scoring and Analytic Traits in a Double-Marking Scenario***

	based on double-marking scenarios			
	G-coefficient		Φ (Index of Dependability)	
	– for relative decisions		– for absolute decisions	
	Voting	Tobacco	Voting	Tobacco
<i>Overall Score Produced by Different Scoring Procedures</i>				
Holistic Score	0.81	0.83	0.74	0.83
Composite Analytic Score	0.85	0.85	0.85	0.85
<i>Analytic Trait</i>				
<i>Content</i>	0.73	0.79	0.73	0.79
<i>Language Use</i>	0.77	0.75	0.76	0.74
<i>Mechanics</i>	0.74	0.68	0.73	0.68
<i>Organisation</i>	0.77	0.73	0.77	0.73
<i>Vocabulary</i>	0.55	0.66	0.54	0.66

The first observation that can be made of the results in Table 8.1 is that, in a double-marking scenario, the G-coefficients for the holistic and the composite analytic scores reach the conventionally desired level of score reliability (i.e., 0.8) (Schoonen, 2005; Shavelson & Webb, 1991). The coefficients ranged from 0.81 to 0.85 for overall scores produced from the two scoring procedures, across both prompts. These coefficients can be interpreted as the expected correlation between an average of two scores provided by two markers with another average of two scores provided by another two markers randomly selected from the same universe of all admissible markers (Schoonen, 2005, p. 15).

Similarly most of the dependability indices (Φ) for the holistic and composite analytic scores are above the acceptable level (i.e., 0.8), ranging from 0.83 to 0.85. The only exception is the comparatively lower Φ value of 0.74 for the holistic scores for the Voting prompt. An explanation for this exception can be found at the variance components tables in Appendices H & I. It can be seen from those tables that, while the main rating effect (i.e., the average inconsistency amongst human markers – σ^2_r) is negligible elsewhere, the main rating effect for holistic scores for the Voting prompt accounts for 14.3% of the total variance in the observed scores. Thus inconsistency in human rating is a sizeable factor contributing to score variability in the case of holistic scoring of essays written to the Voting prompt. It is noted that, as this study uses the final adjudicated scores from the double-marking process for the agreement analysis, the reliability of the final human scores is expected to be better than those reported at Table 8.1, which were estimated from a pure double-marking scenario. As an indication, in the case of the holistic scores for the Voting prompt, the dependability index improves to 0.81 for a triple-marking scenario.

A second observation from the results in Table 8.1 is that the reliability of the composite analytic scores is consistently better than that of the holistic scores. Additionally, the differences in reliability for the overall scores produced from different scoring procedures vary across prompts. While the difference in Φ values is marginal for the Tobacco prompt, the difference in the Φ values is clearly greater for the Voting prompt (i.e., 0.85 versus 0.74).

A third observation from the results in Table 8.1 is that scores related to *Vocabulary* are less generalisable than scores about other traits on both prompts. While the Φ values for scores

about the *Content*, *Language Use*, *Mechanics*, and *Organisation* traits mostly ranged from 0.73 to 0.79 across the two prompts, the Φ value for scores for *Vocabulary* was significantly lower (i.e., an average of 0.60 across the two prompts). This result should be investigated further by larger studies involving more markers in an attempt to determine whether this is a repeatable phenomenon and what may be the systemic causes of this outcome. Such investigations should be useful for targeted professional development to improve inter-rater reliability on particular writing traits.

Overall, the above observations confirm human overall scores that are to be used for the agreement analysis generally reach the nominated level of reliability. The results also confirm findings from other studies that writing scores are affected by facets of the writing assessment unrelated to the person's writing proficiency, such as marker consistency and prompt effects. These effects are mediated by scoring procedures and the traits assessed (Barkaoui, 2007; Schoonen, 2005).

8.2 Reliability of the Human Scores Acquired from the Pearson Field Tests

Since this study also received the original holistic and analytic scores assigned by human markers from the Pearson's field tests for all essays written to the two prompts included in the study, similar G-theory analyses were carried out to investigate the dependability of these scores. The significance of this analysis is that the original Pearson scores acquired from the field tests were used for "training and validating" the automated essay scoring system used for the PTE Academic (Pearson, 2011a, p. 4). Therefore it is worthwhile understanding the level

of dependability from these scores to the expected scores on the universe of generalisation. It was the approximations of these expected scores that were intended to be used to train and calibrate the IEA. Table 8.2 reports the dependability indices for human scores obtained from the Pearson field tests.

Table 8.2

Dependability Index for Analytic and Holistic Scoring and Analytic Traits in a Double-Marking Scenario – Pearson Field Tests

Scoring Procedure/Trait	Φ (Index of Dependability) Based on double-marking scenarios	
	Voting	Tobacco
<i>Overall score produced by different scoring procedure</i>		
Holistic Score	0.69	0.60
Composite Analytic Score	0.86	0.87
<i>Analytic trait</i>		
<i>Content</i>	0.63	0.64
<i>Development, Structure and Coherence</i>	0.55	0.51
<i>General Linguistic Range</i>	0.58	0.69
<i>Grammar Usage and Mechanics</i>	0.72	0.64
<i>Vocabulary Range</i>	0.64	0.57

Figures 8.1 and 8.2 compare these dependability indices to the corresponding ones obtained from this study. Note that the human markers from the field tests did not mark the *Spelling*, or the *Formal Requirement* traits. For ease of interpretation, traits assessed in this study and those in the Pearson field tests that are of similar nature are grouped together in the figures.

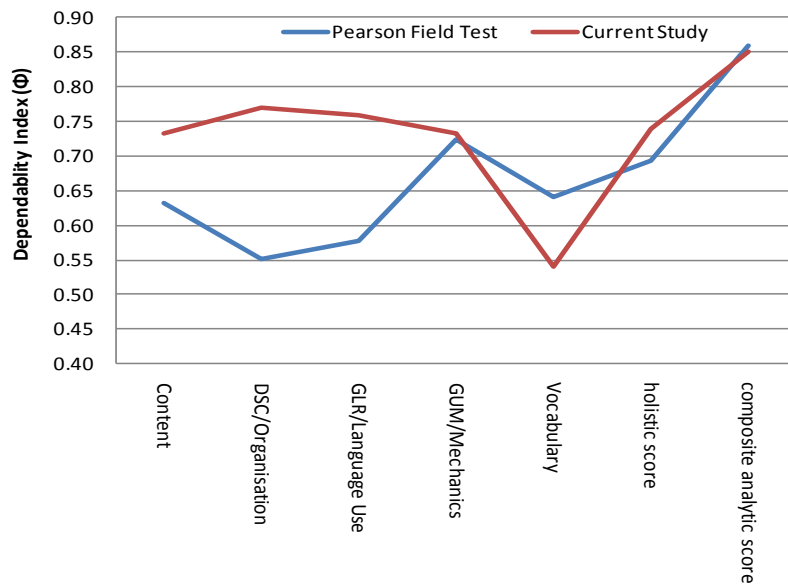


Figure 8.1 Dependability Indices (Based on a Double-Marking Scenario) – Voting

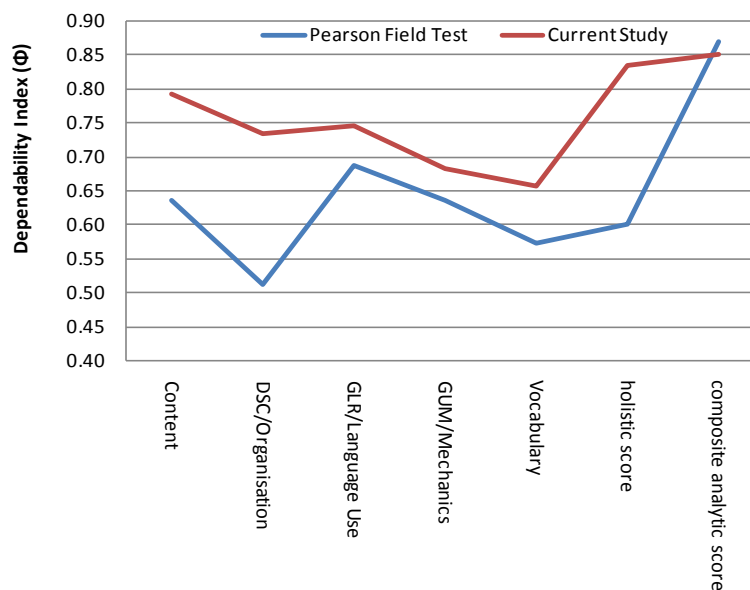


Figure 8.2 Dependability Indices (Based on a Double-Marking Scenario) – Tobacco

Note. 1. The abbreviations of the trait names from the Pearson field tests are as follows:
DSC: *Development, Structure and Coherence*, GUM – *Grammar Usage and Mechanics*, GLR: *General Linguistic Range*.
2. All essays received from Pearson for the two prompts are included in the G-theory analysis.
3. Composite analytic scores from the Pearson field tests are calculated by summing all scores from the five IEA traits human markers marked: *Content, Development, Structure and Coherence, Grammar Usage and Mechanics, Vocabulary Range, and General Linguistic Range*.

The results (Table 8.2 and Figures 8.1 and 8.2) indicate that scores obtained from the human markers of this current study are generally more dependable than those from the Pearson field tests. While the dependability indices (Φ) for the composite analytic scores from the field tests are slightly better than those for corresponding scores acquired through this study (an average of 0.87 across two prompts as compared to 0.85 from the current study), the Φ values for holistic scores and for scores on most analytic traits from the field tests, are considerably worse than the Φ values for the corresponding scores acquired from this current study. For example, the Φ value for the holistic scores obtained from the field tests for essays written to the Tobacco prompt is 0.60 in a double marking scenario, markedly lower than the 0.83 achieved in this current study and considerably lower than the desirable level of reliability for high-stakes tests (i.e., 0.8). Similarly, the Φ values for the two higher order traits (*Content*, *Development*, *Structure and Coherence*) range from 0.51 to 0.64 for the scores from the Pearson's field tests across the two prompts, whereas the corresponding Φ values for scores on the *Content* and *Organisation* traits acquired from this current study are noticeably better, ranging from 0.73 to 0.79.

It is noted that even with the adjudication process (involving a third marker giving a third rating), as employed by the PTE Academic field tests in addition to its double-marking scheme, the reliability of the scores obtained in many rating situations may still fall well short of the desirable level. For example, for the Tobacco prompt, the Φ value for the *Development*, *Structure and Coherence* trait scores may improve to 0.61 in a triple-marking scenario, but this is still below the desired level. The less than desirable reliability exhibited in the Pearson

scores acquired from the field tests, in particular in the trait scores, is an issue, because these scores were used to model the IEA scores (Pearson, 2011a).

Possible explanations for the observed differences in the estimated reliability of human scores obtained from the Pearson field tests and from this current study include the training the two groups of markers received, and the inherent difficulties in training the human markers to be consistent in their rating behaviour. The human markers who marked the same essays in the Pearson field tests were from a pool of 200 international markers from different countries (Pearson, 2011c, p. 3). They may be more heterogeneous in background than the group of markers recruited for this study who were selected to have similar educational backgrounds and similar prior teaching and marking experiences. As rater background factors, such as teaching foci, expectations and perceptions of language proficiency, are known to influence rating behaviour (e.g., Erdosy, 2004, Santos, 1988), it is feasible and likely that the international markers may have been more difficult to train to achieve comparability, than the group of markers selected for this study.

Pearson stated that the reason for using international markers to obtain scores to train the IEA system is so that “the machine is trained on a rich set of international human judgments” which are “person-independent” (Pearson, 2011c, p. 3). However, this goal can only be achieved if markers are carefully selected and consistently trained, or the scores generated are carefully calibrated to minimise any systematic biases that may exist in the rating behaviour amongst markers from different countries.

Some insights into how marking behaviour might differ among markers from different linguistic and cultural background can be obtained from this study. Initially, this current study recruited a second group of five markers from China to mark the same essays across the two prompts using the same rating scales as those used by the Australian markers. All Chinese markers selected met the criteria of being experienced English teachers in senior secondary high schools and having had a number of years of training and marking for high-stakes exams such as the National Higher Education Entrance English Examinations in China. The training was carried out by the same trainer who had trained the Australian team of markers. The main aim of having a group of international markers for this study was to improve the reliability of the human scores which were to be used as the criterion measure later in this study.

However, statistical and distributional analysis revealed significant differences in the rating behaviour between the two groups of markers (Australian and Chinese markers). Tables 8.3 and 8.4 show the mean ratings generated by the two groups of markers for the same essays using the same scales, across the two prompts. Also reported in the table are paired t-test results comparing the ratings from Australian markers to those from Chinese markers for the same essays.

Table 8.3***Mean and Standard Deviation of the Ratings by Two Groups of Markers – Voting***

Marker Group	<i>Content</i>		<i>Organisation</i>		<i>Vocabulary</i>		<i>Language Use</i>		<i>Mechanics</i>		<i>Holistic Scores</i>	
	(0–3)		(0–3)		(0–3)		(0–3)		(0–3)		(0–5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Australia	1.67	0.73	1.82	0.69	1.93	0.59	1.85	0.71	2.18	0.72	2.47	0.94
China	1.45	0.66	1.46	0.64	1.64	0.65	1.57	0.72	1.64	0.76	2.98	0.63
Mean Difference (MD) and paired t test results	MD: 0.22, $t(119)=3.87$, $p<0.001$		MD: 0.35, $t(119)=6.16$, $p<0.001$		MD: 0.29, $t(119)=5.86$, $p<0.001$		MD: 0.28, $t(119)=4.68$, $p<0.001$		MD: 0.54, $t(119)=8.25$, $p<0.001$		MD: -0.51, $t(59)= -5.33$, $p<0.001$	

Table 8.4***Mean and Standard Deviation of the Ratings by Two Groups of Markers – Tobacco***

Marker Group	<i>Content</i>		<i>Organisation</i>		<i>Vocabulary</i>		<i>Language Use</i>		<i>Mechanics</i>		<i>Holistic Scores</i>	
	(0–3)		(0–3)		(0–3)		(0–3)		(0–3)		(0–5)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Australia	2.02	0.69	2.16	0.66	2.27	0.59	2.23	0.64	2.26	0.61	2.76	1.11
China	1.60	0.57	1.58	0.58	1.73	0.51	1.70	0.60	1.81	0.62	3.08	1.29
Mean Difference (MD) and paired t test results	MD: 0.41, $t(119)=9.03$, $p<0.001$		MD: 0.58, $t(119)=12.28$, $p<0.001$		MD: 0.54, $t(119)=11.25$, $p<0.001$		MD: 0.53, $t(119)=9.68$, $p<0.001$		MD: 0.45, $t(119)=8.94$, $p<0.001$		MD: -0.32, $t(59)= -2.78$, $p=0.007$	

Note: p values are two tailed p values. Each observation in the analysis is calculated as the average of the first and second ratings each essay received in each rating situation, by markers from the same country.

Australia: marker group from Australia; China: marker group from China. Due to rounding errors, the mean differences reported might not be exactly the same as the differences in the two means reported.

A consistent pattern from the results in Tables 8.3 and 8.4 is that, for analytic scoring,

Chinese markers were more severe than the Australian markers across all traits and across

both prompts. However, for the holistic scoring, they were consistently more lenient in their

judgements than the Australian markers for both prompts. Paired t-test results show that the analytic scores assigned by Chinese markers for each trait were statistically different (lower) than those assigned by Australian markers, with the mean difference ranging from 0.22 for *Content* on responses to the Voting prompt to 0.58 for the *Organisation* trait on responses to the Tobacco prompt. However, the same group of Chinese markers assigned statistically higher holistic scores than the Australian markers, with a mean difference ranging from 0.32 for the Tobacco to 0.51 for the Voting prompt.

Some of the above results are perhaps not surprising as they are consistent with findings from other studies (e.g., Fayer & Krasinski, 1987; Santos, 1988). For example, Santos (1988) also found that university professors of Non-Native English Speaking background (NNS) were more severe on language errors than professors of Native English Speaking (NS) background. According to him, a possible reason for this difference in severity was because these NNS markers “had invested in learning a language themselves, which led to them to attribute errors to a lack of commitment on the learners’ part” (Santos, 1988, p. 85, as cited in Erdosy, 2004, p. 6).

There seem to be, however, no readily available explanations for why the NNS markers and NS markers in this study used the holistic and analytic rating scales very differently. Although there is some evidence suggesting that markers’ linguistic background might impact on their perceptions of language proficiency and their assumptions of language acquisition which in turn might affect how they use the rating scales (e.g., Erdosy, 2004), more studies are needed

to understand the complex interactions that may exist between markers' linguistic and cultural background, their prior training and marking experience, and their use of rating scales.

Additional analysis points to further differences in the marking behaviour of two groups of markers. The group of Chinese markers was found to be less internally consistent than the Australian group in marking for all analytic traits across both prompts, except for the *Vocabulary* trait. Figures 8.3 and 8.4 report the exact agreement rates (i.e., the proportion of the times two markers marking the same essay on the same trait agreed exactly) amongst Chinese and Australian markers respectively, across the two prompts.

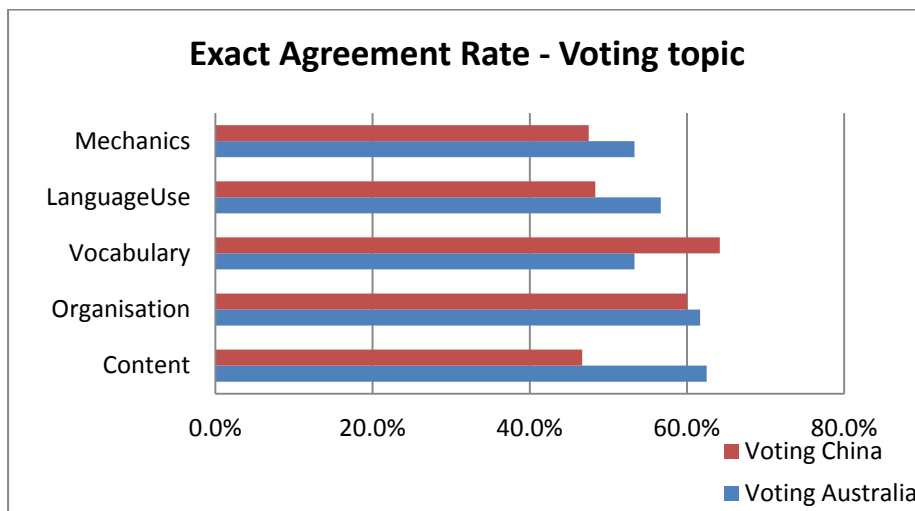


Figure 8.3 Exact Agreement Rates for Australian and Chinese Markers – Voting

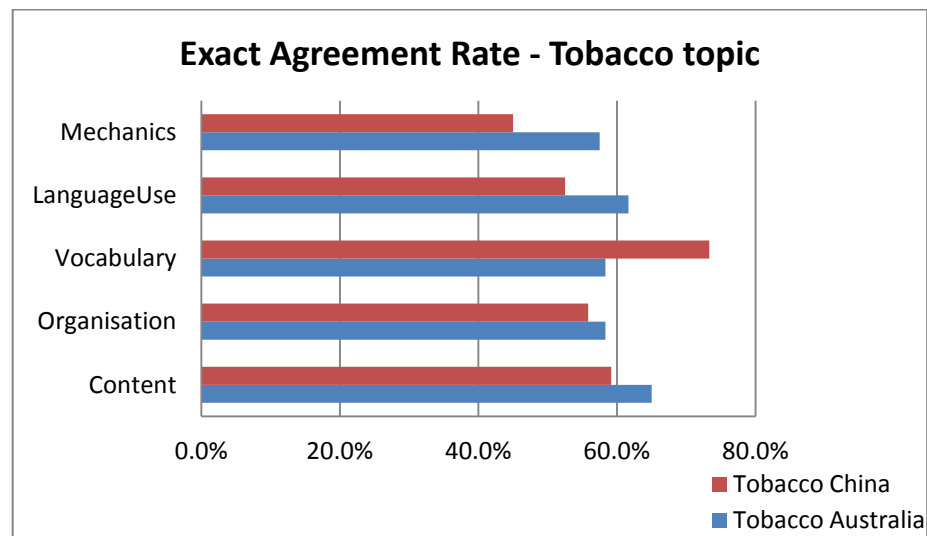


Figure 8.4 Exact agreement Rates for Australian and Chinese Markers – Tobacco

The analysis performed in this section clearly indicates that the differences in the rating behaviour of the two groups of markers (i.e., the Australian and the Chinese markers) are not random. The two groups of markers cannot be considered interchangeable. Consequently taking the average of the raw marks from these two groups will corrupt the meanings of the average scores produced, even though such a practice may improve the ‘reliability’ of the resultant scores. For this reason, all G-theory results presented above and all further validation analysis to be presented in this or later chapters only use scores from the Australian markers. The interpretations of the results from this study, where only scores produced by the Australian human markers are used, are then necessarily only generalisable to the universe of the Australian markers who have levels of education and prior teaching and marking experiences similar to this group of Australian markers.

In summary, the G-theory analysis indicates that the holistic and the composite analytic scores produced by the Australian markers generally reach an acceptable level of reliability that would normally be required for the high-stakes tests. This lends some credibility for using these scores as the external criterion measures for the investigation of the properties of the scores produced by the IEA. However, there is a concern that the data used to train the IEA model may not have reached the desirable level of reliability, which may impact on the quality of scores generated by the IEA.

The next two sections report the correspondence rates between the scores produced by human markers and those by the IEA to help illustrate the utility of the IEA system.

8.3 Correspondence Rates Between Overall Human Scores and the IEA Scores

The analysis in this chapter focuses on the level of correspondence between the overall score produced by the IEA and that produced by the human markers from this study for the same essay, when overall scores are converted to a 0–5 scale. PTE Academic uses the overall score produced by the IEA for the writing task to calculate and report a total score for a test taker on the entire test (Pearson, 2011b, p. 7).

The IEA scores used in these analyses are those calculated in accordance with the formula used by PTE Academic in real rating situations for an independent writing task. In such a rating situation, PTE Academic has two minimum requirements on content and length. If an

essay has fewer than 120 words or more than 380 words (i.e., the score assigned for the *Formal Requirement* trait is 0) or if the essay does not properly deal with the prompt, or the response is not English or irrelevant (i.e., the *Content* score is 0), then the essay will not be scored on any other traits and the overall score allocated will be 0 (Pearson, 2011b, p. 59). When an essay satisfies these two minimum requirements, the sum of the scores on all seven IEA traits is calculated. Since the raw scores on a trait can be negative and can be slightly above the maximum score allowed for the trait,²² the raw scores are first transformed to be within the permissible range for each IEA trait before they are summed. In other words, the negative scores and scores above the maximum score allowed are first set respectively to the minimum and to the maximum scores specified for the trait assessed (Jinshu Li, Pearson, personal communication, January 14, 2010). The summed score is then divided by the maximum possible score. In this case, the maximum possible score is 15, since all IEA traits are scored on a 0–2 scale except for the *Content* trait which is scored on a 0–3 scale. The result (i.e., the overall score represented as a proportion of the maximum possible score) is then transformed to a nominated scale, say a 0–5 scale, by multiplying the proportion by 5. This method of calculating an overall score is similar to that used in the reliability study published by Pearson (Automated Scoring Writing, 2009, p. 2). This study uses this method to

²² The raw IEA-generated trait scores received from Pearson for this study contained negative scores which represented 0.5% of the total number of trait scores received. On the other hand, 0.7% of scores received were above the maximum score allowed for the trait assessed. No explanation is found in the publicly available documents released by Pearson for these out-of-range scores. A possible reason could be the standardisation and normalisation processes undertaken by IEA before trait scores were produced.

transform the IEA trait scores for each essay to an overall score on a 0–5 scale. This score is referred to hereunder as the IEA score.

The corresponding human scores are those assigned by markers in this study using the holistic rating scale (i.e., TOEFL Independent Writing Scale 0–5), and the analytic rating scale (i.e., the modified ESL Composition Profile) respectively. As discussed in Section 5.2.2, both scales chosen are appropriate for use with the PTE Academic writing tests.

Each essay receives two final scores, one each from human markers using the holistic scale and the analytic scale respectively. The final holistic score is calculated as the average of the two holistic scores each essay received, or in the case of an adjudication, the average of the closest two holistic scores. This score is already on a 0–5 scale, and is referred to hereunder as the “human holistic score”. In the case of the human analytic score, a final score for each trait is first calculated by taking the average of the closest two human ratings. The overall analytic score is then calculated by combining the final trait scores using Formula 8.1. The maximum value of the resultant score is 3, since all traits are scored on a scale of 0–3. This score is then converted to be on a 0–5 scale, by dividing it by 3 and multiplying the result by 5. This final score is referred to hereunder as “the human analytic score”.

8.3.1 Descriptive Statistics for the Human Analytic Score, the Human Holistic Score and the IEA Score

Table 8.5 compares the means and standard deviations of the final scores produced by the human markers recruited for this study, using either the analytic scale or the holistic scale, to those of the IEA scores, for the same set of essays (N=60 per prompt).

Table 8.5

Mean and Standard Deviation of the Final Human and IEA Scores

	IEA overall score		Human holistic score		Human analytic score		Effect Size (Cohen's <i>d</i>)	
	Mean	SD	Mean	SD	Mean	SD	IEA-holistic	IEA-analytic
Voting	2.39	1.23	2.47	0.94	3.05	0.91	-0.07	-0.61
Tobacco	2.75	1.29	2.76	1.11	3.57	0.92	-0.01	-0.73

Note: N=60 for each prompt, because only 60 essays per prompt were marked by the IEA and were also double-marked holistically and analytically by human markers. A negative effect size means the IEA mean score is less than the corresponding human mean score. IEA-holistic: comparison between IEA mean score and human mean holistic score. IEA-analytic: comparisons between IEA mean score and human mean analytic score.

It is first noted from Table 8.5 that human markers, on average, assigned higher scores to the essays written to the Tobacco prompt than to those written to the Voting prompt, irrespective of the scoring procedure used to produce the scores. This seems to reflect the views expressed at the group discussions that the Tobacco prompt may be an easier prompt to write to than the Voting prompt for test takers with English as a Second Language background. This pattern is replicated in the IEA scores. In this case, the IEA scores detected the same performance pattern across the two prompts as the human scores, which can be seen as supporting evidence

for the level of accuracy of these IEA scores. It is cautioned that this performance pattern should not be interpreted as evidence of the relativity in prompt difficulty across the two prompts, because different groups of test takers responded to the two prompts. Therefore the differences in the mean scores of the two prompts could be attributable to the differences in the underlying abilities of the test takers who responded to the prompts.

It is also clear from Table 8.5 that the distributional properties of the human scores are influenced by the type of the scoring procedures human markers used to score the essays, and this in turn changes the evaluative outcomes when IEA scores are compared to the human scores. Table 8.5 shows that the mean of the IEA scores is close to the mean of the human holistic scores, but noticeably lower than that of the human analytic scores. Results from the paired *t*-tests (two-tailed) confirm that the IEA scores are not statistically different from the human holistic scores, for both prompts [$t(59) = -0.60, p = 0.55$ for the Voting prompt; $t(59) = -0.05, p = 0.96$ for the Tobacco prompt]. However, the IEA scores are statistically different (lower) than the human analytic scores [mean difference = $-0.66, t(59) = -5.36, p < 0.001$ for the Voting prompt; mean difference = $-0.82, t(59) = -7.81, p < 0.001$ for the Tobacco prompt]. The estimated effect size, Cohen's *d* (1988), is 0.61 for the Voting prompt and 0.73 for Tobacco. By Cohen's (1988) conventional criteria, both can be viewed as a medium effect size. It is noted that the observed similarity or dissimilarity between the IEA scores and the human scores are confounded by the differences in the rating scales used by the human markers and the IEA. Accordingly the above results should be interpreted as what would happen to the means of the scores if a human double-marking process using a particular rating

scale is switched to the IEA scoring process with the IEA scoring rubric. One clear finding is that the choice of the scoring procedures used by the human markers to produce the human scores can have a direct bearing on the human scores produced, and this can affect the outcome of the evaluation when human analytic or holistic scores are used as criterion measures for comparisons with the IEA scores.

8.3.2 Correlations Between Human and IEA Scores

Although different rating scales used in the human and the IEA scoring processes may result in different distributional properties of the scores produced, the expectation is that, overall, human and IEA scores should line up closely, in terms of the rank order of persons produced, considering that these scoring processes are meant to measure the same underlying writing ability. For this purpose, Tables 8.6 and 8.7 report the Spearman's rank order correlation coefficients (r_s) between the final human scores (holistic and analytic scores respectively) and the IEA scores, across both prompts. The inter-marker reliability obtained from the first and second human ratings is also reported in the same tables. The choice of Spearman's correlation coefficients over Pearson product-moment correlation coefficients in this section is to produce rank order correlations that are less skewed by the distributions of the IEA scores (Bachman, 2004). Kolmogorov-Smirnov tests of normality show that the IEA scores deviate significantly from the normal distribution, for both prompts. A main reason for the significant departure in the IEA scores from the normal distribution is the relatively large number of essays that received a 0 total score from the IEA. This is further explained in the

next section. However, as Pearson correlation coefficients are the most commonly used correlation statistics in the AES literature, they are also calculated and reported at Appendix J.

Table 8.6

Spearman's Correlation Coefficients (r_s) Between Human Analytic Scores and the IEA Scores

Prompts	IEA/ Marker 1	IEA/ Marker 2	Marker 1/ Marker 2	IEA/ Human analytic score
Voting	0.71	0.75	0.81	0.77
Tobacco	0.72	0.79	0.75	0.81

Table 8.7

Spearman's Correlation Coefficients (r_s) Between Human Holistic Scores and the IEA Scores

Prompts	IEA/ Marker 1	IEA/ Marker 2	Marker 1/ Marker 2	IEA/ Human holistic score
Voting	0.68	0.66	0.71	0.73
Tobacco	0.71	0.78	0.77	0.80

The following notes apply to both Tables 8.6 and 8.7:

All correlations are significant at the 0.01 level (2-tailed). N=60 for holistic scores; N=120 for analytic scores

IEA/Marker 1: correlation between the IEA scores and the human scores calculated based on the first human ratings;

IEA/Marker 2: correlation between the IEA scores and the human scores calculated based on the second human ratings;

IEA/Human score: correlation between the IEA overall scores and the (adjudicated) final human scores.

Overall, the Spearman correlation coefficient (r_s) indicates there is a relatively strong relationship between the final human scores and the IEA scores, across both prompts,

irrespective of the scoring procedure (analytic or holistic) used by the human markers to generate the human scores. The correlation coefficient ranged from 0.73 (between the human holistic scores and IEA scores for the Voting prompt) to 0.81 (between the human analytic scores and the IEA scores for the Tobacco prompt).

A second observation that can be made from Tables 8.6 and 8.7 is that the IEA scores seem to have a stronger relationship with the human analytic scores than they do with human holistic scores. The correlation between the IEA scores and the human analytic scores is 0.77 for the Voting prompt, and 0.81 for the Tobacco prompt. This is better than the corresponding correlations between the IEA scores and the human holistic scores (0.73 and 0.80 for the Voting and Tobacco prompts respectively), though the difference in correlation coefficients for both pairs is not substantial.

These results seem to suggest that the IEA scores are more consistent with human analytic scores than with holistic scores, when rank ordering of the persons is concerned. This is contrary to the earlier observation of the IEA scores being more similar to the holistic scores than to analytic scores, when the average score of the persons is concerned. More studies with larger sample sizes could be repeated to examine the generalisability of these results.

A third observation is that the IEA scores correlate better with the final human scores than with the scores from the individual markers (i.e., with the scores based on the first and the second ratings respectively). This is a desirable feature of the system.

One further observation from Tables 8.6 and 8.7 is that the IEA's performance in reproducing the same rank order of persons as the human scores varies across the prompts. While the correlations between the human and the IEA scores across the two scoring procedures were 0.73 and 0.77 for the Voting prompt, the corresponding correlations were 0.80 and 0.81 for the Tobacco prompt. This indicates caution in generalising the IEA's performance based on a small number of writing prompts.

The corresponding Pearson product-moment correlations (r) between human scores and IEA scores (Appendix J) indicate that the average Pearson correlation is 0.75 across the two prompts and across the two scoring procedures.²³ All other observations mentioned above remain unchanged if the Pearson correlation statistics were used.

8.3.3 Correlations Between Human and IEA Scores after the Removal of the Two Minimum Requirements

It is noted that the IEA scores used in the previous section have been subjected to the two minimum requirements by IEA, and this may have skewed the results. This is because human markers did not use length as a minimum requirement for producing a total score, irrespective of the type of scoring procedures used during the marking processes. Furthermore, if markers did implement a minimum content requirement during the marking processes in accordance

²³ The average correlation coefficient is calculated using the Fisher's r - z ' transformation. See Fisher (1915, 1921) for the transformation formula.

with the scoring rubrics used, they were likely to have implemented this requirement in a different way to the IEA.²⁴

In total, 46 essays, out of 240 in the sample, across the two prompts were given a score of 0 by the IEA. Forty of them were because of the minimum length and content requirements imposed by the IEA. Of the two minimum requirements, the length requirement was the main reason, which would cause all forty essays to receive a score of 0 from the IEA. Five of these forty essays would have been given a score of 0 under the minimum content requirement alone. The validity issues associated with the two requirements imposed by the PTE Academic, including those arising from the accuracy of content scoring by the IEA for essays at the low end of the achievement scale, are explored more fully in Chapter 12.

Since the two requirements were either not implemented or implemented potentially in a different manner by human markers, the IEA overall scores were recalculated without the imposition of the two minimum requirements. The correlations between the new IEA scores

²⁴ The scoring rubric used by human markers during their holistic marking processes indicated an essay that either “rejects the topic, or is not connected to the topic, (or) is written in a foreign language”, should receive a score of 0 (Appendix E). This rule, however, is potentially different from the content requirement implemented by the IEA which applies to all essays that have been deemed to “not deal properly with the topic” (see Appendix A). The scoring rubric used by human markers during their analytic scoring processes (Appendix D) did not specify a content-related requirement, though the markers agreed at the training session that if an essay was off-topic, it should receive a total score of zero. Again, this treatment is potentially different from the minimum content requirement implemented by the IEA. It is noted that the treatment of off-topic essays is largely dependent on the purpose of the test. This point is further discussed in Section 11.4.4.

and human scores are then re-analysed. In this analysis, Pearson product-moment correlations are used, as there do not seem to be substantial departure in the IEA scores from normal distributions,²⁵ after the removal of the two minimum requirements.

Another focus of the analysis in this section is the partial correlation between human and IEA re-calculated scores, after the influence of essay length on the corresponding sets of human and IEA scores is removed. This type of analysis is important because prior research has repeatedly demonstrated that essay length alone can predict human scores to a great extent (Breland & Jones, 1984; Chodorow & Burstein, 2004; Kaplan et al 1998). It is therefore crucial to collect independent evidence to demonstrate that “the automated scores do not amount to counting words” (Attali, 2007, p. 2) and that the relationship between IEA scores and human scores is to some degree independent of essay length. The removal of the two minimum requirements from the IEA overall scores for the purpose of this analysis has the added benefit of minimising bias in the reporting of correlations as the inclusion of an essay length criterion in the IEA scores would have over-reported the sensitivity of IEA trait scores to the essay length variable. To contrast with the partial correlations, Pearson product-moment correlations are referred to as “zero order” correlations in the following paragraphs.

²⁵ After the removal of the two minimum requirements, Kolmogorov-Smirnov tests of normality (using SPSS 18) indicate that the IEA scores for the Voting prompt, human analytic and holistic scores for Tobacco prompt are all approximately normally distributed (all p values > 0.05). For the other score sets, visual inspections of score distributions (through histograms and normal Q-Q plots) indicate that the departure in the individual datasets from a normal distribution does not seem to be substantial.

Results of both zero-order and partial correlation analysis are presented in Table 8.8. If the raw relationship observed between human and IEA scores, as demonstrated through the zero-order correlations, is completely due to the influence of essay length, the corresponding partial correlations controlling for the word count should show no statistically significant relationships between the human and the IEA scores.

Table 8.8

Correlations Between Human (Analytic and Holistic) Final Scores and the IEA Scores – Calculated Without the Two Minimum Requirements

Prompt	Correlations between	IEA scores calculated without the two minimum requirements	
		Pearson zero-order r	partial r (after the removal of essay length)
Voting	Human Analytic Score	0.79	0.67
	Human Holistic Score	0.64	0.46
Tobacco	Human Analytic Score	0.84	0.66
	Human Holistic Score	0.80	0.62

Note: all zero-order and partial correlations are significant at 0.01 level (two-tailed)
 N=120 per prompt for correlations between IEA and human analytic scores. N=60 per prompt for correlations between IEA and human holistic scores.

Table 8.8 shows that the removal of the two minimum requirements has improved the predictive relationship between the IEA scores and human scores (compare Pearson zero-order correlations in Table 8.8 with the original Pearson correlations in Appendix J). The zero-order correlations between the new IEA scores (i.e., without the two minimum

requirements imposed) and the human scores are at the 0.8 level across two prompts, with one exception – the correlation (i.e., 0.64) the new IEA scores have with the human holistic scores for the Voting prompt. This exception is more likely a reflection of the inconsistency in the human scores, than of the inaccuracy in the IEA scores, because the human scores generated from the holistic scoring process for the Voting prompt were found to be the least dependable of all sets of human scores (see discussions in Section 8.1). The relatively less reliable human scores are likely to have resulted in the three outliers (i.e., the three pairs of human–IEA scores) that are easily identifiable from a scatter plot of the human holistic scores and the IEA scores for the Voting prompt. These outliers would have influenced the Pearson correlation statistic as this type of statistic is sensitive to outliers.

The partial correlation coefficients indicate that after the effect of length is removed, there is still a statistically significant and moderate association between human scores and the IEA scores across the prompts, irrespective of the scoring procedures used to produce the human scores. With the exception of the human holistic scores for the Voting prompt, on average, the IEA scores achieve a sound partial correlation of 0.65 with the human scores. This indicates that the level of agreement the IEA scores (calculated without the two minimum requirements) have with human scores is independent of the influence of the essay length to a great extent.

The same analysis performed in this section was also carried out based on the IEA scores that were recalculated with only the length requirement removed. Results are very similar to the ones presented in Table 8.8. Therefore they are not separately reported.

8.4 Chapter Summary

This chapter has demonstrated the importance of including examinations of the quality of criterion measures used in an AES validation study, as an integral part of the study. It first investigated the error in generalising from observed human overall scores to expected scores on the universe of generalisation, which helped illustrate the potential pitfall of using human scores as external criterion measures to validate machine scores. Results from G-theory analysis indicated that the dependability of the human scores acquired by this study, either from using a holistic or an analytic scoring procedure, generally reached the acceptable level of reliability required for high-stakes tests. The same, however, may not be said of the holistic scores and scores on individual traits produced by markers from the Pearson field tests. Chapters Nine and Ten will continue to explore the measurement and structural qualities of the human scores that are used in this study as external criterion measures.

When comparing the IEA total scores to the human total scores, this chapter found that on average, the IEA scores correlated strongly with human total scores, irrespective of the scoring procedure used by human markers to produce these scores. It further established that the level of agreement the IEA scores had with human scores was independent of essay length to a great extent.

However, there were still some differences in the evaluative outcomes when different sets of human scores were used. For example, while the IEA scores were not statistically different from human holistic scores, they were statistically different from human scores generated from using an analytic scoring procedure. On the other hand, when IEA scores were compared to human scores on the basis of the rank order of person scores, the IEA scores were observed to be more consistent with the human analytic scores than with the human holistic scores. These observations highlighted the necessity to take great care when evaluating AES systems because many factors unrelated to AES systems, such as the scoring procedures used to produce human scores and the types of analysis chosen to compare human scores to the IEA scores, could change the interpretations of the evaluative outcomes.

The next three chapters aim to accumulate more direct evidence supporting or challenging the validity of the machine-generated scores.

Chapter 9 Measurement Properties of the Intelligent Essay Assessor (IEA) and Human Scores

9.1 Introduction

This chapter demonstrates how evidence relevant to the measurement aspect of validity – the third component of the AES (Automated Essay Scoring) framework proposed in Chapter Four – may be collected and evaluated. More particularly the chapter examines the measurement characteristics of scores produced by the IEA (Intelligent Essay Assessor). The importance of this part of the study is that Automatic Essay Scoring (AES) is primarily associated with producing valid measures of constructs (Bennett, 2004) and there has to be evidence that the marks that are produced are in fact measures, and not numbers assigned to objects, as these do not suffice as scientific measurement (Michell, 1997).

