Current Issues in Language Evaluation, Assessment and Testing

Current Issues in Language Evaluation, Assessment and Testing:

Research and Practice

Edited by

Christina Gitsaki and Christine Coombe

Cambridge
Scholars
Publishing

Current Issues in Language Evaluation, Assessment and Testing:
Research and Practice

Edited by Christina Gitsaki and Christine Coombe

This book first published 2016

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-8590-8
ISBN (13): 978-1-4438-8590-4

TABLE OF CONTENTS

List of Tables viii

List of Figures xii

List of Appendices xiv

List of Abbreviations xv

Preface xix
Christina Gitsaki and Christine Coombe

Issues in the Analysis and Modification of Assessment Tools and Tests

Chapter One 2
How Well do Cloze Items Work and Why?
James Dean Brown, Jonathan Trace, Gerriet Janssen, and Liudmila Kozhevnikova

Chapter Two 40
Estimating Absolute Proficiency Levels in Small-Scale Placement Tests with Predefined Item Difficulty Levels
Kazuo Amma

Chapter Three 63
Bilingual Language Assessment in Early Intervention: A Comparison of Single- versus Dual-Language Testing
Caroline A. Larson, Sarah Chabal, and Viorica Marian

Chapter Four 80
Frequency and Confidence in Language Learning Strategy Use by Greek Students of English
Penelope Kambakis-Vougiouklis and Persephone Mamoukari

Chapter Five 98
The Development of a Vocabulary Test to Assess the Breadth
of Knowledge of the Academic Word List
Lee-Yen Wang

Issues in the Creation of Assessment and Evaluation Tools

Chapter Six 118
Assessment for Learning; Assessment for Autonomy
Maria Giovanna Tassinari

Chapter Seven.. 137
Cultivating Learner Autonomy through the Use of English Learning
Portfolios: A Longitudinal Study
Beilei Wang

Chapter Eight.. 158
Sharing Assessment Power to Promote Learning and Autonomy
Carol J. Everhard

Chapter Nine.. 177
Developing a Tool for Assessing English Language Teacher Readiness
in the Middle East Context
Sadiq Midraj, Jessica Midraj, Christina Gitsaki, and Christine Coombe

Chapter Ten 201
Foreign Language Teachers' Proficiency: The Implementation
of the EPPLE Examination in Brazil
Douglas Altamiro Consolo and Vera Lúcia Teixeira da Silva

Issues in Language Assessment and Evaluation

Chapter Eleven 222
Vocabulary Size Assessment as a Predictor of Plagiarism
Marina Dodigovic, Jacob Mlynarski, and Rining Wei

Chapter Twelve 236
What is the Impact of Language Learning Strategies on Tertiary Students'
Academic Writing Skills? A Case Study in Fiji
Zakia Ali Chand

Chapter Thirteen 253
Speaking Practice in Private Classes for the TOEFL iBT Test:
Student Perceptions
Renata Mendes Simões

Chapter Fourteen 272
Assessing the Level of Grammar Proficiency of EFL and ESL Freshman
Students: A Case Study in the Philippines
Selwyn A. Cruz and Romulo P. Villanueva Jr.

Chapter Fifteen 287
Methodology in Washback Studies
Gladys Quevedo-Camargo and Matilde Virginia Ricardi Scaramucci

Contributors 305

Index 314

LIST OF TABLES

Table 1-1: Descriptive statistics for 50 cloze passages and reliability–Japan sample (adapted and expanded from Brown, 1998).
Table 1-2: Descriptive statistics for 50 cloze passages and reliability–Russia sample (adapted and expanded from Brown, 1998).
Table 1-3: Frequencies and percentages of tests that functioned well or poorly in terms of internal consistency reliability.
Table 1-4: Frequencies and percentages of items that functioned well in terms of Item Facility (*IF*).
Table 1-5: Frequencies and percentages of items that functioned well in terms of Item Discrimination (*ID*).
Table 1-6: Summary fit statistics for analyses of 1,496 items by country of origin.
Table 1-7: Vertical rulers for test taker ability, test version difficulty, and test item difficulty for Japan.
Table 1-8: Vertical rulers for test taker ability, test version difficulty, and test item difficulty for Russia.
Table 1-9: Frequencies and percentages of items that functioned well in terms of CTT ID and Rasch Fit.
Table 1-10: Item misfit by linguistic background for different parts of speech for 1,496 items.
Table 1-11: Item misfit for word type (content & function) in the Japan and Russia data for 1,496 items.
Table 1-12: Item misfit for word origin (Latinate & Germanic) in the Russia and Japan data for 1,496 Items.
Table 1-13: Item misfit as a function of word frequency (in the *Brown Corpus* found in Francis & Kučera, 1979) in the Japan and Russia data for 1,496 items.
Table 2-1: Descriptors of overall reading comprehension (Source: Council of Europe, 2001, p. 69).
Table 2-2: Item difficulty levels and required proficiency in the placement test.
Table 2-3: Description of imaginary test items.

