Title
Modeling Speaker Proficiency, Comprehensibility, and Perceived Competence in a Language Use Domain

Author
Schmidgall, Jonathan Edgar

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Modeling Speaker Proficiency, Comprehensibility, and Perceived Competence in a Language Use Domain

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Applied Linguistics

by

Jonathan Edgar Schmidgall

2013

ABSTRACT OF THE DISSERTATION

Modeling Speaker Proficiency, Comprehensibility,
and Perceived Competence in a Language Use Domain

by

Jonathan Edgar Schmidgall
Doctor of Philosophy in Applied Linguistics
University of California, Los Angeles, 2013
Professor Lyle F. Bachman, Chair

Research suggests that listener perceptions of a speaker's oral language use, or a speaker's *comprehensibility*, may be influenced by a variety of speaker-, listener-, and context-related factors. Primary speaker factors include aspects of the speaker's proficiency in the target language such as pronunciation and grammatical accuracy, and domain-related skills (e.g., teaching skills). Listener factors include proficiency in the target language, attitudes towards the speaker, familiarity with the speaker's native language or accent, familiarity with the speaker's topic, and familiarity with non-native speakers or speech in general. Finally, a variety of contextual factors may play a role including the norms of interaction or interpretation, speech form and content, and purpose of communication. Although previous research has identified many of these factors, none of the studies reviewed have attempted to integrate many of these factors into a larger conceptual model. Research that has examined the relationships between

several of these factors has been limited by small sample sizes, constrained or inauthentic speech samples, and homogenous groups of speakers or listeners.

The primary purpose of this study was to examine the relationships among naïve listener perceptions of oral language use and speaker, listener, and contextual factors within a single model, using the framework of structural equation modeling. The conceptual model evaluated by this study may be viewed in several different ways: first, as an interactive model of oral language use centered on listener perceptions of the speaker; second, as a model of the relationships among teacher- and student-related factors that influence a teacher's comprehensibility to students in an academic domain.

The results of this study support an interactional perspective on oral language use which holds that both speaker- and listener-related factors influence comprehensibility. In this study, comprehensibility was predicted by speaker-related factors including components of oral proficiency and teaching effectiveness, and listener-related factors such as perceptions of the speaker's personality, attitude towards the speaker, and interest in the speaker's topic. The results have practical implications for language teaching and education policy, and emphasize an important consideration for oral proficiency assessments in this domain: providing sufficient information about "real world" constructs to make appropriate decisions.

The dissertation of Jonathan Edgar Schmidgall is approved.

Peter M. Bentler

Noreen M. Webb

Lyle F. Bachman, Committee Chair

University of California, Los Angeles

2013

DEDICATION

To my parents, and their dedication to education.

Table of Contents

LIST OF FIGURES xiii
LIST OF TABLES xv
ACKNOWLEDGMENTS xix
VITA xxi
Chapter 1: Statement of the problem 1
1.1 Setting and Importance 1
1.2 Research Questions 4
1.3 Assumptions 5
1.4 Definition of terms 6
1.4.1 Naïve listener 6
1.4.2 Comprehensibility 6
1.4.3 Oral language skills necessary to TA 7
1.5 Delimitations 7
1.6 Importance of the study 9
1.6.1 Theoretical importance 9
1.6.2 Practical importance 10
Chapter 2: Review of relevant literature 13
2.1 Conceptualizations of listener perceptions of oral proficiency 13
2.1.1 Overview 13
2.1.2 Intelligibility 14
2.1.3 Accentedness 15
2.1.4 Comprehensibility 16
2.1.5 Interpretability 17
2.1.6 Summary of conceptualizations of listener perceptions of proficiency 17
2.2 Factors that impact the comprehensibility of speech 19
2.2.1 Overview 19
2.2.2 Speaker-related factors 20
2.2.3 Listener-related factors 20
2.2.4 Context-related factors 22
2.2.5 Summary 23
2.3 Oral language skills necessary to TA 24