In order for valid inferences to be drawn from the Pearson Test of English (PTE) Academic writing scores, it is essential to check that the scores are governed by the basic requirements of scientific measurement.²⁶ To achieve this, the chapter focuses on two key questions:

- 1) Do the traits assessed by the IEA cooperate to define a single construct? and
- 2) Do the rating scales used for the scoring of each of the traits function as intended?

²⁶ This thesis uses the term ‘requirements’, rather than ‘technical assumptions’ as used by Kane (2006) to describe these basic measurement requirements. This is to emphasise the key concept that these requirements are not assumptions that may be taken for granted. There must be empirical evidence to demonstrate these requirements are satisfied before a total score is calculated (Keeves & Alagumalai, 1999).

It has been previously stated that this thesis uses the term “trait” to denote a dimension of writing performance that is evaluated by an AES model²⁷. The first question asks whether the construct being measured exhibits the characteristics of uni-dimensionality. This is required if a single score is being used to summarise the performance of persons. It has been established that the Independent Writing Task component of PTE Academic produces a single score as a measure of each individual’s writing ability. This score is derived essentially by summarising performances on seven writing traits as scored by IEA. Implicit in this practice of using a single score to summarise a person’s overall writing performance is the requirement for the score to be represented on a single uni-dimensional continuum (Tognolini, 1989). Explicitly, this requires evidence of traits measuring the same underlying construct. Only when this requirement is met, can scores on these traits be used to form a single score.

It is noted that uni-dimensionality is a relative concept (Andrich, 1988). It is common that, when developing an achievement construct, there are numerous sets of skills that are closely related to the main skill being assessed, but qualitatively quite different. For example, while the seven writing traits assessed by IEA for the PTE Academic writing tests may be reflections of a single writing ability construct, they are also “multi-dimensional” in the sense that each assesses a unique aspect of writing performance requiring somewhat different

²⁷ This usage of the term ‘trait’ in this thesis is prevalent in the literature about rater effects. Researchers who are familiar with item response theory (IRT), however, will need to bear this usage firmly in mind when reading this thesis (in particular, this chapter), because, in the field of IRT- and Rasch-trained psychometricians, a ‘latent trait’ means “a construct (or variable) that is operationally defined by a set of items (or tasks) designed to elicit a response from an individual” (Myford & Wolfe, 2003, p. 388).

knowledge and skills. In the context of this study, the practical judgement of the uni-dimensionality requirement should be whether the IEA writing construct for a particular prompt is sufficiently uni-dimensional that a single score is useful for the purpose of making university/college admission decisions.

The second question relates to the requirement that the rating scale used by the IEA to measure each writing trait is functioning as intended and is producing observations that contribute to the development of a single achievement construct. Explicitly, this requires that the scale categories on each of the rating scales are used meaningfully, and together these categories serve to define the ability continuum.

Both requirements must be empirically tested before the meaning of the PTE Academic writing scores can be understood and scores accepted for the practical purposes of measurement and comparison. The next section describes the measurement model used in this study to investigate both the uni-dimensionality and the rating scale functionality requirements.

9.2. The Rasch Model Used in this Study

A modern approach to assess uni-dimensionality involves analysing the data according to a uni-dimensional measurement model to determine the extent to which the data conforms to the requirement of the model. A useful model for this type of analysis is the Rasch Model (Rasch, 1960, 1980). Two fundamental principles underpin this model. The first is that a

single ability underlies the developmental sequence as represented by the criteria used to score performance (Bond & Fox, 2001). The second is that a person's likelihood of success on an item is dependent on the interaction between the ability of the person and the difficulty of the item, both of which are measured and calibrated on the same construct (Rasch, 1960).

A proven important feature of the Rasch model is that the total score (e.g., count of correct responses in an achievement test) is a sufficient statistic for calculating the ability of the person achieving it (Choppin, 1982; Rasch, 1960, 1980). This means that when the data conforms to the model, the total score captures the entire profile of the observed scores on the test items.

A special form of the Rasch model is the Rasch Rating Scale Model (Andrich, 1978). This model is highly applicable in achievement tests, such as the PTE Academic writing tests, where performances on a writing trait are judged in ordered categories on the rating scale such as "Poor, Fair, Good, and Excellent", or "0", "1", "2", "3" in accordance with pre-defined scoring criteria.

When a rating scale is used in assessments, an important requirement is that a higher score category on the scale, in general, should imply more of the underlying ability and vice versa. This requirement, referred to as the "inferential property" of a rating scale by Linacre (1999), is consistent with the scale definition and with the intended use of the scale. As an illustration, the definition of the IEA *General Linguistic Range* (GLR) scale (see Figure 9.1) indicates a clear progression of ability through the sequential categories on the GLR rating scale (i.e., 0,

1, 2). When this requirement is not met, doubts would be cast on the meaning of the scale and on the validity of the measurement outcomes (Eckes, 2009).

Score Category	Descriptions
2	The essay exhibits mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No sign that the test taker is restricted in what they want to communicate.
1	The essay shows a sufficient range of language to provide clear descriptions, express viewpoints and develop arguments.
0	The essay contains mainly basic language and lacks precision.

Figure 9.1 IEA *General Linguistic Range* Scale (Adapted from Pearson, 2011b, p. 60)

The Rasch Rating Scale model incorporates this concept of order within a framework of unidimensionality. As visually demonstrated in Figure 9.2, the model conceptualises a functional rating scale as dividing the latent continuum into ordered categories, which qualitatively advance along this continuum (Linacre, 2010).

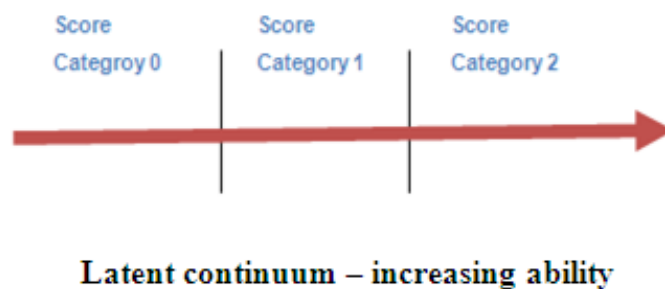


Figure 9.2 *General Linguistic Range* Rating Scale with Ordered Categories

Equation 9.1 is the mathematical expression of the Rasch model used in this study to investigate the measurement properties of scores generated by the IEA. This model is an extension of the general Rating Scale Model (Wright & Masters, 1982).

$$\log (P_{nik}/P_{ni(k-1)}) = B_n - D_i - F_{ik} \quad \text{Equation 9.1}$$

where:

- P_{nik} is the probability of a person n achieving a score category k for a trait i ;
- $P_{ni(k-1)}$ is the probability of this person n achieving an adjacent lower category $(k-1)$ for a trait i ;
- B_n is the ability for person n ;
- D_i is the difficulty for trait i ; and,
- F_{ik} is the impediment to be observed in category k relative to category $(k-1)$, on the particular trait i .

This measurement model calibrates the person ability, trait difficulty and the category thresholds that are specific to each trait in one statistical and measurement framework, with the idea of one single latent continuum along which persons and traits have a unique order. The category threshold (Andrich, 1978), designated F_{ik} , represents “the impediment to be observed in category k relative to category $k-1$ ”, on the particular trait i (Linacre, 1999, p. 103). It is noted that the model, as specified, allows for the rating scale structure to vary across different traits. This specification is appropriate for this study as the IEA uses different forms of rating scales across the seven traits (i.e., *Content* is rated on a 0–3 rating scale, while all other traits are rated on a 0–2 rating scale). The usefulness of the Rasch model is that it offers a number of quality control indicators which can be used to investigate the uni-dimensionality requirement, as well as the degree of empirical consistency in the IEA trait scores to support the intended use of the seven IEA rating scales in generating an overall

writing score in PTE Academic for each person. The next section describes the indicators used in this study to ascertain the degree of accord between the data and the model.

9.3 Fit Indicators Incorporated in Rasch Analysis

Global Model Fit

The overall data-model fit can be investigated by the distribution of standardised residuals (of the writing trait scores). The residuals indicate the difference between the actual observed score and the score value expected by the Rasch model, when a person's performance on a writing trait is observed. If data fits the Rasch model sufficiently well, the standardised residuals should be close to a normal distribution (i.e., $N(0,1)$) (Linacre, 2010). Otherwise, the extent to which the distribution of standardised residuals deviates from a normal distribution is an indication of the extent to which the data does not accord with the requirements of the model: either because the writing traits are not measuring the same construct, or because there are other sources of variance in the data. Satisfactory model fit is indicated when about 5% or fewer of all the responses have (absolute) standardised residuals ≥ 2 and about 1% or fewer have (absolute) standardised residuals ≥ 3 (Linacre, 2008).

Parameter Level Fit Analysis

When the global model fit analysis reveals significant misfit, more detailed parameter level fit analysis can be carried out to investigate where the misfit may originate. The Rasch model provides two types of fit statistics which are indicators of how well an observed response pattern fits the measurement expectations. These two types of fit statistics are referred to as

INFIT and OUTFIT mean-squares (Wright & Stone, 1999). Both are chi-square ratios based on the standardised residuals. While the OUTFIT statistic is an unweighted statistic which is heavily influenced by outlying, off-target, unexpected responses, the INFIT is sensitive to irregular inlying patterns with relatively more impact being given to unexpected responses close to a person's or item's measure (Wright & Masters, 1982; Wright & Stone, 1999). Both mean-square statistics have an expected value of 1.0, and a range from 0 to positive infinity. Values less than 1.0 indicate over-fit; that is, data is too predictable with respect to model expectations, causing summary statistics such as reliability indices, to report inflated results. Values greater than 1.0 indicate under fit; that is, there is more un-modelled noise in the data than expected. High mean-squares are considered a much greater threat to the validity than low mean-square values, because they suggest a possible violation of the uni-dimensionality requirement (Linacre, 2002, 2010; Myford & Wolfe, 2003).

When assessing item fit, there are no fixed rules for determining mean square values that are too large or too small, as the interpretation of mean-square indicators depends on the particular features of the testing situations (Bond & Fox, 2007; Eckes, 2009). For this study, a relatively conservative range defined by a lower-control limit of 0.7 and a higher-control limit of 1.3 is used (Adams & Khoo, 1993). This choice is based on the fact that the PTE Academic tests are high-stakes and they use a criterion-based rating scale scoring rubric.

Mean-square values are also reported in various standardised forms, such as the INFIT and OUTFIT z-standardised t-statistics reported by the Winsteps Rasch computer program. The statistical convention is that when the absolute value of a standardised t-statistic is greater

than 2 (i.e., $p < 0.05$), the null hypothesis that the data fits the Rasch model (perfectly), should be rejected.

9.4 Rasch Analysis Performed in this Study

In this study, the Winsteps computer program (Linacre, 2010) is utilised to provide the psychometric analysis concerning the IEA scores, with the writing traits being treated as items in the model.

The analysis comprises four parts. Part I analyses the overall model fit. Part II focuses on trait-level fit analysis. Part III analyses the functionality of the trait rating scales which are employed by the IEA to score the 7 traits. Part IV investigates the relationships between the Rasch person ability measures constructed from the IEA trait scores and from the human trait scores for the same group of people, in order to address the question as to whether the two scoring methods are measuring the same achievement construct in the same manner. This analysis is considered to provide more useful information about the extent of the differences that may exist in the two different scoring methods than the use of simple correlations between raw scores. This is because the analysis uses person ability estimates that have better measurement qualities than raw scores, and it uses them in the context of the measurement errors that are associated with these ability estimates.

Comparative analysis involving human scores

The same Rasch analyses are also conducted on trait scores generated by human markers, across the two prompts. This is done in order to compare the measurement properties of the human scores generated from the double-marking process to those of the IEA scores and thus provide further evidence of the extent of differences or similarities which may exist in the human scoring and IEA scoring methods.

In the present study, each essay was also scored by two human markers on each of the five traits: *Content*, *Organisation*, *Vocabulary*, *Language Use* and *Mechanics*, using the modified ESL Composition Profile 0–3 rating scales (Appendix D). As two markers were used to mark each essay, the Rasch rating model included additional parameters to account for differences in the severity of the markers. The resultant measurement framework is the Many Facet Rasch Measurement (MFRM) (details see Linacre, 1989; Linacre & Wright, 2002).

Equation 9.2 specifies the MFRM measurement model that is used for all the Rasch analyses involving human data where more than one marker was used per essay:

$$\log (P_{nij k} / P_{nij (k-1)}) = B_n - D_i - C_j - F_{ik} \quad \text{Equation 9.2}$$

where:

- $P_{nij k}$ is the probability of person n being awarded, on trait i by marker j , a rating of category k ;
- $P_{nij (k-1)}$ is the probability of person n being awarded, on trait i by marker j , a rating of category $(k-1)$;
- B_n is ability of person n ;
- D_i is the difficulty of trait i ;
- C_j is the severity of marker j ; and,
- F_{ik} is the threshold of being observed in category k relative to category $(k-1)$, on trait i

The above model conceptualises that the person ability, marker severity, trait difficulty and the way in which the markers apply the rating scales, dominate the scores given to a person's writing performance on a particular trait. It is noted that the model specified above allows the structure of the rating scales to vary across different traits but to hold constant across markers. Therefore the threshold estimates (F_{ik}) relate to how markers, as a group—not as individuals—used the modified ESL Composition rating scales in different ways across different traits. All Rasch analysis involving human data was carried out using the FACETS computer program (Linacre, 2008).

In order to prepare data for the Rasch analysis of the IEA scores, raw continuous trait scores assigned by the IEA were converted to be within the permissible score range specific to each trait, before being rounded to discrete score points on the respective rating scale for the trait. No transformations of human scores were necessary as they were already discrete score points (0–3) on the respective rating scales.

9.5 Results of Rasch Analysis

9.5.1 Part I – Global Model Fit Statistics

Table 9.1 shows the distributional statistics for the standardised residuals generated from Winsteps for the IEA scores and from FACETS for human scores.

Table 9.1

Distribution of Standardised Residuals

		Voting		Tobacco	
		Human	IEA	Human	IEA
Standardised Residuals	Mean	-0.01	0.06	0.00	0.17
	SD	1.01	1.38	1.00	2.63

		Voting		Tobacco	
		Human	IEA	Human	IEA
% of the total responses having (absolute) standardised residuals ≥ 2		4.50%	4.90%	4.90%	3.90%
% of the total responses having (absolute) standardised residuals ≥ 3		0.40%	1.90%	0.40%	2.50%

The IEA scores show that, while the mean of the standardised residuals is close to the expected value of 0, the standard deviation (SD) of the standardised residuals across both prompts is noticeably greater than the expected value of 1.0 (Voting: 1.38; Tobacco: 2.63). This indicates that there is considerably more noise in the IEA trait scores than the Rasch model expects. In addition, the percentage of extremely unexpected responses (i.e., trait scores) associated with the absolute standardised residuals ≥ 3 exceeds the usual limit allowed

for satisfactory model fit (i.e., 1.0%), consistently so across both prompts (Voting: 1.9%, Tobacco: 2.5%). The percentage of trait scores associated with the absolute standardised residuals ≥ 2 , however, is within the usual limit recommended (i.e., 5%), across both prompts (Voting: 4.9%; Tobacco: 3.9%).

On the other hand, when human trait scores for the same samples of essays were analysed by the FACETS model, the mean and the SD of the standardised residuals across both prompts are extremely close to the expected values of 0 and 1.0 (mean and SD of -0.01 and 1.01 for Voting; 0.00 and 1.00 for Tobacco). The percentage of the unexpected responses associated with absolute standardised residuals ≥ 2 in the human data is within the recommended limit for reasonable model fit, across both prompts. Furthermore, percentages of responses that are deemed as extremely unexpected by the Rasch model (absolute standardised residuals ≥ 3) are noticeably less than those detected in the IEA scores, and are contained well within the limit recommended, across both prompts (0.4% for both prompts).

Overall, the distribution patterns of the standardised residuals for these prompts suggest that traits assessed by humans, on a global level, fit a uni-dimensional Rasch model sufficiently well, however, there appear to be some deviations in the IEA data from a uni-dimensional measurement framework. The next section focuses on the fit analysis on a trait level, to determine the genesis of the overall misfit.

9.5.2 Part II – Trait Fit Analysis

Trait Fit Analysis for the IEA scores

Tables 9.2 and 9.3 report trait fit statistics for the IEA scores, across both prompts.

Table 9.2

Fit Statistics for the Voting Prompt

Trait Name	Difficulty Estimate	Error Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Spelling</i>	2.51	0.18	2.25	5.50	8.48	5.79	0.30
<i>Form</i>	-0.19	0.18	1.23	1.67	1.60	2.66	0.65
<i>Content</i>	-0.73	0.19	0.77	-1.99	0.78	-1.68	0.77
<i>DSC</i>	-0.67	0.25	0.82	-1.24	0.77	-0.96	0.67
<i>GUM</i>	-0.07	0.22	0.79	-1.65	0.72	-1.66	0.74
<i>Vocabulary</i>	-0.42	0.23	0.71	-2.28	0.53	-2.85	0.76
<i>GLR</i>	-0.42	0.23	0.55	-3.89	0.42	-3.74	0.81

Table 9.3

Fit Statistics for the Tobacco Prompt

Trait Name	Difficulty Estimate	Error Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Spelling</i>	3.66	0.20	2.18	5.28	9.90	9.91	0.20
<i>Form</i>	-0.72	0.21	1.03	0.23	1.24	1.03	0.78
<i>Content</i>	0.11	0.21	0.75	-2.00	0.85	-0.89	0.81
<i>DSC</i>	-1.14	0.24	0.86	-1.09	0.88	-0.56	0.76
<i>GUM</i>	-0.68	0.24	0.66	-2.92	0.53	-2.68	0.81
<i>Vocabulary</i>	-0.71	0.23	0.66	-2.98	0.51	-2.86	0.82
<i>GLR</i>	-0.53	0.23	0.6	-3.54	0.46	-3.67	0.85

Note: For Tables 9.2 and 9.3, the IEA traits are abbreviated as follows:

Form – Formal Requirement; *DSC* – Development, Structure and Coherence; *GUM* – Grammar Usage and Mechanics; *Vocabulary* – Vocabulary Range; *GLR* – General Linguistic Range.

The z-standardised t-statistics are denoted as “ZSTD” in the above two tables.

The first observation from Tables 9.2 and 9.3 is that *Spelling* has extremely large misfit statistics (i.e., under-fit), with OUTFIT mean-square values ranging from 8.48 for Voting to 9.90 for Tobacco. This indicates that there is, on average, eight times more noise in the IEA scores than would be expected from the governing Rasch model. Furthermore, the z-standardised t-statistics associated with the mean-square values for *Spelling* (OUTFIT ZSTD, 5.79 for Voting and 9.91 for Tobacco) indicate that the misfit is statistically significant and unlikely to be due to chance. The respective INFIT mean-square values for *Spelling* are 2.25 (ZSTD = 5.50) for Voting and 2.18 (ZSTD = 5.28) for Tobacco, which suggest that the large departure from the model expectations in the *Spelling* scores, is not entirely caused by the unexpected patterns of scores associated with persons who are located far from the trait location on the latent continuum (Linacre, 2010).

A second observation is the misfit associated with the *Formal Requirement* trait. For essays written to the Voting prompt, the OUTFIT mean-square value for the *Formal Requirement* trait is 1.60 (ZSTD=2.66), indicating 60% more noise than expected. Though the INFIT statistic is reduced to 1.23 (ZSTD=1.67), when scores from persons closer to the trait difficulty measure are given relatively more weight, the high OUTFIT value suggests that there are segments of data that do not support useful measurement. This trait seems to fit the Rasch model better for the Tobacco prompt [OUTFIT: 1.24 (ZSTD=1.03), INFIT: 1.03 (ZSTD=0.23)]. Approximately 7% of the scores on the *Formal Requirement* trait are considered as unexpected scores by the Rasch model (i.e., those scores with absolute standardised residuals ≥ 2) across the two prompts.

It would appear as though most of the overall misfit to the model is due to the misfit associated with these two traits. All the other traits seem to fit the model fairly well although the language traits do show some signs of over-fitting (i.e., having low mean square values). However, as Linacre (2010) points out, the average of the mean-squares of all traits is usually forced to be around 1.0. Therefore when there is a trait like *Spelling* with very large mean-square values, there would be counter balancing traits of low mean-squares. To investigate more about the possible over-fitting problem, the two traits *Spelling* and *Formal Requirement* were removed from the data before data was reanalysed. The results of the new analysis are contained in Tables 9.4 and 9.5.

Table 9.4

Trait Fit Statistics for the Voting Prompt When the Spelling and Formal Requirement Traits Are Removed

Trait Name	Difficulty Estimate	Error Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Content</i>	-0.46	0.27	1.16	0.93	1.09	0.41	0.83
<i>DSC</i>	-0.12	0.3	1.15	0.98	1.24	0.58	0.69
<i>GLR</i>	0.05	0.29	0.6	-3.13	0.3	-1.83	0.79
<i>GUM</i>	0.47	0.27	1.1	0.72	0.82	-0.22	0.75
<i>Vocabulary</i>	0.05	0.29	0.77	-1.62	0.42	-1.36	0.77

Table 9.5

Trait Fit statistics for the Tobacco Prompt When the Spelling and Formal Requirement Traits Are Removed

Trait Name	Difficulty Estimate	Error Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Content</i>	-1.68	0.25	0.86	-0.87	0.71	-1.43	0.88
<i>DSC</i>	0.02	0.27	1.26	1.65	0.87	-0.30	0.75
<i>GLR</i>	0.63	0.27	0.78	-1.61	0.53	-1.72	0.86
<i>GUM</i>	0.54	0.27	1.12	0.84	0.76	-0.58	0.77
<i>Vocabulary</i>	0.49	0.27	0.72	-2.1	0.47	-1.74	0.85

Note: For Tables 9.4 and 9.5, the IEA traits are abbreviated as follows: *Form* – *Formal Requirement*; *DSC* – *Development, Structure and Coherence*; *GUM* – *Grammar Usage and Mechanics*; *Vocabulary* – *Vocabulary Range*; *GLR* – *General Linguistic Range*

It can be seen from Tables 9.4 and 9.5 that the five remaining traits fit better to a uni-dimensional model than the original seven traits. All mean-square values are now below the upper-control limit of 1.3, indicating that no trait is under-fitting the uni-dimensional model. This is corroborated by separate global model fit analyses which indicate that across both prompts, the distribution of standardised residuals of responses based on five traits is closer to the expected $N(0,1)$ distribution than the distribution based on the seven traits. The mean and standard deviation of the standardised residuals are (-0.01, 0.82) for the Tobacco prompt and (0.00, 0.88) for the Voting prompt, respectively. However, the fact that the standard deviations are 18% and 12% below the expected value of 1.0 indicates that scores on the five traits show signs of a Guttman pattern and are tending to be too predictable. This is further confirmed by the fit statistics reported in Tables 9.4 and 9.5 for the five traits. Two language

traits, *General Linguistic Range* and *Vocabulary Range*, show signs of data-model over-fit, meaning scores on these two traits are being too predictable. Across the two prompts, the OUTFIT mean-square values for both traits range from 0.30 to 0.53, which are noticeably below 0.7, the lower-control limit. A low mean-square value, such as the 0.30 for the *General Linguistic Range* trait for the Voting prompt, suggests that scores for this trait only have 30% of the randomness the model predicts; that is, they only contain 30% of the measurement information that they should have. Consequently, these results indicate that the *General Linguistic Range* and *Vocabulary Range* traits are less efficient and less productive for measurement than desired (Linacre, 2010). As noted by measurement theorists (e.g., Linacre, 2010; Smith, 1996), possible reasons for an item (or a writing trait, in this case) having a low mean square value include the item being redundant (e.g., the item is assessing the same or very similar characteristics as other item(s)) or category range restriction (e.g., some categories on the rating scales are being overused). In the context of this analysis, the possible reasons for observing that the two IEA traits may be over-predictable could then be that these two traits (as assessed by the IEA) are measuring characteristics of writing that are too similar to those already being assessed by other traits; or the two traits in general are too similar to each other; or some score categories on the rating scales are being overused; or a combination of all the above.

The long line of research into human marking behaviour (e.g., Cooper, 1984; Leckie & Baird, 2011; Myford & Wolfe, 2009; Robbins, 1989; Saal, Downey & Lahey, 1980) suggests that human markers have two very similar problems: the preponderance of markers to overuse the

middle category or particular categories of a rating scale; and the propensity of markers to transfer their judgements on one trait of writing performance to another. The latter tends to result in conceptually different traits being more similar than they should be. The following section examines the psychometric quality of the human scores, including looking at whether Rasch analysis reveals similar problems in human scores.

Trait fit analysis for the human data

Equivalent fit analysis at the trait level is also carried out in FACETS for the human data with results being reported in Tables 9.6 and 9.7.

Table 9.6

Trait Fit Statistics for the Voting Prompt – Human Scores

Trait Name	Difficulty Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Content</i>	-0.06	0.96	-0.45	0.95	-0.51	0.51
<i>Organisation</i>	-0.69	0.89	-1.24	0.88	-1.32	0.53
<i>Vocabulary</i>	1.06	0.98	-0.25	0.95	-0.48	0.48
<i>Language Use</i>	1.23	0.94	-0.65	0.98	-0.12	0.5
<i>Mechanics</i>	-1.53	1.17	1.75	1.39	3.19	0.42

Table 9.7***Trait Fit Statistics for the Tobacco Prompt – Human Scores***

Trait Name	Difficulty Estimate	INFIT Mean Square	INFIT ZSTD	OUTFIT Mean Square	OUTFIT ZSTD	Point Measure Correlation
<i>Content</i>	0.67	1.01	0.09	0.97	-0.29	0.51
<i>Organisation</i>	0.26	0.92	-0.86	0.89	-1.12	0.52
<i>Vocabulary</i>	-0.57	0.86	-1.68	0.84	-1.57	0.53
<i>Language Use</i>	-0.11	0.93	-0.74	0.89	-1.06	0.51
<i>Mechanics</i>	-0.25	1.24	2.57	1.37	3.31	0.43

All INFIT and OUTFIT mean square values are under the recommended higher-control limit of 1.3, with only one small exception – the OUTFIT mean square values for the *Mechanics* trait are slightly above 1.30 across both prompts. In terms of z–standardised t statistics, only the *Mechanics* trait has an INFIT or OUTFIT z–standardised t statistic greater than 2. A close observation of the raw data indicates that the slight misfit associated with *Mechanics* is due to a few able test takers not achieving high *Mechanics* scores, while a few less able test takers achieved highest scores on *Mechanics*. This type of misalignment between *Mechanics* scores and scores on other traits such as *Content* and *Organisation* are also observed in other studies of non-native English speaking students’ writing (Lee et al., 2008; Matsuno, 2009).

Taken as a whole, the misfit in the human scores is insignificant. It is also noted that the OUTFIT and INFIT mean square values for all of the traits are above the recommended lower-control limit of 0.7, meaning no trait showing signs of data being too predictable. It is

therefore concluded that, overall, the traits assessed by human markers fit the model relatively well and each of the traits contribute to the measurement process productively and usefully.

Differences in the estimated trait difficulties across the human and IEA scoring methods

One way of gauging whether the IEA scores the traits in the same way as human markers is to compare the order of the trait difficulty, as estimated from the IEA scores, to that from the human scores, for the same traits. The underlying logic is similar to the item invariance principle that is fundamental to sound test construction. In this case, the relative difficulties of the traits, estimated from the same sample of essays, should remain stable across the human and the IEA scoring methods, if the two scoring methods are measuring the same traits in a similar manner.

For this analysis, the difficulty estimates for the five traits assessed by IEA (reported in Tables 9.4 and 9.5), are compared to the difficulty estimates for the five traits assessed by human markers (reported in Tables 9.6 and 9.7). It can be seen from discussions in Section 7.2 that, at the surface level, the assessment coverage of these two sets of five traits is similar, with the only differences being: 1) the grammatical aspect of writing performance is included in the *Language Use* trait assessed by human markers, but included in the *Grammar Usage and Mechanics* trait assessed by the IEA; 2) the spelling trait of the writing performances is included in the *Mechanics* trait assessed by human markers but not included in the IEA *Grammar Usage and Mechanics* trait. It is emphasised that the purpose of this analysis is to demonstrate a new method of detecting potential issues related to trait scoring by an AES

system. It is also noted that the differences in the assessment coverage noted above are not expected to alter significantly the interpretations of the results to be presented.

An interesting observation from Tables 9.4 to 9.7 is that, while the human scores reveal no obvious patterns of the ordering of the trait difficulties across the prompts, there is a consistent pattern in the order of the trait difficulties estimated from the IEA data. The two higher order IEA traits – the *Content* and the *Development, Structure and Coherence* traits – are consistently estimated as being easier to achieve than the IEA language traits, with *Content* the easiest of all traits, regardless of the prompt.

The approximate t statistic as shown in the following equation (Eckes, 2009, p. 19; also in Wright & Masters, 1982) can be used to judge the statistical significance in differences in the difficulty estimates for each pair of traits:

$$t_{j,k} = \frac{\hat{\alpha}_j - \hat{\alpha}_k}{(SE_j^2 + SE_k^2)^{1/2}},$$

where, SE_j and SE_k are the standard errors associated with difficulty measure estimates $\hat{\alpha}_j$ and $\hat{\alpha}_k$ respectively. The statistic is approximately distributed as a t statistic with $df = n_j + n_k - 2$, with n_j and n_k being the number of ratings provided on the j and k traits, respectively (adapted from Eckes, 2009, p. 19).

For the Tobacco prompt, the IEA assessed *Content* and *Development, Structure and Coherence* as being easier to achieve than any of the language traits, with *Content* being statistically easier to achieve than any other traits (p values for all pairs of trait difficulty

involving *Content* <0.001). In fact, *Content* is 2.31 logits easier than the most difficult language trait (*General Linguistic Range*), which is not a trivial difference considering it represents 0.7 of the standard deviation of the person ability measures.

However, a very different pattern of trait difficulty is observed based on the human scores. For the same essays written to the Tobacco prompt, human markers assessed *Content* and *Organisation* as being the two most difficult amongst all five traits, with *Content* being statistically more difficult than any other traits (p values for all pairs of trait difficulty involving *Content* <0.05). *Content* is 1.24 logits more difficult than the easiest trait of all (*Vocabulary*), which is 0.5 of the standard deviation of the person ability measures.

The order of trait difficulty, particularly with regard to the relative difficulty of the *Content* trait, can have substantive importance in both language testing and language teaching. It has long been established that task effect introduces a fair amount of variance in writing scores (e.g., Lee, Kantor & Mollaun, 2002; Moon, Loyd & Hughes, 1996; Schoonen, 2005). One of the causes that induces the task specific variance is the content or topic of the writing task (Benton, Sharp, Corkill, Downey & Khramtsova, 1995; Kellog, 1987; McCutchen, 1986). Schoonen (2005) contends that the topic-induced variance can be caused by “differences in amount of topic knowledge, and the degree of interest in or familiarity with the topic and the rhetorical context” (p. 19). These differences affect one’s ability to generate substantive and relevant ideas which are key assessment criteria for a *Content* trait. Thus, if the *Content* trait has a relatively large (positive) difficulty estimate for the persons tested (i.e., content is more difficult for the testing cohort to achieve relative to other traits), it may indicate task specific

variance. This may then require remedial actions, depending on the purpose of the tests. For general language proficiency tests such as the PTE Academic tests, it would be good practice to monitor and mitigate any systematic variance introduced by the effect of topic knowledge. In this regard, although the analysis performed in this section presents a useful way to detect potential issues around the topic-induced variance, such usefulness depends on the accuracy of the trait scoring by a scoring system.

But perhaps more important is that the relativity of the difficulties among traits provides useful diagnostic information to teachers in relation to the strengths and weaknesses of student's writing in different areas of writing proficiency. This information is particularly useful for those teachers whose students are still learning English as a second language, and who therefore are likely to have non-uniform score profiles. These teachers often rely on scored performance across the different traits to identify areas for improvement by their students (as demonstrated in comments collected from English as a Second Language teachers used for this study).

However, the results in this section indicate that different patterns of trait difficulty could emerge depending on which set of scores is used (i.e., the trait scores generated by the IEA or those by the human markers), even though the samples of writings used are the same. There appear to be no readily available answers to explain the divergence observed, partly because no AES studies so far (to the researcher's knowledge) have analysed differences/similarities in the order of trait difficulty based on AES scores and human scores. However, this issue is important because it is critical to the use of an AES system in classroom for instructional

purposes. More studies are needed to further understand the extent of the differences in the human and AES scoring of traits and its implications for our interpretations of student performance in different areas of writing proficiency.

Item Characteristics Curves for the IEA traits

Item Characteristic Curves (ICC) were next inspected for all the IEA traits, in particular for the two traits (i.e., the *Spelling* and the *Formal Requirement*) which show signs of under-fit. These graphs provide clues as to which person ability group has the greatest difference between actual and expected scores. An example of an ICC is provided at Figure 9.3.

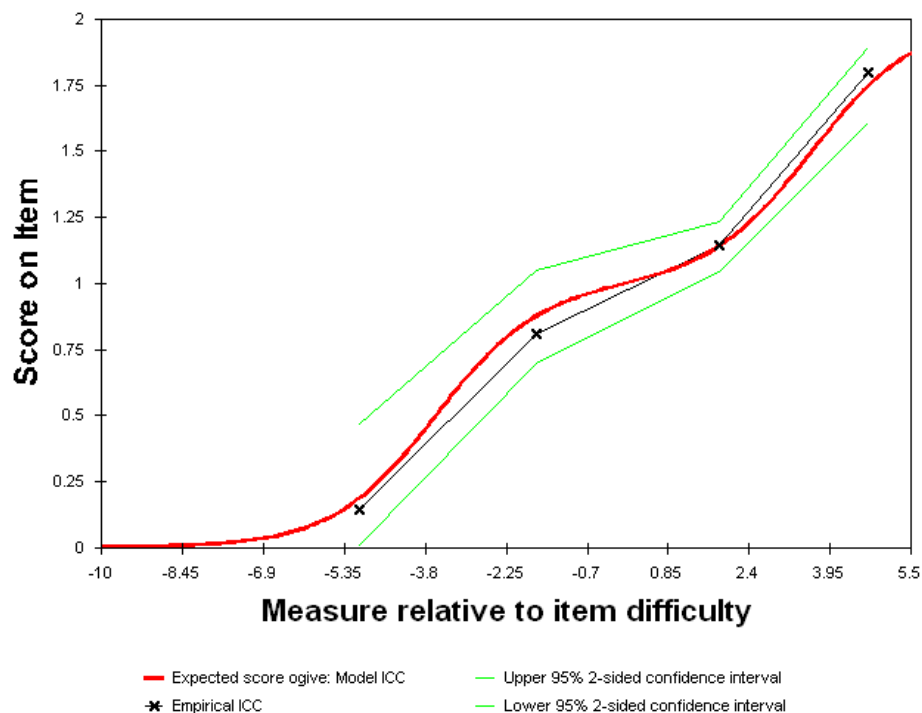


Figure 9.3 ICC Graph Obtained for the *Development, Structure and Coherence* Trait for the Tobacco Prompt

As illustrated on Figure 9.3, the x-axis of an ICC represents the underlying ability continuum and the y-axis shows the expected average (trait) score as determined from the model or the actual average (trait) score for a particular group of persons located somewhere along the ability continuum. The red curve on an ICC shows the expected relationship between the person ability and the score for a trait. The black line (with the crosses) depicts the empirical curve. The two green lines represent the 95% confidence interval around the expected ICC (Linacre, 2010). Where a group – represented by a cross on the black empirical line – falls outside the 95% confidence interval, it is an indication that this group of responses does not fit the Rasch model well. When this happens, it could be due to chance because some empirical groups may have few persons. The purpose of this analysis is to identify the groups for particular traits that do not conform to the Rasch model and to use this information to guide the investigation of the trait level analysis in Chapter Eleven. Figure 9.3 is an example of an ICC graph for a trait that shows evidence of conformity with the Rasch model; that is, there are no unexpected groups falling outside the green lines.

ICC graphs obtained for the *Spelling* Trait

Figures 9.4 and 9.5 are the ICC graphs for the *Spelling* trait.

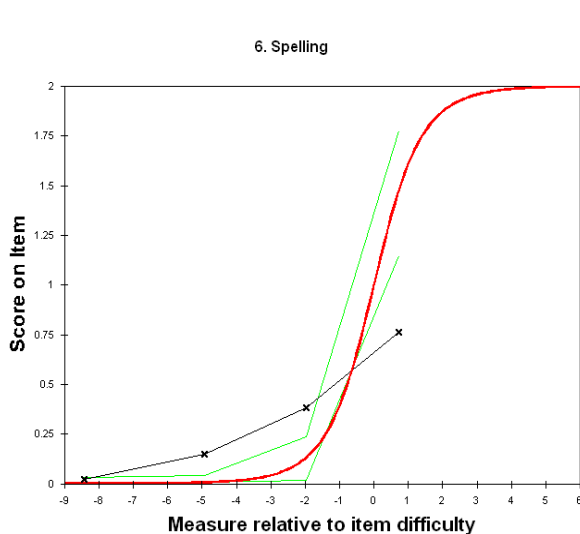


Figure 9.4 ICC for *Spelling* – Voting

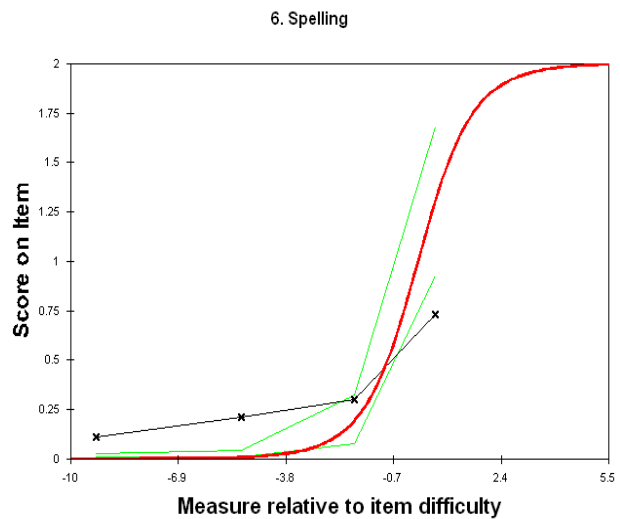


Figure 9.5 ICC for *Spelling* – Tobacco

These two graphs reveal that, across both prompts, the majority of the groups (i.e., three out of four ability groups displayed) fall outside the confidence intervals of the expected scores, indicating significant misfit in the *Spelling* scores. It also seems that misfit in the *Spelling* scores occurs across the range of the ability continuum.

Another observation from the above two ICC curves is that, for the *Spelling* trait, all four ability groups have an average score less than 1 (on a rating scale 0–2), an indication that the criteria for the trait may be too difficult for the persons tested. Examination of the raw statistics confirms this observation. For the Tobacco prompt, 78% of the persons received a score of 0, 13% received a score of 1 and only 9% of the persons scored 2. Similar proportions of persons in each score category were also observed for the Voting prompt (i.e., 74% received a score of 0, 12% a score of 1, 14% a score of 2). Item difficulty estimates

reported in previous tables (9.2 and 9.3) confirm that the estimated item difficulties for the *Spelling* trait (2.51 logits for the Voting, 3.66 logits for the Tobacco) are well above the average person ability (0.35 logits for Voting; and 0.65 logits for Tobacco). As a result, both ICC graphs show that the curves discriminate over a narrow ability range on the ability continuum. This suggests that the trait almost functions as a switch, no longer providing useful measurement information to discriminate amongst persons of different ability levels.

ICC graphs obtained for the *Formal Requirement* trait

Figures 9.6 and 9.7 are the ICC graphs obtained for the *Formal Requirement* trait.