Table 2-4: Summary report of a candidate (Adapted from Linacre 1987, p. 25).
Table 2-5: Simulation 1.
Table 2-6: Simulation 2.
Table 2-7: Item difficulty and responses (arranged in the order of difficulty levels).
Table 2-8: Statistics of the candidates' proficiency levels and confidence intervals.
Table 2-9: Item difficulty and responses by exact scoring.
Table 2-10: Adjusted responses for Candidate A (part).
Table 2-11: Adjusted responses for Candidate B (part).
Table 2-12: Comparison of estimated proficiency levels and confidence intervals.
Table 3-1: Demographic information for study participants.
Table 3-2: Receptive language ability as indexed by *The Rossetti* (1990) *Language Comprehension* subtest.
Table 3-3: Expressive language ability as indexed by *The Rossetti* (1990) *Language Expression* subtest.
Table 5-1: Participant information.
Table 5-2: Distribution of the sample words across the AWL Levels.
Table 5-3: Omitted AWL words with Level information.
Table 5-4: Pseudowords with Level information and the original AWL words.
Table 5-5: Descriptive statistics (freshmen = 50, seniors = 39).
Table 5-6: Comparison between freshmen and seniors in their knowledge of non-omitted/omitted academic words using ANOVA.
Table 5-7: Reliability analysis.
Table 5-8: Descriptive statistics for the 52 and 38 omitted items.
Table 5-9: Omitted academic words and their COCA ranks and frequencies.
Table 6-1: Reasons for choice of components (more than one answer possible).
Table 6-2: Effects of the self-assessment on participants (more than one answer possible).
Table 6-3: Comparison of studies on learner autonomy.
Table 7-1: Student participants in the study.
Table 7-2: Demographic data of case study participants.
Table 7-3: Phases of treatment in the junior high school.
Table 7-4: Independent samples test of LA and its subscales.

Table 7-5: Planning and reflecting.
Table 7-6: Alice's monthly learning plan (A-P1-9).
Table 7-7: David's weekly plan.
Table 8-1: AARP assessment overview (Writing 1 and Writing 2).
Table 8-2: AARP assessment overview (Oral Presentation).
Table 8-3: AARP mean scores for Writing 2, for Groups C and D, with ANOVA results for self-assessment variations.
Table 8-4: Paired *t*-test results for Writing 2 self-assessment Variations (AY2).
Table 8-5: Responses to questions about peer- and self-assessment.
Table 8-6: Modified extract from the AARP model showing the relationship between assessment and degrees of autonomy (Based on Everhard 2014, 2015a; Harris & Bell, 1990).
Table 9-1: Example performance indicator and achievement indicators (Source: TESOL 2010, p. 69).
Table 9-2: Criteria for the external review of the indicators and the MCIs.
Table 10-1: Marks in the FCE and in the TEPOLI (Source: Baffi-Bonvino, 2007, pp. 276-277).
Table 10-2: Grammatical accuracy as measured in TEPOLI (Source: Borges-Almeida & Consolo, 2010).
Table 10-3: Grammatical complexity as measured in TEPOLI (Source: Borges-Almeida & Consolo, 2010).
Table 11-1: Correlation coefficient (effect size) values for lexical errors.
Table 12-1: Correlations between strategy use, gender and ethnicity.
Table 12-2: Results of the error analysis.
Table 12-3: Correlation between strategy use and language proficiency.
Table 13-1: Classification of production variables in oral communication tasks (Adapted from Ellis, 2003, p. 117).
Table 13-2: Data collection procedures in each phase of the study.
Table 13-3: Student perceptions of their language skills before the course.
Table 13-4: Change in students' perceptions ($N = 17$).
Table 13-5: Average perceived overall improvement in language skills.
Table 13-6: Students' final TOEFL iBT score.
Table 13-7: Comparative table of the TOEFL iBT scores.
Table 14-1: A scale for measuring grammar proficiency.

Table 14-2: Mean and standard deviation of participants' total score based on nationality.

Table 14-3: Means and standard deviations for the ESL participants.

Table 14-4: Means and standard deviations for the EFL participants.

Table 15-1: Studies distributed by continent.

Table 15-2: Number of instruments used in the studies.

LIST OF FIGURES

Figure 2-1: Sample representation of logistic regression analysis of a test item (Adapted from Amma, 2001).

Figure 2-2: Sample representation of logistic regression analysis of a candidate (Adapted from Amma, 1990).