2.3.1 The construct of the oral language skills necessary to TA: ITA assessments 24
2.3.2 Stakeholder perceptions of the oral skills necessary to TA 28
2.4 Teaching effectiveness 29
2.4.1 Components of teaching effectiveness 30
2.4.2 Dimensionality of teaching effectiveness scales 34
2.5 A preliminary conceptual model of listener perceptions of the oral language skills necessary to TA 36
Chapter 3: Methodology 42
3.1 Description of the research approach used 42
3.2 Participants 42
3.2.1 Speakers (international graduate students) 42
3.2.2 Listeners (undergraduate students) 43
3.2.3 Raters 45
3.2.3.1 Raters of TOP pronunciation, lexical-grammar, rhetorical organization, and question handling 45
3.2.3.2 Raters of teaching effectiveness 45
3.3 Measurement instruments 46
3.3.1 Test of Oral Proficiency (TOP) tasks 46
3.3.2 Measures of speaker-related factors 47
3.3.2.1 Pronunciation, lexical-grammar, rhetorical organization, question handling 48
3.3.2.2 Teaching effectiveness component measures 49
3.3.3 Measures of comprehensibility and listener-related factors 50
3.3.3.1 Comprehensibility 50
3.3.3.2 Oral proficiency to TA 52
3.3.3.3 Familiarity with speaker's accent and native language 52
3.3.3.4 Familiarity with non-native speakers of English 53
3.3.3.5 Familiarity with ITAs 53
3.3.3.6 Familiarity with, complexity of, and interest in speaker's topic 53
3.3.3.7 Attitude homophily 54
3.3.3.8 Teacher personality 54
3.4 Data collection procedure 54
3.4.1 Small-scale pilot studies 55
3.4.2 Main data collection 57

3.5 Data analysis for the main study 59
3.5.1 Exploratory data analysis 59
3.5.2 Cross-validation data analysis 62
3.6 Software used for data analyses 63
Chapter 4: Results 64
4.1 Pilot studies 64
4.1.1 Pilot study 1 64
4.1.1.1 Pilot study 1: Dataset 64
4.1.1.2 Pilot study 1: Comprehensibility scale 65
4.1.1.2.1 Descriptive statistics 65
4.1.1.2.2 Internal consistency 66
4.1.1.2.3 Interrater consistency 67
4.1.1.2.4 Correlations with criterion variables 68
4.1.1.3 Pilot study 1: Oral skills to TA scale 69
4.1.1.3.1 Descriptive statistics 69
4.1.1.3.2 Internal consistency 69
4.1.1.3.3 Interrater consistency 71
4.1.1.3.4 Correlations with criterion variables 71
4.1.1.4 Pilot study 1: Attitude homophily scale 72
4.1.1.4.1 Descriptive statistics 72
4.1.1.4.2 Internal consistency 73
4.1.1.4.3 Interrater consistency 75
4.1.1.4.4 Correlations with criterion variables 75
4.1.1.5 Pilot study 1: Discussion 76
4.1.2 Pilot study 2 76
4.1.2.1 Pilot study 2: Dataset 77
4.1.2.2 Pilot study 2: Revised Comprehensibility scale 77
4.1.2.2.1 Descriptive statistics 77
4.1.2.2.2 Internal consistency 78
4.1.2.2.3 Interrater consistency 80
4.1.2.2.4 Correlations with criterion variables 80
4.1.2.3 Pilot study 2: Teacher personality scale 81
4.1.2.3.1 Descriptive statistics 81

4.1.2.3.2 Internal consistency 82
4.1.2.3.3 Interrater consistency 84
4.1.2.3.4 Correlations with criterion variables 84
4.1.2.4 Pilot study 2: Discussion 85
4.2 Main study 87
4.2.1 Exploratory phase 87
4.2.1.1 Dataset 87
4.2.1.2 Data cleaning 87
4.2.1.3 Descriptive statistics 88
4.2.1.3.1 Speaker-based components 88
4.2.1.3.1.1 TOP oral proficiency measures 88
4.2.1.3.1.2 Teaching effectiveness measures 91
4.2.1.3.2 Listener-based components 95
4.2.1.3.2.1 Comprehensibility 95
4.2.1.3.2.2 Oral skills to TA 96
4.2.1.3.2.3 Attitude homophily 97
4.2.1.3.2.4 Teacher personality 98
4.2.1.3.2.5 Other listener-based measures 99
4.2.1.4 Measurement models 105
4.2.1.4.1 Teaching effectiveness components 105
4.2.1.4.1.1 Teaching effectiveness components: Exploratory data analysis 106
4.2.1.4.1.2 Teaching effectiveness components: Cross-validation data analysis 112
4.2.1.4.2 Comprehensibility 113
4.2.1.4.4 Attitude homophily 114
4.2.1.4.5 Teacher personality 115
4.2.1.5 Structural model 116
4.2.1.5.1 Parameter estimates 118
4.2.1.5.2 Model fit 120
4.2.1.5.3 Specification search 121
4.2.1.5.4 Recommendations for a revised model 123
4.2.1.5.4.1 Comprehensibility and Oral skills to TA: One construct or two? 123
4.2.1.5.4.2 Pronunciation as a predictor of Comprehensibility 125
4.2.1.6 Undergraduate interviews 130