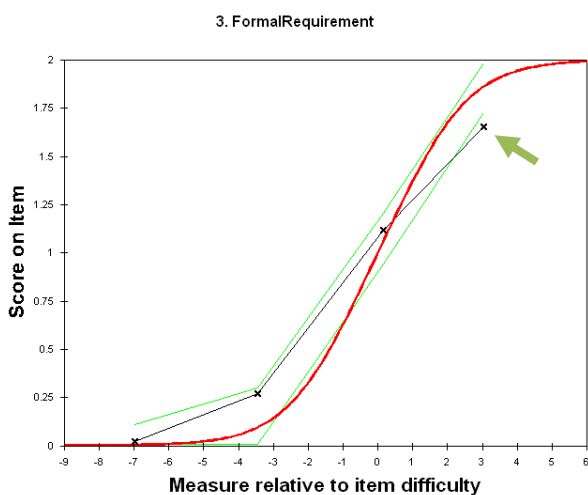


Figure 9.6 ICC Graph for Voting

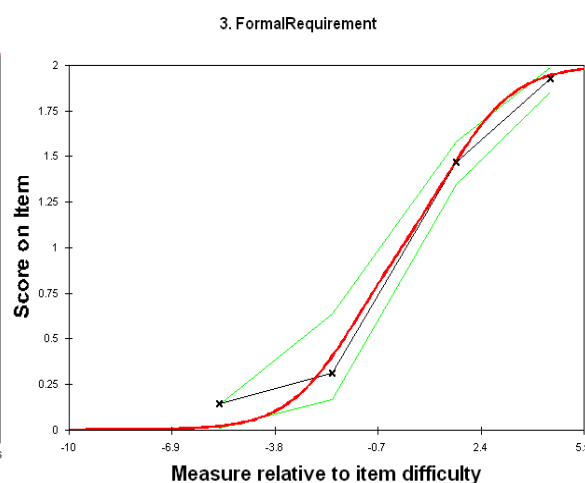


Figure 9.7 ICC Graph for Tobacco

All groups show evidence of data-model conformity, except the one group on the ICC graph for the Voting prompt, which falls slightly outside the confidence interval (as pointed out by the green arrow on Figure 9.6). Since the OUTFIT mean-square value for *Formal Requirement* for the Voting prompt is 1.60 (Table 9.2), indicating that segments of the data

might not support useful measurement, it is worthwhile investigating this outlying group of scores further to ascertain whether the unexpectedness in the data is due to randomness or to a substantive reason related to the IEA scoring method. Consequently a second ICC graph for the Voting prompt was obtained, increasing the ability groups from the original four to five groups (Figure 9.8). The aim was to help identify which persons along the ability continuum had a significant difference between the observed score and the model estimate. In other words, they did not fit the model on the *Formal Requirement* trait.

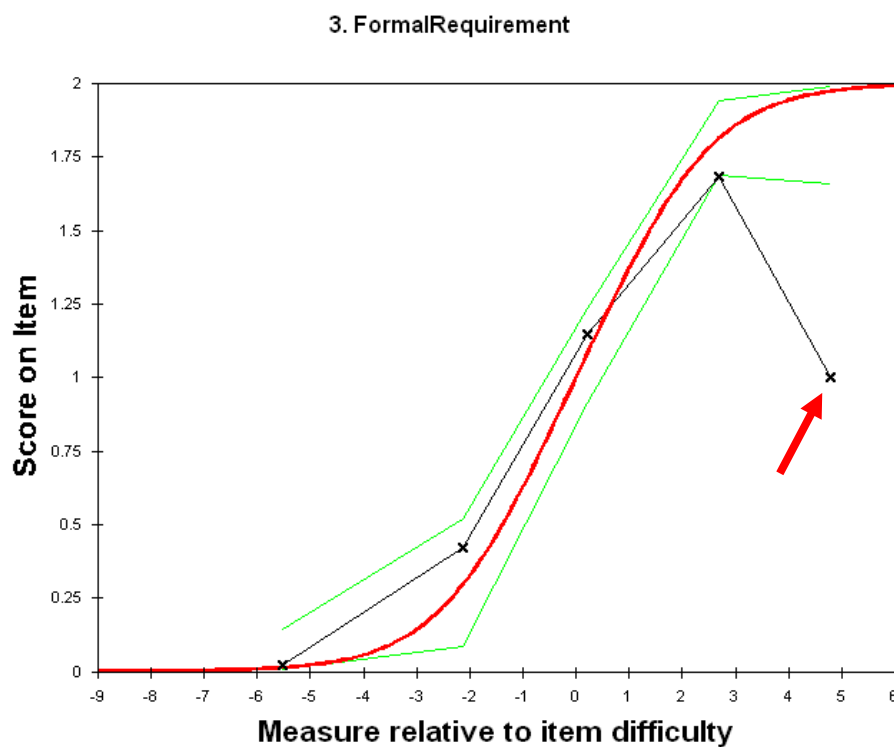


Figure 9.8 ICC Graph for the *Formal Requirement* Trait for Voting

The new graph (Figure 9.8) reveals that the most able group of persons (or essays) is clearly outside the confidence interval. The model expectation was that this ability group would

achieve the highest score of 2 for this trait. Essays contained in this group were identified and were separately examined. The cause of the misfit is not entirely due to the variability of person performance across traits. Rather there is a substantive reason, which is related to the way this trait is scored. This will be described in full detail in Chapter 11.

Apart from *Spelling* and *Formal Requirement* traits, no significant misfit is observed from the ICC graphs obtained for any other IEA traits. The next section focuses on empirical evidence concerning the effectiveness of the rating scales employed by the IEA to evaluate the traits.

9.5.3 Part III – Effectiveness of the IEA Trait Rating Scales

Prior to presenting the analysis results, it is stressed that the focus of this analysis is exclusively on the empirical consistency in the trait rating scale data that confirms or contradicts the intended use of these scales. It is not the intention of this analysis to scrutinise the appropriateness of the categorisation of the rating scales on theoretical grounds. Rather, it is assumed that the categories of the IEA trait rating scales meet the essential requirements of desirable rating scale design (i.e., they are ordered categories that are clearly defined, unequivocal, substantively relevant, and exhaustive) (Guilford, 1965; Linacre, 1999).

Table 9.8 reports a number of useful diagnostic indicators as recommended by measurement theorists (e.g., Andrich, 1996; Linacre, 1999; Lopez, 1996) for the IEA rating scales.

Table 9.8

Category Frequencies, Average Ability Measures, OUTFIT Statistics and Rasch-Andrich Threshold Measures

TRAIT	Category	Cnt	Voting			Tobacco			
			Average Ability Measure	OUTFIT MNSQ	Rasch-Andrich Threshold Measure	Cnt	Average Ability Measure	OUTFIT MNSQ	Rasch-Andrich Threshold Measures
<i>Content</i>	0	3	-4.18	1.8	NONE	2	-6.78	0	NONE
	1	43	-1.42	0.64	-5.7	18	-3.44	0.75	*
	2	63	1.3	0.71	-0.3	65	0.63	1.05	-2.34
	3	11	3.11	0.81	3.81	35	3.22	0.73	2.56
<i>DSC</i>	0	7	-3.51	1.35	NONE	9	-5.12	1.74	NONE
	1	91	0.04	0.78	-4.19	74	0.12	0.97	-4.74
	2	22	2.86	0.49	2.85	37	3.11	0.53	2.45
<i>Form</i>	0	27	-1.89	1.4	NONE	19	-3.6	1.04	NONE
	1	45	0.18	2.18	-1.37	41	0	1.53	-2.35
	2	48	1.78	1	0.99	60	2.44	0.8	0.92
<i>GLR</i>	0	10	-3.91	0.52	NONE	13	-5.06	0.27	NONE
	1	86	0.08	0.49	-3.53	72	0.37	0.55	-3.64
	2	24	3.09	0.33	2.68	35	3.34	0.48	2.58
<i>GUM</i>	0	14	-3.1	0.83	NONE	11	-5.24	0.6	NONE
	1	83	0.26	0.74	-2.89	77	0.35	0.59	-4.14
	2	23	2.8	0.64	2.75	32	3.4	0.47	2.79
<i>Spelling</i>	0	89	0.01	2.14	NONE	93	0.29	1.47	NONE
	1	14	0.62	11.1	2.75	16	2.19	44.95	3.53
	2	17	1.91	8.31	2.26	11	1.42	72.68	3.8
<i>Vocab</i>	0	10	-4	0.43	NONE	11	-5.33	0.31	NONE
	1	86	0.18	0.75	-3.53	76	0.34	0.62	-4.14
	2	24	2.79	0.48	2.68	33	3.35	0.52	2.72

Note:

Abbreviations:

Cnt: Count.

DSC: Development, Structure and Coherence; *Form*: Formal Requirement; *GLR*: General Linguistic Range;

GUM: Grammar Usage and Mechanics; *Vocab*: Vocabulary Range

*: Rasch-Andrich threshold not estimated by Winsteps.

Category Frequencies

One observation of Table 9.8 is that score category 0 for the *Content* trait is severely under-utilised. A very small proportion of essays in both samples (2% to 3% across the two prompts) were scored zero for the *Content* trait. Low numbers of responses in the categories (if replicated in future larger studies) may alert the test constructors of the need to examine the use of the category and/or the need to combine adjacent categories into a single category. It is cautioned that, where responses are fewer than ten in a category, other category statistics (e.g., the associated category thresholds) estimated from the Rasch analysis may be relatively unreliable (Linacre, 1999).

Category Fit Statistics

Attention is first paid to those categories with large OUTFIT mean-square values (i.e., greater than 2), as per the guidelines provided by Linacre (1999). Score categories 1 and 2 for the *Spelling* trait have excessively high mean-square values (i.e., 11.1 and 8.3 for category 1 and 2 for the Voting prompt, and 45.0 and 72.7 for the Tobacco prompt). This is followed by category 1 on the *Formal Requirement* trait for the Voting prompt. Its OUTFIT mean-square value is 2.2, indicating there is 1.2 times more un-modelled noise in the data than the modelled stochasticity, which means the data doesn't fit the model well.

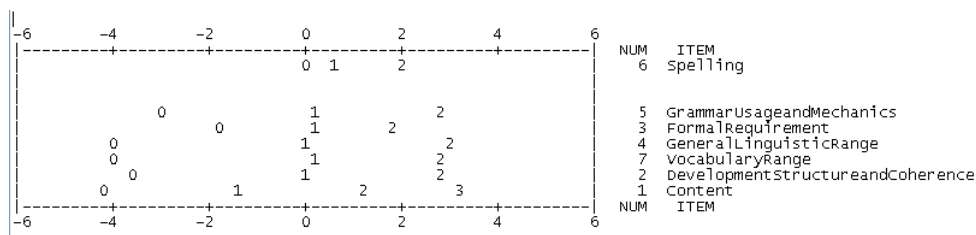
On the other hand, consistently across both prompts, almost all categories for the two language traits – *General Linguistic Range* and *Vocabulary Range* – show significant signs of data-model over-predictability. This indicates that the score categories on the rating scales for

the two language traits have considerably less measurement information than desired, and consequently yield little new information to help define the ability continuum. Since almost all the categories on these two rating scales are unproductive categories, the specificity of the scoring of these two traits (e.g., what micro text features are attended to by the two language traits and how are they scored by the IEA) should be examined more closely by future studies.

Average ability measures

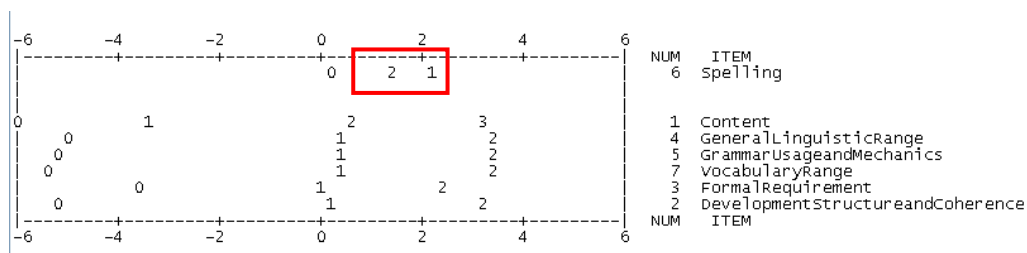
The expectation of a functional rating scale is that in general, persons with higher ability produce observed results in higher score categories, and vice versa (Linacre, 1999). One useful way to check whether the empirical data adheres to this expectation is to examine the average ability measure estimated for each score category. The average ability measure is defined as the average of the ability estimates for all persons who were rated in this particular category for this trait.

Table 9.8 shows that, apart from the *Spelling* trait, the average person ability measures advance with the sequential score categories on all trait rating scales, as expected. The only exception is score category of 1 and 2 for the *Spelling* rating scale for the Tobacco prompt, where the average ability measure for category 2 is noticeably lower than that for category 1. This exception can best be seen in Figure 9.10, where the locations of the observed average measures (in logits) for each rating scale category are plotted against the horizontal axis – the latent ability continuum. As a comparison, Figure 9.9 is an equivalent graph for the Voting prompt.



Latent Continuum – increasing ability (logit)

Figure 9.9 Observed Average Measures for Score Categories – Voting



Latent Continuum – increasing ability (logit)

Figure 9.10 Observed Average Measures for Score Categories – Tobacco

The “disordered categories”, which are marked in the red box (on Figure 9.10), contradict the intention that a higher category on a rating scale indicates more of the underlying ability. This represents a serious threat to the interpretability of the *Spelling* scores. Taken together with the large mean square fit statistics reported in the last section for *Spelling* categories, it indicates that the scoring of this trait needs to be further investigated.

It is also noted from Figures 9.9 and 9.10 that, apart from the *Spelling* trait, there seems to be an ideal spread of the ordered categories on the single ability continuum for all other IEA traits, although the *Formal Requirement* trait has a shorter rating scale than others (i.e., score categories of this trait are located more closely to each other on the continuum than categories of other traits).

Category Probability Curves and threshold estimates

Category Probability Curves simplify inferences about which category is most likely to be observed at any point along the ability continuum by visually presenting the category boundaries (Linacre, 2010). On these graphs, the horizontal axis represents the ability continuum, whereas the vertical axis shows the probability of being rated in each category, for each trait. There is one probability curve for each category. Thresholds (reported as Rasch-Andrich thresholds in Table 9.8) are located at the intersections of adjacent probability curves. Figure 9.11 shows a typical graph for a trait that conforms to the model expectations.

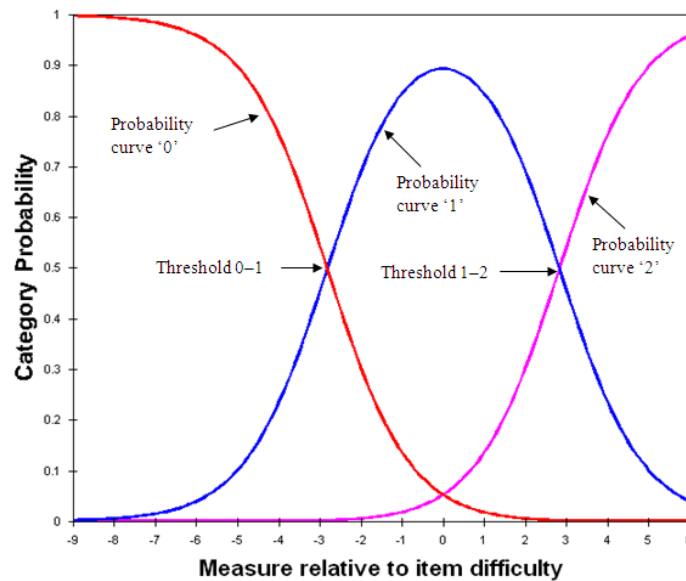


Figure 9.11 Category Probability Curve for the *Grammar Usage and Mechanics Trait – Voting*

Figure 9.11 (obtained from Winsteps for the IEA *Grammar Usage and Mechanics* trait for the Voting prompt) shows that all score categories on the rating scale are “modal”; that is, each category is the most probable response category for some portion of the latent construct. The thresholds are spread across the latent continuum; they are neither too close nor too far apart. Collectively, all the categories help in defining distinct points on the latent construct being measured. No disordered thresholds (i.e., where a higher threshold such as the 1–2 threshold has a lower measure on the latent continuum than a lower threshold such as the 0–1 threshold) are observed. Higher ability persons are more likely to score in a higher category than lower ability persons, across the continuum, as expected.

However, the Category Probability Curves obtained for the *Spelling* trait for the Voting prompt (see Figure 9.12) reveal a problem.

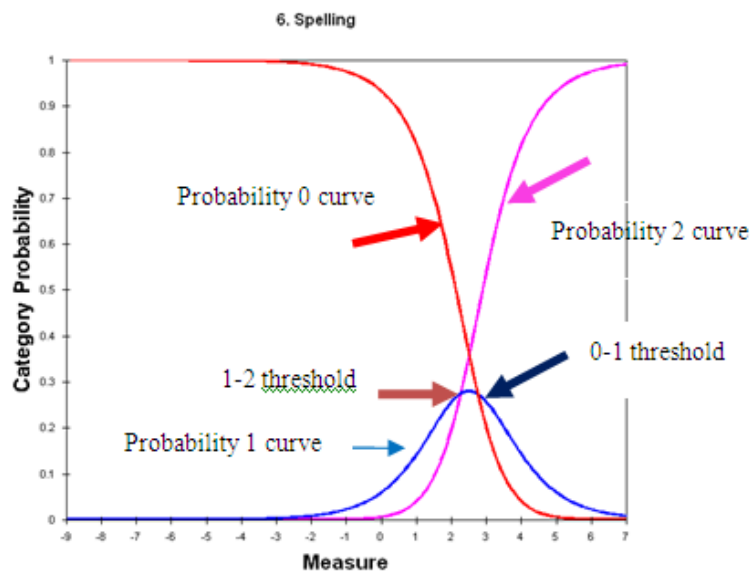


Figure 9.12 Category Probability Curve for *Spelling* – Voting

Figure 9.12 shows that only two out of the three categories are “modal” (i.e., they are the most likely of the categories to be observed), depending on persons’ locations on the ability continuum. Rasch-Andrich thresholds are disordered in Figure 9.12; that is, the 0–1 intersection (threshold) occurs at an ability level higher than required to pass the 1–2 intersection (threshold). The score category of “1” therefore does not emerge as the most likely outcome, for any location on the continuum. As noted by Linacre (2010), disordered thresholds imply less frequently used intermediate categories, in this instance, the score category 1. This warrants further investigation, as any potential “irregularity in observation

frequency across categories may signal aberrant category usage” (Linacre, 1999, p. 110). The use of the score categories for the *Spelling* trait will be extensively examined in Chapter Eleven.

In summary, evidence from analysis of the IEA rating scale functionalities further suggests that the IEA’s *Spelling* scale, and to a lesser degree the *Formal Requirement* rating scale, function differently from the other five rating scales. The most serious problem detected in this section of the analyses relates to the *Spelling* scale, in that the categories on this scale demonstrate signs of not measuring the same construct as measured by the other categories, and that a higher category on the *Spelling* scale does not always correspond to an increase in the underlying ability being measured. This detracts from the meaning and the interpretability of a total score that is derived by adding scores from all seven traits.

Analysis of rating scale effectiveness for the human scores

Appendix K reports results from equivalent rating scale analysis for the human scores, generated from the Many Facet Rasch Measurement (MFRM) analysis using FACETS, as described in Section 9.4. Across the five rating scales and for each of the prompts, the average ability measures increase monotonically with sequential rating scale categories, which confirm the scale developers’ intention that higher rating scale categories manifest higher performance levels. For each trait, there is also a clear progression of scale category thresholds along the ability continuum, indicating that each category is in turn the most likely category along the continuum.

All categories except one have desirable OUTFIT mean-square values (i.e., mean-square values less than 2). The one exception (category ‘0’ on the *Content* trait for the Tobacco prompt) could be due to randomness in the data as the fit statistic is estimated based on only three observations in that category. Only one score category for rating scales used by human markers shows any sign of data being over predictable. Score category 0 for the *Vocabulary* scale for Tobacco is the one. However, its fit statistic is estimated based on one observation only. In terms of the rating scale structure across five human traits, this can be seen more clearly from the two variable maps produced from FACET (Figures 9.13 and 9.14). Figure 9.13 displays the variable map for the Voting prompt which represents the calibrations of all measurement facets (i.e., persons, raters, traits, rating scales) in one single frame of reference. The logit scale appears as the first column in the map. All measures of persons, raters, traits, as well as the category boundaries, are positioned on this scale.²⁸ Figure 9.14 is the equivalent map for the Tobacco prompt.

²⁸ In the variable maps (Figure 9.13 and 9.14), the second column (labelled “Trait”) displays writing traits in terms of their relative difficulties. The more difficult is the trait, the higher it appears in the column. The third column (labelled “Rater”) compares the markers in terms of their relative level of severity or leniency. More severe markers appear higher in the column, and more lenient ones appear lower in the column. The fourth column (labelled “Candidate”) displays the estimated logit measures of writing proficiency for persons. While each star represents one person in Figure 9.13, each star represents two persons and each dot represents one person in Figure 9.14. The higher the estimated ability measure, the higher the person appears in the column.

Measr	-Trait	-Rater	+Candidate	S.1	S.2	S.3	S.4	S.5
7	+	+	*****	(3)	(3)	(3)	(3)	(3)
			W					
			W					
6	+	+	+	+	+	+	+	+
			W					
5	+	+	*****	+	+	+	+	+
			W					
			W					
			.					
4	+	+	W	---		---		
			W		---		---	---
			W					
			W					
			W					
3	+	+	*****	+	+	+	+	+
			W					
			W					
			W					
2	+	+	*****	2	2	2	2	2
			W					
			W					
			.					
			W					
1	+	+	W					
	Content		W					
			W					
0	Organisation	Rater 3	W	---	---			
	Language Use	Rater 4	W					
	Mechanics	Rater 5	W					
	Vocabulary		W					
-1	+	+	W					
			.					
-2	+	+	W	1	1	1	1	1
			.					
-3	+	+						
-4	+	+	.	(0)	(0)	(0)	(0)	(0)
Measr	-Trait	-Rater	* = 2	S.1	S.2	S.3	S.4	S.5

Figure 9.14 Variable Map for Tobacco Prompt

Note: the notations for the rating scales for individual traits in the last five columns, are as follows:
S.1: *Content*; S.2: *Organisation*; S.3: *Vocabulary*; S.4: *Language Use*; S.5: *Mechanics*

The last five columns in each figure (9.13 and 9.14) map the individual rating scales to the equal-interval logit scale. The column ‘S.1’, for example, denotes the rating scale as it is used by the markers for the *Content* trait. In this column, each number represents a category value

and each horizontal dashed line is positioned at the Rasch-half-score-point thresholds of this category value; that is, at the locations where the average expected score on the rating scale is 0.5 score points above and below the category value. These thresholds illustrate the boundaries between categories, when they are conceptualised as average performances (Linacre, 2010, p. 242). Extreme categories (0 and 3) are shown in parentheses only. This is because the boundaries of the two extreme categories are $-\infty$ (for the lowest category 0) and $+\infty$ (for the highest category 3) (Eckes, 2009).

Figure 9.13 shows that, for the Voting prompt, while human markers used the rating scale in the same manner for the two higher order traits (i.e., *Content* and *Organisation*), they used the scales for the three language traits slightly differently from each other and from the higher order traits. For the Tobacco prompt, Figure 9.14 demonstrates that the five rating scales used by the human markers share a very similar structure, both in terms of the threshold locations on the logit scale and distances between adjacent threshold measures.

Overall the findings from this section's analysis suggest that the rating scales used by human markers are functioning as intended and score categories are properly ordered. It must be noted that the spelling and formal length traits are not isolated for marking by the human markers. They include spelling within the *Mechanics* trait and do not assess length at all directly.

9.5.4 Part IV – Relationship Between Rasch Ability Measures Estimated from Human Trait Scores and from the IEA Trait Scores

This section collects evidence to examine whether the two sets of scores (i.e., the human and the IEA scores) are producing statistically equivalent person ability measures. This is achieved by plotting pairs of the Rasch-modelled ability measures (in logits) for the same persons on one graph, a technique commonly used for testing the invariance of the item difficulty or person ability estimates for test equating purposes. Figures 9.15 and 9.16²⁹ show the scatter plots for the pairs of person ability measures, for the Voting and Tobacco prompts respectively. On both figures, the y-axis represents the Rasch person ability measures based on the human trait scores generated from the double-marking process, and the x-axis represents the person measures based on the IEA scores on all seven IEA traits. Each dot represents a person in the sample, who has two independent ability estimates, one estimated from the human scores and the other from the IEA scores.

²⁹ Figures 9.15 to 9.18 show the person ability measures, estimated based on the human scores, have many more discriminating levels than those based on the IEA scores, in each of the figures. This is because the ability measures based on the human scores are estimated from the double-marking process (i.e., two human ratings for each trait), which result in a more discriminating instrument.

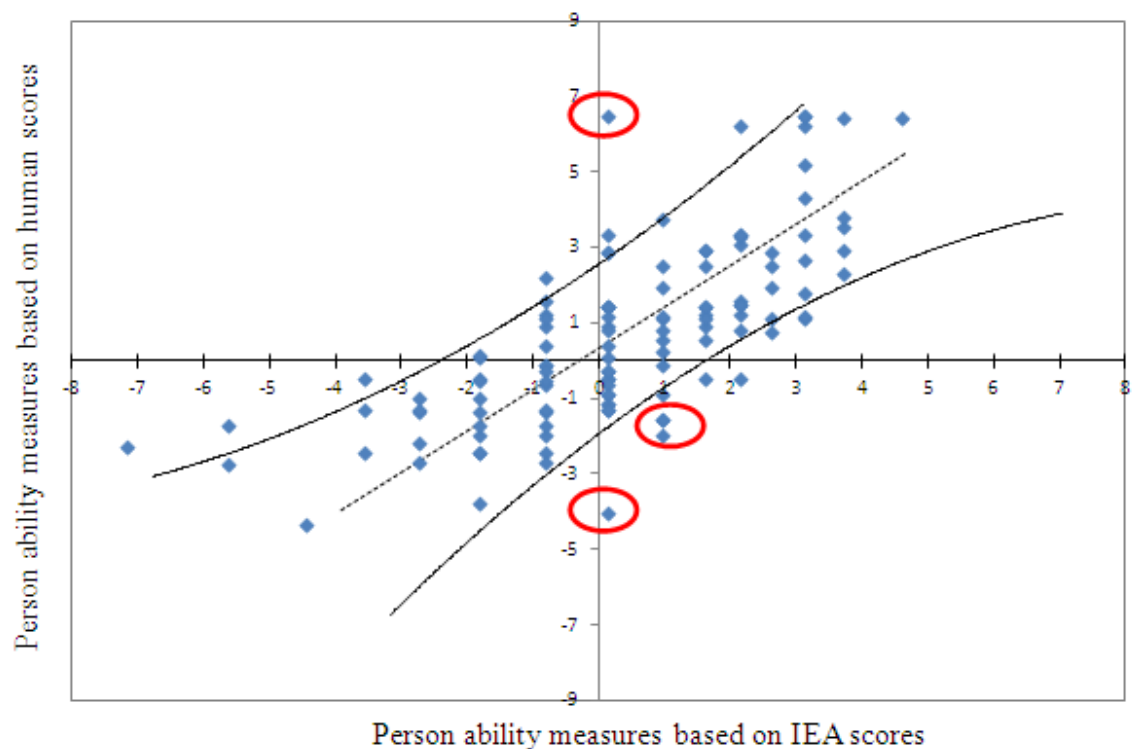


Figure 9.15 Scatter Plot of Person Measures – Voting

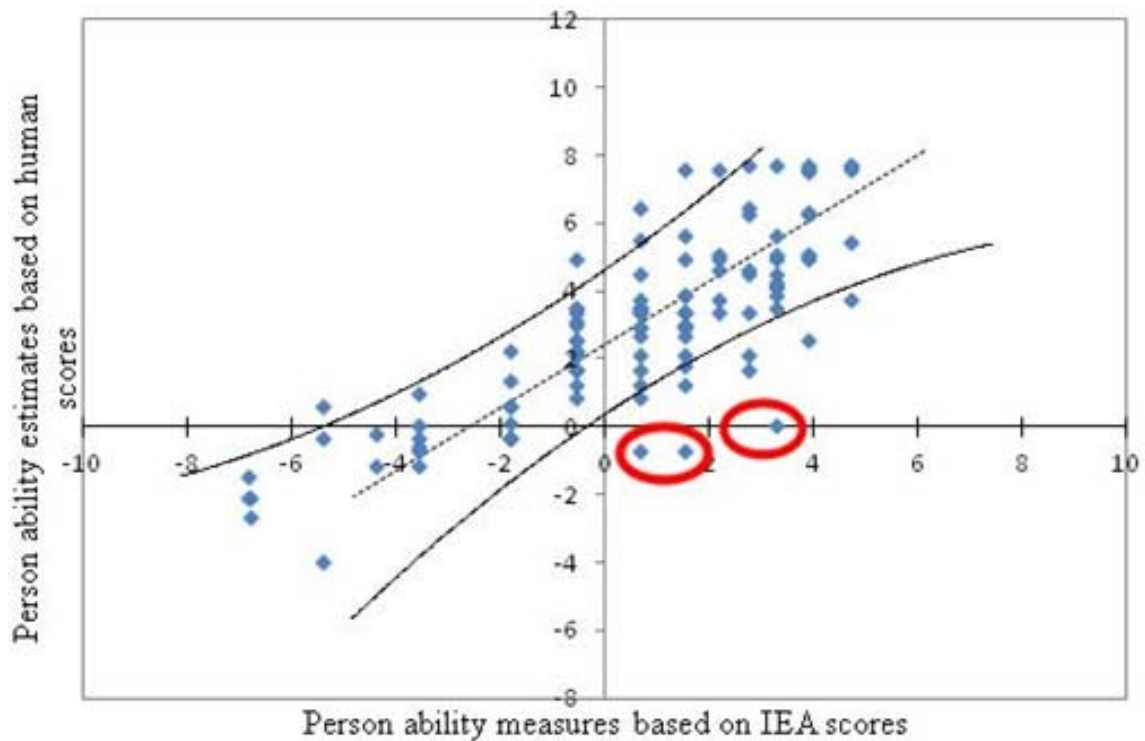


Figure 9.16 Scatter Plot of Person Measures – Tobacco Prompt

Since the same essays are scored by the human markers and the IEA, the underlying writing ability of each person being measured is common to both scoring situations. Hence if there were no measurement error associated with the person ability estimates, and if the human markers were measuring the same underlying ability in the same manner as the IEA, each dot representing the two ability estimates of the same person should lie along the diagonal line which goes through the origins of the plots and has a slope of 1. This line is referred to as the “identity line” in the Rasch literature (Wright & Stone, 1999).

However, the Rasch person ability measures estimated from the human and the IEA scores are likely to be different, partly due to the differences in the rating scales used in the respective marking processes. In this case, empirical best-fit lines are more useful representations of the expected co-relations between the two sets of person ability measures, for the purpose of this investigation. These best-fit lines, calculated to adjust for the differences in the means and in the dispersions of the two sets of measures, are therefore drawn as the dotted lines on Figures 9.15 and 9.16 (Wright & Stone, 1999).

In order to take into account the measurement errors associated with ability measure estimates, confidence intervals (represented by two solid lines) around the empirical best-fit lines are also constructed based on the error estimates provided by the Rasch models. These are calculated as the approximate 95% two-sided confidence bands around the dotted lines.³⁰

³⁰ See Wright and Masters (1982, pp. 115–117) for the mathematical specification of the confidence bands.

If the two independent scoring methods (i.e., the IEA and the human scoring methods) measure the same ability in the same fashion, the expectation is that 95% or more of the persons should have statistically equivalent ability estimates (i.e., they should fall within the confidence intervals) (Bond & Fox, 2001).

Results and analysis

Figures 9.15 and 9.16 indicate that, the majority of the data points (i.e., persons) have statistically comparable ability measures; that is, they either fall within the confidence intervals or lie very close to the confidence interval lines. On average, 9% of the persons across the two prompts deviate from the expected relation. This result implies that there are no significant differences in the overall patterns that exist in the human and IEA scores across the traits since these scores are used to summarise person performances and to produce person ability measures.

On the other hand, since the proportion of the persons falling outside the confidence intervals exceeds the expectation (i.e., 5%), the hypothesis that the scores produced by IEA and the human markers are measuring the same achievement ability in the same manner is rejected. Data points which represent the most significant contradiction to the expected co-relation between the two sets of ability measures are identified in red circles on both figures. For these data points or persons, writing ability estimated from the human scores and from the IEA scores differ significantly from each other. The essays in question provide a good source for further investigations in order to understand the potential differences that exist in the human

and IEA scoring methods. They are individually tagged in the data files and will be examined as part of the trait-level scoring analysis in Chapter Eleven.

Since the two IEA traits – *Spelling* and *Formal Requirement* – do not fit the uni-dimensional model well, it is worthwhile examining the empirical relations between the two sets of ability measures without the influence of these two IEA traits. Two additional graphs (Figures 9.17 and 9.18) show the ability measures estimated from the IEA scores for the remaining five traits, for the Voting and Tobacco prompts, respectively.

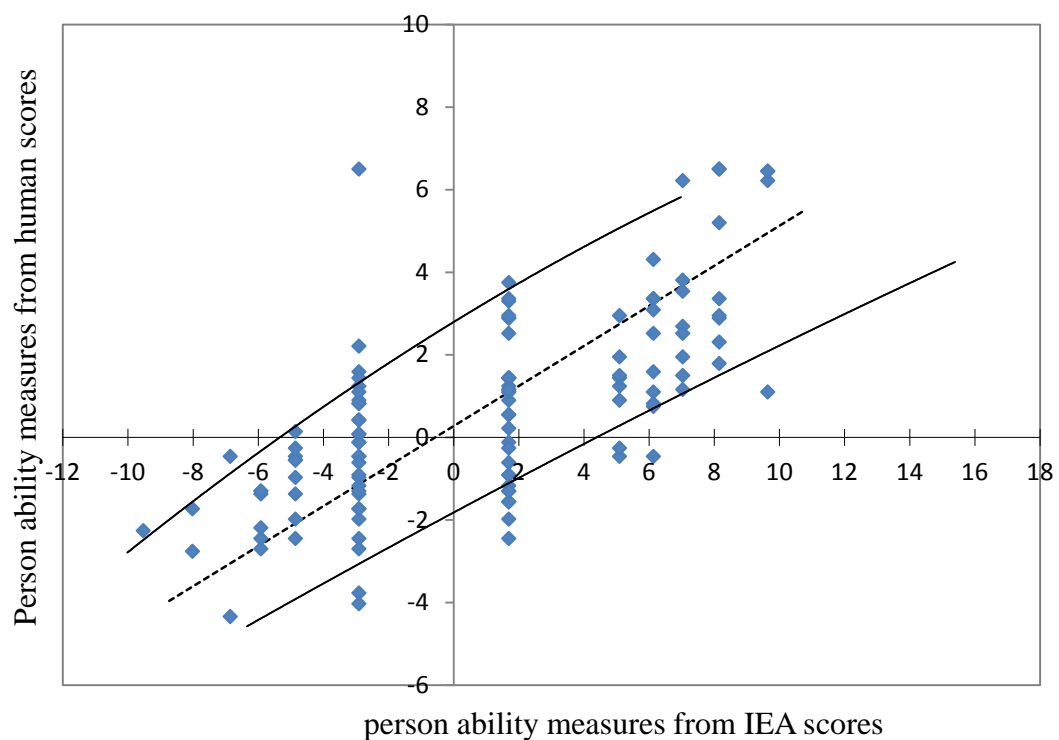


Figure 9.17 Scatter Plot of Person Measures (Five Traits) – Voting Prompt

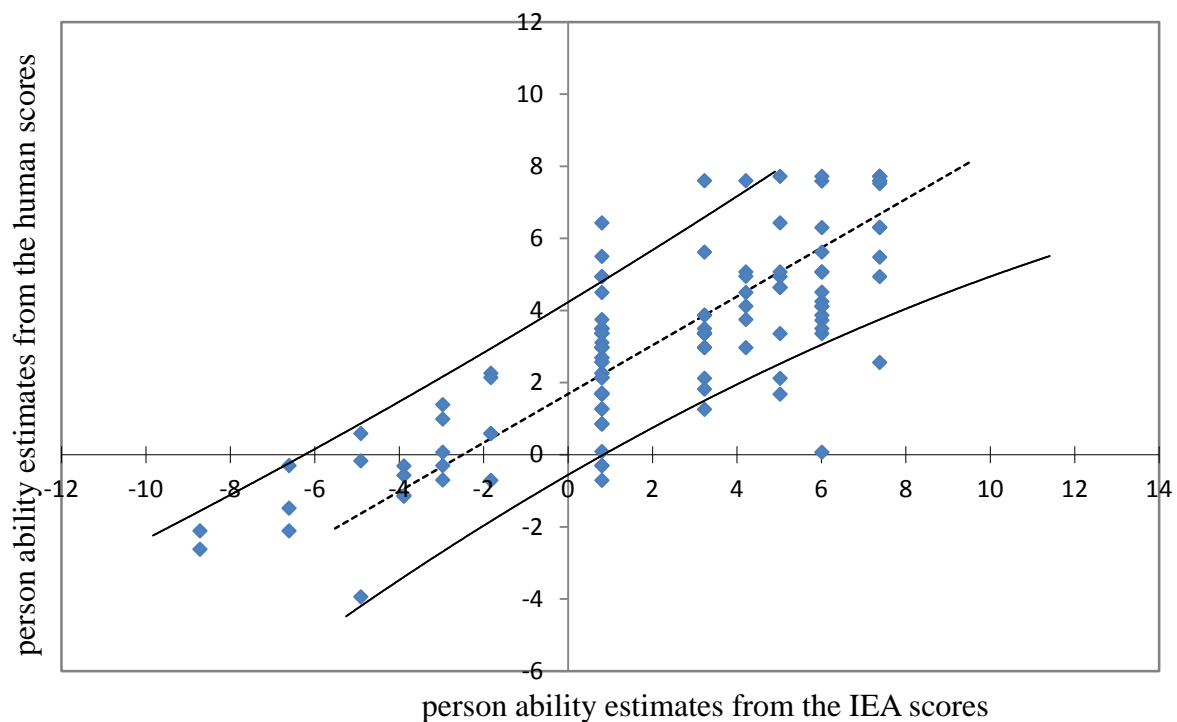


Figure 9.18 Scatter Plot of Person Measures (Five Traits) – Tobacco Prompt

Figures 9.17 and 9.18 indicate that the removal of the two worst fitting IEA traits from the analysis did not produce substantially different patterns than those observed in the previous Figures 9.15 and 9.16. Overall, approximately the same proportion of persons fell outside of the confidence intervals, after the two worst fitting IEA traits were removed.

9.6 Chapter Summary

This chapter has demonstrated psychometric techniques that can be used to address two key validity questions concerning the measurement component of the AES validation framework:

1) the degree to which the writing traits assessed by an AES system conform to a uni-

dimensional model; and 2) whether the rating scales used by an AES system to score the different traits of writing performance function as intended.

There is strong evidence from Rasch analysis to suggest that the IEA *Spelling* trait, and to a lesser degree, the *Formal Requirement* trait, do not fit well with a uni-dimensional model. In addition, a higher score category on the *Spelling* rating scale does not always indicate more of the underlying ability, which makes it hazardous to interpret the meaning of the *Spelling* scores, and in turn the meaning of the overall PTE Academic writing score. When the two worst fitting traits are removed from the Rasch analysis, the remaining five traits function well to support the development of a single construct. The only concern is that the *General Linguistic Range* and the *Vocabulary Range* traits show signs of over-fitting the model, which indicate they are less productive traits that do not bring as much new information to the measurement system as desired.

On the other hand, human scores demonstrate sufficient empirical consistency in the data to support the argument that the five traits are measuring one latent construct, and that all five rating scales as used by human markers to score the traits are functioning as intended. These results help attest to the quality of the human scores obtained for this study and lends credibility to their use as criterion measures in this study.

This chapter has also demonstrated that analysis of the order of trait difficulty based on the human and IEA scores could be a useful way of detecting potential issues related to trait

scoring by an AES system. The accuracy of trait difficulty information derived from trait scores has important implications for both language testing and language teaching.

Analysis of the relationship between the two sets of person ability measures, estimated from the IEA scores and human scores respectively, reveals that for the majority of persons, human scoring and IEA scoring produce statistically comparable ability measures. However, the proportion of persons falling outside the confidence intervals provides further evidence to challenge the assertion that the two scoring methods are measuring the same construct in the same manner. The essays identified from this analysis, and potential anomalies identified from other analysis contained in this chapter, are pursued in Chapter Eleven.