Figure 2-3: Logistic regression analysis of Simulation 1.

Figure 2-4: Logistic regression analysis of Simulation 2.

Figure 2-5: Logistic fit of Candidate A with confidence interval. Bottom layer = ‘fail’.

Figure 2-6: Logistic fit of Candidate B with confidence interval. Bottom layer = ‘fail’.

Figure 3-1: Participants’ receptive language assessment results using The Rossetti (1990) Language Comprehension subtest. Error bars represent standard errors and asterisks indicate significant differences at $p < .05$.

Figure 3-2: Participants’ expressive language assessment results using The Rossetti (1990) Language Expression subtest. Error bars represent standard errors and asterisks indicate significant differences at $p < .05$.

Figure 3-3: Percent language delay in primary-language-only assessment and in dual-language assessment using The Rossetti (1990) Language Comprehension and Language Expression subtests. Error bars represent standard errors and asterisks indicate significant differences at $p < .05$.

Figure 4-1: An example from the SILL questionnaire employing the [01] bar for frequency and confidence.

Figure 6-1: The dynamic model of learner autonomy (Source: Tassinari, 2010, p. 203).

Figure 7-1: The process of the ELP practice.

Figure 9-1: Project stages.

Figure 10-1: The EPPLE examination, Oral Test, Part 1.

Figure 10-2: The EPPLE examination, Oral Test, Part 3.

Figure 13-1: Difficulties with language at the course ($N = 17$).

Figure 13-2: Student perceptions of their language ability at the end of the course ($N = 17$).

Figure 13-3: Student self-perceived learning ($N = 17$).

Figure 13-4: Students' perceptions of their language skills before the course (*N* = 17).
Figure 13-5: Students' perceptions of their language skills after the course (*N* = 17).
Figure 13-6: Student improvement in Task 1 and Task 2 (*N* = 17).
Figure 15-1: Studies per year of publication.
Figure 15-2: Data-collection instruments in the reviewed studies.

LIST OF APPENDICES

Appendix 1-A: List of participating universities.
Appendix 1-B: Example cloze test (Adapted from Brown, 1989).
Appendix 1-C: Actual word frequencies.
Appendix 1-D: FACETS input file for 50 cloze tests with 10 anchor items.

LIST OF ABBREVIATIONS

A	Self-Assessor (in Tukey-Kramer formulae)
AARP	Assessment for Autonomy Research Project
ANOVA	One-way Analysis of Variance
AS-unit	Analysis of Speech Unit
ASC	Ascendente (rising intonation)
ASHA	American Speech-Language-Hearing Association
AUTh	Aristotle University of Thessaloniki
AVL	Academic Vocabulary List
AWL	Academic Word List
AY	Academic Year
B	Peer-Assessor (in Tukey-Kramer formulae)
BNC	British National Corpus
C	Teacher-Assessor (in Tukey-Kramer formulae)
CAEP	Council for the Accreditation of Educator Preparation
CAPES	Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior - Brazilian Agency for the Development of Graduate Studies
CEEC	College Entrance and Examination Center
CEFR	Common European Framework of Reference
CEPA	Common Educational Proficiency Assessment
CG	Control Group
CI	Confidence Interval
CILL	Centre for Independent Language Learning
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COCA	Corpus of Contemporary American English
Comm Arts 1	Communication Arts and Skills 1
Corr. Coeff.	Correlation Coefficient
CRAPEL	Centre de Recherches et d'Applications Pédagogiques en Langues
CTT	Classical Test Theory
EAL	English as an additional language
EFL	English as a Foreign Language
EG	Experimental Group

ELF	English as a Lingua Franca
ELL	English Language Learners/Learning
ELP	English Learning Portfolio
ELT	English Language Teaching
EMI	English Medium Instruction
ENAPLE-CCC	Ensino e Aprendizagem de Língua Estrangeira: Crenças, construtos e competências
Eng AN	Introduction to Language Arts
EPPLE	Exame de Proficiência para Professores de Língua Estrangeira
ERWL	English Reference Word List
ESL	English as a Second Language
ESOL	English for Speakers of Other Languages
ESP	English for Specific Purposes
ETS	Educational Testing Service
F	Female
FATEC	Faculdades de Tecnologia
FCE	First Certificate in English
FEU	Far Eastern University
FL	Foreign Language
FUB	Freie Universität Berlin
GSEAT	General Scholastic English Ability Test
GSL	General Service List
HE	Higher Education
ID	Item Discrimination
IELTS	International English Language Testing System
IF	Item Facility
INCOMP	Incomprehensible
IRT	Item Response Theory
L2	Second Language
LA	Learner Autonomy
LLS	Language Learning Strategies
LPFLT	Language Proficiency of Foreign Language Teachers
M	Male
Max	Maximum
MCI	Multiple-Choice Item
MENA	Middle East and North Africa
MFRM	Many Facet Rasch Measurement
Min	Minimum
MOE	Ministry of Education