4.2.1.6.1 Data and procedure 130
4.2.1.6.2 Results 134
4.2.1.6.2.1 Interview research question 1: Do naïve listeners consider different aspects of performance when rating Comprehensibility versus Oral skills to TA? 134
4.2.1.6.2.2 Interview research question 2: Do naïve listeners who are more familiar with the topic, and the speaker's native language and accent find the speaker more comprehensible? 137
4.2.1.6.2.3 Summary of results 137
4.2.1.7 Revised conceptual model 138
4.2.1.7.1 Revised Comprehensibility construct 138
4.2.1.7.2 Addition of Teacher personality measure to the conceptual model 141
4.2.1.7.3 Relationship between familiarity with ITAs and familiarity with NNS 143
4.2.1.7.4 Revised conceptual model 144
4.2.2 Cross-validation phase 148
4.2.2.1 Dataset 148
4.2.2.2 Data cleaning 148
4.2.2.3 Descriptive statistics 148
4.2.2.3.1 Speaker-based components 149
4.2.2.3.1.1 TOP oral proficiency measures 149
4.2.2.3.1.2 Teaching effectiveness measures 151
4.2.2.3.2 Listener-based components 153
4.2.2.3.2.1 Comprehensibility as an ITA 153
4.2.2.3.2.2 Attitude homophily 153
4.2.2.3.2.3 Teacher personality 154
4.2.2.3.2.4 Other listener-based measures 155
4.2.2.4 Multivariate normality 159
4.2.2.5 Measurement models 159
4.2.2.5.1 Teaching effectiveness components 159
4.2.2.5.2 Comprehensibility as an ITA 161
4.2.2.5.3 Attitude homophily 161
4.2.2.5.4 Teacher personality 162
4.2.2.6 Structural model 163
4.2.2.6.1 Parameter estimates 163

4.2.2.6.2 Model fit 165
4.3 Summary of results for research questions 167
4.3.1 Research question 1 167
4.3.2 Research question 2 167
4.3.3 Research question 3 168
Chapter 5: Discussion 170
5.1 Key findings 170
5.2 Implications of this study 173
5.2.1 Implications for the assessment of oral proficiency 173
5.2.2 Implications for language teaching 175
5.2.3 Implications for educational policy 176
5.3 Limitations of the study 178
5.3.1 Methodological limitations: Validity and consistency of measures 178
5.3.2 Conceptual limitations: Generalizing the conceptual model to other domains 179
5.4 Future directions 180
5.4.1 Exploring the conditions under which oral proficiency measures better predict comprehensibility 180
5.4.2 Using more objective measures of speech in the model 181
5.4.3 Availability of visual information, length of interaction, listener accommodation, and comprehensibility 182
5.4.4 The impact of the construct underrepresentation on decision errors for the TOP 183
5.5 Concluding comments 183
Appendix A – TOP Pronunciation, Lexical-grammar, Rhetorical organization, and Question handling rating scales 185
Appendix B – Teaching skills rating scales 186
Appendix C – Teaching effectiveness holistic rating items 187
Appendix D – Comprehensibility rating scale 188
Appendix E – Oral skills to TA rating scales 189
Appendix F – Familiarity with speaker's accent 190
Appendix G – Familiarity with speaker's native language (L1) 191
Appendix H – Attitude homophily scale 192
Appendix I – Teacher Personality scale 193
Appendix J – Overview and instructions for participants in the main study 194

References 196

LIST OF FIGURES

Figure 2.1	Primary speaker-, listener-, and context-related factors expected to influence comprehensibility	24
Figure 2.2	Primary speaker- and listener-related factors expected to influence naïve listener judgments regarding whether a speaker has the oral skills necessary to TA	29
Figure 2.3	Final model of components of speaker oral proficiency and listener-related factors and perceptions of oral language use for the TOP (Schmidgall, 2012)	38
Figure 2.4	An integrated model of components of speaker oral proficiency, speaker-related factors, listener-related factors, and listener perceptions of oral proficiency for the TOP	40
Figure 3.1	A diagram of the procedure followed in the main data collection	57
Figure 3.2	Preliminary SEM model specifying the relationships between listener perceptions of oral proficiency, and speaker- and listener-related factors	61
Figure 4.1	Histograms for the original and transformed values of ITA familiarity for the exploratory dataset	102
Figure 4.2	Measurement model for the three-factor correlated traits model of components of teaching effectiveness for the exploratory dataset	107
Figure 4.3	Measurement model for the revised three-factor correlated traits model of components of teaching effectiveness for the exploratory dataset	111
Figure 4.4	Measurement model for Comprehensibility for the exploratory dataset	113
Figure 4.5	Measurement model for Oral skills to TA for the exploratory dataset	114
Figure 4.6	Measurement model for Attitude homophily for the exploratory dataset	115
Figure 4.7	Measurement model for Teacher personality for the exploratory dataset	115
Figure 4.8	Revised preliminary SEM model specifying the relationships between listener perceptions of oral proficiency, and speaker- and listener-related factors	117
Figure 4.9	Parameter estimates for the preliminary structural model	118
Figure 4.10	A path model of the relationships between speaker oral proficiency variables and listener perceptions of the speaker (Main study, exploratory sample)	126