The next chapter considers the structural aspect of validity for scores produced by IEA.

Chapter 10 Structural Properties of the IEA (Intelligent Essay Assessor) and Human Scores

10.1 Introduction

This chapter demonstrates how evidence pertinent to the structural component of the proposed AES (Automatic Essay Scoring) validation framework can be collected and examined.

Explicitly, it will investigate the structural properties of the Pearson Test of English (PTE) Academic writing scores produced by the Intelligent Essay Assessor (IEA). This component of the framework requires scores produced by an AES system to exhibit internal patterns that are rationally consistent with what is known about the structural relations inherent in behavioural manifestations of the underlying construct (Loevinger, 1957; Messick, 1996). The fundamental idea is that if a scoring system is developed based on the theory of the construct domain and is measuring the underlying ability in an appropriate manner, the internal structure of the trait scores should be consistent with what is known about the internal structure of the construct domain (Messick, 1996).

Knowledge about the structure of the construct domain can be developed from either domain theory or from empirical observations. Both approaches are used in this study. The first part of this chapter considers whether both sets of scores (i.e., the IEA trait scores and the human trait scores) reproduce the internal structural relations that are consistent with those hypothesised from the writing domain theories. This is followed by an investigation of

whether both sets of scores exhibit similar structural patterns as those from earlier empirical studies on manifested writing performance across population subgroups.

10.2 Dimensional Structure in Analytic Writing Scores

Domain theories in Chapter Six established that writing ability in an academic setting was influenced by two distinct competencies – *language competence* and *strategic competence*. This led to a hypothesis that language writing traits (such as vocabulary, sentence structure, and language convention) and higher order writing traits (such as content/rhetoric aspects of writing) should exhibit a level of discriminant evidence because these two sets of traits are manifestations of the two conceptually distinct competencies. This hypothesis is now further refined, taking into account the measurement theories discussed in the previous chapter.

It is hypothesised that a two-factor structure exists within the analytic trait scores. The first factor is conceived to be the primary factor. This factor drives the scores of the various writing traits to function coherently with each other and to conform to the expectations of a uni-dimensional model, because these traits are thought to be manifestations of the same underlying ability. The second and less influential factor impels scores on the higher order traits to be independent of those on the language traits, due to the theorised difference in the competencies which underlie test-takers' performances on these traits. In Chapter Six the point has been made that the confirmation of this hypothesised two-factor structure within empirical data depends on the nature of the writing tasks and the characteristics of the test takers, as a result of the potentially close interaction between the two competencies of

language competence and strategic competence during the writing process. This section further pursues the issue of the appropriateness of the IEA scoring of writing traits by comparing the internal structure of the IEA trait scores to the hypothesised two-factor structure, and to the corresponding internal structure of the human scores, for the same group of essays.

The significance of the first factor is determined through an analysis of how much of the total variance in the trait scores can be explained by a uni-dimensional model – the Rasch Model. This first factor is also referred to in the Rasch literature as “the Rasch dimension” (Linacre, 2010). Principal Component Analysis (PCA) is then used to extract the “second factor” from the inter-trait residual correlation matrix, after the influence of the first factor has been removed from the raw data. It is noted that the PCA analysis may extract a number of secondary factors from the inter-trait residual correlation matrix. This section reports the first secondary factor that explains the most residual variance (under the hypothesis that there is such a factor). This secondary factor is referred to in this thesis as the “second factor”.³¹ If the human or the IEA scores conform to the expectations of a uni-dimensional stochastic model perfectly, there should be no patterns amongst the correlations of the standardised residuals across traits. Therefore PCA should reveal no meaningful secondary factors in the residual data. However, where there are two sets of contrasting traits “that share most strongly some

³¹ For a description of the PCA extraction method being used to detect secondary factors in the data, see Linacre (2010, p. 319).

substantive off-Rasch dimension attribute”, the contrast between the two sets of traits becomes the second factor whose meaning may then be interpreted (Linacre, 2010, p. 439).

With regard to the Rasch-PCA analysis of the IEA scores, the same Rasch Rating model and the same Rasch program (Winsteps) as used in the preceding chapter were used. The analysis started with all seven IEA traits. However, it was found that the *Spelling* and the *Formal Requirement* traits dominated the meaning of the second factor extracted from the trait residual data. It had already been established that the *Spelling* trait did not behave in accordance with the requirements of a uni-dimensional model. It will be further established in the next chapter that both the *Spelling* and the *Formal Requirement* traits are prone to be influenced by external attributes unrelated to the writing ability being measured. Consequently these two traits created relatively substantial variance that remained unexplained by the Rasch model, which caused both traits to have comparatively higher loadings on the second factor (both have loadings ranging from 0.6 to 0.7, across the two prompts). The results reported in this chapter are those based on analysis from scores for the five remaining (i.e., with *Formal Requirement* and *Spelling* removed) IEA traits.

Each essay was marked twice by human markers across the five traits using the modified ESL Composition Profile (Appendix D). The Rasch-PCA analysis using human scores was carried out on the first and second sets of scores using the Winsteps program separately. The first and second sets of scores consisted of the first and second ratings given by human markers on each trait, respectively. Results were similar for both sets of scores. Hence only the results from the first set of scores are used in the remainder of this chapter.

Before structural patterns from the human and IEA trait scores are reported, the differences in the assessment coverage of the five traits measured by the IEA and those measured by human markers are first noted, as these differences will need to be borne in mind when results are interpreted. From discussions in Section 7.2, it is clear that, at least at the surface level, the assessment coverage of the five IEA traits used in this section's analysis is similar to that of the five traits assessed by the human markers, with the two higher order IEA traits and the three IEA language traits being able to be respectively mapped to the two higher order traits and three language traits assessed by the human markers. The only differences in the assessment coverage between the two sets of traits are: 1) the grammatical aspect of writing performance is included in the *Language Use* trait assessed by human markers, but included in the *Grammar Usage and Mechanics* trait assessed by IEA; 2) the spelling trait of the writing performances is included in the *Mechanics* trait assessed by human markers but not included in any of the five IEA traits. These differences, however, are not expected to significantly alter the main interpretations of the findings to be reported in this section.

Table 10.1 reports the significance of the first and the second factor for both human and IEA-generated scores, across the two prompts.

Table 10.1***The First and Second Factors in the Human and the IEA Analytic Trait Scores***

			Voting		Tobacco	
			Human	IEA	Human	IEA
% of variance explained by the first factor (i.e., Rasch Dimension)	Person ability	Empirical	51.5%	55.2%	53.7%	56.3%
		Modelled	50.9%	54.0%	53.6%	54.5%
	Trait difficulty	Empirical	13.9%	8.3%	6.8%	7.5%
		Modelled	13.7%	8.1%	6.7%	7.3%
	Total	Empirical	65.4%	63.5%	60.4%	63.8%
		Modelled	64.6%	62.1%	60.3%	61.7%
% of variance explained by the 2 nd Factor (Eigenvalue)			14.1% (2.0)	12.6% (1.7)	16.2% (2.1)	13.8% (1.9)

Note: Results are based on five IEA traits (i.e., excluding the *Formal Requirement* and *Spelling* traits from the full complement of seven traits).

In Winsteps, the reported variance explained by traits and persons is normalised to equal the variance explained by all the Rasch measures. This apportions the variance explained by the rating scale structures.

Empirical: Figures in the “Empirical” rows display the variance in the observed data that are explained.

Modelled: Figures in the “Modelled” rows display the variance that would be explained if the data fit the Rasch model exactly (see Linacre, 2010, p. 319).

It can be seen from Table 10.1 that the first factor (i.e., the Rasch Dimension) explains the majority of the variance in the raw data for both the human and the IEA scores. It explains 65.4% of the total variance in the human scores obtained for the Voting prompt and 60.4% for the Tobacco prompt. It also accounts for 63.5% and 63.8% of the total variance in the IEA scores for the Voting and the Tobacco prompt respectively. The first factor in both sets of scores (i.e., scores produced by the human markers and the IEA) can be regarded as a dominant factor. Additionally, for both sets of scores, variance explained in the raw data (i.e.,

figures in the “Empirical” rows) is close to the variance that would be explained if the data accorded exactly with the Rasch definition of uni-dimensionality (i.e., those figures in the “Modelled” rows). These results are somewhat expected, based on the fit analyses conducted in the previous chapter.

From Table 10.1, it is also noted that the Eigenvalues for the second factor are approximately 2 in the human scores, and 1.8 in the IEA scores, across the prompts. They are greater than the critical Eigenvalue (i.e., 1.5) expected by chance for similar data simulated to fit the Rasch model³² (Linacre & Tennant, 2009). This suggests that the second factors detected in both sets of scores are unlikely to be due to random noise in the data. Furthermore, the second factor extracted from the trait residuals across both prompts explains more variance in the human and in the IEA scores than do the Rasch trait difficulty measures. For example, for the Tobacco prompt, the second factor explains 16.2% of variance in the trait scores produced by the human markers. This is more than double the total variance explained by the trait difficulties, which is 6.8%. Similarly, the second factor in the IEA scores explains 13% to 14% of the total variance in the IEA scores across both prompts. This is once again more than the total variance attributable to the IEA trait difficulties (i.e., approximately 8%).

Accordingly, the existence of this second factor in both sets of scores is worthwhile investigating.

³² Separate simulation processes were conducted in this study to generate data similar to those used in this study but are simulated to fit the Rasch model. These data files are then used to acquire the critical Eigenvalue of the second factor expected by chance. The expected value is consistently at 1.5.

A final observation of Table 10.1 is that the strength of this second factor seems to be slightly stronger for the human scores than for the IEA scores. On average, variance explained by the second factor in the human scores is nearly one-quarter of the primary factor. The second factor in the IEA scores is about 21% of the primary factor. The second factor is investigated in the next section, with the results for the human scores reported first.

The Second Factor in the Human Data

Figures 10.1 and 10.2 show the factor loadings from the PCA to make clear the meaning of this second factor in the human scores, after the first factor has been accounted for by the Rasch model.

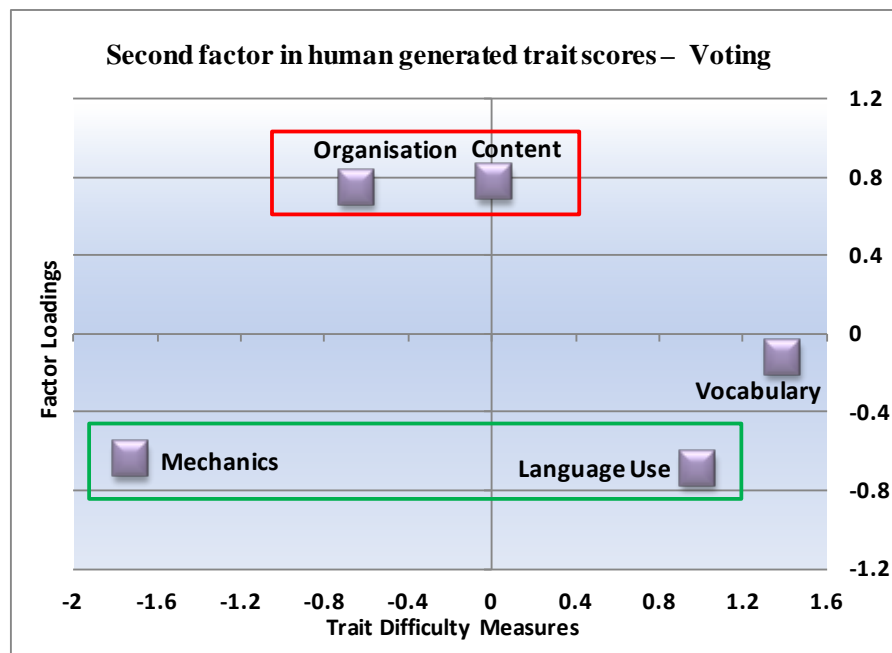


Figure 10.1 Plot of the Traits with Contrasting Loadings on the Second Factor in the Human Data – Voting Prompt

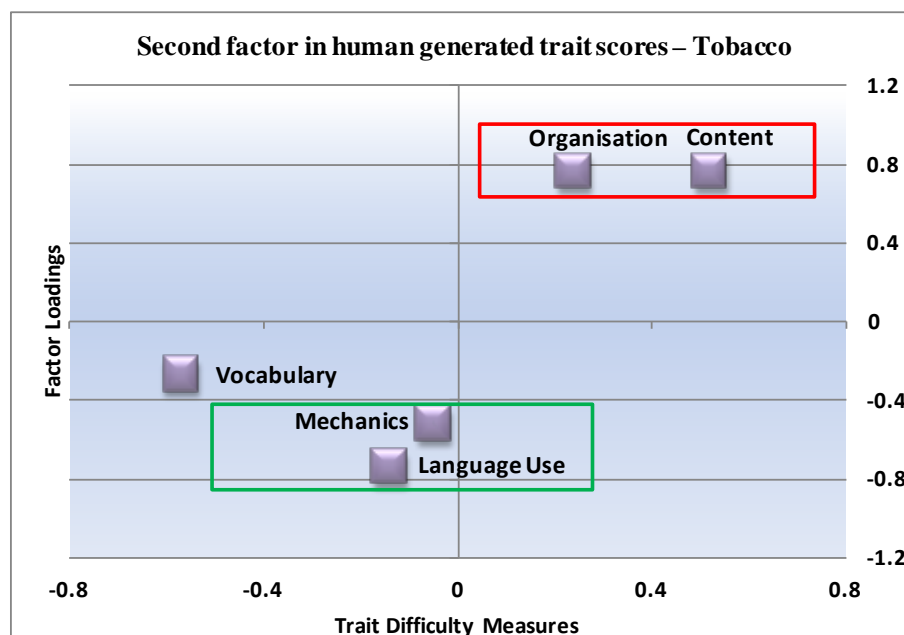


Figure 10.2 Plot of the Traits with Contrasting Loadings on the Second Factor in the Human Data – Tobacco Prompt

In both figures, the x-axis represents the underlying ability continuum (in logits). Trait difficulties and person abilities are both represented along this continuum. The traits are located on both figures based on their estimated difficulties – easier traits with negative logits are on the left and the more difficult traits with positive logits are on the right.³³ The y-axis represents the second factor within the residual data. The numbers on the y-axis indicate the factor loadings; that is, the correlations the traits have with this second factor. The sign of the loadings itself is arbitrarily set by the Rasch Program – Winsteps.

³³ As the Rasch analysis for human scores conducted in this section is based on the first ratings given by the human markers on each trait, the trait difficulty measures estimated can be different from those estimated from two ratings on each trait, as reported in Table 9.6 and Table 9.7.

If data is uni-dimensional, the distribution of the standardised residuals should resemble that of a normal random deviate, and the expectation is that the second factor would have a large factor loading on one trait and small loadings on other traits (Linacre, 2010). If there are clusters of traits which have significant loadings (i.e., those situated at the top or the bottom of the plots), then the meaning of the second factor is interpreted by contrasting the traits with opposite signs of loadings (Linacre, 2010). This study uses a factor loading of 0.5—as used by other researchers such as Daftaripard & Lange (2009)—as the cut-off for identifying traits with substantial loadings on the second factor. In both figures, traits with loadings that are of the same sign and are equal to or greater than 0.5 are grouped together in boxes, for easy interpretation of the meaning of the second factor.

Figures 10.1 and 10.2 show that consistently across both prompts, the *Content* and *Organisation* traits have a large positive loading (i.e., with a loading of 0.5 or greater). This is contrasted to *Language Use* and *Mechanics* traits each having a large negative loading (i.e., with a loading of -0.5 or less). The *Vocabulary* trait has a small negative loading (i.e., the size of the loading being less than 0.3 across both prompts). This indicates that the second factor, extracted from the human residual data after the influence of the primary factor has been accounted for by a uni-dimensional model, is characterised by the contrast between the *Content* and *Organisation* traits and the language traits *Mechanics* and *Language Use*.

The discovery and interpretation of this latent second factor lends weight to the hypothesis of the distinction between the higher order traits and language traits, formulated from writing ability theories. The existence and the strength of this second factor demonstrates that for

these two samples of essays, the conceptual distinction between the higher order and language traits is measurable. The human markers from this study were not only able to appreciate the subtle difference in test-takers' performance across different traits, but were also able to reflect this difference in their scores, resulting in a statistically discernible structural pattern that was consistent with the domain theories.

This finding corroborates earlier research of ESL (English as a Second Language) writing (e.g., Cumming et al., 2002; Lee et al., 2008; Santos, 1988). These studies also showed that human markers, when evaluating non-native speakers' essays at college level, were both willing and able to judge higher order traits and language traits independently, to the extent possible.

A word of clarification is in order regarding the impact of the second factor on the quality of the measures produced by the human markers. The discovery of the second factor in the human scores does not invalidate the findings from the previous chapter that the human trait scores are useful for the practical purpose of comparing a single ability. Several reasons support this statement. First, the human traits fit the uni-dimensional model sufficiently well. Secondly, the Rasch measure is the dominant dimension explaining the majority of the variance in the raw scores. However, the most important reason is that there is a theoretically-driven explanation for the second factor in the residual data. The nature of this factor reflects the theoretical distinction between the two sub strands that are within the same general writing construct. This is a common phenomenon with most of the achievement constructs.

Consequently it is still productive to think of the human scores as manifesting one dimension and that they are useful for the practical purpose of measuring a single ability.

Another noteworthy observation from Figures 10.1 and 10.2 is that the *Vocabulary* trait has a considerably smaller loading on the second factor than the other two language traits – *Language Use* and *Mechanics*. The small loading associated with the *Vocabulary* trait indicates this trait has relatively little significance in the interpretation of the second factor. The proximity of the *Vocabulary* trait to the higher order traits is not surprising given the well-understood strong relationship between choice of words and the meaning of the text. This structural pattern also confirms findings from an earlier study conducted to investigate college professors' reactions to the academic writing of non-native English speaking students. In that study, Santos (1988) reported that the abilities of the college professors to judge content and language independently were severely constrained when serious lexical errors were present in the essays. Santos' explanation was as follows: "it is precisely with this type of error (lexical error) that language impinges directly on content; when the wrong word is used, the meaning is very likely to be obscured" (p. 84).

Overall, the observed structural patterns in human analytic scores (i.e., the two-factor structure and the proximity of the vocabulary trait to the higher order traits in this two-factor structure), are consistent with the theory of the construct domain and consistent with writing experts' understanding of the construct domain. This strengthens the structural aspect of the validity evidence for scores generated by human markers.

The Second Factor in the IEA-generated Scores

Factor loadings of the five IEA traits are visually presented in Figures 10.3 and 10.4, in the same way that they were presented for the scores generated by the human markers. Both figures show that the contrast between the language traits and the two higher order traits does not define the second factor extracted from the IEA trait-residual analysis. This is in contrast to the human scores. Additionally the meaning of the second factor in the IEA residual data varies across two prompts. This is again contrary to the human scores which revealed a consistent structural pattern for the second factor across both prompts.

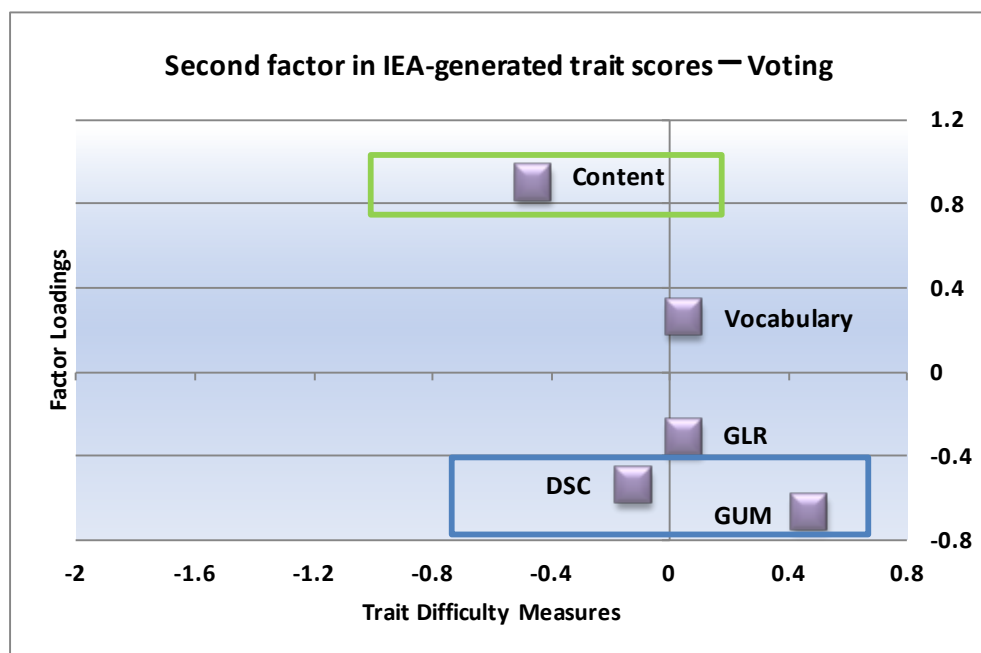


Figure 10.3 Plot of Loadings on the Second Factor – Voting (Five IEA Traits)

Note: The following notations apply to this figure:

DSC – *Development, Structure and Coherence*; GUM – *Grammar Usage and Mechanics*

GLR – *General Linguistic Range*; Vocabulary – *Vocabulary Range*.

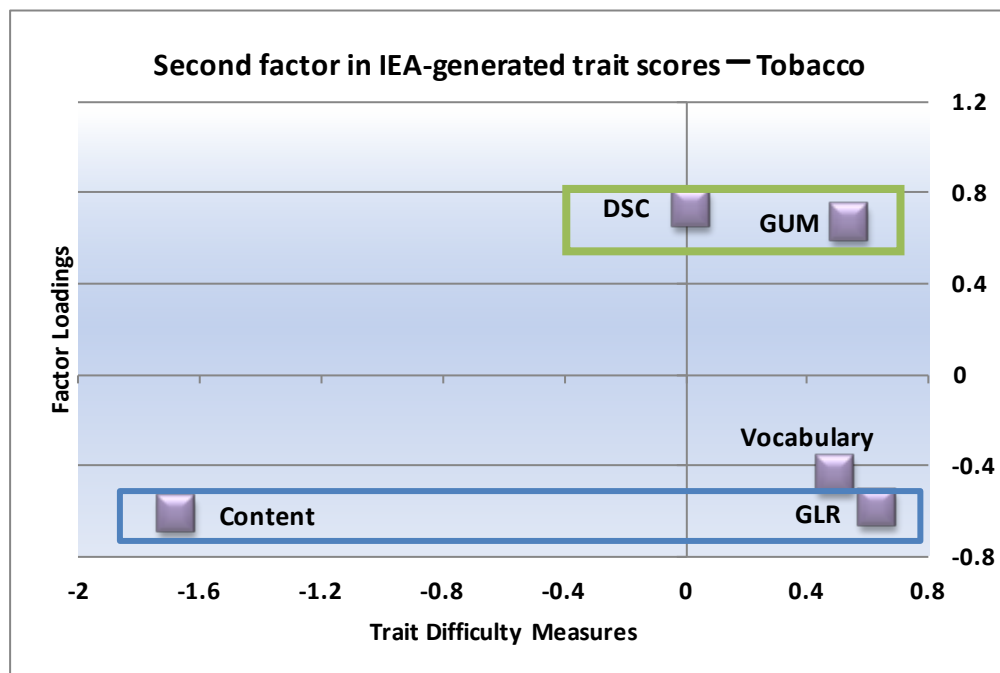


Figure 10.4 Plot of Loadings on the Second Factor – Tobacco (Five IEA Traits)

Note: The following notations apply to this figure:

DSC – *Development, Structure and Coherence*; GUM – *Grammar Usage and Mechanics*

GLR – *General Linguistic Range*; Vocabulary – *Vocabulary Range*.

What is intriguing is the distinction between the *Content* and the *Development, Structure and Coherence* (DSC) traits which has consistently been identified as a feature of the second factor in the IEA data, after the influence of the first factor (i.e., the Rasch dimension) has been removed. This result is difficult to explain considering that performance on these two higher order traits is thought to be a reflection of the same competence – strategic competence. So in theory, these two traits should be more homogeneous than any other pairs of the writing traits involving one of these two traits. In real testing situations though, task and/or person specific factors can change the intensity of the interaction between language knowledge and strategic competence (as pointed out in Chapter Six) which may then alter the inter-relationships in writing performance across different traits. However, the fact that the

human scores for the same essays revealed a clear distinction between the higher order traits and the language traits as theorised, indicates that this distinction most likely did exist in the actual performance demonstrated in sample essays produced for these two prompts and that human marking was sophisticated enough to detect this distinction. It must also be noted that the structural patterns discovered in this study's human scores are similar to those discovered by other ESL researchers studying the same type of essays written by similar kinds of students to this study's (i.e., non-native English speaking college or prospective college students) (Lee et al., 2008; Santos, 1988). For example, using a different statistical method, the Lee et al. (2008) study also showed that there is a similar distinction between two content/rhetoric traits (i.e., development, organization) and three language-related traits (i.e., grammar/usage, sentence variety/construction, vocabulary). Their study used the human judgements acquired for 930 TOEFL argumentative type of essays. Consequently it is plausible that the results from this analysis are a reflection of the IEA's inability to discriminate appropriately between the various conceptually distinct writing traits.

It is acknowledged that the above results could be affected (to an extent) by the small sample size this study used and by the differences in the assessment coverage of the five traits assessed by human markers and those by the IEA. It is recommended that further studies with larger sample sizes and with a robustly measured *Spelling* trait included in the IEA set of traits, be repeated to confirm the structural patterns in the IEA trait scores.

The next section demonstrates a different way of examining the structural aspect of validity, by investigating the structural patterns in writing scores for persons from different population subgroups.

10.3 Structural Patterns in the Analytic Scores as a Result of Gender Effect

This part of the analysis focuses on the structural patterns in the human or the IEA scores that may arise from the effect of gender on writing performance. There is already a significant body of research that identifies the effect of gender as impacting on test-takers' ability to produce quality written prose (e.g., Engelhard, 1992; Engelhard, Gordon, Walker & Gabrielson, 1994; Gyagenda & Engelhard, 2009). Studies have found that gender difference on writing tests tends to favour females, but the magnitude of the difference varies across different populations of the examinees. When the test populations are college bound international students whose first language is not English, gender difference, while still favouring females, is much smaller than that observed for school students, ranging from one-tenth to about one-third of a standard deviation (see Breland, Lee, Najarian & Muraki, 2004, for a comprehensive review of gender difference studies).

The intention of the analysis performed here is to determine whether the same performance patterns across the two gender groups exist in essay scores produced by human markers and by the IEA for this study. The rationale is that the human and the IEA scores should reveal the same performance differences across genders for the same set of essays, if the two scoring methods are measuring the writing performance in the same manner.

In a more general sense, this method of exploring the structural aspect of validity for an AES system can be expanded to analysing any patterns of performance that may exist in any sub-demographic population, such as test takers of different ethnicity or of different socio-economic status. This can be achieved by comparing patterns of performance across different groups based on AES scores to a pre-formulated hypothesis or to corresponding patterns detected in human scores, for the same set of essays.

In this study, gender difference on writing performance is examined through statistical analysis of the differences in the mean person ability measures for the two gender groups, obtained separately through the IEA scores and the human scores, for the same sample of essays. For each prompt, and for each scoring method, the average person measures from the Rasch analysis (in logits) for males and females are estimated and then tested statistically (Welch's two-sided *t*-test) to determine if there is any statistically significant gender difference. All seven IEA traits are included in the Rasch analysis for the IEA scores.

Performance patterns across gender groups – analysis and results

Tables 10.2 and 10.3 show the Rasch mean person ability measures (in logits) for the female and male groups and outcomes of statistical tests for differences in each pair of the means for the Voting and Tobacco prompts, when the human and the IEA trait scores are used in the Rasch estimation process respectively.

Table 10.2***Gender Difference Statistics – Voting Prompt***

		count	mean measure (in logits)	measure	Mean Difference		
					t^*	d.f.	prob.
IEA	Female	64	0.26	-0.19	-0.47	114	0.64
	Male	56	0.45				
Human Markers	Female	64	0.54	-0.29	-0.49	113	0.62
	Male	56	0.83				

* Welch's two-sided t -test of statistical difference between the average abilities of the two subgroups.

Table 10.3***Gender Difference Statistics – Tobacco Prompt***

		count	mean measure (in logits)	Mean Difference			
				measure	t^*	d.f.	prob.
IEA	Female	78	0.2	-1.28	-2.81	113	<0.01
	Male	42	1.48				
Human markers	Female	78	2.94	-1.53	-2.9	104	<0.01
	Male	42	4.47				

* Welch's two-sided t -test of statistical difference between the average abilities of the two subgroups.

Table 10.2 shows that, when essays written to the Voting prompt are analysed, both the IEA and human scores reveal that females on average perform marginally worse than the males on this writing prompt, but then the difference in the mean ability measure for the two groups is not statistically significant [IEA: $t(114) = -0.47$, $p = 0.64$; human: $t(113) = -0.49$, $p = 0.62$].

For the Tobacco prompt (Table 10.3), the human and IEA scores reveal the same performance pattern across the two gender groups; that is, males perform significantly better than females on this prompt. Based on the IEA scores, the average ability measure for the female group is 0.2 logits, which is 1.28 logits lower than the average ability measure estimated for the male group. This gender difference (favouring males) is statistically significant [$t(113) = -2.81$, $p < 0.01$]. The standardised mean difference, Cohen's d , calculated using the pooled standard deviation, is 0.48. This can be viewed as a medium effect size, using Cohen's (1988) conventional criteria for *small*, *medium* and *large* effect sizes.

Correspondingly, when human scores are used to generate the Rasch ability measures, a similar gender difference on the writing performance is observed. The average person ability measure for the female group is 2.94 logits, which is 1.53 logits lower than that for the male group. This difference in the mean ability (favouring males) is again statistically significant [$t(104) = -2.9$, $p < 0.01$]. The effect size, Cohen's d , is 0.52 which can also be viewed as a medium effect size.

The fact that male group is observed to perform significantly better than female group for the Tobacco prompt does not necessarily indicate a "gender effect" for this prompt (i.e., this prompt is favouring a particular gender group). This is because differences in the performance across the two gender groups could simply be due to the differences in the abilities of the female and male persons in the two groups. This study does not have the necessary data to separate the effects of gender from the effects attributable to differences in the underlying

abilities of the two gender groups³⁴, therefore it is not possible to verify if this prompt does favour males or not.

However, if “gender effect” had been detected in a prompt, it would have warranted further investigation, particularly for high-stakes tests like PTE Academic, in order to ensure the fairness of tests. In general, there could be a number of reasons why a particular prompt might favour or disadvantage a particular demographic sub-population. One could be that the population sub-group has not had the same level of exposure to the subject matter as other groups, either through curriculum design or life experience. Or it could be that the particular population sub-group finds it more difficult to write about a topic because of its cultural background.

One possible way to further investigate the differences in the human and the IEA scores, as well as to investigate the genesis of a gender effect, if it exists, is to identify the traits which do not perform the same way for the different gender groups using Differential Item Functioning analysis. Differential Item Functioning (DIF) refers to the situation where one group of persons is scoring better than another group on a trait, after adjusting for the overall scores of the persons (Linacre, 2010). If the IEA is scoring the same trait in the same manner as the human markers do, the DIF pattern that might be detected in the same trait for the two

³⁴ One way of measuring gender effect more accurately is to compare residual-based effect sizes after the differences in English language ability across the two gender groups have been controlled. See Breland, Lee, Najarian & Muraki (2004) for how this type of analysis can be carried out.

gender groups should be the same across the two scoring methods, when the same essays are analysed. Though not performed in this study due to small sample sizes, contrasting DIF patterns can be a useful way of examining the differences in the human and the IEA scores at a finer level.

Notwithstanding this, the fact that the IEA has detected the same gender performance pattern across the two prompts as did human markers, can be seen as supporting evidence for the level of accuracy of these IEA scores. It also indicates that the use of the machine scoring system might be further extended to quality control tools for test validation purposes. Breland et al. (2004), in their study of prompt difficulty and gender difference on a large number of TOEFL prompts, recommended routine implementation of statistical quality control tools to identify prompts that might be less comparable than others or biased towards certain groups of populations, as reviews by human experts in this regard were not always efficient. The AES systems can be a very cost effective quality control tool for this purpose, if it can be demonstrated through statistical analysis that these systems have the capacity to accurately detect subtle differences in different groups of students' performance.

The final structural analysis to be performed in this study is to examine IEA traits and human traits together in a multi-dimensional space. The aim is to explore and discover any defining characteristics concerning the similarities or dissimilarities amongst all the analytic traits. One hypothesis is that traits that are measuring the same or similar aspects of writing performance, such as the *Content* trait measured by the human markers and the same trait measured by the IEA should be closer together in the space, than they are to other traits which are measuring

conceptually different aspects of writing performance. This analysis is described in the next section.

10.4 Analysing the IEA and Human Trait Scores in One Two-Dimensional Space

This section uses Multi-Dimensional Scaling (MDS) to represent the similarity/dissimilarity among pairs of traits as distances among points in a low-dimensional, geometric space (see Borg & Groenen 1997, 2005, for descriptions of this technique). Essentially, the closer the traits are in an n-dimensional space, the closer the observed values for the traits. Likewise, the further the traits are apart in this n-dimensional space, the greater the difference in these traits, or the more independent scores for these traits are.

The MDS analysis was conducted for both prompts separately using SPSS version 18.0. The human scores were the averages of the two available human scores for each of the five analytic traits. Five IEA traits (excluding the *Spelling* and *Formal Requirements* from the full complement of seven traits) were used in this analysis, to remove any influence those two traits would have on the structural patterns to be observed.³⁵ For each analysis, the IEA scores

³⁵ For both prompts, the Multi-Dimensional Scaling analysis performed on the seven IEA traits showed that, in a 2-dimensional space, the *Spelling* and *Formal Requirement* traits were found to be the most distinct from the other five traits, and distant from each other as well. Therefore these two traits dominated the meanings of the two dimensions. Chapter Nine provided evidence that scores on the *Spelling* trait did not support the development of a single construct. Chapter Eleven will provide additional evidence that both the *Spelling* and the *Formal Requirement* traits are prone to be influenced by external attributes unrelated to the writing ability being measured. These two traits were therefore removed before the MDS analysis was performed for this section.

as well as the human scores were standardised before the distance matrices of all pairs of the traits were computed. The scores were standardised to take account of the differences in the maximum scores for different traits. The distance matrices consisting of pair-wise dissimilarity measures between traits were then used as inputs to the subsequent MDS analysis.

The ALSCAL (Alternating Least Square Scaling) algorithms (Young & Lewyckyj, 1979) implemented in SPSS were used for the MDS optimisation process. The Euclidean model was used as a basis to compute the optimal distances between objects in an n-dimensional space. Because of the small number of traits and small sample sizes, the number of dimensions in the conceptual space was set at 2. The maximum number of iterations and the convergence criterion for changes in Young's S-stress (Takane, Young & de Leeuw, 1977) were set at 30 and 0.0001, respectively.

Figure 10.5 is the plot representing the traits assessed by the human markers (referred to as "human traits") and those assessed by IEA (referred to as "IEA traits") in a two-dimensional space for the Voting prompt. Figure 10.6 is the equivalent plot for the Tobacco prompt. To differentiate the traits assessed by the two different scoring methods, trait names with a suffix of "_Human" in both graphs refer to the traits assessed by the human markers; and trait names with a suffix of "_IEA" refer to the traits assessed by the IEA.

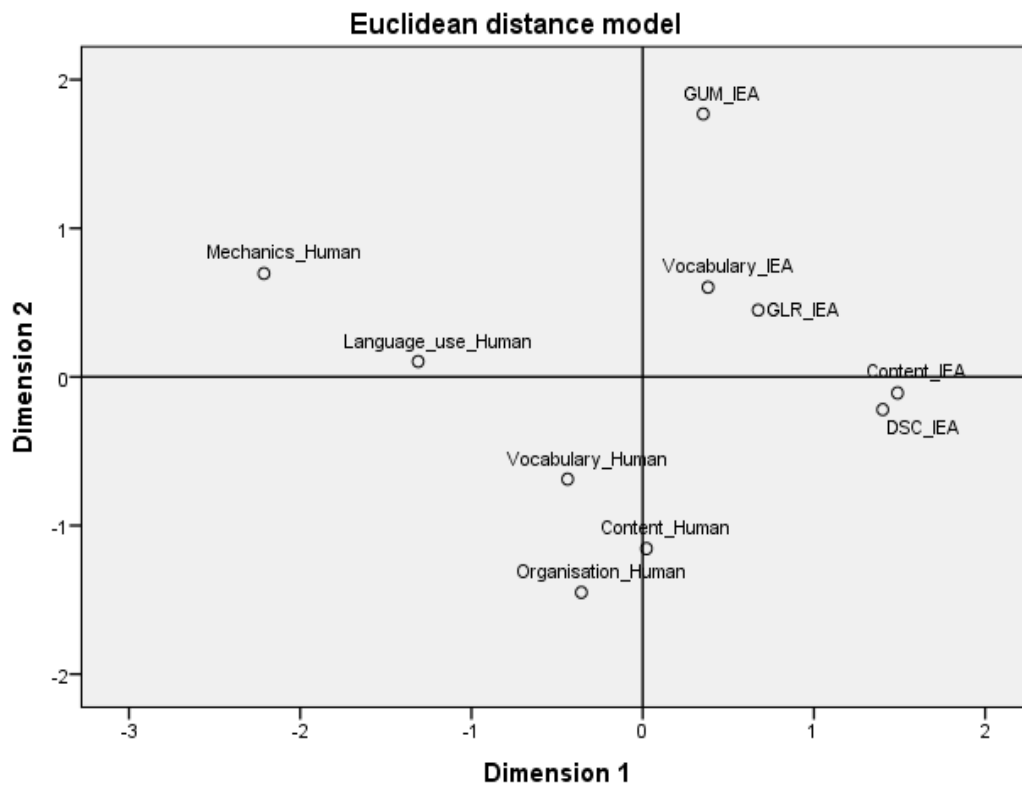


Figure 10.5 Representation of Five Human Traits and Five IEA Traits in a Two-Dimensional Space –Voting

Note: Trait names with a suffix of “_Human” in the graph refer to the traits assessed by the human markers. Trait names with a suffix of “_IEA” in the graph refer to the traits assessed by IEA.

Abbreviations:

“GLR_IEA”: IEA *General Linguistic Range* trait; “DSC_IEA”: IEA *Development, Structure and Coherence* trait; “Vocabulary_IEA”: IEA *Vocabulary Range* trait; “GUM_IEA”: IEA *Grammar Usage and Mechanics* trait.