N	Number (of participants)
n.s.	Non-significant differences in One-way ANOVA
NAPO	National Admissions and Placement Office
NBPTS	National Board for Professional Teaching Standards
NCATE	National Council for Accreditation of Teacher Education
NECS	National English Curriculum Standards
NNS	Non-Native Speakers
NRT	Norm-Referenced Test
NS	Native Speaker
OPI	Oral Proficiency Interview
OPT	Oxford Placement Test
P-A	Peer-Assessment
P-T	Peer-Assessment and Teacher Assessment
r	Reliability
S-A	Self-Assessment
S-T	Self-Assessment and Teacher Assessment
SAT	Scholastic Assessment Test
SD	Standard Deviation
SEM	Standard Error of Measurement
Sig.	Significance
SILL	Strategy Inventory for Language Learning
SOE	School of English
SPSS	Statistical Package for the Social Sciences
SVA	Subject Verb Agreement
T-A	Teacher Assessment
T-TRI	TESOL Teacher Readiness Inventory
TECOLI	Teste de Compreensão Oral em Língua Italiana
TEPOLI	Teste de Proficiência Oral em Língua Inglesa
TESOL	Teaching English to Speakers of Other Languages
The Rossetti	*The Rossetti Infant-Toddler Language Scales*
TOEFL	Test of English as a Foreign Language
TOEFL-iBT	Test of English as a Foreign Language–Internet-based Test
TOEIC	Test of English for International Communication
UEM	Universidade Estadual de Maringá
UERJ	Universidade Estadual do Rio de Janeiro
UnB	Universidade de Brasília
UNESP	Universidade Estadual Paulista
UNIP	Universidade Paulista

UNISEB/Estacio	Centro Universitário do Sistema Educacional Brasileiro
VST	Vocabulary Size Test
Y/N	Yes/No

PREFACE

Language assessment, whether formative or summative, plays an important role in second language learners' educational experience and learning outcomes. Whether assessment is used for student initial screening, placement, or progression in a language course, it always involves gathering, interpreting and evaluating evidence of learning. Such information collected through the different assessment and evaluation tools allows educators to identify student needs and plan a course of action to address these needs, provides feedback about the effectiveness of teaching practice, guides instruction and curriculum design, and provides accountability for the system.

For language educators, assessment is perhaps one of the most difficult and demanding tasks they have to perform given that designing valid and reliable assessment tools requires specialized skills, while decisions about assessment, especially 'high-stakes' exams, can have a lasting impact on students' progress and life.

Furthermore, in order for assessment to be useful, it must align itself with the mandated standards and academic expectations of the specific context where it occurs. Since no single type of assessment can provide all the information that is necessary to gauge students' progress and language proficiency levels, educators need to incorporate a variety of assessment techniques into their practice and be aware of approaches and methods that can help provide valid and reliable evidence of student learning.

The edited volume presented here, *Current Issues in Language Evaluation, Assessment and Testing: Research and Practice*, is a collection of papers that address relevant issues in language assessment from a variety of contexts and perspectives. The book is divided into three major sections. The first section addresses *Issues in the Analysis and Modification of Assessment Tools and Tests.* In Chapter One, JD Brown, Jonathan Trace, Gerriet Janssen, and Liudmila Kozhevnikova discuss a comparative study of analyzing cloze tests using Classical Theory Test (CTT) item analysis and multifaceted Rasch analysis. Through the examination and analysis of almost 7,500 cloze tests from university students studying English as a foreign language (EFL) in Japan and Russia, the Rasch analyses proved to be more appropriate than CCT. In Chapter Two, Kazuo Amma proposes using a logistic regression analysis with predefined item difficulty levels in order to properly assess a student's true proficiency level and the confidence interval, avoiding the pitfalls that may occur from estimating proficiency based on the total test score. In Chapter Three, Caroline Larson, Sarah Chabal, and Viorica Marian examine the use of *The Rossetti Infant-Toddler Language Scale* with Spanish-English speaking children. Their findings suggest that when the instrument is used in the children's primary language only, their language skills are underestimated and their language delay is overestimated leading to inappropriate Early Intervention referrals. In order to maximize the efficacy, reliability and validity of *The Rossetti*, the researchers recommend administering it in both the primary and the secondary languages of the child. In Chapter Four, Penelope Kambakis-Vougiouklis and Persephone Mamoukari present the results of their investigation of language learning strategy use of Greek EFL students. Their study is a pilot of a modified version of the Strategies Inventory for Language Learning (SILL) using a bar instead of a Likert scale, measuring frequency of strategy use as well as student confidence in the effectiveness of the different strategies, and administering the instrument orally rather than in writing. Their modifications of the SILL allowed them to detect discrepancies in student understanding of language learning strategies and their effectiveness which would not have been otherwise evident, as well as problematic items within the SILL that need further modification prior to future administrations of the instrument. In the last chapter for this section, Chapter Five, Lee-Yen Wang describes the use of a Yes/No test to measure University EFL students' vocabulary acquisition of academic words that were left out of the Ministry-defined vocabulary list for schools in Taiwan. The analyses of the data highlighted the limitations of having a centrally controlled national wordlist.

The next part of the volume, addresses *Issues in the Creation of Assessment and Evaluation Tools.* In Chapter Six, Maria Giovanna Tassinari discusses the creation of a dynamic model for assessing language learner autonomy and provides evidence of the use of the model with foreign language learners in a German University context. Her findings indicate that the model initiated and maintained pedagogic dialogue between the students and their teachers, raised students' awareness of the different dimensions of learner autonomy, and enhanced their reflexive learning. In Chapter Seven, Beilei Wang describes the use of a three-dimensional learner autonomy scale with junior high school EFL students in China. Findings revealed that the use of the English Language Portfolios was conducive to helping students gain learner autonomy. Learner autonomy was also the primary interest of Carol Everhard's study in Chapter Eight. Greek EFL students in a university context participated in self- and peer-assessment activities of their oral and writing skills over the course of the 5-year study. Findings suggest that the use of such formative assessment techniques activated students' criterial thinking and metacognitive awareness of their learning process. The next two chapters in this section describe the creation of assessment and evaluation tools for measuring teacher proficiency. In Chapter Nine, Sadiq Midraj, Jessica Midraj, Christina Gitsaki, and Christine Coombe describe the process of compiling a contextually relevant resource for independent learning and self-assessment in order to strengthen EFL teachers' content knowledge, pedagogical knowledge, and professional dispositions. The resource was created for teachers in the Gulf Region using internationally accepted professional standards in teaching English to speakers of other languages (TESOL). In Chapter Ten, Douglas Altamiro Consolo and Vera Lúcia Teixeira da Silva discuss a meta-analysis of a string of research studies conducted within the framework of designing assessment tools for evaluating foreign language teachers' proficiency in the foreign language they teach. Their meta-analysis is motivated by the need to revise and improve the EPPLE examination of foreign language teachers in Brazil.

The third and final section of the volume, *Issues in Language Assessment and Evaluation*, comprises studies that implemented different instruments to measure learner proficiency in different language skills. In Chapter Eleven, Marina Dodigovic, Jacob Mlynarski, and Rining Wei describe how they used instruments such as *Grammarly* and the Vocabulary Size Test (VST) to investigate possible correlations between academic plagiarism and vocabulary knowledge in the academic writing of University EFL students in China. Through their investigation poor vocabulary command emerged as a major cause of plagiarism. In Chapter Twelve, Zakia Ali Chand investigated whether there is a correlation between language strategy use and academic writing proficiency, using the SILL. The study involved University ESL students in Fiji, and the preliminary results showed that students were moderate users of language learning strategies and their academic writing was not influenced by their use of language learning strategies indicating the need for more strategy training in the language classroom. In Chapter Thirteen, Renata Mendes Simões addresses an area of language teaching that has not received much attention in research, that of one-to-one tutorials focusing on candidate preparation for high stakes exams. The study involved Brazilian students preparing for the Test of English as a Foreign Language (TOEFL). The process of assessing students' progress over the course of several weeks is described and enhanced by the use of qualitative and quantitative data in order to provide a deeper understanding of the effectiveness of English for Special Purposes (ESP) courses for exam preparation. In Chapter Fourteen, Selwyn Cruz and Romulo Villanueva describe the development and administration of a grammar proficiency test in order to investigate issues in the grammatical proficiency of Korean and Filipino students studying in an English Medium University in the Philippines. The final paper in the volume, Chapter Fifteen, addresses issues of washback from language assessment. Gladys Quevedo-Camargo and Matilde Virginia Ricardi Scaramucci discuss the results of a meta-analysis of research studies on washback from around the globe and their respective methodologies and provide an account of the diverse instruments used to investigate washback.