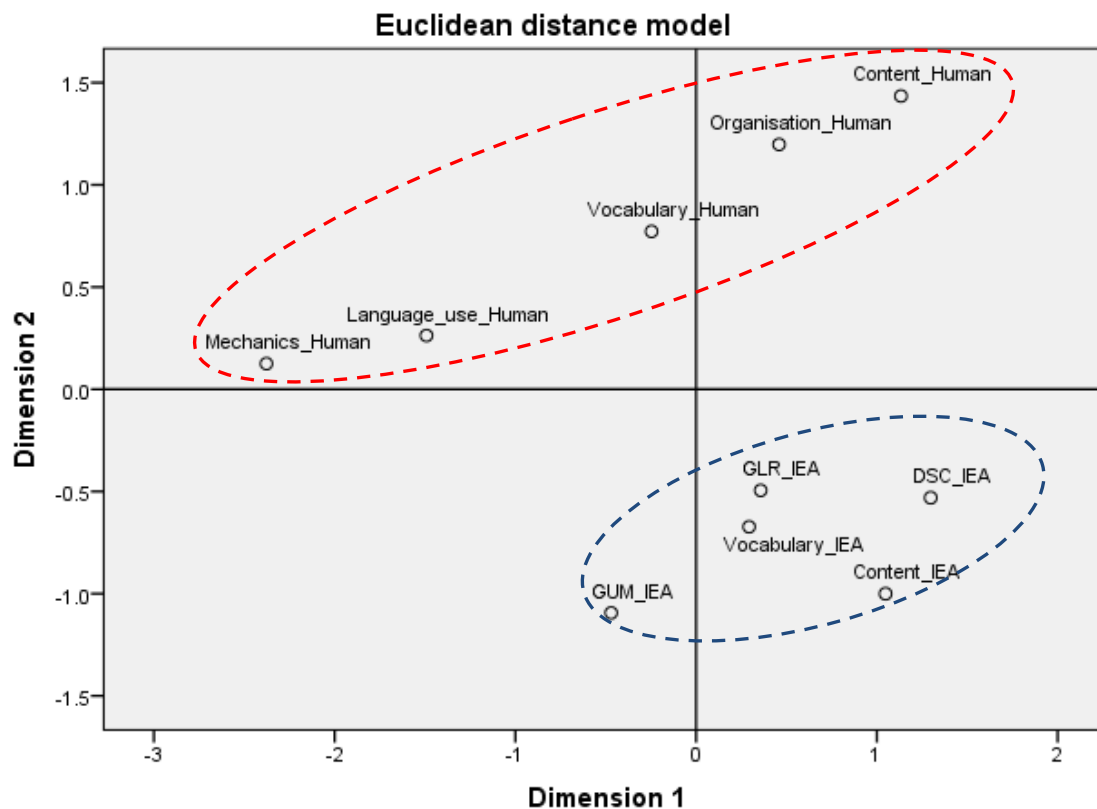


Figure 10.6 Representation of Five Human Traits and Five IEA Traits in a Two-Dimensional Space –Tobacco

Note: Trait names with a suffix of “_Human” in the graph refer to the traits assessed by the human markers. Trait names with a suffix of “_IEA” in the graph refer to the traits assessed by IEA.

Abbreviations:

“GLR_IEA”: IEA *General Linguistic Range* trait; “DSC_IEA”: IEA *Development, Structure and Coherence* trait; “Vocabulary_IEA”: IEA *Vocabulary Range* trait; “GUM_IEA”: IEA *Grammar Usage and Mechanics* trait.

Two observations can be made from these figures. First, in a two-dimensional space, the traits assessed by human markers, as a group, are clearly separated from those traits assessed by IEA, on a single dimension. In Figure 10.5, the two sets of traits are separated on Dimension 1, and in Figure 10.6, they are separated on Dimension 2. Note that the names of the dimensions and the orientation of the plots are arbitrary.

This separation potentially has great implications for AES research, because it could provide supporting evidence for fundamental differences between IEA and human scoring. The most plausible interpretations of this separation, taking human marking as the benchmark, include: 1) IEA is not capturing all of the writing characteristics that are relevant to the construct; 2) IEA is measuring extraneous features irrelevant to the construct; 3) IEA is measuring the writing traits differently to the human markers, in a systematic way.

The one dimension that separates IEA traits from human traits in a two-dimensional space could be reflective of those qualities in written prose where the critics believe that the machine is never able to appreciate, let alone to judge, such aspects as the socio-cultural aspect of the writing, complex logical reasoning, creativity and originality, evidence of abstract concepts, use of irony, or the extended use of metaphor or allusion.

Further studies are needed to determine whether this separation between human and IEA traits along a single dimension is a repeatable phenomenon and if so, to further explore the reasons for separation.

The second observation is that traits assessed by the human markers seem to be more independent of each other (i.e., they are more dispersed) than the IEA traits. This is particularly evident for essays written to the Tobacco prompt. In the two-dimensional space for this prompt (Figure 10.6), the IEA traits are closer to each other in distance (thus forming a smaller cluster in the space, as shown in the figure) than those traits assessed by the human markers. For this prompt, scores assigned to the analytic traits by the IEA scoring system are

evidently more similar (or less discriminant) amongst themselves than those awarded by the human markers.

The analysis performed here can be used to assess the potential impact of a common problem in marking—“halo effect” —on scores produced. Though there are a number of definitions in the literature, this chapter adopts the definition put forward by Robbins (1989) which refers to the halo effect as “the tendency for an evaluator to let the assessment of an individual on one trait influence his or her evaluation of that person on other traits” (Robbins, 1989, p. 444).

The propensity of markers to transfer their judgements on one aspect of writing performance to another is long recognised as an issue in the human marking behaviour (e.g., Cooper, 1984; Robbins, 1989; Saal et al., 1980). Compared to other classical problems in human marking, the halo effect seemed to have received the most attention in the research literature (Myford & Wolfe, 2003). For example, Cooper (1984, p. 7) noted that “readers are much more likely to seek out and emphasise errors in poorly developed essays than in well-developed ones”; and this halo effect works the other way as well. A well-organised and interesting paper “prompt[s] readers to overlook or minimise errors in spelling, mechanics, usage, and even sentence structure”. The effect impacts on marker’s ability to adequately discriminate among conceptually distinct aspects of an individual’s writing performance. This causes scores assigned by a marker on the analytic writing traits for the same essay to be more homogeneous than they should be.

Traditionally, the halo effect has been studied in the literature through the following methods: inter-correlations among scores on traits; factor analysis of the trait inter-correlation matrix;

variances or standard deviations associated with scores of a given examinee across all traits (Saal et al., 1980). The MDS analysis conducted in this section presents another way of studying halo effects, particularly in terms of comparing the effect of the halo across different scoring methods.

Results from the MDS analysis conducted in this section indicate that, at least for the Tobacco prompt, the IEA traits are clearly more homogeneous than human traits. This seems to indicate that the halo effect is more conspicuous in the IEA scores than in the human scores. It is noted that, as spelling is not included in the IEA traits, and it is included in the human traits, this difference would have impacted the spread of the IEA traits to an extent. However, taken together with the measurement evidence from Section 9.5.2, which suggested that scores on two IEA language traits contributed less distinct information than desired to the measurement process, it is quite likely that these results reflect the IEA's inability to discriminate the various conceptually different traits as well as the human markers. It is recommended that future studies, with larger sample sizes and with a more robustly measured spelling trait included into the set of IEA traits, be repeated to confirm these patterns.

A plausible explanation for the patterns observed in this section is that the IEA models are trained on human scores, hence the inherent problems in human judgements (e.g., errors and biases arising from halo effect) may also have been modelled into the IEA scores. It would seem that, when this happens, problems in the human scores may have been exacerbated, rather than simply being reproduced. These results therefore lend further support to the

importance of evaluating the adequacy of human marking processes and the quality of the resultant scores as a prerequisite for using these scores to train and develop AES models.

10.5 Chapter Summary

This chapter demonstrated a number of new statistical methods that could be used to expose and compare the structural patterns in the analytic scores awarded for the different writing traits by the IEA and the human scoring methods, as a means to accumulate evidence for the structural aspect of IEA validity. Results from these analyses were mixed. On the one hand, there was evidence to suggest that there was similarity in the two scoring methods, as both produced scores that revealed the same gender difference in the test-takers' performance.

On the other hand, the characteristics of the second factor detected in the IEA and human residual data after the influence of the dominant Rasch measures were removed, were inconsistent. Although the human trait scores confirmed the conceptual distinction between the higher order traits and language traits as hypothesised from the domain theory, the IEA trait scores did not reveal the same pattern. In addition, separate Multi-dimensional Scaling analysis (MDS) revealed that traits assessed by the IEA seemed to be less independent of each other than the similar set of traits measured by markers, indicating that the IEA may not discriminate among the various traits as well as human markers. These results, if repeated in future studies, are likely to weaken the argument for the structural fidelity of the IEA scores. Such results also have implications for the educational benefits of this AES system. Being able to accurately identify essays that have non-uniform score profiles (i.e., essays that exhibit

different levels of performance across different traits) enhances the instructional value of an AES system, because such information could be used to generate feedback information about students' strengths and weaknesses in particular aspects of writing.

The MDS analysis also showed that, human traits were clearly differentiated from the IEA traits as a group in a two-dimensional space. The existence of one single dimension that separated the two sets of traits in this space potentially has far reaching implications as it could be an indication of the IEA capturing systematic construct-irrelevant variance; or it could be reflective of IEA not measuring all important parts of the construct. The reason for the separation between IEA and human traits will need to be investigated further by future research studies.

The new approaches used in this chapter to analyse and present differences in internal structures of the human and machine scores, in themselves, have implications for AES research studies. These approaches (i.e., methods used to investigate the nature of the second factor in writing trait scores and patterns of similarity/dissimilarity amongst traits in a two-dimensional space) have proven to be not only sensitive to scoring differences, but also effective in demonstrating these differences in visual ways that are easily grasped by non-technical audience. Consequently, these approaches have the potential to help stakeholders to better appreciate the implications of AES use. The next chapter considers issues associated with the IEA trait-level scoring.

Chapter 11 Examination of the Individual IEA (Intelligent Essay Assessor) Traits

11.1 Introduction

This chapter collects and considers evidence in relation to the accuracy and the reliability of the IEA (Intelligent Essay Assessor) scoring at the trait level. The investigative process undertaken in this chapter illustrates how the following validity question of the proposed validation framework can be investigated: *Can the writing traits be accurately and reliably assessed by an Automated Essay Scoring system?* It is noted that the analysis conducted in this chapter encompasses the identifications of any aspects in the IEA scoring of individual traits which may compromise or strengthen the validity of the overall scores. This includes discussing validity implications arising from the scoring criteria used for the scoring of the IEA traits. Scoring criteria are a part of the IEA scoring process, and their appropriateness is a validity question for the scoring procedure component of the AES (Automated Essay Scoring) validation framework.

In the course of the investigations, this chapter makes use of the anomalies identified through the Rasch analysis for the IEA trait scores in Chapter Nine. These anomalies epitomise trait scores that did not meet the expectations of a uni-dimensional model; that is, they did not align well with the overall patterns of the scores across all traits and across all test takers. They present as good examples for close scrutiny to verify whether the underlying cause of the unexpected score profiles observed is due to problems in the IEA scoring of particular

traits or is attributable to the individuals having different levels of proficiency on different writing dimensions.

Table 11.1 reports the top three traits that have the greatest number of unexpected IEA trait scores across the two prompts, as identified by the Rasch program – Winsteps. The unexpected scores are defined as those scores which have the absolute standardised residuals greater than or equal to 2.

Table 11.1

Top Three Traits with the Greatest Number of Unexpected IEA Trait Scores Across the Two Prompts

Trait	Count
<i>Spelling</i>	29
<i>Formal Requirement</i>	16
<i>Content</i>	12
All other traits	17
Total	74

Note: The unexpected responses are those that have absolute standardised residuals ≥ 2 .

It is noted from Table 11.1 that overall, *Spelling* scores account for 39% of the unexpected scores. This is followed by the *Formal Requirement* scores (22%) and the *Content* trait scores (16%). This chapter therefore examines the IEA scoring of each trait, with particular emphasis on these three traits – *Spelling*, *Formal Requirement* and *Content*. Throughout the chapter, where examples are sourced from the essays used in this study, these example essays are

denoted by an essay sequence identifier (seq#), to prevent identification of the individuals who produced the essays.

It should be noted that the results from the analysis in this chapter are attributable only to the IEA scoring capabilities at the time that the Pearson Test of English (PTE) Academic field tests were conducted in May and June 2008. Since then, there have been continuous improvements made by the test developers to the IEA scales and to the measurement capabilities of the system. Where such information is available, these improvements are noted in the relevant parts of the chapter to add currency to the discussions.

11.2 The IEA Scoring of the *Spelling* Trait

Rasch analysis in Chapter Nine provided strong evidence that the IEA *Spelling* trait was not functioning as well as expected, relative to a uni-dimensional model (the Rasch model). This section examines the underlying cause of this observation by first inspecting the IEA scoring criteria for the *Spelling* trait which can influence the validity and interpretability of the scores produced. Investigations of the broad factors that may influence the capabilities of the automated spelling checking technologies are considered subsequently.

11.2.1 *Scoring Criteria for the Spelling Trait*

Table 11.2 describes the scoring criteria used for the IEA *Spelling* trait; that is, how ordered score points on the *Spelling* rating scale are assigned by the IEA in accordance with the IEA scoring criteria.

Table 11.2

Descriptions of the IEA Spelling Rating Scale Used for the PTE Academic

Score Category	Descriptions
2	Correct spelling, but there may be one typing error
1	Contains one spelling error and/or more than one typing error
0	More than one spelling error and/or numerous typing errors

Source: Pearson Academic Score Guide, November, 2011 (Pearson, 2011b, p. 60). More details see Appendix A

There are a couple of issues worth noting from Table 11.2. First, IEA uses the number of spelling errors detected in an essay (including genuine spelling and typographical errors) as a basis for assigning scores on the *Spelling* scale. This means that one spelling error in an essay of 50 words would be given the same *Spelling* score of ‘1’ as one spelling error in an essay of 300 words, assuming no typographical errors are made in both instances. Although the two candidates in this example have demonstrated different levels of spelling competency by way of different spelling accuracy ratios, the IEA nonetheless assigns the same spelling score to both candidates. A consequence of this is that, given equivalent spelling capability, the test taker who is more productive (a general sign of greater writing ability) and writes more words is likely to generate more spelling errors, and as a result, receives a lower score on the *Spelling* trait. This compromises the link between the IEA spelling score, the underlying spelling competency, and the writing ability. Two example essays with *Spelling* scores identified as having large standardised residuals from the Rasch analysis (attached at Appendix L), help further illustrate this point.

The first example is a well written essay of 228 words (seq# V2105), which received the highest score of 3 from the 2 human markers recruited for this study for all 5 traits. However, as the essay had 2 *Spelling* mistakes, it received a '0' spelling score from the IEA. The second essay (seq# V287) is a considerably shorter essay of 63 words, which received an average score of 1.1 out of 3 across the 5 writing traits from 2 human markers recruited for this study. This less well written essay received a spelling score of 1 because it had one spelling error. Even though the longer essay demonstrated a slightly higher spelling accuracy ratio ($99.1\% = 100 * (1 - (2/228))$) than the shorter one (98.4%), it received a lower score on the *Spelling* trait despite being nearly three times longer than the one produced by a less proficient test taker, with potentially more opportunity for error.

One recommendation to improve the validity of the IEA scoring of the spelling for PTE Academic is to judge the spelling competency by assessing the impact of the spelling errors on the communicability of essays. This can be partially achieved by using spelling accuracy ratios such as $100 * (1 - \frac{\text{number of spelling errors}}{\text{total number of words}})$ (Formula 11.1), rather than using the absolute number of spelling errors, which is unduly influenced by essay length.

The second point that arises from the scoring criteria is that the IEA spelling score is dependent on a count of errors made, without taking into account the severity of the errors made. As noted by Mitton (1996), simply marking spelling as right or wrong is a very crude way of assessing spelling competency. The coarseness in such a method does not reflect the

complexity of the spelling process in a test, nor does it enable the resulting scores to discriminate the competency being measured, in a robust manner.

Mitton (1996) observed that “the poorer spellers did not just make more errors; they also made worse errors” (p. 51). His studies (1996, 1987) on the misspelling patterns across spellers of different competency levels indicated that, while the competent spellers generally made spelling mistakes that involved single letter violations (i.e., the misspelt word is one letter different from the correct word or involved the transposition of two adjacent letters), the poorer spellers made errors that differed more substantially from the correct words.

Researchers assert that the seriousness of the misspellings are reflections of spellers’ spelling proficiency including knowledge of the complex English spelling rules, comprehension of odd features in English orthography, as well as the ability to perform phonemic analyses correctly and to use the analyses results to guide the spelling of a word (Baron, Treiman, Wilf & Kellman, 1980; Ellis, 1994; Mitton, 1987, 1996; Perin, 1983;).

Since the seriousness of the spelling errors reflects the level of spelling proficiency, a further recommendation is that the IEA *Spelling* scale be refined to incorporate measures of error seriousness, such as the Levenshtein edit distance measure which calculates the number of error-spots in a misspelling compared to the intended word (Hayes, 2010). However, if such a measure is implemented in a real rating scenario, a challenge for the machine scoring system will be to work out what the intended word is, since the essays written to independent writing tasks are unconstrained texts.

An additional method to strengthen the substantive link between the IEA spelling scores and the competency being measured is to adjust the spelling scores by the difficulty of the words misspelt. Everything else being equal (e.g., same number of error-spots in misspelt words), errors made on more frequently used words should, in general, indicate a lower proficiency level than errors made on more difficult or rarely used words, provided that these errors are genuine competency errors, not typographical errors.

A third point from Table 11.2 is that the scoring criteria for some score points on the *Spelling* scale are somewhat ambiguously specified. For example, the separation of score point 1 from 0 on the scale partially hinges on the difference between “numerous typing errors” and “more than one typing error”, definitions of which are not included in the scoring criteria for the IEA *Spelling* trait.

Although this type of ambiguity is not uncommon in human marking systems with scoring criteria used by human markers often containing words such as “some errors”, “numerous errors” and so on, supplementary training packages for the human markers more often than not provide guidance and benchmark papers to explain and illustrate how the criteria should be implemented. However, with the IEA *Spelling* scale, there does not seem to be any published supplementary material to explain how the IEA scoring system makes the difference between “more than 1 error” and “numerous errors”. It may be speculated that the IEA learns to make such a difference through its training on a batch of pre-scored essays. Nonetheless, the lack of information about this in the public domain not only increases the difficulty for an independent evaluator to make an informed judgement of the validity of the

IEA scoring method, but it also impacts on the interpretability of the scores produced from this scale, which is an aspect of score validity.

A fourth point concerning the scoring criteria for the IEA *Spelling* trait is related to the identification and the treatment of the typographical errors in the scoring of the *Spelling* trait. Typographical errors are influenced by the test takers' typing skills, which are not related to the spelling competency that this trait is designed to measure. Unless the definition of the construct intended to be measured by the PTE Academic writing tasks is changed to "the ability to write on computer", the extraneous factor can introduce construct-irrelevant variance and make the spelling scores less aligned with the underlying writing ability.

The scoring criteria for the IEA *Spelling* trait attempt to acknowledge the undue influence of typing ability on the validity of the spelling scores by trying to separate genuine spelling competency errors from typographical errors and by giving more importance to spelling errors than to the typographical errors in assigning the spelling scores. The intent is a valid one and should enhance the substantive link between spelling scores and the spelling competency that these scores are intended to reflect. However, the analysis conducted in this study suggests that there is some inaccuracy in the IEA scoring process to differentiate typographical errors from genuine spelling competency errors.

This study analysed all 12 essays which were identified through human scores as having demonstrated a high level of English academic writing proficiency but which were assigned a score of 0 for the *Spelling* by IEA. All these essays received full marks unanimously from the

two human markers recruited for this study across all five analytic traits including the *Mechanics* trait. Analyses reveal that, although a human marker can reasonably judge the small number of misspellings made in each of these essays as mainly typographical errors, thus not penalising the errors by marking down on the *Mechanics* trait, the IEA seems less capable of making the same judgement. In fact the inability of the IEA to accurately discriminate between typographical errors and spelling errors is the main factor explaining why these 12 essays received a 0 score from the IEA for the *Spelling* trait.

One example from this pool of essays is provided in Appendix M to support this finding. In this example (essay seq#T12), there are only two misspelling tokens (instances of misspelling): gouvernement and Goverment (both underlined in the essay). It is safe to assume that both are typographical errors for two reasons. The main reason is that the writer was able to spell the word *government* six times accurately somewhere else in the same essay (highlighted in green in the essay). Another reason is that the respective misspelling patterns in the two error tokens, one involving a single letter insertion error and the other involving a single letter omission error, are some of the common categories of mistypings as identified by various researchers (Mitton, 1996; Pollock & Zamora, 1983). It can thus be rationally inferred that the IEA assigned a spelling score of 0, an undeservedly low score, because it erroneously categorised both error tokens as genuine competency errors, rather than typographical errors.

The demonstrated deficiency in the IEA measurement capability helps explain why three quarters of all the essays in the sample received a spelling score of 0 from the IEA, as many

essays have more than one typographical error. As a result, the measurement properties of the *Spelling* scale were seriously compromised.

It is fair to say that it is sometimes difficult for human markers to separate the genuine competency errors which are reflections of knowledge deficiencies in the English orthographical system from performance slips which can be attributable to external factors such as carelessness, fatigue and typing ability. However, the evidence presented above suggests that human markers seem to be more capable of making a holistic judgment of the spelling competency being measured than the IEA, which tends to just isolate and judge the errors as they appear.

The next section expands on the issue of the accuracy in the IEA scoring of the *Spelling* trait by discussing generic challenges that any automated spelling checking technologies (referred to hereunder as “spelling checking programs”) face in order to identify spelling errors accurately and reliably.

11.2.2 Inherent Challenges in the Automated Scoring of Spelling

Analysis of the spelling error corpora from the sample data used for this study and reviews of relevant literature reveal that there remain significant gaps in the accuracy of machine scoring of spelling, contrary to the common perception that machine scoring of spelling is accurate and should be identical to human scores “if the human raters work error-free” (Pearson, 2009, p. 2).

This section lists the challenges that are common to all spelling checking programs, since the exact technology implemented by the IEA to check the spellings is not disclosed in the public domain. However, whenever possible, the misspelling examples used in the discussions are sourced from the sample data used in this study. Again, these examples are denoted by the essay sequence identifier (seq#) from the sample.

The first significant challenge that the spelling checking programs face in order to identify spelling errors accurately relates to “real-word errors” (i.e., when words are correctly spelt but they are not the intended ones) (Mitton, 1996). For example, when “to” is written for the intended word “too” (e.g., “the Governments all *to* often take the sides of the companys” – seq# T122) or “*were*” for “*where*” (e.g., “if there *where* less smokers” – seq# T18). Italics were inserted by the present author.

While spelling checking programs can flag with accuracy and objectivity the non-word errors (i.e., words which are misspelt and can’t be found in the dictionary), it is challenging for them to identify real-word errors because the misspelt words exist in the dictionary and there is no certain way for the computers to know what the intended words are.

This problem is exacerbated by the fact that real-word errors can be quite common both in handwritten and in typed essays. Studies have demonstrated that one quarter to one third of the spelling errors in the handwritten essays can be real-word errors; more if word division problems (i.e., words incorrectly divided into two in such a way that both parts are real words) are counted towards spelling errors (Brooks, Gorman & Kendall, 1993; Mitton, 1996;

Sterling, 1983; Wing & Baddeley, 1980). There is also evidence to suggest that typographical errors can lead to more real-word errors, which in turn increases the presence of real-word errors in the typed responses (Damerau, 1964; Grudin, 1983; Peterson, 1986; Pollock & Zamora 1984). The prevalence of real-word errors in the essays thus presents a significant challenge for spelling checking programs.

The second challenge that the spelling checking programs face is related to spellings which may not be in a dictionary but are in fact correct. This mostly happens with proper nouns, but also with neologisms and foreign words. Inaccurate reporting of proper nouns as spelling errors is acknowledged by *e-rater* developers as a problem (see Quinlan et al. 2009, p. 19). Examples of this problem found in the sample data used for this study include: Soeharto (seq# V285), an acceptable spelling of the former Indonesia's president, is identified by the spelling checking program implemented in Microsoft Word 2010 as a spelling mistake. In a similar vein, “*vis-a-vis*” (seq# V11) or “*lefko*” – Greek word for “white” as explained in an essay (seq# V2107), are all picked up by this spelling checking program as spelling mistakes. Although an obvious solution to this problem is to expand the dictionary used by these programs to contain common proper nouns, this has to be balanced with the risk of letting a legitimate spelling error slip without detection due to a larger than necessary dictionary. For example, including “baht” as the Thai currency in the dictionary may prevent the spelling “baht” from being identified as a spelling error, but it would also allow a legitimate misspelling intended for words such as “bat” or “bath” to be accepted.

The complexity of spelling errors is readily acknowledged by test developers including those of the PTE Academic and various ways to improve the performances of spelling checking programs are being explored (John De Jong from Pearson, personal communication, December 2010). Efforts being made in the spelling checking technology field to address these challenges include the incorporation of syntactical/grammatical rules in spelling checking programs, development of context-sensitive programs and the use of probability models to predict intended words (*Google Wave*, n.d.; Mitton, 1996). Though progress is being made, there still exist gaps in the capabilities of these programs to appropriately measure the underlying spelling competency.

11.3 IEA Scoring of the *Formal Requirement* Trait

11.3.1 Introduction

This section focuses on the Formal Requirement trait. Table 11.3 describes the scoring criteria used for the *Formal Requirement* trait.

Table 11.3

Description of the Formal Requirement Scale

Score Category	Descriptions
2	Length is between 200 and 300 words
1	Length is between 120 and 199 or between 301 and 380 words
0	Length is less than 120 or more than 380 words. Essay is written in capital letters, contains no punctuation or only consists of bullet points or very short sentences.

Source: PTE Academic Score Guide, November, 2011, (Pearson, 2011b), p. 60. More details see Appendix A

As can be seen from Table 11.3, the *Formal Requirement* trait is primarily scored according to the length of an essay. In rare cases, when an essay does not meet other formal requirements (e.g., it is written in capital letters, contains no punctuations, or consists only of bullet points), the essay is given a score of 0 for this trait. The appropriateness of having a length-based criterion in the IEA scoring model will be the focus of this section.

The instructions accompanying the prompts for the PTE Academic Independent Writing Tasks ask the test takers to write 200 – 300 words in their essays. During the test, PTE Academic provides a counting box on the screen to inform the test taker of the number of words written.

According to the PTE Academic (Pearson, 2011b, p. 59), this trait also acts as a “minimum requirement” to the calculation of an overall score for an essay written to an independent

writing task for PTE Academic. If the score on this trait is 0, then none of the other traits will be assessed and the essay will receive a score of 0 as the overall score.

Accuracy is not an issue with regard to the measurement of this trait as the machine is reliable in evaluating essay length. The key question from a validity point of view is “Does the trait contribute to the construct?” For the question to be answered, there is a need for both empirical evidence and theoretical arguments to support the substantive link between measurements from this trait and the underlying writing ability construct. The next section collects the evidence supporting or challenging such a link.

11.3.2 Impact of the Scores from the Formal Requirement on the Validity of the Overall Scores

Analysis of the Item Characteristic Curves (ICC) for this trait conducted in Chapter Nine identified a cluster of responses on the ICC for the Voting prompt that warrants further investigation (See Figure 9.8 in Section 9.5.2). This cluster consists of essays at the highest writing ability level but with an average score of 1 on the *Formal Requirement* trait. The expected score from the Rasch model on this cluster of essays for the trait is 2, given these essays’ average location on the developmental continuum.

An examination of the essays in this cluster identifies one type of anomaly with one essay example illustrating the point. This essay (seq#V276, see Appendix N) demonstrated a high level of overall writing ability and received the highest score across all five analytic traits

from the human markers recruited for this study unanimously. Though the IEA awarded similarly high scores for all other traits, it assigned a score of 1 to the *Formal Requirement* trait because the word count is 301, just one word over the upper limit for the highest point of 2 for this trait. According to the calculation formula for the overall score, this 1 point loss carries the same significance towards the overall score as a one point loss on other far more salient writing traits (such as *General Linguistic Range*, *Grammar Usage and Mechanics*, *Vocabulary Range and Development*, *Structure and Coherence*). In this example, writing one more word should not have been penalised in such a disproportionate manner. It would appear in this instance that the score produced for this trait does not contribute to the overall measurement process; rather it weakens the interpretability and inferential property of the overall score generated by the IEA for the PTE Academic writing task.

Theoretically this issue can affect all essays that approximate the word limits for each of the score points for this trait. Figure 11.1 shows the essays that are on the margin of the word limit for the score point 2 for the *Formal Requirement* trait. The y-axis in this figure represents the calculated average of the marks assigned by 2 human markers over the five analytic traits of the ESL Composition Profile, for each essay. The average mark ranges from 0.0 to 3.0, and is used here as a proxy measure for the overall writing ability as assessed by the human markers for each essay. Green dots in the figure represent those essays which were 10 words less than the lower word limit (i.e., 200) and red dots represent those which were 10 words over the upper word limit (i.e., 300). In total, 16 (or 6%) of all essays in the samples across both prompts are in the neighbourhood of the word count range specified for the score

point 2. Drawing a horizontal line, such as the one in the graph, makes it easy to locate those essays which demonstrate similar levels of writing ability to the essays in green or red colours, but are advantaged by the scoring of this trait because they have produced a few more or less words. It can thus be argued that essays that are on the margin of the word limits for the score points of this scale are likely to be unduly disadvantaged or advantaged by such a rigid cut off based purely on the number of words.

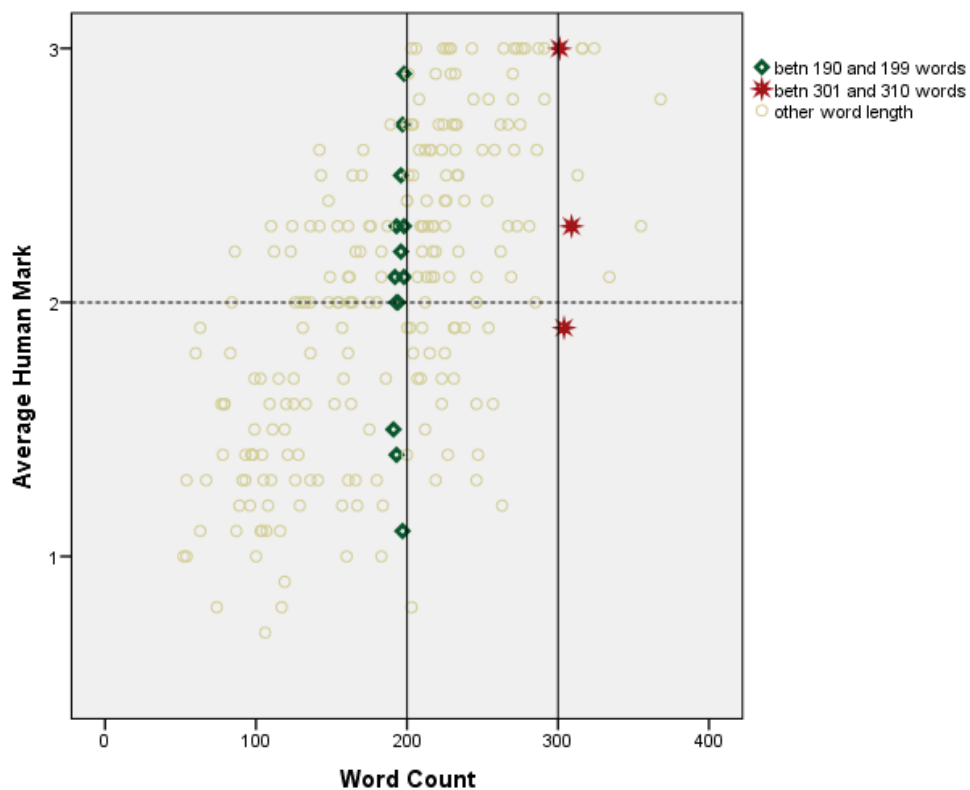


Figure 11.1 Average Trait Marks by Word Count

Note: Each symbol represents an essay. The y-axis shows the average of the human marks, assigned by 2 human markers, over the five analytic traits of the ESL Composition Profile, for each essay. The maximum score is 3.0. The average mark is used here as a proxy measure for the overall writing ability as assessed by the human markers. Essays written to both prompts are combined in this analysis.

The problem is exacerbated for those essays which are on the margin of the word count limit for the score point 0; that is, just a few words shorter than 120 words or a few words longer than 380 words. These essays would receive a score of 0 for the *Formal Requirement* trait and with this trait being a requirement, none of the other traits would be scored, resulting in a total score of 0 for these essays. It is argued here that this requirement can potentially disadvantage particular types of essays, such as those that are short but to-the-point, and when that happens, the IEA-generated overall score provides a distorted picture of the underlying writing ability. This point is best illustrated in Figure 11.2.

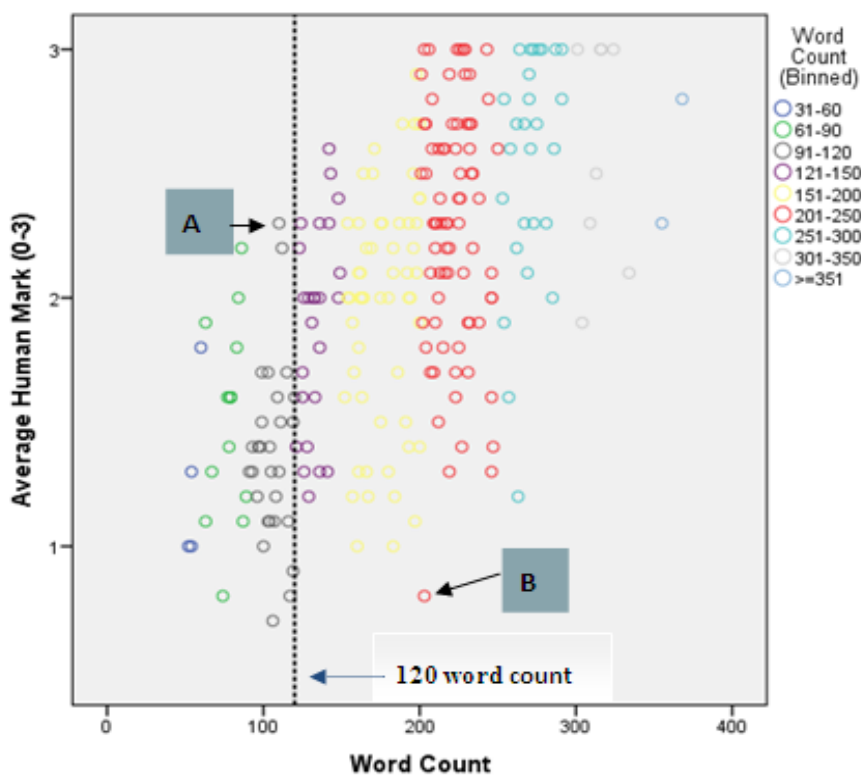


Figure 11.2 Examples of Potential Anomalies Arising from the *Formal Requirement* Trait Scoring

The y-axis shows the average of the human marks, assigned by 2 human markers, over the five analytic traits of the ESL Composition Profile, for each essay. The average mark is used in this analysis as a proxy measure for the overall writing quality as assessed by the human markers for each essay.

As in the previous figure, the y-axis in Figure 11.2 shows the average mark given by human markers over the five analytic traits, for each essay. The average mark is used here again as an indication of the overall writing ability, as assessed by human markers, for each essay. It is on a scale of 0–3. For easy reference, this scale is referred to hereunder as the ‘ability scale’.

Figure 11.2 shows that, for the group of essays under 120 words across both prompts, there is wide variation in the overall writing ability as assessed by the human markers, with the average marks ranging from under 0.7 to 2.3 on the 0–3 ability scale. This means that there is still a great deal of pertinent information on the overall quality of these essays.

Notwithstanding this, the IEA scores would provide no measurement information for this group of essays because of the word limit. Furthermore, it is observed that human markers assessed some of the essays which were shorter than the minimum requirement as demonstrating considerably higher writing ability than some of the others who have met the desirable word range requirement. Two examples (marked A and B in the graph with corresponding essays being provided at Appendix O) are provided to illustrate the point.

Essay A, though 10 words short of the 120 word limit, received an average score of 2.3 on the 0–3 ability scale from human markers recruited for this study. It was considered by the markers as having communicated one main point related to the topic in a clear and effective manner. In particular, human markers were unanimous in assigning the highest score for the two language traits – *Vocabulary* and *Language Use*. This is in recognition that the writer has demonstrated a high-level ability to use the syntactical and lexical elements of the language to

deliver the message effectively. However, since Essay A is less than 120 words, none of these traits was assessed by IEA. The essay automatically received a total score of 0 from the IEA.

On the other hand, Essay B (203 words) received a considerably lower average mark, 0.8 out of the maximum of 3, from human markers, as the essay demonstrated less communicative effectiveness. Compared to Essay A, Essay B was more difficult to understand because of the severity of the language errors it had. It is also likely that the writer's limited linguistic skills had restricted his/her ability to develop the central thesis in a succinct manner. However, as Essay B was within the most desirable range of the word limit for the *Formal Requirement* trait, it was assessed on all seven traits by IEA and was assigned a total score of 7 out of the maximum score of 15. A noticeable contributor to the total score was the score of 2 that came from the *Formal Requirement* trait. As a consequence of using the *Formal Requirement* trait as a minimum requirement, the respective overall scores produced by the IEA clearly did not paint an accurate picture of the relative order of the writing ability of these two writers, as assessed by experienced markers.

11.3.3 Other Validity Concerns Relating to the Formal Requirement Trait

While the preceding paragraphs discussed where the use of this trait might disadvantage or advantage certain types of essays, a potentially more serious concern of having an essay length criterion in an automated scoring system is the undesirable consequences of this practice. Such a practice can encourage unhealthy and unproductive learning or teaching behaviour, as “students can simply be coached to write essays to a certain length and to write

more or less about the subject specified by the item” in order to gain higher scores in the tests (Kaplan et al., 1998, p. 10).

The direct inclusion of surface features such as a length criterion in a scoring model also makes the scoring model less educationally defensible. The following questions may be asked by assessment and education professionals in relation to the thresholds set for word limits: “how are the cut off sizes (e.g., 120 words, 380 words etc) determined?” “Is there any strong evidence linking the nominated word limits to the different levels of writing ability?” In writing rubrics used by human markers in similar testing contexts such as TOEFL (Appendix E) and IELTS (IELTS, n.d.), essay length is not an explicit criterion. That is because, conceptually, length is not a criterion that can be used to directly define writing proficiency, even though there is a strong correlation between length and human scores, mediated possibly through variables such as organisation, development and sentence fluency.

The main rationale for having the *Formal Requirement* trait included in the PTE Academic writing ability construct is the observation that academic writers are often required to “provide highly relevant and sometimes quite complex information, while observing a strict (and sometimes quite tight) formal requirement relating to the number of words” (John De Jong, Pearson, personal communication, May 19, 2010). However, it can be argued that counting the number of words in an essay is a fairly crude and unreliable way of measuring the underlying competencies required to summarise complex information in a written product with a word length limit. Unless the machine scoring engine has very sophisticated semantic models to analyse the content for superfluosity or relevance, including a length criterion in

the scoring model will increase the IEA's vulnerability to test-taking strategies and bad-faith writing. In turn this will heighten the risks of the IEA providing distorted writing ability scores that are unduly influenced by a feature that is not directly related to the underlying ability construct.

As a result of the validity implications mentioned above, researchers such as Sheehan (2001), Chodorow and Burstein (2004), and Landauer et al. (2003) have all argued that the development of an automated essay marking system should avoid using any direct measures of essay length.

11.4 The IEA Scoring of the *Content* Trait

This section examines the accuracy of the IEA scoring of content by comparing the content scores assigned by IEA to those assigned by the human markers, analysing discrepancy rates at a score point level, and examining those essays which have seriously discrepant content scores between the human markers and IEA.

For essays written to the writing tasks of PTE Academic, the accuracy in the IEA scoring of content is crucial, because *Content* is the second trait which acts as a requirement to the calculation of a total score by IEA. If the *Content* score as assessed by the IEA is less than 0.5, then no other traits would be assessed and IEA would automatically assign a total score of 0 to the essay. Thus, if there is evidence of the IEA significantly under-estimating the quality and quantity of the content for some types of essays, the scoring of the content would

have additional implications for the validity of the total scores generated by the IEA.

Additionally, it is necessary for IEA to have the capacity to detect superfluous, repetitive, or irrelevant content to ensure that the test taker is not just filling the words to get the maximum score on the length requirement trait, in order to boost the overall score.

For analysis in this section and in the subsequent sections concerning the remaining four traits, human scores used are those from the field test data supplied by Pearson. The main reason for the choice of this source data is the desire to examine rates of agreement between human scores and IEA scores at the score point level. A previous section (Section 3.2) argued for the importance of this type of analysis as AES systems' performance can vary significantly across different writing proficiency levels. As the scales used by the human markers in the field tests are the same as those used by IEA, this choice of the source data facilitates the calculation of agreement rates, in particular agreement rates at a micro-level, for each of the five traits. In these analyses, it is assumed that the scoring criteria used by both IEA and human markers to score these traits are appropriate. To improve the reliability of the results, data from all 391 essays as supplied by Pearson are used. For each trait, the final human scores are calculated as the averages of the two scores (or in the case of an adjudication, the average of the closest two scores) each essay received on each trait. These scores are simply referred to hereunder in the following section as "human scores".