All fifteen papers included in this volume underwent a rigorous selection process through a double-blind peer review process that involved a number of notable academics. The papers underwent further review and editing before being published in this book. Below is the list of academics who were involved in the double blind review process:

Thomaï Alexiou	Aristotle University, Greece
Ramin Akbari	Modares Tarbiat, Iran
Deena Boraie	American University of Cairo, Egypt
Helene Demirci	Higher Colleges of Technology, UAE
Aymen Elsheikh	New York Institute of Technology, UAE
Atta Gebril	American University of Cairo, Egypt
Melanie Gobert	Higher Colleges of Technology, UAE
Tony Green	University of Bedfordshire, UK
Sahbi Hidri	Sultan Qaboos University, Oman
Elisabeth Jones	Zayed University, UAE
Mary Lou McCloskey	EDUCO, USA
Josephine O'Brien	Zayed University, UAE
Sufian Abu Rmaileh	UAE University, UAE

The volume presents research studies conducted in a variety of contexts (from early childhood to University and post-graduate studies) from around the world covering an equally diverse range of issues in language assessment and evaluation. It is hoped that it will be of use to both new and seasoned researchers in the field of Applied Linguistics and TESOL as well as teacher educators, language teachers, curriculum and assessment designers.

Christina Gitsaki and Christine Coombe

ISSUES IN THE ANALYSIS AND MODIFICATION OF ASSESSMENT TOOLS AND TESTS

CHAPTER ONE

HOW WELL DO CLOZE ITEMS WORK AND WHY?

JAMES DEAN BROWN, JONATHAN TRACE, GERRIET JANSSEN, AND LIUDMILA KOZHEVNIKOVA

Abstract

This study examined item-level data from fifty 30-item cloze tests that were randomly administered to university-level examinees from Japan (*N* = 2,298) and Russia (*N* = 5,170). A single 10-item anchor cloze test was also administered to all students. The analyses investigated differences between the two nationalities in terms of both classical test theory (CTT) item analysis and multifaceted Rasch analysis (the latter allowed us to estimate test-taker ability and item difficulty measures and fit statistics simultaneously across 50 cloze tests separately and combined for the two nationalities). The results indicated that considerably larger proportions of items functioned well in the Rasch item analyses than in the traditional CTT item analysis. Rasch analyses also turned out to be more appropriate for our cloze test analysis and revision purposes than did traditional CTT item analyses. Linguistic analyses of items that fit the Rasch model revealed that blanks representing certain categories of words (i.e., function words rather than content words, and Germanic-origin words rather than Latin-origin words), and to a greater extent relatively high frequency words were more likely to work well for norm-referenced test (NRT) purposes. In addition, this study found that different items were functioning well for the two nationalities.

Introduction

Taylor (1953) first proposed the use of cloze tests for evaluating the readability of reading materials in US elementary schools. In the 60s and 70s, a number of studies appeared on the usefulness of cloze for English as a second language (ESL) proficiency or placement testing (see Alderson, 1979 for a summary of this early ESL research). Since then, as Brown (2013) noted, this research on using cloze in ESL proficiency or placement testing has continued, but has been inconsistent at best with reported reliability estimates ranging from .31 to .96 and criterion-related validity coefficients ranging from .43 to .91.

While the literature has focused predominantly on fixed interval (i.e., every *n*th word) deletion cloze tests, other bases have been used for developing cloze tests. For example, *rational deletion* cloze was developed by selecting blanks based on word classes (cf., Bachman, 1982, 1985; Markham, 1985). *Tailored cloze* involved using classical test theory (CTT) item analysis techniques to select items and thereby create cloze tests tailored to a particular group of students (cf. Brown, 1988, 1989, 2013; Brown, Yamashiro, & Ogane, 1999, 2001; Revard, 1990).

For the most part, cloze studies have been based on CTT. However, Item Response Theory (IRT), including Rasch analysis, has been applied to cloze in a few cases. Baker (1987) used Rasch analysis to examine a dichotomously scored cloze test and found that "observed and expected item characteristic curves show reasonable conformity, though with some instances of serious misfit…no evidence for departure from unidimensionality is found for the cloze data…" (p. iv). Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) found that IRT provided stable estimates for cloze in their study of the degree to which groups of cloze items related to different subparts of the overall Test of English as a Foreign Language (TOEFL). Hoshino and Nakagawa (2008) used IRT in developing a multiple-choice cloze test authoring system. Lee-Ellis (2009) used Rasch analysis in developing and validating a Korean C-test. However, Rasch analysis has not been used to study the effectiveness of individual items.

The Study

Certainly, no work has investigated the degree to which cloze items function well when analyzed using both CTT and IRT frameworks, and little research has examined the functioning of cloze items in terms of their linguistic characteristics. To address these issues and others, the following research questions were posed, all focusing on the individual items involved in 50 cloze tests that were administered to university-level examinees from two linguistically different backgrounds:

1. How do the CTT descriptive statistics, reliability, and item analyses differ for test-taker groups from different linguistic backgrounds?
2. How do Rasch item difficulty measures differ for the test-takers from different linguistic backgrounds?
3. How do the proportions of functioning cloze items differ between the CTT and IRT analyses, when based on the test-taker responses from different linguistic backgrounds?
4. In what ways do Rasch item fit patterns differ in terms of factors such as linguistic background and four cloze item linguistic features: parts of speech, word type, word origin, and word frequency?
5. Which linguistic characteristics will increase the probability of cloze items functioning well during piloting?