11.4.1 Overall Agreement Between IEA Content Scores and Human Scores (Obtained from the PTE Academic Field Tests)

This section compares the *Content* scores assigned by IEA to those by the human markers using the same IEA 0–3 *Content* scale. The scoring criteria for *Content* are shown at Appendix A. Table 11.4 displays the distributional properties of the IEA and human *Content* scores for the same essays.

Table 11.4
Means and Standard Deviations of Human and IEA Content Scores

Prompts	No of Essays	IEA		Marker 1		Marker 2		Human (final) Score		paired <i>t</i> -test (Human–IEA)
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Voting	187	1.75	0.65	1.72	0.96	1.78	0.98	1.78	0.9	$t(186) = 0.39$, $p = 0.70$
Tobacco	204	2.07	0.71	2.01	0.94	2.02	0.95	2.12	0.89	$t(203) = 1.02$, $p = 0.31$

Note: SD: Standard Deviation. *p* values for paired *t*-tests are two tailed *p* values. Before calculating these statistics, the IEA scores are first rounded to the nearest whole numbers.
Human–IEA: comparisons between human (final) scores and IEA scores.

As indicated on Table 11.4, paired *t*-tests comparing the two sets of scores confirm that the IEA *Content* scores are not statistically different from human (final) scores across the two prompts [$t(186) = 0.39$, $p = 0.70$ for the Voting prompt and $t(203) = 1.02$, $p = 0.31$ for the Tobacco prompt]. Similarly, the means of the IEA scores are not substantially different from those of the individual markers. However, human content scores, either those from individual markers or final scores, are noticeably more variable than the IEA scores. The standard deviations for scores assigned by Markers 1 and 2 are 0.96 and 0.98 for the Voting prompt, and 0.94 and 0.95 for Tobacco, greater than the corresponding standard deviations of the

scores assigned by the IEA (0.65 for Voting and 0.71 for Tobacco respectively). Although the standard deviations are smaller for human (final) scores than for the scores from the individual markers, they are still statistically greater than the standard deviations of the IEA scores across both prompts [$F(186,186) = 1.92 > F_c(186,186) = 1.27$ for Voting, $F(203,203) = 1.57 > F_c(203,203) = 1.26$ for Tobacco, at $\alpha = 0.05$ level].

Table 11.5 reports the exact agreement and exact + adjacent agreement rates between the IEA and human *Content* scores. While the exact agreement rates show the proportions of the times the IEA-generated *Content* scores matched exactly with the human scores, the exact + adjacent rates indicate the proportions of the times the IEA scores are within one score point of the corresponding human scores. In order to get a better understanding of the true level of agreement between two sets of scores, the Kappa rates (Cohen, 1960) which adjust the raw agreement rates by taking out the amount of agreement that would be expected by chance alone, are also reported³⁶. To add a perspective to these IEA-human correspondence rates, the inter-rater reliability in the form of the agreement between two human markers is also included in the same table.

³⁶ Kappa is a measure of the difference between the observed agreement and the expected agreement by chance alone, standardised to be on a -1 to +1 scale. A Kappa rate of +1 indicates perfect agreement, while a rate of 0 is what would be expected by chance alone. Negative Kappa rates indicate agreement less than chance; that is, potential systematic disagreement by observers (Viera & Garrett, 2005, p. 361).

Table 11.5***Agreement Rates Between the IEA Content Scores and Human Content Scores***

		Exact Agreement (Kappa)	Exact + Adjacent Agreement (Kappa)
Voting	IEA/H1	0.46 (0.22)	0.91(0.85)
	IEA/H2	0.43 (0.20)	0.96 (0.93)
	IEA/Human Score	0.52 (0.31)	0.97 (0.95)
	H1/H2	0.47 (0.25)	0.88 (0.83)
Tobacco	IEA/H1	0.50 (0.27)	0.97 (0.95)
	IEA/H2	0.46 (0.22)	0.95 (0.92)
	IEA/Human Score	0.56 (0.34)	0.99 (0.98)
	H1/H2	0.45 (0.20)	0.88 (0.82)
Combined	IEA/H1	0.48 (0.26)	0.94 (0.91)
	IEA/H2	0.45 (0.22)	0.95 (0.92)
	IEA/Human Score	0.54 (0.34)	0.98 (0.97)
	H1/H2	0.46 (0.23)	0.88 (0.83)

Note:

IEA/H1: Agreement rates between the IEA and human marker 1; IEA/H2: Agreement rates between the IEA and human marker 2;

IEA/Human Score: Agreement rates between the IEA and human (final) scores; H1/H2: Agreement rates between marker 1 and 2;

Combined: All essays across the two prompts are analysed together.

Before calculating agreement statistics, IEA scores are first rounded to the nearest whole numbers.

Overall, across the two prompts, the IEA scores agreed exactly with the human scores 54% of the time, and agreed with the human scores within one score point 98% of the time. It is also observed that the degree of agreement between the IEA and the human scores, both in terms of the exact and exact + adjacent rates, is better than that between the IEA and an individual marker. This phenomenon is partly due to the fact that the IEA model is trained to model around the resolved human scores, not scores from individual markers, which is a desirable feature.

A further observation is that, while the IEA agreed exactly with an individual marker at a similar rate to that between two human markers, the IEA performed better than human markers in terms of the exact + adjacent agreement rate. Across the two prompts, the IEA agreed with an individual marker within one score point 95% of the time, while two individual markers did so 88% of the time. However these results should be interpreted with care as they are sensitive to the reliability of the human scores used. The lower the reliability of human scores, the more likely that an AES system will compare favourably to individual human markers, in terms of inter-scorer consistency.

Taking out the probability of chance agreement, the Kappa rate of 0.34 across the two prompts suggests that the IEA scores have a fair level of agreement with human scores, when the analysis focus is the rate of exact match between the IEA scores and human scores (using criteria recommended by Landis and Koch, 1977).³⁷ When the analysis focus is the rate of agreement within one score point, the Kappa rate of 0.97 indicates almost perfect agreement between human scores and IEA scores.

It is cautioned that Kappa rates are influenced by factors other than agreement, such as the prevalence (i.e., whether or not the rating categories are equally probable to be observed in

³⁷ Landis & Koch (1977) provided the following guidelines for interpreting the Kappa rates: Kappa < 0 indicating no agreement, 0–0.20 as slight agreement, 0.21–0.40 as fair agreement, 0.41–0.60 as moderate agreement, 0.61–0.80 as substantial agreement, and 0.81–1.0 as almost perfect agreement. This set of guidelines is not universally accepted in the literature.

the population under study) and the number of categories a rating scale has (Bakeman, Quera, McArthur & Robinson, 1997; Sim & Wright, 2005). Everything else being equal, the greater the number of rating categories, or the more equiprobable the categories are, the higher the Kappa rates. The Kappa rates reported in this study therefore should not be directly compared to those reported in other studies, unless these factors are taken into account.

11.4.2 Agreement Between the IEA Scores and Human Scores (from the PTE Academic Field Tests) at the Score Point Level

Though the analysis of overall agreement rates reveal encouraging results, to fully address the validity question, it is necessary to inspect the accuracy of the IEA scoring of *Content* at the score point level. This analysis uses the human (final) scores obtained from the Pearson field tests, assuming they reflect the underlying competency being measured. Table 11.6 reports the distribution of the IEA scores against the scores assigned by human markers, at each score point, using all 391 essays from both prompts.

Table 11.6

Matrix of the Frequency of Occurrence of Human and the IEA Content Scores Across the Two Prompts

IEA Content Score	Human Content Score				Total
	0	1	2	3	
0	5	4	1	0	10
1	11	60	7	4	82
2	1	55	94	79	229
3	0	1	16	53	70
Total	17	120	118	136	391

Table 11.6 provides a variety of information. First, overall it can be seen that the IEA assigned fewer scores that were either very low or very high, when compared to human markers, who used the same scoring criteria to score content. While human markers assigned a score of 3 (the highest score) to 136 essays, the IEA only did so for 70 essays. While human markers assigned 17 essays a score of 0, the IEA only assigned 10 essays a score of 0. These results indicate that when the IEA is used to score essay content, it seems to exhibit a greater “central tendency” than those human markers used in this analysis. “Central tendency” is a well-established phenomenon in human marking (Leckie & Baird, 2011). It refers to the propensity of markers to avoid using extreme categories of a rating scale or a preponderance to overuse the middle categories of a rating scale (Landy & Farr, 1983). This issue has been documented in various contexts including in the assessment of Advanced Placement English Literature and Composition essays (Myford & Wolfe, 2009) and in the national assessments of school student writing (Leckie & Baird, 2011). This tendency of avoiding the extreme categories has been suggested by some researchers as inherently difficult for some markers to

overcome (Linn & Gronlund, 2000). When markers avoid using the extreme categories, it leads to a reduction in the effective width of the scale and results in less discriminating scores (Anastasi, 1988). The fact that the IEA has assigned less extreme *Content* scores to the same essays indicates that when assessing content, this problem seems to be more pronounced in the IEA scoring than in human scoring. The above results also help explain why the IEA scores are observed to be more compact than human scores (e.g., the IEA scores have smaller standard deviations compared to human scores as reported in Table 11.4).

A related observation is that the discrepancy between the human scores and the IEA scores is greatest at the very high and very low ends of the *Content* scale. While the discrepancy rates between the IEA and human scores for the middle two score points (1 and 2) are 50% and 20% respectively, the same rates for the extreme score points 0 and 3 are 71% and 61%. These results indicate that, while the IEA's overall agreement rates seem to be satisfactory, these rates are not uniform across the score points. It would seem that the IEA agrees more frequently with the human markers on mediocre essays than on essays at the two ends of the achievement scale.

The third observation from Table 11.6 is that overall, the IEA has roughly the same tendency to either under-estimate or over-estimate the quality of essay content, if human scores are regarded as true approximations of the underlying proficiencies. Across the two prompts, the IEA assigned a higher score than the human score for 21.5% (or 84) of the essays and a lower score for 24.3% (or 95) of the essays.

The fourth observation from Table 11.6 is that there are very few essays (seven out of 391) which have seriously discrepant scores between the IEA and human scores (i.e., scores are different by more than 1 score point). The next section pays close attention to issues around the scoring of the content for these seven essays.

11.4.3 Investigation of Essays with Seriously Discrepant Content Scores

Close examination of the seven essays for which the IEA assigned a content score that was more than 1 score point different from the human score obtained from Pearson field tests suggests that for some of these essays, the IEA score might not be significantly different from the true score once the unreliability of human scores has been taken into account. However, one example which can demonstrate the IEA under-valuing the content is attached at Appendix P (SEQ# V2119). This essay received the highest score for *Content* (i.e., 3) unanimously from all four human markers (two from the field tests and two from this study, who used the ESL Composition Profile *Content* 0–3 scale). The high score is in recognition of the test taker developing the main points of the argument with sufficient supporting details and clear reasoning. In addition, the test taker demonstrated the ability to evaluate different points of view. Overall the human markers believed that the test taker had substantively and adequately dealt with the topic.

However the IEA scoring system assigned a score of 1.46 (out of 3) for *Content* for this essay. Although it is extremely difficult to identify the exact reason(s) for the discrepancy, it may be suggested that the machine judged this essay to have used less “content-relevant” words than

desired for the highest score. It is plausible that the IEA failed to recognise the reasoning and the arguments developed by the writer to support the main thesis, which reflected the writer's higher order critical thinking skills. It is also possible that the IEA failed to recognise the relevance of an *example* (highlighted in Italics in the text at Appendix P) provided in this essay to the writing topic, and consequently, marked down the essay on the quality of the content. This is a reasonable guess considering that the Latent Semantic Analysis (LSA) technique, used by the IEA to evaluate essay content, assumes that the meaning of an amount of text is simply the sum of the meaning of the individual words it contains. Consequently, when assessing the content, the IEA does not try to deduce the logical relevance of an example given in an essay to the arguments being developed. Rather it calculates the "semantic" link between chunks of words contained in the example and the topic based on word to word relationships in a condensed space constructed from a mathematical model. When the *example* in this case did not contain many content-relevant words, the IEA might regard the *example* as having very little relevance to the topic to which the essay was required to respond. A confounding factor here is that this *example*, provided by this writer to support his/her thesis development, can be considered original and unique, hence the chance of co-occurrences of words used in the *example* and in the other essays written to the same topic was minimal. Consequently, developing a topic-specific semantic space using pre-scored training essays would not have helped for the scoring of content for this essay.

While the IEA uses a simplified mathematical model which relies on word co-occurrences alone to establish semantic relatedness between text units, humans use a much more

complicated multi-dimensional cognitive process to “presuppose and analyse and conjecture and conclude” in order to make sense of the meaning (Berthoff, 1981, p. 43). In this regard, the level of common intelligence and sophistication embodied in this process ensured that human markers who scored this essay had no difficulty in following the reasoning and the arguments provided by the writer and appreciating the relevance of the examples given to support the main ideas.

This difference in the processes used by the human markers and the IEA to understand the meaning of an essay is likely to affect the IEA’s judgements on those essays which contain original and creative content or those which make extensive use of abstract concepts, irony, metaphor or allusion. The impact of this difference on the validity of the IEA content scores is expected to increase when the purpose of the writing shifts from “regurgitating specific content-knowledge information in a predetermined form”, to discovering, interpreting and eventually generating new meaning and transforming knowledge (Ericsson, 2006, p. 30).

A contrary example is where the IEA may provide unfair advantage to other types of essays, because of its “bag of words” approach to assessing content. The LSA technique ignores how words are arranged within a sentence or how sentences are arranged within a text when it assesses the meaning of the content. This, of course, is different from the process used by human markers who rely upon arrangement, cohesion and coherence to deduce the meaning of texts. The nature of the process used by LSA to derive meaning from the texts explains why it is plausible for the IEA to assign undeservedly high scores to essays which have gibberish, ludicrous or obscure content, as has been demonstrated by various successful

attempts to fool the IEA (e.g., Anson, 2006; McGee, 2006). An example from this study illustrating this particular problem is displayed below.

SEQ# T113 – Prompt: Tobacco

tobacco ingoures to helth it is in large quantty of people have bad habit of smoking & do a some kind of tobacco taking habits they dont know that ingouries them selves but also harm to othters aiso it creates a serviour problems like cancer, blood caecer, mouth cancer , etc much more it means that they cheat themselves they dont know that laught of life is depanding on them . they do unhealdy to their family mambers but also to the smallers childrens . it is to be stop if the person want to change the life & he/she want to change habit of smoking most of the good government hospitels help them to change th

For the content in this example essay, the average human score across five available human scores is 0.6,³⁸ out of the maximum score of 3. This low score recognises that this essay did not discuss in detail the role of the government to legislate to protect citizens from the harmful effects of smoking in any great detail, which was the requirement of the prompt. In addition, the meaning intended to be conveyed by this essay was nearly obscure because of the writer's limited knowledge of key aspects of the English language. Despite limited development of the topic, this essay was awarded a score of 2.09 out of 3 by the IEA. When comparing this essay to the previous example (in Appendix P), it is hard to understand why

³⁸ This essay received three marks from markers at the Pearson field tests. They were 0, 1, 1 respectively. Two human markers from this study, who used the ESL Composition Profile *Content* 0–3 scale gave a score of 0 and 1 for the content of this essay.

the IEA assigned a higher score for this essay than the earlier essay which had developed its main points thoroughly and clearly, apart from speculating that the IEA detected more content-relevant words in this essay than in the other. Words in this essay such as “habits”, “harm”, “cancer”, “blood cancer”, “mouth cancer” are assumed to have contributed to the IEA’s calculation of the content score.

The next section looks at the appropriateness of the treatment that the IEA gives to the off-topic essays.

11.4.4 Treatment of Off-topic Essays

The IEA scoring criteria for *Content* indicate that any essays which receive a less than 0.5 *Content* score from the IEA would automatically receive a total score of 0. If scored appropriately, these essays should be mainly off-topic essays (i.e., essays not pertinent to the topic) or those essays which are too short to evaluate.

In order to understand how similar off-topic type essays are treated in other testing situations, the researcher held discussions with the experienced markers recruited for this study. These markers commented that for the NSW Higher School Certificate (HSC) marking, the meaning of the essay and the fulfilment of the task were considered to be the foremost criteria in determining the quality of an essay. As a result, an essay that was completely off the topic or too short to understand was scored 0. This same scoring rule also applies in the Test of English as a Foreign Language (TOEFL) where its rubric makes it clear that an essay which

“rejects the topic, or is otherwise not connected to the topic” should be scored 0 (TOEFL Independent Writing Rubric – Appendix E). However, when assessing persuasive writing tasks for the Australian National Assessment Program of Literacy and Numeracy (NAPLAN) tests, an off-topic essay does not prevent human markers from marking other analytic traits. It would seem that the treatment of off-topic essays is one which depends on the purpose of the testing. As one of the purposes of the NAPLAN is to provide teachers with useful diagnostic information, it is important to score all the language traits even if the essay may have been off-topic. On the other hand, the main purpose of HSC, TOEFL and PTE Academic is for college/university admission decisions, therefore it is quite appropriate to give particular importance to the meaning of the essay and the fulfilment of the task.

Of the 391 essays that were scored by markers from the Pearson field tests, 10 received an IEA *Content* score less than 0.5. Of these 10 essays, half received a final *Content* score of 0 from human markers, and half received a (final) score of 1 from human markers. Although human scores and machine scores were close for these essays, had the IEA assigned a score of 1 for those same five essays as did the human markers (assuming accuracy of human scoring), it would have had to score all six other traits for these essays, resulting in higher overall scores. This indicates that any measurement inaccuracy in the IEA scoring of content for those essays that are at the low end of the *Content* achievement scale could have greater implications for the validity of the resulting overall scores.

11.5 The IEA Scoring of the Other Four Traits

The remaining four traits in the IEA scoring model for the PTE Academic are *Development*, *Structure and Coherence*, *General Linguistic Range*, *Grammar Usage and Mechanics* and *Vocabulary Range*, all of which are scored on 0–2 scales. There is very little, if any information publicly available on how the IEA scores each of these traits. Details of the micro-text features included in these traits and the manner in which they are included are considered proprietary information (Karen E. Lochbaum, Pearson, personal communication, March 31, 2010).

In the absence of more detailed scoring process information which would have facilitated a more substantive investigation, this section relies on the agreement rates and correlation statistics to investigate the reliability and the accuracy of the IEA scoring of these traits, assuming human (final) scores used in these analyses are approximations of the true scores. Human scores used are those assigned by markers from the Pearson field tests using the same IEA trait scales.

11.5.1 Distributional Statistics for Human and IEA Scores

Table 11.7 provides a comparison of the means and standard deviations of the scores from human markers and from the IEA for the fore-mentioned four traits, across the two prompts.

Table 11.7***Descriptive Statistics of the Trait Scores by the IEA and by Human Markers***

Prompt	Trait	Marker 1		Marker 2		Human Score		IEA		Effect Size (<i>d</i>)* IEA /human
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	
Voting	<i>Development, Structure and Coherence</i>	1.16	0.68	1.21	0.74	1.14	0.66	1.16	0.47	0.04
	<i>General Linguistic Range</i>	1.16	0.71	1.07	0.68	1.14	0.68	1.12	0.5	-0.03
	<i>Grammar Usage and Mechanics</i>	1.09	0.71	1.07	0.7	1.05	0.7	1.08	0.53	0.05
	<i>Vocabulary Range</i>	1.19	0.66	1.13	0.67	1.17	0.62	1.14	0.5	-0.05
Tobacco	<i>Development, Structure and Coherence</i>	1.18	0.69	1.23	0.72	1.15	0.67	1.19	0.56	0.06
	<i>General Linguistic Range</i>	1.15	0.68	1.17	0.64	1.18	0.64	1.14	0.57	-0.07
	<i>Grammar Usage and Mechanics</i>	1.15	0.69	1.07	0.68	1.13	0.68	1.11	0.53	-0.03
	<i>Vocabulary Range</i>	1.14	0.65	1.17	0.68	1.11	0.64	1.18	0.57	0.12

Note: SD: Standard Deviation. Before calculating these statistics, the IEA scores are first rounded to the nearest whole numbers.

*: A positive effect size (Cohen's *d*) denotes where the IEA means are greater than the human means.

A negative effect size (Cohen's *d*) denotes where the IEA means are less than the human means.

Human score: human (final) score, which is calculated as the average of two scores from human markers (or in the case of adjudication, the average of the closest two scores), each essay received on each trait. This score is rounded to the nearest whole number, before being used in this analysis.

As a baseline comparison across the two prompts and across the four traits, the means of the scores provided by two markers are very similar. Effect sizes (calculated using Cohens' *d*), comparing mean scores assigned by individual markers, are small and range from 0.03 to 0.12. Similarly, as shown in Table 11.7, the mean IEA scores across the four traits and across

both prompts are not meaningfully different from the means of scores provided by individual markers, nor are they meaningfully different from the means of the human (final) scores. The effect sizes, when comparing the IEA mean scores to human means scores, are small and also range from 0.03 to 0.12. In addition, there is no evidence of a systematic pattern in the direction of the effective sizes, indicating no systematic upward or downward bias in the means of the scores produced by the IEA, as compared to the means from the human scores.

The only issue is that, consistently across all traits and across both prompts, the standard deviation of the IEA scores in each category is smaller than the corresponding standard deviation of scores from human markers. For the Voting prompt, the variation in the IEA-generated scores is statistically smaller than the variation in the human scores, at $\alpha = 0.05$ level, for all four traits. For the Tobacco prompt, the variation in the IEA scores is statistically smaller than that in human scores for two traits at $\alpha = 0.05$ level; and statistically smaller for the remaining two traits at $\alpha = 0.1$ level.³⁹ This result is consistent with what is observed for the *Content* scoring; that is, the IEA scores are more compact than human scores at the individual trait level.

³⁹ For the Voting prompt across the four traits, testing the significance of difference between standard deviations in human and IEA scores, the smallest $F(186, 186)$ is 1.53, the largest $F(186, 186)$ is 1.95. They are all greater than the critical value $F_c(0.05, 186, 186) = 1.27$. For the Tobacco prompt, across the four traits, the smallest $F(203, 203) = 1.25$, the largest $F(203, 203)$ is 1.64. The critical value $F_c(0.05, 203, 203) = 1.26$ and $F_c(0.1, 203, 203) = 1.2$.

An explanation for the smaller standard deviations in the IEA trait scores is the propensity of the IEA to award scores to the middle point of the rating scales. Appendix Q shows the matrix of frequency of occurrences of the IEA scores and human trait scores at each score point level, across the four traits. For each trait, the IEA assigned a lesser number of essays to the lowest and highest score points (i.e., 0 and 2), than did human markers. For example, for the *Development, Structure and Coherence* trait, the IEA assigned 24 essays a score of 0 and 93 essays a score of 2, out of a total of 391 essays across the two prompts. These are compared to 62 essays receiving a score of 0, and 118 receiving a score of 2 from human markers. As with the scoring of content, the problem of “central tendency” (the tendency of a scorer to avoid extreme score categories) seems to be more conspicuous in the IEA scoring than in human scoring, across all four traits.

11.5.2 Influence of Essay Length in IEA Trait Performance

Before various correspondence rates across the four IEA traits are examined, an attempt is first made in this section to understand how sensitive the four traits are to the essay length (i.e., how strongly scores for these traits are related to essay length). The need for such an investigation originates from the observation that there is inadequate information in the public domain about how exactly IEA measures these four traits. Though the IEA developers explicitly state that “essay length has been expressly excluded from any of the IEA component measures” (Landauer et al., 2003, p. 102), it could still be possible for the length variable to be indirectly introduced into the scoring process through the method of scoring. Quinlan et al. (2009) provides a good example of how the length variable may creep back into

the scoring process through transformation algorithms for calculating linguistic accuracy ratios. For the IEA, the method of scoring these analytic traits could involve counting the number of words that meet various criteria (e.g., number of argument development verbs ending in *-ing*, or number of auxiliary verbs). Since the count of words relates directly to essay length, and the length alone has been demonstrated to be a strong predictor of human scores, the method of scoring (rather than the machine measurement capabilities) could be the sole contributing factor to any observed associations between the IEA and human trait scores (see Chodorow & Burstein, 2004, p. 17, for discussion).

Based on the scoring criteria for the four IEA traits, it is expected that the trait *Development, Structure and Coherence* would be the most sensitive of the four to the essay length, because key criteria for this trait, such as “a good development of” the thesis, require the ideas to be developed with examples and with supporting statements, which go hand in hand with essay length (see Appendix A for the IEA scoring criteria). On the other hand, the *Grammar Usage and Mechanics* trait is hypothesised to have the weakest relationship with the number of words because the key scoring criteria for this trait refer to whether the language errors are rare or easily detected. This suggests that a language accuracy ratio, which already adjusts for the influence of essay length, is used in the scoring of this trait. Table 11.8 provides the Pearson correlations between word count and the IEA scores, and between word count and the human scores, across the four traits.

Table 11.8

Pearson Correlations Between Word count and the IEA Scores and Between Word Count and the Human Scores

Trait Name	IEA	Human Scores
<i>Development, Structure and Coherence</i>	0.71	0.53
<i>General Linguistic Range</i>	0.67	0.50
<i>Grammar Usage and Mechanics</i>	0.43	0.43
<i>Vocabulary Range</i>	0.60	0.46

Note: The IEA scores used in the analyses are raw continuous scores produced by IEA.
All correlations are significant at the 0.01 level (two-tailed). Number of cases for each trait: 391

Table 11.8 indicates that the IEA scores follow the human score patterns in terms of the relative sensitivity that each of the traits exhibits to the essay length. The hypothesised patterns of the relative sensitivity across the four traits are observed in both sets of the human and the IEA trait scores. For example, the *Development, Structure and Coherence* trait is confirmed to be the most sensitive (i.e., scores for this trait correlate highest with essay length) and the *Grammar Usage and Mechanics* is the trait least sensitive to the essay length. In this regard, the IEA scores have demonstrated characteristics that are consistent with the expectations from the scoring criteria for each of the traits and these characteristics are consistent with those of the human scores.

However, it is also noted that the IEA scores are considerably and statistically more sensitive to the essay length than are the human scores, for all traits except for the *Grammar Usage and*

Mechanics trait. The difference in the correlations ($r_{IEA} - r_{Human}$) for these three traits ranged from 0.14 to 0.18 (the largest one-tailed p value for $r_{IEA} - r_{Human} = 0.003 < 0.01$).⁴⁰ This difference in the sensitivity of the IEA and the human traits to the essay length variable indicates subtle disparity in the scoring process that is used by the human markers and by the IEA for these traits. As a result of the observed moderate to strong relationships between essay length and the IEA scores, it is worthwhile examining the partial correlations after the removal of the influence of essay length from both the human and the IEA scores, to gain a better understanding of the level of correspondence between the two sets of scores.

11.5.3 Correspondence Rates Between Human and IEA Scores Across the Four Traits

Table 11.9 provides a summary of various correspondence rates including partial correlations between the IEA scores and human scores, across four traits. As there is little difference in these statistics across prompts, results displayed are for the combined set of essays.

⁴⁰ The significance of the difference in two correlation coefficients is calculated using Fisher's r - z transformation and testing the statistical difference in the z values. See Fisher (1915, 1921) for the transformation formula.

Table 11.9***Agreement Rates and Correlations Statistics Across the Traits – Between the IEA and the Human scores***

Trait	Exact Agreement Rates (Kappa)	Exact + Adjacent Agreement Rates (Kappa)	Pearson zero-order r	partial r (after removing the effect of length)
<i>Development, Structure and Coherence</i>	0.68 (0.40)	0.997 (0.994)	0.65	0.47
<i>General Linguistic Range</i>	0.69 (0.44)	1.00 (1.00)	0.66	0.51
<i>Grammar Usage and Mechanics</i>	0.67 (0.41)	1.00 (1.00)	0.66	0.58
<i>Vocabulary Range</i>	0.74 (0.50)	1.00 (1.00)	0.69	0.59

Note: All correlations are significant at the 0.01 level (two-tailed). Number of cases for each trait: 391. For correlation analysis, the IEA continuous scores were used. For agreement rates analysis, the IEA continuous scores were rounded to the nearest score points. Human scores used in the analysis were those adjudicated scores from the Pearson field tests.

Table 11.9 indicates that, across the four traits, the IEA scores agreed exactly with the human scores 67% to 74% of the time, and agreed with the human scores within one score point nearly all the time. In addition, with regard to the rate of exact match between the IEA scores and human scores, the IEA has demonstrated a fair to moderate level of agreement with human scores (using criteria recommended by Landis and Koch, 1977), after such a rate has been adjusted for chance agreement.

Across the four traits, the Pearson zero order correlation between the IEA scores and human scores is similar, ranging from 0.65 for *Development, Structure and Coherence*, to 0.69 for *Vocabulary Range*. In fact, the correlations between the IEA scores and human scores are not

statistically different across the four traits (using Fisher's r - z transformation and testing the statistical difference in the z values, the lowest p value is $0.31 > 0.05$), indicating that the IEA's performance in predicting human scores is similar across the four traits.

A last observation from Table 11.9 is that after the effect of length is removed, there is still a statistically significant and moderate association between human scores and the IEA scores, for each of the four traits. This suggests that the IEA, in its scoring of these traits, does use deeper and richer information other than word count to predict human scores.

It is noted that the agreement rates reported in Table 11.9 are sensitive to the number of score categories on the rating scales. In this case, there are only three categories for each trait. In order to see if the agreement rates are uniform across the score points, the discrepancy rates (i.e., the proportion of occasions when the IEA scores do not agree with the human scores exactly) at each of the three score points are reported in Table 11.10.

Table 11.10

Discrepancy Rates at the Score Point Level

Trait	Score point on the scale		
	0	1	2
<i>Development, Structure and Coherence</i>	66%	15%	46%
<i>General Linguistic Range</i>	64%	14%	43%
<i>Grammar Usage and Mechanics</i>	61%	13%	50%
<i>Vocabulary Range</i>	52%	14%	40%

The above table indicates that while the IEA exhibited a satisfactory level of agreement with human markers across the four traits, this performance differed across the score points. Consistent with the findings from the analysis performed for the *Content* trait, the IEA had significantly higher discrepancy rates at the two tail ends of the scales (i.e., score points 0 and 2) than the middle score point (i.e., score point 1). The lowest score point had the highest discrepancy rates of all score points, consistently so across the four traits.

11.6 Inter-relations amongst Human and IEA Traits

To further examine the convergent and discriminant evidence necessary for the establishment of construct validity for the IEA scoring method, this section focuses on the inter-relationships of the traits assessed by the IEA and those by the human markers. This section uses a matrix of inter-correlations, a type of the multitrait-multimethod matrix developed by Campbell and Fiske (1959), to conduct this investigation.

The rationale of the investigation is that if the IEA's scoring method is similar to the human scoring method, measures of the same trait (e.g., human scores on the *Content* trait and the IEA scores on the same *Content* trait) should show sufficiently large and positive correlations. This requirement is evidence of convergent validity (Campbell & Fiske, 1959, p. 82). Furthermore, a measure of one trait using one scoring method should correlate more highly with another measure of the same trait using an alternate scoring method, than it does with measures of any other traits that happen to employ the same alternate scoring method. For example, the IEA scores on the *Content* trait should correlate higher with human scores on the

same trait than they do with human scores on any other traits, because scores on the same traits are intended to capture the same aspects of writing that are not covered by any other traits. This requirement provides evidence of discriminant validity (Campbell & Fiske, 1959, p. 83).⁴¹

Table 11.11 provides the matrix of all inter-correlations between scores on traits assessed by the IEA and those assessed by the human markers from the Pearson field tests using the same scoring criteria across the two prompts. Correlations of the measures of the same traits (i.e., the diagonal values) are highlighted for easy comparison.

⁴¹ See also Kane (2006, p. 40) for the interpretations of the multitrait-multimethod matrices.

Table 11.11***Inter-correlations Amongst Traits As Assessed by Human Markers and the IEA***

Prompt	Traits assessed by the IEA	Traits assessed by Human Markers				
		<i>DSC</i>	<i>GLR</i>	<i>GUM</i>	<i>Vocabulary</i>	<i>Content</i>
Voting	<i>DSC</i>	0.63	0.57	0.53	0.58	0.52
	<i>GLR</i>	0.55	0.62	0.63	0.64	0.57
	<i>GUM</i>	0.40	0.52	0.62	0.60	0.44
	<i>Vocabulary</i>	0.57	0.59	0.65	0.66	0.55
	<i>Content</i>	0.52	0.60	0.54	0.58	0.66
Tobacco	<i>DSC</i>	0.67	0.65	0.57	0.66	0.57
	<i>GLR</i>	0.61	0.70	0.62	0.68	0.61
	<i>GUM</i>	0.57	0.63	0.70	0.63	0.54
	<i>Vocabulary</i>	0.63	0.68	0.61	0.72	0.58
	<i>Content</i>	0.58	0.63	0.55	0.64	0.68
Combined	<i>DSC</i>	0.65	0.61	0.55	0.62	0.54
	<i>GLR</i>	0.58	0.66	0.63	0.66	0.59
	<i>GUM</i>	0.49	0.57	0.66	0.61	0.50
	<i>Vocabulary</i>	0.60	0.64	0.63	0.69	0.55
	<i>Content</i>	0.53	0.60	0.54	0.58	0.69

Note: Pearson product-moment correlations are used in this analysis.

The following abbreviations are applied to the IEA trait names:

DSC: Development, Structure and Coherence; *GUM*: Grammar Usage and Mechanics; *Vocabulary*: Vocabulary Range; *GLR*: General Linguistic Range

All correlations are significant at 0.01 level (two tailed). N=187 for the Voting prompt; N=204 for the Tobacco prompt.

Several observations can be made from Table 11.11. First, correlations of the measures of the same traits (i.e., the diagonal values) are sufficiently large and range from 0.62 to 0.72, across

traits and across prompts. The sound relationships between scores on the traits measuring the same aspects of writing provide convergent evidence necessary to support the validity of the IEA scoring of these traits.

The second observation is that there seems to be a certain level of discriminant evidence for the IEA traits. When essays written to the Tobacco prompt are assessed, measures of the same trait correlate better with each other than they do with measures of any other traits that employ the same scoring method, for all five traits. This can be verified from the table, as each diagonal value is higher than any other values lying in its row or column.

When essays in response to the Voting prompt are assessed, the IEA scores on three traits – *Content, Development, Structure and Coherence* and *Vocabulary Range* – exhibit the expected pattern providing evidence of discriminant validity. Two small exceptions exist. IEA scores for the *General Linguistic Range* correlate best with human scores for the *Vocabulary Range* ($r = 0.64$), instead of with human scores on the same trait ($r = 0.62$). Human scores on *Grammar Usage and Mechanics* correlate best with the IEA scores on the *Vocabulary Range* trait ($r = 0.65$), instead of with IEA scores for the same trait ($r = 0.62$). However, the differences in the two pairs of correlations being compared are very small, and are not

statistically different⁴² ($z = 0.32$, $p = 0.75$ for the former pair; $z = 0.48$, $p = 0.63$ for the latter pair).

In summary, this section has provided convergent evidence to support the observation that there appear to be certain commonalities in the characteristics of the writing that are being captured by the same traits measured by both the IEA and human markers. In addition, there is also a degree of discriminant evidence to support the claim that the IEA traits are measuring what they are supposed to be measuring.

11.7 Chapter Summary

This chapter has demonstrated various ways in which the accuracy and validity of the IEA scoring at the trait level can be investigated. This type of investigation addresses validity questions related to the writing trait and to the scoring procedure (i.e., the first and the second) components of the AES validation framework. This investigation is important because it forms an essential part of the evidentiary argument supporting or challenging the machine scoring.

The analyses first identified several issues concerning the IEA's scoring of the spelling, particularly those arising from the scoring criteria used by the IEA for PTE Academic. In

⁴² The significance of the difference in two correlation coefficients is calculated using Fisher's r - z transformation and testing the statistical difference in the z values. See Fisher (1915, 1921) for the transformation formula.

addition, there are a number of factors that can adversely impact on the accuracy of the automated scoring of the *Spelling* trait, such as real-word problems and recognition of proper nouns.

Though there is no concern regarding accuracy for the scoring of the *Formal Requirement*, it is found that the specificity of the scoring criteria used for this trait can advantage or disadvantage particular types of essays. There are also issues arising from the consequential aspects of the validity as a result of the IEA directly including a highly coachable surface feature in the scoring model.

The accuracy investigations for the remaining five traits revealed that the IEA performed satisfactorily on the scoring of these traits with regard to overall agreement rates and correlation statistics. However, when the agreement rates were disaggregated at a score point level, it was found that they were worst at the two tail ends of the achievement scales, consistently so across all five traits. This is a concern, particularly for high-stakes tests that require accurate discriminations across the whole achievement scale.

This chapter also discussed the limitations of the technology that the IEA used to assess essay content. On the one hand, the IEA's semantic model might disadvantage creative essays reflecting high level critical thinking skills because of its tendency to respond to the stimuli of words, rather than to the underlying interconnected concepts and ideas. On the other hand, it was also probable for the IEA to provide unfair advantage to those essays which had obscure or incoherent content due to its "bag-of-words" approach. For general English proficiency

tests such as PTE Academic, the proportion of essays that may receive a significantly undeservedly high or low content score due to these limitations is expected to be small. However, when the writing to be scored involves more knowledge transformation (e.g., making of new meanings and new ideas), the impact these limitations can have on the validity of IEA scores is likely to increase. In the latter situations, semantic models built based on existing knowledge may not allow for the adequate measuring of the conceptual relevance and significance of the new meanings.

Chapter 12 Discussions and Conclusions

12.1 The AES Validation Framework

This thesis identifies the need for a structured and coherent approach to establishing the validity of AES (Automated Essay Scoring) systems. In order to address this need, it presents an AES validation framework to facilitate the systematic collection and examination of empirical evidence, as well as the theoretical rationale, in support of claims regarding the validity of outputs from AES systems.

This framework identifies key areas where AES validation efforts should focus. Since the framework also specifies the critical assumptions supporting the intended interpretation and use of test scores, it is by nature, an interpretative argument, which can be adapted for use in any AES validation process. The assumptions specified are phrased as key validity questions within the framework to emphasise the need for the examination of these assumptions.

The AES framework proposes that, in order for a convincing argument to be made for the validity of an AES system, evidence for the following five components of the framework must be collected and examined together: 1) the writing traits scored by an AES system; 2) the type of scoring procedure used by an AES system to derive an overall score; 3) the structural, 4) the measurement and 5) the consequential aspects of score validity.

The first two components of the framework address key aspects of an AES scoring process, specifically, the writing traits assessed by an AES system, how well they are assessed, and

how they contribute to the overall scores. Evidence from these two components is critical to the understanding of the meaning of the scores generated by an AES system, as well as being essential to score defensibility. This is because the credibility and validity of any scoring system depends upon its capacity to rationally explain how scores are determined.

The third and fourth components of the framework focus on the measurement and structural aspects of validity for scores produced from AES systems. The measurement component makes clear the essential requirements that AES scores must meet empirically in order for them to be used for meaningful comparisons of test-taker performance along a single ability continuum. These requirements include evidence that the writing construct produced by an AES system is sufficiently uni-dimensional to render it useful for the purpose at hand, and that the rating scales used to score the various traits are functioning as expected.

The structural component refers to AES scores demonstrating the property of structural fidelity. This component requires evidence that the AES scores exhibit internal structural patterns that are consistent with expected relations among the various writing traits. Evidence from the above two components provides essential backing for the claim that an AES system has been developed in a rational manner and that the resultant scores reflect appropriately the underlying writing abilities being measured.

The last component, the consequential component of the framework, identifies the types of evidence required for analysis of the potential positive and negative impacts AES can have on learning, instruction and writing curricula. These types of evidence are in addition to those

related to the more immediate consequences, such as direct benefits and costs, of AES use. Evidence collected for this component forms an integral part of the overall argument for or against the appropriateness and legitimacy of the use of AES in a particular context.

In order to construct a convincing validity argument, different forms of evidence collected according to the AES framework need to be evaluated together, in terms of their combined effects on the meaning of the score and the implications of score use. The appropriate weight to be carried by each form of validity evidence relative to the overall evaluative judgement is dependent on the purpose and nature of the test use.

The utility of this framework is then tested in Chapters Five to Eleven using a particular automated essay scoring system – the Intelligent Essay Assessor (IEA) in conjunction with writing tasks from the Pearson Test of English (PTE) Academic. In the course of applying this framework, this current study has made particular efforts to illustrate how evidence for different components of the framework can be collected and examined through a combination of theoretical arguments and statistical methods.

12.2 A Validity Argument for IEA

Table 12.1 now summarises all the evidence collected from applying the proposed AES framework in the current study. It also links evidence collected against the validity questions that were set out at the beginning of this thesis for the IEA. Those validity questions are listed

in the first column. The second column of the table lists the components of the AES framework and the associated validity issues examined in the study.

The third and fourth columns list evidence collected. Each piece of evidence is denoted by an alpha-numerical code (e.g., E1, E2). The sections of the thesis within which the evidence has been discussed are indicated in the brackets after the descriptions of the evidence. The third column lists evidence (coloured in green) that potentially supports the validity argument of the scores produced by the IEA (e.g., E1, E2, etc). The fourth column lists evidence (coloured in black) that provides the backing for potential rebuttals to the validity claim (e.g., E3, E5, etc).

Table 12.1 Evidence Collected from Using the Proposed AES Framework

Validity questions identified in this study	AES Framework	Validity Evidence Collected from Using the Framework	
1) How do the IEA writing traits relate to the writing ability being assessed?	<p>Writing Traits Component:</p> <p>* Are the writing traits assessed by an AES model representative of the construct of interest?</p> <p>* Are the traits assessed by an AES model relevant to the construct of interest?</p>	<p>E1: At the very high level, the writing traits assessed by IEA seem to cover all important parts of the target writing construct (Section 7.2).</p> <p>E2: The writing traits assessed by IEA reflect the main dimensions of writing quality that are emphasised by experienced human markers in the target writing construct, with the exception of the <i>Formal Requirement</i> trait (7.2).</p>	<p>E3: The inclusion of the <i>Formal Requirement</i> trait in the scoring model has the potential to introduce construct-irrelevant variance to the measurement process (7.2).</p>
2) How well does IEA assess the traits?	<p>* Can these writing traits be accurately assessed by the AES model?</p> <p>* How well do the overall scores align with those assigned by experienced human markers and with independent measures?</p> <p>* Can AES scores generalise to scores that would be expected to be obtained under different but parallel conditions?</p>	<p>E4: For each of the five IEA traits (excluding the <i>Formal Requirement</i> and the <i>Spelling</i> traits from the full complement of seven traits), the average agreement rates between IEA-generated trait scores and human-generated trait scores are satisfactory (11.4.1, 11.5.3).</p> <p>E6: For each of the five traits analysed, the means of the IEA scores are not meaningfully different from the means of the human scores (11.4.1, 11.5.1).</p> <p>E8: The IEA total scores have a relatively strong relationship with human total scores. The correlation between human and IEA total scores ranged from 0.73 to 0.81, across the two prompts, and across the two scoring procedures used to generate human scores (8.3.2).</p> <p>E10: After the effect of length is removed, there is still a statistically significant and moderate association between human and IEA overall scores (8.3.3).</p> <p>No generalisability evidence collected.</p>	<p>E5: For each of the five IEA traits, IEA was less in agreement with human markers at the two tail ends of achievement scale, with discrepancies the most significant at the lowest assessed level for each trait (11.5.3, 11.4.2).</p> <p>E7: For each of the five traits, the dispersion in the scores generated by IEA on the same sample of essays is smaller than that in the scores generated by human markers, indicating IEA scores are less discriminating than the corresponding human scores (11.4.1, 11.5.1).</p> <p>E9: There are significant questions about the appropriateness of the <i>Spelling</i> scores produced by the IEA, attributable to factors not only associated with the measurement capability of IEA, but also with the scoring criteria used (11.2).</p> <p>E11: The two minimum requirements (<i>Content</i> and <i>Formal Requirement</i>) implemented by IEA in the calculation of an overall score can have adverse implications on the validity of the overall score produced, in addition to those already associated with the scoring of these two traits (11.3.2, 11.4.4).</p>

Continued to next page

Validity questions identified in this study	AES Framework	Validity Evidence Collected from Using the Framework	
3) Empirically, do these (IEA) traits behave as expected?	<u>Measurement Component:</u> * Is there empirical evidence of various traits measuring the same construct? * Are the rating scales used to score the individual traits functioning as intended?	E13: When the two misfitting traits – <i>Spelling</i> and <i>Formal Requirement</i> are removed, the remaining five IEA traits function well together to support the development of a single construct (9.5.2). E14: The rating scales for the five IEA traits seem to function as expected; that is, a higher score category on a scale indicates more of the property being assessed by the construct (9.5.3).	E12: The <i>Spelling</i> trait significantly underfit a uni-dimensional model. Some segments of the <i>Formal Requirement</i> trait scores do not fit the uni-dimensional model well (9.5.2). E15: The two language traits (<i>General Linguistic Range</i> and the <i>Vocabulary Range</i>) show signs of data-model over-fit, indicating that the two traits are less efficient and less productive for measurement than desired (9.5.2).
	<u>Structure Component:</u> * Does the internal structure of AES scores confirm the theoretical distinctions about the construct? * Is the internal structure of AES scores consistent with that in the human expert scores? * Is the internal structure of AES scores consistent with a theoretical view of writing as a number of intercorrelated yet conceptually distinct dimensions?	E16: The IEA overall scores have successfully replicated human score differences between prompts and between gender (8.3.1, 10.3). E17: The IEA scores have revealed the same pattern in the relative sensitivity of the analytic traits to the essay length as the human scores have. The pattern observed is also in accord with the expectations based on the scoring criteria for these traits (11.5.2). E18: There is a certain level of convergent and discriminant evidence collected from using the multitrait-multimethod comparisons which support the argument that the IEA traits seem to be measuring what they are supposed to be measuring, and that each of them seems to be capturing some common characteristics of the writing that are also captured by the same trait as assessed by human markers (11.6). E19: The IEA and the human trait scores have produced statistically equivalent person ability measures for the majority of the persons (9.5.4).	E20: While scores on the traits assessed by human markers reveal a two-factor structure that is consistent with the structure of the construct domain, the corresponding IEA scores fail to exhibit the same structure (10.2). E21: The traits assessed by the IEA are less independent of the essay length than those assessed by human markers (11.5.2). E22: When the two sets of five analytic traits, one set assessed by the IEA and the other set assessed by the human markers are analysed together in one 2-dimensional space, the IEA analytic traits are clearly distinguished from the human analytic traits as a group (10.4). E23: In the same 2-dimensional space, traits assessed by IEA are shown to be less independent of each other than the traits assessed by human markers (10.4). E24: The IEA scores yield a different order of the difficulty amongst the writing traits, as compared with the order detected in corresponding human scores based on the same set of essays (9.5.2).
4) What is the validity implication of the procedure used by IEA to produce the overall scores?	<u>Scoring Procedure Component:</u> * What are the validity and reliability implications of the AES scoring procedure? * Are the scoring criteria used by an AES model appropriate?	E25: For different writing prompts, the IEA uses the same scoring procedure to combine the trait scores to overall scores. This ensures the consistency and comparability in the meaning of the overall scores produced across the different writing prompts (7.3).	E26: The scoring procedure used by IEA to derive an overall score seems to associate the <i>Spelling</i> trait with a level of importance to the overall score that is inconsistent with the understandings of experienced markers of writing effectiveness at the college/university level (7.3.3). E9: There are questions about the appropriateness of the scoring criteria used to score the <i>Spelling</i> trait (11.2).
	<u>Consequential Component:</u> * Would the use of AES unfairly disadvantage or advantage certain groups of students? * Would the use of AES have positive impact on teaching, learning and curricula?	E27: Feedback and reports can be delivered to test takers within days, according to Pearson (2011b). The quick turn-around of results can increase the chances that the information will be used and lead to improved learning.	E28: The <i>Formal Requirement</i> trait can unfairly disadvantage or advantage certain types of essays; as well as make the scoring model less educational defensible and more susceptible to cheating and test-taking strategies (11.3.2, 11.3.3). E29: The deficiency in the IEA scoring of content quality may disadvantage or advantage certain types of writing such as those which contain innovative thoughts or unusual ideas (11.4.3).

Before making a validity argument for the IEA, it should be noted that not all validity questions contained within the framework were examined in this study, even though the study did collect a relatively wide range of evidence. Questions for which evidence was not collected were highlighted in red in Table 12.1. For example, because the data used in this study was part of the training data, there was no evidence collected regarding how generalisable the IEA scoring model was to new PTE Academic writing prompts or to other independent groups of test takers. In this regard, while there is some evidence in the public domain about the generalisability of the IEA models to independent sets of test takers (Pearson, 2011a), there is comparatively less information about the generalisability of the IEA scoring models to new and parallel tasks. Another piece of evidence the study did not collect was the empirical evidence of the usefulness of the writing measures produced by IEA for making educational decisions about the test takers and any consequences or impact of these decisions on these test takers. Evidence of this nature concerning PTE Academic test scores is also lacking in the public domain. Part of the reason could be that PTE Academic is still a relatively new international language test.

Taking into account the lack of evidence in these areas, this thesis makes the following argument, having weighed up evidence collected in this study, and having assumed IEA's generalisability to other parallel writing prompts. There seems to be sufficient evidence to claim that, for the majority of essays written to the PTE Academic writing tests, the IEA is able to assign a score that reflects the quality of writing relatively accurately, and because of that, the score is likely to be useful for making decisions concerning university/college

program admissions. The overriding pieces of evidence collected in this study that support this claim include: 1) the IEA total scores have a relatively strong relationship with those produced by human markers using independent marking processes that are appropriate for the PTE Academic test context. Additionally, the IEA scores for the five traits analysed also have a satisfactory level of agreement with scores produced by human markers using the IEA scoring criteria (E4, E6 and E8 in Table 12.1); 2) there is a certain level of convergent and discriminant evidence supporting the argument that the IEA traits seem to be measuring what they are purporting to be measuring, and that each trait seems to be capturing some common characteristics of quality writing that is being captured by the same trait as assessed by human markers (E18); and 3) trait scores obtained from the IEA and from human markers have produced statistically equivalent person ability measures for the majority of the persons, implying no significant differences in the overall patterns that exist in the trait scores generated by human markers and by the IEA (E19).

However, for a very small number of essays written for the PTE Academic writing tests, there is a real possibility that the total score awarded by the IEA for an essay does not appropriately reflect the underlying writing proficiency demonstrated in the written product. Key pieces of validity evidence that underlie the rationale for this claim include: 1) there is relatively strong evidence to suggest that spelling is not yet robustly measured by the IEA (E9 and E12). This could impact on the interpretability of the overall scores, which are derived from scores on all seven traits including *Spelling*; 2) for each of the five IEA traits analysed, the IEA aligned less well with human markers for the two tail ends of each achievement scale than it did with

human markers for the middle point(s) of the scale (E5); 3) there is evidence that the scoring of the *Formal Requirement* and the *Content* traits may result in inappropriate and unfair scores being awarded to certain types of essays. The types of essays that may be problematic for the IEA include those that simply string together concepts without presenting a well-developed argument; those that have superfluous language; those that are short but concise; and those that are inventive and have individual styles (E28 and E29).

Based on the claims made above, when writing scores produced by the IEA are used for high-stakes university/college admission decisions, it seems necessary to have a safeguard measure built into the IEA scoring process to ensure that the validity of scores produced by the IEA holds across different testing cohorts or across different types of essays. Such a measure can be in the form of a human marker scoring essays in tandem with the IEA or a human marker checking the IEA scores for some types of essays (e.g., those that receive a 0 total score from the IEA; or those that receive significantly dissimilar scores across traits).

12.3 Future Work to Strengthen the Validity of Scores Produced by IEA

The evidence collected in this study helps identify priorities for future development of the IEA to further strengthen the validity of the scores produced by the IEA.

The most serious threat to the validity of the total scores generated by the IEA is the *Spelling* trait as assessed by the IEA for PTE Academic. Psychometric analyses in Chapter Nine provide strong evidence that the *Spelling* trait is not measuring the same writing construct as

the other traits. Further analyses in Chapter Eleven identify multiple factors that impact on the accuracy and appropriateness of the scores assigned to the *Spelling* trait. As a consequence, a priority to further strengthen the validity of the IEA writing scores is to improve the robustness of the IEA scoring of the spelling competency.

To this end, this study has recommended the following measures to improve the alignment of the IEA *Spelling* scores with the underlying competency being investigated: 1) using an accuracy ratio rather than the absolute number of spelling errors made in an essay as a basis for scoring; 2) incorporating measures of error seriousness in order to discriminate the competency being measured more appropriately (e.g., using a Levenshtein edit distance measure and/or using an error severity measure that takes into account the difficulties of the words misspelt); and 3) enhancing the IEA's capacity to deal with complexities in English orthographic errors (e.g., genuine competency errors versus typographical errors; identification of "real-word errors" and appropriate recognition of proper nouns). Most of these identified improvements should also apply to all AES systems using a spelling checking program because the difficulties identified in this thesis in accurately scoring spelling are generic to all such programs.

A second priority for future development is to improve the IEA's capacity to measure those characteristics of writing that are intended to be captured by the *Formal Requirement* trait, such as succinctness of the writing and task fulfilment, in a more robust and substantive manner. As demonstrated in this study (Section 11.3.2), this trait currently is a blunt instrument for what it purports to measure. Scores awarded for this trait can weaken the

interpretability and inferential property of the overall scores generated by the IEA, particularly for those essays whose (word count) lengths happen to be around the upper and lower limits for each of the score points for this trait. In addition, including the *Formal Requirement* trait in a scoring model can have unintended negative impact on learning and teaching, as well as making the model more susceptible to cheating and test-taking strategies. Consequences arising from the use of the IEA scores are a part of the consequential aspect of score validity and should not be overlooked. Though this aspect of validity is not extensively examined in this study, evidence elsewhere suggests that the way an AES system scores an essay for a high-stakes test can shape teaching and learning focus (*How to tackle the Analytic Writing Assessment?*, n.d.).

A third priority for enhancing the validity of the scores generated from the IEA is to develop a larger and a more sophisticated array of linguistic and rhetorical features that are not yet measured by the IEA, in order to “capture more of the richness and diversity of human language” (Chodorow & Burstein, 2004, p. 31). Evidence from this study suggests that human markers used in this study seem to discriminate among the various conceptually distinct writing traits better than the IEA. For example, traits assessed by the IEA were found to be less independent of each other than the similar set of traits assessed by human markers (E23). Additionally, whereas no trait assessed by human markers exhibited evidence of scores for the trait being too predictable, the two language traits (i.e., the *General Linguistic Range* and *Vocabulary Range* traits) assessed by the IEA had less variability in the data than the stochastic Rasch model predicted. This indicates that the textual features captured by these

two traits may be overlapping with, or too similar to, those already assessed by other traits (E15). Furthermore, the distinction between the higher order traits and the language-related traits was not observed in the trait scores assigned by the IEA, but clearly present in the scores assigned by the human markers. Such a distinction was consistent with writing domain theory and with the findings from other research of similar types of writing by non-native English speaking test takers (E20).

The finding that the IEA may not discriminate among the writing traits as well as the human markers suggests that more textual features need to be further developed to provide new and useful measurement information to the assessment of each trait. This should help the traits to be more accurately measured, and hence more appropriately distinguished, by the IEA. Such an enhancement is particularly important to the wider realisation of educational benefits of the IEA. The value of the IEA as an instructional tool providing diagnostic feedback on the writing traits depends on the accuracy of trait scoring by the IEA.

A further enhancement to the IEA is to improve its ability to score accurately those essays that fall into the two tail ends of the achievement distribution. This is based on the evidence which suggests that the IEA discriminates less satisfactorily for the lowest and highest score points for each of the five traits investigated (E5). This enhancement is necessary if the IEA is to be used in high-stakes scoring scenarios where there is a requirement that the scoring tool demonstrates an acceptable level of measurement precision and accuracy across all ability levels. For example, for the writing tests of the Australian National Assessment Programs for Literacy and Numeracy (NAPLAN), it is essential that the scoring tool chosen can

discriminate students of low writing ability with sufficient precision. This is not only because the results from NAPLAN tests are used for important funding and policy decisions, but it is also because a sizeable proportion of the Australian school students achieve below the national minimum standard for writing (Australian Curriculum, Assessment and Reporting Authority, 2011). For this group of students, the scoring tool used needs to be able to discriminate their proficiency levels adequately so that scores can be used for instructional and accountability purposes.

12.4 Implication of this Study for Future AES Research and Development

12.4.1 Implications Arising from the AES Framework Developed

A major contribution of this study has been the development of an AES validation framework to assist the systematic collection and examination of validity evidence for AES systems. The utility of the proposed AES framework has been demonstrated through this study.

The framework has proven effective in providing direction as to where to focus evaluative efforts when undertaking a validation process. In addition, applying the framework has allowed for an overall validity judgement to be made in an integrated manner, on the basis of a relatively comprehensive set of evidence collected. In contrast to most of other AES studies carried out to date, which have relied on a limited range of evidence to make an evaluative judgement, this study has collected and examined a comparatively wide range of evidence (28 bundles altogether), all of which are essential to the interpretation of the score meaning and to the justification of score use in a particular context. The wide range of evidence collected has

also provided a sound basis for the prioritisation of future developments of AES capabilities to improve the validity of AES scores.

The completeness and usefulness of the proposed framework is further substantiated by Table 12.1 demonstrating how this framework can be applied and actioned to answer the four questions listed in Chapter One with respect to the IEA. These four questions represent the kinds of common questions to which key stakeholders would want answers in order to make a decision about AES use. A related practical advantage of this framework is that it supports the exploration of broad questions by identifying explicit, specific examples for investigations. For instance, this study has illustrated that a broad question like “*Empirically, do these (IEA) traits behave as expected?*” may be investigated by examining whether IEA scores are in accord with the expectations of a uni-dimensional model and/or whether IEA trait scores exhibit internal structural patterns that are consistent with theoretically or empirically derived expectations.

There are other benefits of this framework. For example, it can provide a basis for evaluating the adequacy of the validation efforts made. The framework (summarised in Chapter Four, Figure 4.1) lists the key validity questions that need to be adequately addressed in order for a convincing validity argument to be made about an AES system. Gaps can therefore be identified by comparing all the validity questions that are identified in the framework with the ones that have been investigated (to varying levels) in available studies. When using this method to gauge the adequacy of the validation efforts already made in available studies, an immediate observation is that whereas there is a significant amount of evidence for any one

main AES system with respect to the validity question: “*How well do the overall scores produced by an AES system align with those obtained from human experts?*”, evidence for other types of validity questions, in particular those concerning the generalisability, measurement and consequential aspects of validity, is comparatively lacking across the AES systems. For example, across all the main AES systems, there is very little evidence concerning the generalisability of AES models across different but parallel model development processes, or generalisability of AES models across new populations of test cohorts. These types of evidence are needed to establish the validity of the AES scores as they support or challenge the credibility of the inference drawn from the AES scores to scores that would be expected to be obtained under different but parallel scoring processes. This framework thus is a useful structure for identifying additional evidence that needs to be collected for future validation studies.

Another benefit of this framework is that the evidence collected from applying this framework helps improve the transparency of an automated scoring system, which in turn can result in positive consequences of testing. Transparency of a test (including transparency of the scoring process) is an important validity criterion for high-stakes tests. The key stakeholders (e.g., students, teachers, test administrators) of tests are expected to want to know how the writing scores are derived, what characteristics of good performances are taken into account, and how these characteristics are assessed. They also expect clear explanations for any significant anomalies in test scores, before they can have faith in the results and use them for important decisions. As discussed earlier, the focus of this framework is on the key aspects of a scoring

process and on the important properties of the resultant trait scores, both of which help clarify the writing construct produced by an AES system. Therefore evidence collected by utilising this framework helps make these systems more transparent and meaningful to the stakeholders, as well as helps the stakeholders to appraise the implications of the use of AES scoring.

A further benefit of this framework is that it can be tailored and adapted for use in validation studies of other types of automated scoring systems, such as those developed to score open-ended architectural design problems (Bennett, 2004); or the Qualrus-based SA Grader designed to assess discipline-based substantive knowledge and reasoning through short and focused papers (Brent & Townsend, 2006). This is because the framework is developed on the fundamental principles and concepts of validation that are generalisable to all types of scoring systems.

In essence, when validating any type of scoring system or instrument, the approach encapsulated in the proposed AES validation framework, which represents a structured, coherent and integrated approach to examining validity, should be favoured over any siloed, piecemeal, or opaque approaches that could potentially provide misleading results. In a similar vein, the validation process undertaken in this study should apply to validation of any type of direct performance assessment, which aims to present a strong and coherent evidence-based validity argument. Such a validation process should consist of specifying an interpretative argument by making clear the assumptions and inferences for which evidence must be collected, and evaluating this interpretative argument in a particular context by

collecting and examining empirical evidence and theoretical rationale in an integrated manner.

12.4.2 Limitations of the Framework and Possible Future Enhancements

It should be noted that the framework has been developed by identifying those validity questions that directly relate to the AES. Therefore, it does not, in its current form, take into consideration the complex relationships a scoring system can have with other components of the test design, such as examinee interface, task design, and construct definition. These complex relationships may also have an impact on the validity of test scores. For example, a decision to use an AES system as a scoring tool may prompt the test administrators to change the form of the writing test from pen-and-paper to computer-based. The new testing mode may prohibit some testing cohorts from displaying their full competencies because of their “unfamiliarity with the computer interface in general” or their “anxiety about computer-testing mode” (Association of Test Publishers, 2002, as cited in Yang et al., p. 405). In this case, even if automated scoring may be perfectly accurate, its accuracy matters little if the writing responses collected are not representative samples of the underlying abilities.

Future work to enhance the proposed AES validation framework should consider expanding it to include the validity implications arising from the complex interactions between the requirements of automated scoring and other parts of test designs. Another future direction for this area of research work is to reconcile this framework with a recently proposed alternate validation framework for automated scoring systems (Williamson, Xi & Breyer, 2012), with

an aim to produce one unified model that consolidates and incorporates different elements from the two frameworks. Such a unified model should also recommend a set of standards that can be implemented by different AES vendors for selecting AES models for use in different contexts. The implementation of such a model will no doubt instil a level of standardisation and consistency into the model evaluation and selection processes, which are currently being carried out in an idiosyncratic manner across different AES vendors.

12.4.3 Rethinking of the Role of Human Judgements in the Evaluation and Development of AES Systems

Another important implication of the findings from this study for future AES evaluation and development is that the current study has provided further evidence to support Bennett's (2004) view that scores obtained from human markers should not be used as the sole validity criterion for evaluating AES systems, nor should they be used as the sole basis for future refinements of these systems.

This study has demonstrated that various aspects of the human marking process can change both the dependability and the distributional properties of the human scores produced. The differences in the human scores can result in quite different judgements with respect to the accuracy of the IEA scoring when IEA scores are evaluated against the human scores. It is therefore important, when human scores are used to validate AES scores, to include the evaluation of the quality of human scores as an integral part of the validation process.

In addition, human scores should not be the only criterion measure on which the validity of AES scores is assessed, considering the seemingly perennial nature of errors and biases that exist in human judgements. In this regard, this study has demonstrated various methods to assess AES score validity that are independent of scores assigned by human markers (e.g., through investigating the internal structure and the measurement properties of the AES scores).

Equally important is the view that future developments of AES systems should not rely solely on scores from human markers either. When AES systems are trained on human scores, which is the current standard practice of system building, systematic biases in human judgements can be modelled into these systems. This result is at odds with the great potential of AES technology, which is to remove biases in human scores. Evidence from this study provides strong backing for this position. The current study has found that when errors and biases are transferred to the AES systems, they tend to be exacerbated, rather than being simply reproduced.

For example, multiple pieces of evidence from this study point out that the two typical types of errors in human scores – “central tendency” and “halo effects” – are more pronounced in the IEA scores than in human scores (see discussions in Sections 9.5.2, 10.4, 11.4.1, 11.4.2 and 11.5.1). Evidence from other studies suggests that this same issue would appear to exist across other AES systems as well (e.g., Maddox, 2006; Rudner et al., 2006; Wang & Brown,

2007; Ziegler, 2006).⁴³ In order to realise the AES's full potential to improve assessment and measurement, it is therefore important for the future development of the AES to be more construct-driven; that is, to be based more directly on domain theory than on the scores assigned by human markers. Predicting human scores should not be the ultimate target of the AES systems. Rather the target should be producing valid and reproducible writing measures that are directly linked to writing proficiency. In this regard, refinements to AES should capitalise on theories of writing cognition (such as works of Bereiter & Scardamalia, 1987; Hayes, 1996), which are well articulated theories attempting to capture the differences between novice and expert writers (e.g., Bennett, 2004; Ben-Simon & Bennett, 2007; Quinlan et al., 2009).

12.4.4 Other Implications of this Study for Future AES Research and Study

Two other implications of this study's findings are also noted. First, this study has provided relatively strong evidence to support the view that the validation of AES systems needs to go beyond the agreement rates at the overall score level. The evidence collected in this study points out the necessity of investigating the appropriateness of the machine scoring at the trait level, as well as at the level of micro-textual features which contribute to the scoring of traits,

⁴³ For example, some studies have reported that scores from other AES systems are also more likely to be clustering around the midpoint on the scale than the corresponding human scores (e.g., Maddox, 2006; Ziegler, 2006). Other studies (e.g., Rudner et al., 2006; Wang & Brown, 2007) reported smaller dispersions in the AES scores than in the human scores, which were consistent with the finding from this study (E7 in Table 12.1). Smaller dispersion of scores could be a reflective indicator of either the "central tendency" effect, or "restriction of range" effect (i.e., the tendency of scores clustering around any points of the scale). The latter is another typical problem long recognised in human marking (Myford & Wolfe, 2003).

in order to unveil important validity implications that can otherwise be masked at the overall score level. The second implication is that, in the course of applying the framework, this study has demonstrated a number of new methods which can be used by future studies to investigate different aspects of validity for AES scores. These methods include: 1) the use of Principal Component Analysis technique to investigate the dimensional structure in the AES trait scores, to confirm the theoretical distinction about the construct; 2) the use of Multi-Dimensional Scaling to examine the internal structure of the trait scores produced by an AES system; and 3) the use of the multitrait-multimethod technique to collect the convergent and discriminant validity evidence to help establish the validity of the writing construct produced by an AES system.

12.5 Concluding Remarks

Automated Essay Scoring technology has great potential to improve writing assessment and instruction, as well as to reduce costs and improve marking efficiency. However, to realise the full extent of this potential, the development of AES will need to be more directly related to theories of good writing, and the evaluation of these systems will need to be more thorough and rigorous. The AES validation framework in this study is proposed as a useful mechanism enabling the assessment of the validity of these systems to be conducted in a systematic, robust and comprehensive manner. This should contribute to the further developments of AES systems, and to the wider realisation of the educational benefits of these systems.

As pointed out in Chapter One, the realisation of the educational benefits of the AES systems will also be contingent on the trust the stakeholders have in AES systems. One element underpinning trust is the level of understanding educational professionals have of these systems. When teachers better understand how AES systems assess the quality of written products and the strengths and limitations of the new technologies, it is anticipated that they will become more willing to engage with the new technology, including being more willing to experiment with these systems, and more willing to integrate AES technology with the existing teaching curriculum. Though this study made an attempt to describe the theoretical frameworks and innovative technologies underpinning the various AES systems, these efforts should be continued by more studies to further elucidate the characteristics, the internal structure, the strength and the limitations of these systems. It will be through the combined and sustained efforts of both AES system developers and test validators to make the AES systems more transparent, and the continual strengthening of the validity of the scores generated by the AES systems, that the potential educational benefits of these systems can be fully realised.

References

- Abney, S. (1996). Part of speech tagging and partial parsing. In K. Church, S. Young & G. Bloothoof (Eds.), *Corpus-based methods in language and speech* (pp.118-136). Dordrecht: Kluwer.
- Adams, R., & Khoo, S. T. (1993). *Quest: The interactive test analysis system* [Computer software]. Camberwell, Victoria: Australian Council for Educational Research.
- Alamargot, D., & Andriessen, J. (2002). The ‘power’ of text production activity in collaborative modelling: Nine recommendations to make a computer supported situation work. In P. Brna, M. Baker, K. Stenning & A. Tiberghien (Eds.), *The role of communication in learning to model* (pp.275–300). Mahwah, NJ: Lawrence Erlbaum Associates.
- Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. Dordrecht: Kluwer Academic Publishers.
- American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.
- Andrich, D. A. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Andrich, D. A. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

- Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C. C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp.3–35). Orlando, FL: Academic Press.
- Anson, C. M. (2006). Can't touch this. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.39–56). Logan, UT: Utah State University Press.
- Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.
- Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS Research Rep. No RR-07-21). Princeton, NJ: ETS.
- Attali, Y. (2009, April). *Evaluating automated scoring for operational use in consequential language assessment—the ETS experience*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, 4(3). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>
- Attali, Y., & Powers, D. (2008). *A developmental writing scale*. (ETS Research Rep. No RR-08-19.) Princeton, NJ: ETS.
- Australian Curriculum, Assessment and Reporting Authority. (2011). *NAPLAN summary results 2011*. Retrieved from http://www.nap.edu.au/_Documents/PDF/2011%20NAPLAN%20Summary%20Report.pdf

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: 12 Cambridge University Press.

Bakeman, R., Quera, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2, 357–370.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing*, 12, 86–107.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education*, 4, 305–318.

Baron, J., Treiman, R., Wilf, J. F., & Kellman, P. (1980). Spelling and reading by rules. In U. Frith (Ed.), *Cognitive processes in spelling* (pp.159–194). New York, NY: Academic Press.

Becker, A. (2006). A review of writing model research based on cognitive processes. In A. Horning & A. Becker (Eds.), *Revision: history, theory, and practice (reference guides to rhetoric and composition)* (pp.25–49). West Lafayette, IN: Parlor Press.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay score. *Journal of Technology, Learning, and Assessment*, 6(1). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

- Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring*. (ETS Research Memorandum RM-04-01.) Princeton, NJ: ETS.
- Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4), 9–17.
- Benton, S. L, Sharp, J. M., Corkill, A. J., Downey, R. G., & Khramtsova, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology*, 87, 66–79.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Berthoff, A. E. (1981). *The making of meaning: Metaphors, models, and maxims for writing teachers*. Portsmouth, NH: Boynton/Cook Heinemann.
- Bissett, K., & McDougall, B. (2008, January 21). Record number of pupils drop out. *Daily Telegraph*, p. 4.
- Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139–148.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bond, T. G, & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York, NY: Springer.

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York, NY: Springer-Verlag.

Braungart-Bloom, D. S. (1986, April). *Assessing holistic raters' perceptions of writing qualities: An examination of a hierarchical framework following pre-post training and live readings*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Breland, H. M., & Jones, R. J. (1982). *Perceptions of writing skills* (College Board Report No. 82-4, ETS Research Report No. 82-47). New York, NY: College Entrance Examination Board.

Breland, H., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, 1(1), 101–119.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Report No. 94-4, Educational Testing Service Research Report No. 94-26). New York, NY: College Entrance Examination Board.

Breland, H., Lee, Y., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (TOEFL Research Report No 76). Princeton, NJ: ETS.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.), (pp.1–16). Westport, CT: American Council on Education.

Brent, E., & Townsend, M. (2006). Automated essay grading in the sociology classroom. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.177–198). Logan, UT: Utah State University Press.

Brewer, C. A. (2004). Near real-time assessment of student learning and understanding in biology courses. *Bioscience*, 54, 1034–1039.

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April). *Considering fairness and validity in evaluating automated scoring*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA. Retrieved from http://www.ets.org/Media/Conferences_and_Events/AERA_2009_pdfs/AERA_NCME_2009_Bridgeman.pdf

Bridgeman, B., Trapani, C., & Williamson, D. M. (2011, April). *The question of validity of automated essay scores and differentially valued evidence*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Broad, B. (2006). More work for teacher? Possible futures of teaching writing in the age of computerized assessment. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.221–233). Logan, UT: Utah State University Press.

Brooks, G., Gorman, T., & Kendall, L. (1993). *Spelling it out: the spelling abilities of 11- and 15-year-olds*. Berkshire, England: National Foundation for Educational Research.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34(4), 21–42.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics. Minimising the effect of mean differences. *Written Communication*, 8, 533–556.

Burstein, J. (2003). The E-rater® scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp.112–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Higgins, D. (2007). *Evaluation of Criterion feedback codes for sentence checking in FMI's ProofWriter*. Unpublished manuscript.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA. Retrieved from <http://www.ets.org/research/dload/ncmefinal.pdf>

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (2001). *Enriching automated essay scoring using discourse marking*. (ERIC reproduction service no ED 458 267).

Burstein, J., Marcu, D., Andreyev, S., & Chodorow, M. (2001). Towards automatic classification of discourse elements in essays. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, 98–105.
doi:10.3115/1073012.1073026.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing*, 18(1), 32–39.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105.

- Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics*, 11(2), 207–241.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
- Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal*, 93(4), 47–52.
- Chodorow, M., & Burstein, J. (2004). *Beyond essay length: evaluating e-rater's performance on TOEFL essays* (Research Reports, Report 73, ETS). Princeton, NJ: ETS.
- Choppin, B. (1982). The Rasch model for item analysis. In B. Choppin, D. L. McArthur, K. A. Sirotnik, R. K. Hambleton, R. R. Wilcox, N. Webb, ... J. W. Keesling, *A critical comparison of psychometric models for measuring achievement. Methodology project*. California University, LA: Centre for the Study of Evaluation: 1-279 (ERIC reproduction service no ED 224 823).
- Chung, K. W. K., & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays*. (ERIC reproduction service no ED 418 101).
- Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, 24(4), 310–324.
- Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4), 413-432.
- Cohen, R. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of 1984 International Computational Linguistics Conference* (pp.251–255). Stroudsburg, PA: Associations for Computational Linguistics.

Cohen, R. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163–178.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp.141–160). Boston, MA: Heinle & Heinle.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research* (GRE Board Research Report 84-12). Princeton, NJ: Educational Testing Service.

Crick, G. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalised analysis of variance system* (ACT Technical Bulletin No 43). Iowa City, IA: American College Testing Program.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.), (pp.443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalisability of scores and profiles*. New York, NY: John Wiley.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58, 438–481.

Cumming, A. (1990). Expertise in evaluating second language composition. *Language Testing*, 7, 31–51.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype writing tasks for new TOEFL* (TOEFL Monograph No. MS-30). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making, and development of a preliminary analytic framework* (TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86(1), 67–96.

Cumming, A., & Mellow, D. (1996). An investigation into the validity of written indicators of second language proficiency. In A. Cumming & R. Berwick (Eds.), *Validation in language testing. Modern languages in practice 2* (pp.72–93). Bristol,

PA: Multilingual Matters.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp.621–694). Washington, DC: American Council on Education.

Daftaripard, P., & Lange, R. (2009). Theoretical complexity vs. Rasch item difficulty in reading tests. *Rasch Measurement Transactions*, 2009, 23(2), 1212–1213.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7, 171–176.

Davies, B., & Gralton, T. (2009). *Automated essay scoring – 2008 Australian trial*. Retrieved from http://www.education.tas.gov.au/__data/assets/pdf_file/0003/299145/Automated-Essay-Scoring.pdf

Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R.A. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6), 391–407.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in the judgements of writing ability (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.
- Drechsel, J. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two Year College*, 26, 380–387.
- Dumais, S. T., Furnas, G. W., Landauer, T. K., & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*, 281–285, New York, NY: ACM.
- DuPont, S. (2002). *Employers, professors rate high school grads as computer whizzes, but just 'fair' or 'poor' on their writing, grammar, arithmetic*. Retrieved from <http://www.publicagenda.org/press-releases/what-happened-three-rs>
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155–185.
- Eckes, T. (2009). Many-facet Rasch measurement. In *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Language Policy Division, Council of Europe.
- Educational Testing Service (2006). *The official guide to the new TOEFL iBT*. New York: McGraw-Hill.
- Educational Testing Service (2007). *TOEFL iBT reliability and generalisability of scores*. Princeton, NJ: Educational Testing Service.

Educational Testing Service (2011). *Validity evidence supporting the interpretation and use of TOEFL iBT™ scores*. Princeton, NJ: Educational Testing Service. Retrieved from <http://www.toeflgoanywhere.org/enewsletter/april2011/New-TOEFL-iBT-Research-Insight.html>

Elbow, P. (1981). *Writing with power*. Oxford: Oxford University Press.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37–64.

Elliot, S. (2003). Intellimetric™: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp.43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Elliot, S., & Mikulas, C. (2004, April). *The impact of MyAccess!™ use on student writing performance: A technology overview and four studies from across the nation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Ellis, N. C. (1994). Longitudinal studies of spelling development. In G. D. A. Brown & N. C. Ellis (Eds.), *Handbook of spelling: Theory, process and intervention* (pp.155–178). New York, NY: John Wiley and Sons.

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171–191.

Engelhard, G., Gordon, B., Walker, E. V., & Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research*, 87, 197–209.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgement of essays written by English language learners with e-rater scoring. *Language testing*, 27(3), 317–334.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Rep. No. 70). Princeton, NJ: ETS.

Ericsson, P. F. (2006). The meaning of meaning. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.29–38). Logan, UT: Utah State University Press.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37, 313–326.

Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, 31(8), 4–14.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10(4), 507–521.

Fisher, R. A. (1921). On the ‘probable error’ of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3–32.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behaviour Research Methods, Instruments and Computers*, 28(2), 197–202.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3), 285–307.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27–32.

Freedman, S. W. (1977). *Influences on the evaluators of student writing*. (Unpublished doctoral dissertation). Stanford University, Stanford, CA.

Freedman, S. W. (1979a). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology*, 71, 328–338.

Freedman, S. W. (1979b). Why do teachers give the grades they do? *College Composition and Communication*, 30, 161–164.

Galbraith, D. (2009). Cognitive models of writing. *German as a Foreign Language (GFL)*, 2-3, 7–22.

Golub, G. H., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 2(2), 205–224.

Google Wave, Microsoft Office and Ghotit contextual spell checker comparison (n.d.). Retrieved from <http://www.ghotit.com/context-spell-check.shtml>

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, 41(3), 258–269.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education*, 27, 73–82.

Grimes, D. (2008). *Middle school use of automated writing evaluation*. (Unpublished doctoral dissertation). University of California, Irvine, Irvine, CA. Retrieved from <http://douglasgrimes.com/windocs/Grimes--Middle%20School%20Use%20of%20AWE--Final%20Dissertation%20.doc>

Grimes, D., & Warschauer, M. (2008, March). *Middle school use of automated writing evaluation*. Paper presented at the annual convention of the American Educational Research Association. New York, NY. Retrieved from <http://douglasgrimes.com/windocs/Grimes+Warschauer--AERA%202008--Middle%20School%20Use%20of%20AWE.doc>

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: a multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, 8(6). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

Grishman, R., MacLeod, C., & Meyers, A. (1994). COMPLEX syntax: Building a computational lexicon. *Proceedings of Coling, Kyoto, Japan*. Retrieved from <http://cs.nyu.edu/cs/projects/proteus/complex/>

Grudin, J. T. (1983). Error patterns in novice and skilled transcription typing. In W. E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp.121–144). New York, NY: Springer-Verlag.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York, NY: McGraw-Hill.

Guion, R. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist*, 29, 287–296.

- Gyagenda, I. S., & Engelhard, G., Jr. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, 10(3), 225–246.
- Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp.69–87). Cambridge: University of Cambridge Press.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second-language writing in academic contexts* (pp.241–276). Norwood, NJ: Ablex.
- Hamp-Lyons, L. (1995). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly*, 29(4), 759–762.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337–373.
- Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community and assessment*. TOEFL Monograph Series MS-5. Princeton, NJ: Educational Testing Service.
- Hansen, R. S., & Hansen, K. (1997). *Write your way to a higher GPA*. Berkeley, CA: Ten Speed Press.
- Hardy, R. A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education*, 8, 121–134.

Harris, W. H. (1977). Teacher response to student writing: a study of the response patterns of high school English teachers to determine the basis for teacher judgement of student writing. *Research in the Teaching of English*, 11, 175–185.

Haswell, R. H. (2006). Automations and automated scoring. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.56–78). Logan, UT: Utah State University Press.