Participants

A total of 7,468 English as a foreign language (EFL) students participated in this study: 2,298 of these EFL students were studying at 18 different universities in Japan as part of their normal classroom activities; the remaining 5,170 EFL students were studying at 38 universities in Russia (see Appendix 1-A for a list of the participating universities in both countries). In Japan, about 38.3% of the participants were women, and 61.7% of the participants were men; in Russia, 71.7% of the participants were women, and 28.0% were men, with the remaining 0.3% giving no response. The participants in Japan were between 18-24 years old, while in Russia they were between 14-46 years old. The data from Japan were collected as part of Brown (1993 & 1998); the data from Russia were collected in 2012-2013 and served as the basis of Brown, Janssen, Trace, and Kozhevnikova (2013). Though these samples were convenience samples (i.e., not randomly selected), they were relatively large, which is important as this sample size permits robust analyses of these cloze data.

It is critically important to stress that in this study we are interested in how linguistic background affects different analyses; we do not make any claims for the generalizability of these results to the EFL populations of all undergraduate students in university-level institutions in these countries. In fact, we want to stress that the samples from Japan and Russia cannot be said to be comparable given the sampling procedures, the very different proportions of university seats per million people available in the two countries, the proportions of young people who go to university, and so forth. Thus, any interpretations of these data to indicate that the English proficiency of students in either country is higher than in the other country are unwarranted and indefensible.

Measures

The 50 cloze tests used in this study were first created and used for Brown (1993). The 50 passages were randomly selected from among the adult-level books at a public library in Florida. Passages were chosen from each book by randomly selecting a page then working backwards for a reasonable starting place. Passages were between 366-478 words long with a mean of 412.1 words. Each passage contained 30 items, and the deletion pattern was every 12th word, which created a fairly high degree of item independence relative to the more typical 7th-word deletion pattern. The first and last sentences of all passages were left intact to provide context. Appendix 1-B shows the layout of the directions, example items, and answer key.

A 10-item cloze passage was also administered to all participants to act as anchor items (i.e., items that provide a common metric for making comparisons across tests and examinee samples). This anchor-item cloze was first created in a study by Brown (1989), wherein it was found that these 10 items were functioning effectively.

To check the degree to which the English in the cloze passages was representative of typical written English, the lexical frequencies for all 50 passages combined were calculated (see Appendix 1-C) and compared to the frequencies reported for the same words in the well-known *Brown Corpus* (Francis & Kučera, 1979, 1982; Kučera & Francis, 1967). We felt justified in comparing the 50 passages to this particular corpus for two reasons. First, following Stubbs (2004), though the *Brown Corpus* is relatively small, it is

> “still useful because of their careful design ... one million words of written American English, sampled from texts published in 1961: both informative prose, from different text types (e.g., press and academic writing), and different topics (e.g., religion and hobbies); and imaginative prose (e.g., detective fiction and romance).” (p. 111)

Then, too, we found that the logarithmically transformed word frequencies of the cloze test items (to normalize the Zipfian nature of vocabulary distributions) and the logarithmically transformed frequencies of these same words in the *Brown Corpus* correlated strongly at .93

(Brown, 1998). Thus, we felt reasonably certain that these passages and cloze items were representative of the written English language, or at a minimum the genres of English found in US public library books.

Procedures

The 50 cloze tests were distributed to intact classes by teachers such that every student had an equal chance of receiving each of the 50 cloze test passages. In Japan, 42-50 participants completed each cloze test, with a mean of 46.0 participants completing each passage. In Russia, 90-122 completed each cloze test (Mean = 103.4). All examinees in both countries completed the 10-item anchor cloze. Twenty-five minutes were allowed for completing the tests. Exact-answer scoring was used (i.e., only the word found in the original text was counted as correct). This was done for two reasons: (a) we wanted each item to be interpretable as fillable by a single lexical item for analysis purposes; and (b) with the hundreds of items and thousands of examinees in this study, using an acceptable-answer scoring or any other of the available scoring schemes would clearly have been beyond our resources.