Hayes, C. (2010). *Testing spelling: an investigation into the misspellings of non-native English speakers on the Pearson Test of English Academic*. (Unpublished master's thesis). University of London, UK.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organisation of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp.31–50). Hillsdale, NJ: Lawrence Erlbaum Associates.

Henly, D. C. (2003). Use of web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education*, 7, 116–122.

Herrington, A., & Moran, C. (2006). Writeplacer Plus in place. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.115–127). Logan, UT: Utah State University Press.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145–159.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 18(1), 87–107.

How to tackle the analytic writing assessment? Retrieved from <http://800score.com/gmat-essay.html>

Hunter, D. M., Jones, R. M., & Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation*, 11(2), 61–85.

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk-aloud protocols of expert and novice holistic raters*. (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Philadelphia, PA.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, 41, 201–213.

Huot, B. (1990b). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, 60, 237–263.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp.206–236). Cresskill, NJ: Hampton Press.

IELTS (n.d). *Task 1 writing band descriptors (public version)*.Retrieved from http://www.ielts.org/pdf/UOBDs_WritingT1.pdf

- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House Publishers.
- Jones, B. E. W. (1978). Marking of student writing by high school teachers in Virginia during 1976. *Dissertation Abstracts International*, 38, 3911A.
- Jones, E. (2006). ACCUPLACER'S essay-scoring technology. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.78–92). Logan, UT: Utah State University Press.
- Justham D., & Timmons, S. (2005). An evaluation of using a web-based statistics test to teach statistics to postregistration nursing students. *Nurse Education Today*, 25, 156–163.
- Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31–41.
- Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, 2(3), 135–170.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.), (pp.17–64). Westport, CT: American Council on Education.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
- Kaplan, R. M., Wolff, S. E., Burstein, J. C., Lu, C., Rock, D. A., & Kaplan, B. (1998). *Scoring essays automatically using surface features*. (GRE Board Professional Report No. 94-21P). Princeton, NJ: ETS.

- Keeves, J. P. & Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 23–42). Amsterdam, Netherlands: Pergamon/Elsevier Science.
- Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp.147–167). Mahwah, NJ: Erlbaum.
- Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition*, 15, 256–266.
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp.57-71). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kelly, P. A. (2006). Review of the book *Automated essay scoring: A cross-disciplinary perspective*. *Applied Psychological Measurement*, 30(1), 66-68.
- Kemp, F. (1992). Who programmed this? Examining the instructional attitudes of writing support software. *Computers and Composition*, 10(1), 9–24.
- Kiniry, M., & Strenski, E. (1985). Sequencing expository writing: A recursive approach. *College Composition and Communication*, 36(2), 191–202.
- Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12(1), 26–43.
- Lado, R. (1961). *Language Testing*. New York, NY: McGraw-Hill.

- Landauer, T. K., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.
- Landauer, T. K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia*, 3(11), 4356.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259–284.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2001). *Automatic essay assessment with latent semantic analysis*. Unpublished manuscript.
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary perspective* (pp.87–112). Mahwah, NJ: Lawrence Erlbaum Associates.
- Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. *Proceedings of the 19th Annual Conference of the Cognitive Science Society* (pp.412–417). Mahwah, NJ: Erlbaum.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. San Diego, CA: Academic Press.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418.

Lee, Y., Gentile, C., & Kantor, R. (2008). *Analytic Scoring of TOEFL CBE essays: Scores from humans and e-rater* (TOEFL Research Rep. No. RR-81). Princeton, NJ: ETS.

Lee, Y. W., Kantor, R., & Mollaun, P. (2002, April). *Score reliability as an essential prerequisite for validating new writing and speaking tasks for TOEFL*. Paper presented at the annual meeting of Teachers of English to the Speakers of Other Languages (TESOL). Salt Lake City, UT.

Lee, Y. W., & Kong, N. (2004). *A preliminary investigation of feature organisation frameworks for automated essay scoring and feedback*. Unpublished manuscript.

Lim, G. (2009). *Prompt and rater effects in second language writing performance assessment*. (Unpublished doctoral dissertation). Retrieved from <http://hdl.handle.net/2027.42/64665>

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, 3(2), 103–122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16, 878.

Linacre, J. M. (2008). *Facets Rasch model computer program* [Software program manual]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2010). *A user's guide to Winsteps Ministep Rasch-Model computer programs* [Software program manual 3.70.0]. Chicago, IL:Winsteps.com

Linacre, J. M., & Tennant A. (2009). More about critical Eigenvalue sizes in standardised-residual Principal Component Analysis (PCA). *Rasch Measurement Transactions*, 23(3), 1228.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, 3(4), 484–509.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Columbus, OH: Merrill.

Litman, D. (1996). Cue phrase classification using machine learning. *Artificial Intelligence*, 5, 53–94.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, (Monograph Supplement 9), 635–694.

Lopez, W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions*, 10(1), 482.

Maddox, T. (2006). Piloting the compass e-write software at Jackson State Community College. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.147–153). Logan, UT: Utah State University Press.

Markham, L. R. (1976). Influence of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13(4), 277–283.

Marsh, H. W., & Ireland, R. (1984). *Multidimensional evaluations of writing effectiveness*. [microform] Washington, D.C.: Distributed by ERIC Clearinghouse.
<http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED242785>

Masters, G. N. (2002). *Fair and meaningful measures? A review of examination procedures in the NSW Higher School Certificate*. Retrieved from
http://www.boardofstudies.nsw.edu.au/manuals/pdf_doc/masters_review.pdf

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26(1), 75–100.

Matthews, J. (2004, August 1). Computers weighing in on the elements of essay programs critique structure, not ideas. *The Washington Post*, p. A01.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research*, 64, 148–156.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language*, 25, 431–444.

McGee, T. (2006). Taking a spin on the Intelligent Essay Assessor. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.78–92). Logan, UT: Utah State University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 28(2), 5–11.

Messick, S. (1995). Standards of validity and the validity of and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp.1–18). Washington, DC: National Centre for Educational Statistics.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–383.

Mikulas, C., & Kern, K. (2006, April). *A comparison of the accuracy of automated essay scoring using prompt-specific and prompt-independent training*. Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Mitton, R. (1987). Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management*, 23(5), 495–505.

Mitton, R. (1996). *English spelling and the computer*. Harlow, Essex: Longman.

Moon, T. R., Loyd, B. H., & Hughes, K. R. (1996, April). *Generalisability analyses of a large-scale writing assessment*. Paper presented at the annual meeting of American Educational Research Association, New York, NY.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grade 4, 7 and 10. *Journal of Educational Measurement*, 19, 37–47.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386–422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46, 371–389.

Nichols, P. D. (2004, April). *Evidence for the interpretation and use of scores from an automated essay scorer*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: an illustration for collections of student writing. *The Journal of Educational Research*, 89, 220–233.

Noyes, E. S. (1963). Essays and objective tests in English. *College Board Review*, 49, 7–10.

Odell, L. (1981). Defining and assessing competence in writing. In C. R. Cooper (Ed.), *The nature and measurement of competency in English* (pp.95–138). Urbana, IL: National Council of Teachers of English.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5), 238–243.

Page, E. B. (1968). Analyzing student essays by computer. *International Review of Education*, 14, 210–225.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62, 127–142.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp.43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76(6), 561–566.

Pearson. (2009). Automated scoring writing. Retrieved from http://www.pteacademic.kr/download/US_Automated_Scoring_Writing_V4.pdf

Pearson. (2011a). *Validity and reliability in PTE Academic*. Retrieved from http://www.pearsonpte.com/research/Documents/PTEA_Test%20Validity_Reliability.pdf

Pearson. (2011b). *PTE Academic score guide, November 2011*. Retrieved from http://pearsonpte.com/PTEAcademic/scores/Documents/PTEA_Score_Guide.pdf

Pearson. (2011c). *Pearson Test of English Academic: Automated scoring*. Retrieved from http://www.pearsonpte.com/research/Documents/PTEA_Automated_Scoring.pdf

- Peat, M., & Franklin, S. (2002). Supporting student learning: The use of computer-based formative assessment modules. *British Journal of Educational Technology*, 33(5), 515–523.
- Perin, D. (1983). Phonemic segmentation and spelling. *British Journal of Psychology*, 74, 129–144.
- Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly*, 14(4), 651–671.
- Peterson, J. L. (1986). A note on undetected typing errors. *Communications of the ACM*, 29(7), 633–637.
- Petersten, N. S. (1997, March). *Automated scoring of written essays: Can such scores be valid?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, 25(2), 237–272.
- Pollock, J. J., & Zamora, A. (1983). Collection and characterization of spelling errors in scientific and scholarly text. *Journal of the American Society for Information Science*, 34(1), 51–58.
- Pollock, J. J., & Zamora, A. (1984). Automatic spelling correction in scientific and scholarly text. *Communications of ACM*, 27(4), 358–368.
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping E-rater: Challenging the validity of automated essay scoring* (GRE Board Professional Report No. 98-08bP). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, 26, 407–425.

PTE Academic Australia (n.d). Retrieved from <http://www.pearsonpte.com/australia/Documents/AustraliaRecoPoster.pdf>

PTE Academic UK (n.d). Retrieved from <http://pearsonpte.com/TestMe/About/Pages/ukba.aspx>

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp.237–265). Cresskill, NJ: Hampton Press.

Purves, A. C., Soter, A., Takala, S., & Vahapassi, A. (1984). Towards a domain-referenced system for classifying assignments. *Research in the Teaching of English*, 18(4), 385–416.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (ETS Research Report RR-09-01). Princeton, NJ: ETS.

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). *A comprehensive grammar of the English language*. New York, NY: Longman.

Raforth, B. A., & Rubin, D. L. (1984). The impact of content and mechanics on judgements of writing quality. *Written Communication*, 1, 446–458.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published 1960).

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. *Proceedings of the Empirical Methods in Natural Language Processing Conference, USA, 19*, 1133–1141.

Rijlaarsdam, G., & van den Bergh, H. (1996). The dynamics of composing – an agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp.107–126) Mahwah, NJ: Erlbaum.

Robbins, S. P. (1989). *Organisational behaviour* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Rothermel, B. (2006). Automated writing instruction: Computer-assisted or computer-driven pedagogies? In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.39–56). Logan, UT: Utah State University Press.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *Journal of Technology, Learning, and Assessment*, 4(4). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413–428.

- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18, 613–620.
- Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly*, 22(1), 69–90.
- Scharber, C., Dexter, S., & Riedel, E. (2008). Students' experiences with an automated essay scorer. *Journal of Technology, Learning, and Assessment*, 7(1).
Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>
- Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modelling. *Language Testing*, 22(1), 1–30.
- Shavelson, R. J., Baxter, G. P., & Gao, X. 1993. Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215–232.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.
- Sheehan, K. (2001). Discrepancies in human and computer generated essay score for TOEFL-CBT essays. Unpublished manuscript.
- Shermis, M. D., Koch, C. M., Page, E. B, Keith, T. Z., & Harrington, S. (2002). Trait rating for automated essay scoring. *Educational and Psychological Measurement*, 62, 5–18.
- Shih, M. (1986). Content-based approaches to teaching academic writing. *TESOL Quarterly*, 20(4), 617–648.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, 27(4), 657–677.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, 10(3), 516–517.

Spandel, V., & Stiggins, R. J. (1990). *Creating writers: Linking assessment and writing instruction*. New York, NY: Longman.

Sterling, C. M. (1983). Spelling errors in context. *British Journal of Psychology*, 74, 353–364.

Stevens, J. J., & Clauser, P. (1996, April). *Longitudinal examination of writing portfolio and the ITBS*. Paper presented at the annual meeting of American Educational Research Association, New York, NY. (ERIC Document Reproduction Service No. ED 397 116)

Swartz, R., Patience, W., & Whitney, D. R. (1985). *Adding an essay to the GED writing skills test: reliability and validity issues* (GED Testing Service Research Studies, No.7). (ERIC Document Reproduction Service No. ED 266 288)

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 7–67.

Talento-Miller, E., Siegert, K. O., & Taliaferro, H. (2011). *Evaluating analytical writing for admission to graduate business programs* (GMAC Research Report No. RR-11-03). Reston, VA: Graduate Management Admission Council.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York, NY: John Wiley and Sons.

Tognolini, J. (1989). *Psychometric profiling and aggregating of public examinations at the level of test scores*. (Unpublished doctoral dissertation). Murdoch University, Perth, Western Australia, Australia.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. *The Elementary School Journal*, 90(1), 87–95.

Vann, R. J., Meyer, D. E., & Lorenz, F. O. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly*, 18(3), 427–440.

Vantage Learning. (2001). *RB 612 – WritePlacer research summary*. Yardley, PA: Vantage Learning.

Vantage Learning. (2003a). *Assessing the accuracy of IntelliMetric for scoring a district-wide writing assessment (RB-806)*. Newtown, PA: Vantage Learning.

Vantage Learning. (2003b). *How does IntelliMetric score essay responses? (RB-929)*. Newtown, PA: Vantage Learning.

Vantage Learning. (2003c). *A true score study of 11th grade student writing responses using IntelliMetric Version 9.0 (RB-786)*. Newtown, PA: Vantage Learning.

Vantage Learning. (n.d.). *IntelliMetric: How it works?* Retrieved from <http://www.vantagelearning.com/products/intellimetric/intellimetric-how-it-works/>

Vaughan, C. (1987, March). *What affects raters' judgements?* Paper presented at the meeting of the Conference on College Composition and Communication, Atlanta, GA.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large scale evaluation of writing. *Research in the Teaching of English*, 17, 285–296.

Viera, A., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine*, 37(5), 360–363.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6(2), 103–118.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, 6(2). Available from <http://ejournals.bc.edu/ojs/index.php/jtla/>

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies*, 3(1), 52–67.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2010). Validation of automated scoring of TOEFL iBT tasks against non-test indicators of writing. *Language Testing*, 27(3), 335–353.

- Weigle, S. C., & Lynch, B. (1996). Hypothesis testing in construct validation. In A. Cumming & R. Berwick (Eds.), *Validation in language testing. Modern languages in practice 2* (pp.58–71). Bristol, PA: Multilingual Matters.
- Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educational and Psychological Measurement*, 40, 19–29.
- White, E. M. (1984). Holisticism. *College Composition and Communication*, 35, 400–409.
- Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, 75(3), 200–214.
- Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). ‘Mental model’ comparison of automated and human scoring. *Journal of Educational Measurement*, 36, 158–184.
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2-13.
- Wing, A. M., & Baddeley, A. D. (1980). Spelling errors in handwriting: a corpus and a distributional analysis. In U. Frith (Ed.), *Cognitive processes in spelling* (pp.251–285). London: Academic Press.
- Wohlpart, J., Lindsey, C., & Rademacher, C. (2008). The reliability of computer software to score essays: Innovations in a humanities course. *Science Direct Computers and Composition*, 25, 203–223.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

WriteToLearn 7.0 goes international with increased support for English language learners. August 2011. Retrieved from <http://www.writetolearn.net/news/08222011.php>

Xi, X. (2007). Methods of test validation. In N. H. Hornberger (Series Ed.) & E. Shohamy & N. H. Hornberger (Eds.), *Encyclopaedia of language and education: Vol. 7*, (2nd ed.), (pp.177–196). Boston, MA: Springer.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater v1.0* (ETS Research Report No. RR-08-62). Princeton, NJ: ETS.

Yang, Y., Buckendahl, C., Juskiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4), 391–412.

Young, F. W., & Lewycky, R. (1979). *ALSCAL-4 user's guide*. Carrboro, NC: Data Analysis and Theory Associates.

Ziegler, W. W. (2006). Computerised writing assessment. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays, truth and consequences* (pp.138–146). Logan, UT: Utah State University Press.

Appendix A PTE Academic Writing Scoring Rubric

Content	Formal Requirement	Development, structure and coherence	Grammar Usage and Mechanics*
3: Adequately deals with the prompt			
2: Deals with the prompt but does not deal with one minor aspect	2: Length is between 200 and 300 words	2: Shows good development and logical structure	2: Shows consistent grammatical control of complex language. Errors are rare and difficult to spot
1: Deals with the prompt but omits one major aspect or more than one minor aspect	1: Length is between 120 and 199 or between 301 and 380 words	1: Is incidentally less well structured, and some elements or paragraphs are poorly linked	1: Shows a relatively high degree of grammatical control. No mistakes which would lead to misunderstandings
0: Does not deal properly with the prompt	0: Length is less than 120 or more than 380 words. Essay is written in capital letters, contains no punctuation or only consists of bullet points or very short sentences	0: Lacks coherence and mainly consists of lists or loose elements	0: Contains mainly simple structures and/or several basic mistakes
General linguistic range		Vocabulary range	Spelling
2: Exhibits mastery of a wide range of language to formulate thoughts precisely, give emphasis, differentiate and eliminate ambiguity. No sign that the test taker is restricted in what they want to communicate		2: Good command of a broad lexical repertoire, idiomatic expressions and colloquialisms	2: Correct spelling, but there may be one typing error
1: Sufficient range of language to provide clear descriptions, express viewpoints and develop arguments		1: Shows a good range of vocabulary for matters connected to general academic topics. Lexical shortcomings lead to circumlocution or some imprecision	1: One spelling error and/or more than one typing error
0: Contains mainly basic language and lacks precision		0: Contains mainly basic vocabulary insufficient to deal with the topic at the required level	0: More than one spelling error and/or numerous typing errors

Source: Adapted from PTE Academic Score Guide, November, 2011 (Pearson, 2011b, p. 60).

Note*: This scoring rubric is essentially the same as the rubric this researcher received in 2009. The only noticeable difference is the name for the “*Grammar Usage and Mechanics*” trait in the original rubric was changed to “*Grammar*” in the current published rubric (Pearson, 2011b), although the scoring criteria for this trait has not changed. For this thesis, the original name “*Grammar Usage and Mechanics*” is used.

Appendix B Scoring Rubrics for the Common European Framework (CEF) Scale

Score Point	CEF-W (CEF Written response © 2001 Council of Europe)
4	Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
3	Can write clear, well-structured expositions on complex subjects, underlining the relevant salient issues. Can expand and support points of view at some length with subsidiary points, reasons and relevant examples.
2	Can write an essay or report that develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem.
1	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on familiar routine and non-routine matters within his field with some confidence. Can write very brief, reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
0	I can write a series of simple phrases and sentences possibly linked with simple connectors like „and“, „but“ and „because“. Or uses some unconnected phrases or isolated words.
9	There is no response, response is not English or irrelevant.

Note: Scoring rubric provided by Pearson for this study in 2009

Appendix C Scoring Rubrics for the ESL Composition Profile (Original)

CONTENT	30–27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
	26–22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail
	21–17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic
	16–13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate
ORGANIZATION	20–18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive
	17–14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing
	13–10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development
	9–7	VERY POOR: does not communicate • no organization • OR not enough to evaluate
VOCABULARY	20–18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
	17–14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage but meaning not obscured
	13–10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured
	9–7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate

LANGUAGE USE	25–22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
	21–18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
	17–11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured
	10–5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate
MECHANICS	5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing
	4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured
	3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • meaning confused or obscured
	2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate

Copyright © 1981 by Holly L. Jacobs, V. Faye Hartfiel, Jane B. Hughey, and Deanna R. Wormuth. Newbury House Publisher. All rights reserved. See Jacobs, Zinkgraf, Wormuth, Hartfiel & Hughey (1981), pp. 92–96 for full details of the rubric.

**Appendix D Scoring Rubrics for the Modified ESL Composition Profile
– the Analytic Rating Scale Used by Markers in this Study**

CONTENT	3	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
	2	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail
	1	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic
	0	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate
ORGANIZATION	3	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive
	2	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing
	1	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development
	0	VERY POOR: does not communicate • no organization • OR not enough to evaluate
VOCABULARY	3	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register
	2	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage but meaning not obscured
	1	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • meaning confused or obscured
	0	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate

LANGUAGE USE	3	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
	2	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions but meaning seldom obscured
	1	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • meaning confused or obscured
	0	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate
MECHANICS	3	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing
	2	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing but meaning not obscured
	1	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • meaning confused or obscured
	0	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate

Adapted from Copyright © 1981 by Holly L. Jacobs, V. Faye Hartfiel, Jane B. Hughey, and Deanna R. Wormuth. Newbury House Publisher.

Appendix E Scoring Rubrics for TOEFL Independent Writing Tasks



iBT/Next Generation TOEFL Test Independent Writing Rubrics (Scoring Standards)

Score	Task Description
5	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> effectively addresses the topic and task is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details displays unity, progression, and coherence displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> addresses the topic and task well, though some points may not be fully elaborated is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	<p>An essay at this level is marked by one or more of the following:</p> <ul style="list-style-type: none"> addresses the topic and task using somewhat developed explanations, exemplifications, and/or details displays unity, progression, and coherence, though connection of ideas may be occasionally obscured may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning may display accurate but limited range of syntactic structures and vocabulary
2	<p>An essay at this level may reveal one or more of the following weaknesses:</p> <ul style="list-style-type: none"> limited development in response to the topic and task inadequate organization or connection of ideas inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task a noticeably inappropriate choice of words or word forms an accumulation of errors in sentence structure and/or usage
1	<p>An essay at this level is seriously flawed by one or more of the following weaknesses:</p> <ul style="list-style-type: none"> serious disorganization or underdevelopment little or no detail, or irrelevant specifics, or questionable responsiveness to the task serious and frequent errors in sentence structure or usage
0	<p>An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.</p>

Copyright © 2004 by Educational Testing Service. All rights reserved.

Appendix F Scoring Scheme Used in this Study

		Voting			
essay seq#	Analytic			Holistic	
	R1	R2	R3	R5	R4
10	40	40		40	40
20					
30					
40					
50	40		40	20	20
60					
70					
80					
90		40	40		
100					
110					
120					

		Tobacco			
essay	Holistic		Analytic		
seq#	R1	R2	R3	R4	R5
10	40	40		40	40
20					
30					
40					
50	20	20	40	40	
60					
70					
80					
90			40		40
100					
110					
120				40	

R1: Marker 1 R2: Marker 2 R3: Marker 3 R4: Marker 4 R5: Marker 5
Seq#: sequence number in the randomly ordered essays for a particular prompt.
The cell values indicate the sample sizes.

Appendix G Background Questionnaire of the Markers

Markers' Background Survey

Thank you for participating in this study.

Please complete the following survey and email it back to the researcher: lucy.lu@det.nsw.edu.au or fax to (02) 9561 8055.

The researcher will contact you as soon as possible to let you know of arrangements.

a) First Name _____ Surname _____

b) Contact Phone No. _____ Email (optional) _____

c) Gender _____

d) Academic Qualification/s _____

e) What subjects have you marked in HSC examinations and for how many years have you marked these subjects? (e.g., HSC English Standard, 5 years; History, 3 years)

f) What subjects do you teach at school, and for how many years have you taught these subjects? (e.g., Year 10 English, 5 years; Year 10, History, 3 years)

g) Have you heard about Automatic Essay Scoring Systems? Y/N _____

If yes, what do you think about them?

Appendix H G-Study Results – Estimated Variance Components (σ^2) in Holistic and Analytic Human Ratings – Voting

Source of Variation	Person (σ_p^2)			Rating ($\sigma_{r'}^2$)			Residual ($\sigma_{pxr',e}^2$)		
Trait	Estimated Variance Component	% of total variance	SE	Estimated Variance Component	% of total variance	SE	Estimated Variance Component	% of total variance	SE
<i>Overall Scores Produced by Different Rating Method</i>									
Holistic Score	0.64	58.70%	0.15	0.16	14.30%	0.13	0.29	27.00%	0.05
Composite Analytic Score	0.3	74.50%	0.05	0.00	0.00%	0.00	0.10	25.50%	0.01
<i>Analytic Trait</i>									
<i>Content</i>	0.36	57.90%	0.07	0.00	0.00%	0.00	0.26	42.10%	0.03
<i>Language Use</i>	0.38	61.00%	0.07	0.02	2.40%	0.01	0.23	36.60%	0.03
<i>Mechanics</i>	0.37	57.90%	0.07	0.01	0.90%	0.01	0.27	41.20%	0.03
<i>Organisation</i>	0.36	62.60%	0.06	0.00	0.10%	0.00	0.22	37.40%	0.03
<i>Vocabulary</i>	0.18	37.00%	0.05	0.01	1.30%	0.01	0.30	61.70%	0.04

Note: The composite analytic scores are calculated using Formula 8.1 to combine the ratings each essay received on the 5 analytic traits.

Appendix I G-Study Results – Estimated Variance Components (σ^2) in Holistic and Analytic Human Ratings – Tobacco

Source of Variation	Person (σ_p^2)			Rating ($\sigma_{r'}^2$)			Residual ($\sigma_{pxr'e}^2$)		
Method/Trait	Estimated Variance Component	% of total variance	SE	Estimated Variance Component	% of total variance	SE	Estimated Variance Component	% of total variance	SE
<i>Overall Scores Produced by Different Rating Methods</i>									
Holistic Score	1.00	71.60%	0.22	0.00	0.00%	0.00	0.40	28.40%	0.07
Composite Analytic Score	0.28	74.10%	0.04	0.00	0.00%	0.00	0.10	25.90%	0.01
<i>Scores on Analytic Traits</i>									
<i>Content</i>	0.39	65.70%	0.06	0.00	0.00%	0.00	0.20	34.30%	0.03
<i>Language Use</i>	0.31	59.30%	0.06	0.00	0.80%	0.01	0.21	39.90%	0.03
<i>Mechanics</i>	0.26	51.70%	0.05	0.00	0.00%	0.00	0.24	48.30%	0.03
<i>Organisation</i>	0.32	58.00%	0.06	0.00	0.00%	0.00	0.24	42.00%	0.03
<i>Vocabulary</i>	0.23	49.00%	0.05	0.00	0.00%	0.00	0.24	51.00%	0.03

Note: The composite analytic scores are calculated using Formula 8.1 to combine the ratings each essay received on the 5 analytic traits.

Appendix J Pearson Correlation (r) Between Human Scores and IEA Scores

Pearson correlation (r) between Human Analytic Scores and IEA Scores

Prompts	IEA/ Marker 1	IEA/ Marker 2	Marker 1/ Marker 2	IEA/ Human analytic score
Voting	0.66	0.71	0.80	0.72
Tobacco	0.72	0.82	0.77	0.82

Pearson correlation (r) between Human Holistic Scores and IEA Scores

Prompts	IEA/ Marker 1	IEA/ Marker 2	Marker 1/ Marker 2	IEA/ Human holistic score
Voting	0.60	0.59	0.76	0.63
Tobacco	0.70	0.78	0.79	0.78

Note: The following notes apply to the above two tables.

All correlations are significant at the 0.001 level (2-tailed). N=60 for holistic scores; N=120 for analytic scores

IEA/Marker 1: correlation between the IEA scores and the human scores calculated based on the first human ratings;

IEA/Marker 2: correlation between the IEA scores and the human scores calculated based on the second human ratings;

IEA/Human Score: correlation between IEA overall scores and the (adjudicated) final human scores

Appendix K Category Statistics for Rating Scales Used by Human Markers

		Voting				Tobacco			
Trait Name	Cat*	Cnt	Average Ability Measure	OUTFIT MNSQ	Threshold Measures	Cnt	Average Ability Measure	OUTFIT MNSQ	Threshold Measures
<i>Content</i>	0	10	-2.31	1.1	NONE	3	-0.58	2.2	NONE
	1	99	-0.8	1	-3.91	60	-0.44	0.8	-4.59
	2	93	1.39	0.9	0.36	109	2.36	0.8	0.54
	3	24	2.96	1	3.55	46	3.99	1	4.05
<i>Organisation</i>	0	5	-2.12	1	NONE	3	-3.25	0.7	NONE
	1	80	-0.55	0.8	-4.06	43	-0.26	0.8	-4.09
	2	109	1.68	0.8	0.26	111	2.38	0.9	0.16
	3	32	3.48	0.9	3.79	61	4.23	1	3.93
<i>Vocabulary</i>	0	0	*	*	*	1	-3.45	0.5	NONE
	1	70	-2.46	1	NONE	29	-0.18	0.7	-4.46
	2	122	-0.33	0.9	-1.97	114	2.88	0.9	0.08
	3	34	1.66	1	1.97	73	4.93	0.9	4.38
<i>Language Use</i>	0	0	*	*	*	2	-2.09	1	NONE
	1	95	-2.31	1	NONE	36	-0.21	0.9	-4.17
	2	88	-0.47	1.1	-1.26	110	2.55	0.8	0.12
	3	43	1.54	0.8	1.26	69	4.53	0.9	4.05
<i>Mechanics</i>	0	6	-1.18	1.1	NONE	2	-1.95	1.1	NONE
	1	42	0.05	1.3	-2.7	30	-0.03	1.1	-3.96
	2	99	1.67	1.8	-0.13	112	2.75	1.5	-0.11
	3	79	3.41	1.1	2.83	73	4.23	1.4	4.07

Cat*: Score Category; Cnt: Count

Appendix L Two Examples of the *Spelling* Score Anomalies

Seq# V2105 - Prompt: Voting

Response:

The right to vote is considered one of the greatest freedoms allowed a citizen of any country. So why is it that in some democratic countries, it is deemed compulsory?

Democratic countries - at least - the majority of them, are also relatively peaceful countries with high living standards and comfortable lifestyles. It is easy to become complacent in such environments. If people have no serious concerns about their society and economy, then it usually indicates that the country is being run reasonably well. It is easy for people not to feel any pressing need to make any changes. This is reflected in poor voter turnout at elections. Some theorists believe that everyone should vote, believing if they don't, the election results don't truly reflect the wishes of the people. All people should be involved in the running of the country by participating in the election process.

At the other end of the theoretical spectrum, there are those who believe that voting should be voluntary. If people truly are unhappy about their government, then they will turn out in large numbers to force the required changes.

The basic premise of democracy is "government of the people, by the people, for the people". It is also about freedom.

Should voting be compulsory? I believe yes, it should be. Complacency allows corruption to creep into the system, and could advers

Appendix L Two Examples of the *Spelling* Score Anomalies (Continued)

Seq# V287 – Prompt: Voting

Response:

I am agree with the notion of compulsory voting.

Because if it is not, it is easy to imagine that people become less interested in politics. If politicians do not listen to our opinion, then we do not listen to them.

Furthermore, politics can be controlled as politicians like. The state of country can easily be chaos and our lives will be harder.

Appendix M An Example of *Spelling* Scoring Anomalies

SEQ# T12 - Prompt: Tobacco

The use of tobacco which brings dangerous effects has been known for all this time. Yet, there are still some people who ignore this matter. Such decisions are up to the individual itself. The **government** does not have any rights to forbid tobacco to be banned as there will be drawbacks when the **government** does it. It is clearly proven in the tobacco producing countries. When tobacco is banned, the country will lose the income that it has been usually making from tobacco. In that way, it will also affect the farmers who rely their lives on planting and harvesting tobacco. They will be jobless and it can cause problems to the **government** as gouvernement is expected to take actions regarding the unemployment rate.

In that way, once again **government** cannot have a legitimate role to ban the tobacco in its country. However, there are still some things that can be taken in order to warn the smoker. **Government** requires all the tobacco packaging stamped with a sticker of showing what the consequences are when smoker smoke tobaccos, such as cancer, impotency, heart problem and many more. Government may also impose higher tax on tobacco products so the retail price will also be higher as well. It all depends on the **government** policy. For some countries in Asia like Indonesia or Singapore, tobaccos are sold in relatively cheap prices. While in Australia the price can be ten times higher as the ones in those Asian countries.

Appendix N An Example of Potential Anomaly Arising from the *Formal Requirement* Trait

SEQ# V276 - Prompt: Voting

Response:

In a democratic country a hard-working family man faces the choice between 30 days in prison or a £250 fine. In third world country men and women are beaten up by a gang of party supporters whilst the police look on and do nothing to prevent this. Why? In both situations the answer is the same: these people have, for reasons of personal conscience, chosen not to vote for a political party in a country's election. Is such treatment a demonstration of true democracy? Democracy is, theoretically, government of the people based on the people's choice expressed via the ballot box or other form of voting. If the right to individual choice on the question of government should be expressed, surely the right of individuals to exercise their own conscience by not voting should also be recognised. An objection may be raised to this on the basis that a low percentage of people voting can return a non-representative government in a country. This may be true, but responsible members of society will exercise their right to vote if they so wish, whereas it is unfortunately also true that if people are compelled by law to vote, those who feel no real interest in the political system may take 'the line of least resistance' or be unduly influenced by extremist parties rather than by seriously considering their use of their vote. The result of this would not necessarily benefit the country. On the other hand, should a person choose not to vote, they should be prepared to accept the government elected. If voting is made compulsory in a democratic country, then democracy could become something which limits personal freedom, rather than giving people the freedom to exercise their own conscience and express their views, which surely was the original purpose of democracy.

Appendix O Two Additional Examples of Potential Anomalies Arising from the *Formal Requirement Trait*

SEQ#T291 Prompt: Tobacco – Marked A in Figure 11.2

Response:

Yes, governments do have a legitimate role to legislate to protect citizens from the harmful effects of their own decisions to smoke. When the decision is left to the individual, particularly one who is under the influence of an addiction, they will continue to smoke regardless of the detrimental effect on their health or the health of others. For example, there has always been an option to have non smoking establishments such as restaurants and bars but until there was a threat of an actual government ban on smoking in public places, the majority of establishments continued to allow smoking on their premises. Economic considerations outweighed the potential health issues.

SEQ# V117 Prompt: Voting – Marked B in Figure 11.2

Response:

voting for democracy is something u have to do, it can changes the future of the enviroment. Democracy help the world get through a lot, it can set the right rules and help people to make thing right. It will change for the better for everyone in the planet, that is why we should support the nature of democracy. And should get invole in the democracy by voting for the right candidates to save you trouble for later, and you have the right to vote for anyone you can say whatever you feel right. The nature of democracy is something people should pay more tention on, just in case it goes the wrong way. And u should get to know all the candidate before you make the vote or else you may vote for the wrong candidate and it may affect some other people voting, dont make wrong decision. If you dont know you should ask advise from some one that know well in that aspect. If you have time learn more about the

democracy to get more experience to know what it like, it is good for you in later in the future or get involle with some people to learn more

Appendix P An Example of *Content Scoring Anomaly*

SEQ# V2119 – prompt Voting

The idea of compulsory voting is a poor one, forcing those who do not wish to or feel they don't have the knowledge on the issues to participate. Instead, the willing and educated should be the ones that decide the fates of nations. People who feel neither one way or the other should have the right to abstain from voting, and those who feel they are not educated in the issues enough should also have that privilege. *For example, if a person was told to vote "yes or no on Proposition 1" and was given no details as to what the proposition was pertaining to, they could feel backed into a corner. What if it were to turn out to be a law requiring that a person of their exact same size and build would be forced to give up their possessions? Sure, this example is hyperbolic, but it points out a flaw within the system that forces voting.* Additionally, what would be the consequence to a person who doesn't vote? Would it be jailtime? Or simply people would be tracked down and then forced to vote? That's not a comfortable path. The foundation of democracy is "one man, one vote", but while that sounds like a good idea in theory, the forced execution of such seems overbearing and totalitarian. Totalitarian democracy, not a common occurrence!

(Italics inserted by the present author)

Appendix Q Matrix of Frequency of Occurrences of the IEA Scores and Human Trait Scores Across the Four Traits

		IEA Trait Score			
Trait	Human Trait Score	0	1	2	Total No of Essays
<i>Development, Structure and Coherence</i>	0	21	40	1	62
	1	3	180	28	211
	2	0	54	64	118
	Total	24	274	93	391
<i>General Linguistic Range</i>	0	21	38	0	59
	1	13	181	17	211
	2	0	52	69	121
	Total	34	271	86	391
<i>Grammar Usage and Mechanics</i>	0	30	46	0	76
	1	7	177	19	203
	2	0	56	56	112
	Total	37	279	75	391
<i>Vocabulary Range</i>	0	26	28	0	54
	1	4	197	28	229
	2	0	43	65	108
	Total	30	268	93	391

Appendix R Ethics Approval from the Human Research Ethics Committee

COPY

University of Wollongong



INITIAL APPLICATION APPROVAL

In reply please quote: HE09/130
Further Enquiries Phone: 4221 4457

19 May 2009

Ms Lucy Lu
Data Collections
Planning and Innovation, Dept of Education and Training
Level 5, 35 Bridge St
Sydney NSW 2000

Dear Ms Lu

Thank you for your response dated 11 May 2009 to the HREC review of the application detailed below. I am pleased to advise that the application has been **approved**.

Ethics Number: HE09/130
Project Title: Explore validity evidence on automatic essay scoring system from the perspective of rating processes
Researchers: Professor Jim Tognolini, A/Prof Lori Lockyer, Dr Juho Looveer, Ms Lucy Lu
Approval Date: 19 May 2009
Expiry Date: 18 May 2010

The University of Wollongong/SESIAHS Humanities, Social Science and Behavioural HREC is constituted and functions in accordance with the NHMRC *National Statement on Ethical Conduct in Human Research*. The HREC has reviewed the research proposal for compliance with the *National Statement* and approval of this project is conditional upon your continuing compliance with this document. As evidence of continuing compliance, the Human Research Ethics Committee requires that researchers immediately report:

- proposed changes to the protocol including changes to investigators involved
- serious or unexpected adverse effects on participants
- unforeseen events that might affect continued ethical acceptability of the project.

You are also required to complete monitoring reports annually and at the end of your project. These reports are sent out approximately 6 weeks prior to the date your ethics approval expires. The reports must be completed, signed by the appropriate Head of School, and returned to the Research Services Office prior to the expiry date.

Yours sincerely

A/Professor Steven Roodenrys
Chair, Human Research Ethics Committee

cc Professor Jim Tognolini, Director Pearson Research and Assessment
A/Prof. L. Lockyer, Education

Appendix S Participant Information Sheet

University of Wollongong



PARTICIPATION INFORMATION SHEET FOR MARKERS

TITLE: A Validation Framework for Automated Essay Scoring Systems

PURPOSE OF THE RESEARCH

This is an invitation to participate in a project conducted by researchers at the University of Wollongong. The purpose of the study is to explore the validity of an Automated Essay Scoring system from the perspective of the rating process. The study will focus on collecting evidence on three key questions: 1) what is this system really assessing? 2) how does it assess these? and 3) what compromises are being made within the system in order to achieve high agreement rates with those of human markers.

INVESTIGATORS

Prof. Jim Tognolini	A/Prof. Lori Lockyer	Dr. Juho Looveer	Lucy Lu
Faculty of Education	Faculty of Education	EduMetrics	Faculty of Education
02-9467 6600	02 4221 5511	02 9653 2871	02 9561 8691
jim.tognolini@pearson.com	llockyer@uow.edu.au	juho.looveer@gmail.com	lucy.lu@det.nsw.edu.au

METHOD AND DEMANDS ON PARTICIPANTS

If you decide to take part, you will be asked to complete a background questionnaire on your qualifications, teaching and marking experience. You will also be asked to participate in group discussions concerning the appropriateness of the rating scales chosen by the researcher for the holistic and analytic marking, as well as participate in the marking moderation processes for consistent interpretation and use of these rating scales. You will then be asked to mark a total of between 160 and 180 essays using the two different marking methods: holistic and analytic.

POSSIBLE RISKS, INCONVENIENCES AND DISCOMFORTS

Apart from your time for the marking and participation in group discussions, we can foresee no risks for you. Your involvement in the study is voluntary and you may withdraw your participation from the study at any time and withdraw any data that you have provided to that point. Refusal to participate in the study will not affect your relationship with the University of Wollongong.

FUNDING AND BENEFITS OF THE RESEARCH

This study is not funded by any research grant. It is part of the doctoral work for the research student (Lucy Lu). Findings from the study will only be reported in the student's final thesis. Confidentiality is assured, and you will not be identified in any part of the research.

ETHICS REVIEW AND COMPLAINTS

This study has been reviewed by the Human Research Ethics Committee (Social Science, Humanities and Behavioural Science) of the University of Wollongong. If you have any concerns or complaints regarding the way this research has been conducted, you can contact the UoW Ethics Officer on (02) 4221 4457.

If you are willing to take part in this study, could you please complete the sheet on the next page and fax it to the number given. If you wish to talk to the researcher for more information, please call Lucy on 02 9561 8691.

Thank you for your time.

Lucy Lu.