Analyses

Initially, CTT statistics were used to analyze the cloze test data. These statistics included: the mean, standard deviation, minimum and maximum scores, reliability, item facility, and item discrimination. Rasch analyses were also used in this study to calculate item difficulty measures and to identify misfitting test items. We used FACETS (Linacre, 2014a) analysis rather than WINSTEPS because the former allowed us to easily analyze our nested design (i.e., multiple tests administered to different groups of examinees). Or as Linacre put it, “Use Winsteps if the data can be formatted as a rectangle, e.g., persons-items … Use Facets when Winsteps won’t do the job” (Linacre, 2014b, np).

We performed the analyses in several steps. Initially, we needed to determine anchor values through a separate FACETS analysis of only the 10 anchor items that were administered across all groups of participants. Then, we created a FACETS input file to link our 50 cloze tests by using our 10 anchor items (see Appendix 1-D for a description of the actual code that was used). There were three facets in this analysis: test-takers, test version, and test items. By using the FACETS program, we were able to combine the 50 different cloze procedures for both nationalities into a single analysis using anchor items, and put all of the items onto the same true interval scale for ease of comparison (see e.g., Bond & Fox, 2007, pp. 75-90). Four of the total 1,500 items had blanks that were either missing or made no sense, thus the total number of valid cloze items was 1,496.

Appendix 1-D also shows how we coded the data for the analysis. In order to analyze separate tests in a single analysis using a common set of anchor items, each examinee required two lines of response data. The first line corresponds to the set of items for the particular cloze procedure, set up by examinee ID, test version, the range of applicable items (e.g., 101-130 for items 1-30 on Test 1), followed by the observed response for each item. An additional line was also needed for examinee performance on anchor items, with the same coding format as above except for a common range of items for all examinees (31-40). The series of commas within the data indicates items that were removed as explained above. Using the same setup, we were able to run the program separately for the samples in Russia, Japanese, and Combined (i.e., with the two samples analyzed together as one).

Results

Classical Test Theory

Descriptive Statistics. As most previous item analyses of cloze tests have been based on CTT, we began our analysis by focusing on the CTT characteristics of our cloze tests and their items. Tables 1-1 and 1-2 show the descriptive statistics and internal consistency reliability estimates for our 50 cloze tests in test number order for each nationality. In general, the means are low for the 30-item cloze tests, indicating that the items (scored for exact answers) were quite difficult for the students. Tables 1-1 and 1-2 indicate that the Russia sample generally produced higher means and standard deviations than the Japan sample.

Reliability. Tables 1-1 and 1-2 also show how the reliability estimates of the various cloze passages were for the two nationalities. These cloze tests functioned somewhat less reliably with the Japan sample (ranging from .17 to .87) than with the Russia sample (ranging from .65 to .92). This pattern could be a consequence of the greater variation and perhaps the larger sample sizes in Russia. A synthesis of the cloze passages’ reliability estimates is shown in Table 1-3.

Table 1-1: Descriptive statistics for 50 cloze passages and reliability – Japan sample (adapted and expanded from Brown, 1998).

Japan Test	Mean	*SD*	Min	Max	*N*	*r*
1	5.23	3.16	0	15	48	0.71
2	4.21	3.42	0	13	47	0.86
3	2.02	2.13	0	10	48	0.74
4	7.54	3.87	2	16	46	0.80
5	3.98	2.79	0	13	47	0.73
6	5.11	3.23	0	14	47	0.80
7	6.14	3.41	0	16	43	0.83
8	3.16	2.27	0	8	45	0.46
9	2.85	2.46	0	11	46	0.77
10	2.54	2.31	0	8	46	0.83
11	5.94	3.36	0	16	46	0.74
12	8.98	3.97	0	21	47	0.79
13	2.87	1.71	0	8	46	0.50
14	3.23	2.50	0	9	47	0.68
15	9.18	3.42	4	18	49	0.68
16	1.36	1.41	0	6	48	0.65
17	1.38	1.25	0	5	46	0.35
18	1.02	1.09	0	3	50	0.50
19	4.76	2.88	0	10	50	0.70
20	4.38	3.24	0	15	47	0.86
21	9.92	4.44	0	19	48	0.84
22	3.70	2.86	0	11	47	0.84
23	3.64	2.40	0	11	43	0.65
24	2.96	2.26	0	9	47	0.44
25	5.36	2.74	0	12	46	0.63
26	2.68	1.56	0	5	47	0.17
27	2.34	2.72	0	13	47	0.87
28	2.58	2.17	0	8	43	0.57
29	2.32	1.77	0	7	44	0.64
30	9.56	3.28	3	16	48	0.72
31	3.78	3.08	0	15	46	0.83
32	3.83	2.53	0	9	42	0.77
33	2.14	1.87	0	6	44	0.63
34	5.87	2.92	0	13	45	0.82
35	6.63	3.66	0	17	45	0.72
36	5.00	2.05	0	9	46	0.51