Figure 4.11	A structural model of the relationships between speaker oral proficiency variables and listener perceptions of the speaker (Main study, exploratory sample)	127
Figure 4.12	Measurement model for the construct of Comprehensibility as an ITA for the exploratory dataset	140
Figure 4.13	Revised conceptual model specifying the relationships between Comprehensibility as an ITA, Teacher personality, and speaker- and listener-related factors	144
Figure 4.14	Parameter estimates for the revised structural model	145
Figure 4.14	Measurement model for Teaching effectiveness components for the cross-validation dataset	160
Figure 4.15	Measurement model for Comprehensibility as an ITA for the cross-validation dataset	161
Figure 4.16	Measurement model for Attitude homophily for the cross-validation dataset	162
Figure 4.17	Measurement model for Teacher personality for the cross-validation dataset	163
Figure 4.18	Parameter estimates for the revised structural model with cross-validation dataset	164

LIST OF TABLES

Table 2.1	Summary of the Constructs Evaluated by Selected U.S. Universities' Local ITA Assessments	26
Table 2.2	A Summary of Components of Scales or Frameworks for Teaching Effectiveness	32
Table 4.1	Descriptive Statistics for Items in the Comprehensibility Scale for Pilot Study 1	65
Table 4.2	Correlations among Items in the Comprehensibility Scale for Pilot Study 1	66
Table 4.3	Item-total Correlations for Items in the Comprehensibility Scale for Pilot Study 1	67
Table 4.4	ICCs for Items in the Comprehensibility Scale for Pilot Study 1	67
Table 4.5	Correlations between Comprehensibility Total Scores and Criterion Variables for Pilot Study 1	68
Table 4.6	Descriptive Statistics for Items in the Oral Skills to TA Scale for Pilot Study 1	69
Table 4.7	Correlations among Items in the Oral Skills to TA Scale for Pilot Study 1	70
Table 4.8	Item-total Correlations for Items in the Oral Skills to TA Scale for Pilot Study 1	70
Table 4.9	ICCs for items in the Oral Skills to TA Scale for Pilot Study 1	71
Table 4.10	Correlations between Oral Skills to TA Total Scores and Criterion Variables for Pilot Study 1	72
Table 4.11	Descriptive Statistics for Items in the Attitude Homophily Scale for Pilot Study 1	73
Table 4.12	Correlations among Items in the Attitude Homophily Scale for Pilot Study 1	74
Table 4.13	Item-total Correlations for Items in the Attitude Homophily Scale for Pilot Study 1	74
Table 4.14	ICCs for Items in the Attitude Homophily Scale for Pilot Study 1	75
Table 4.15	Correlations between Attitude Homophily Total Scores and Criterion Variables for Pilot Study 1	75

Table 4.16	Descriptive Statistics for Items in the Revised Comprehensibility Scale for Pilot Study 2	78
Table 4.17	Correlations among Items in the Revised Comprehensibility Scale for Pilot Study 2	79
Table 4.18	Item-total Correlations for Items in the Revised Comprehensibility Scale for Pilot Study 2	79
Table 4.19	ICCs for Items in the Revised Comprehensibility Scale for Pilot Study 2	80
Table 4.20	Correlations between Comprehensibility Total Scores and Criterion Variables for Pilot Study 2	81
Table 4.21	Descriptive Statistics for Items in the Teacher Personality Scale for Pilot Study 2	82
Table 4.22	Correlations among Items in the Teacher Personality Scale for Pilot Study 2	83
Table 4.23	Item-total Correlations for Items in the Teacher Personality Scale for Pilot Study 2	83
Table 4.24	ICCs for Items in the Teacher Personality Scale for Pilot Study 2	84
Table 4.25	Correlations between Teacher Personality Total Scores and Criterion Variables for Pilot Study 2	85
Table 4.26	Descriptive Statistics for TOP Oral Proficiency Measures in the Exploratory Dataset	89
Table 4.27	Correlations between TOP Oral Proficiency Variables, Comprehensibility Total Scores, and Oral Skills to TA Total Scores	90
Table 4.28	Descriptive Statistics for Teacher Personality Measures in the Exploratory Dataset	92
Table 4.29	Correlations among Teaching Effectiveness Measures and Oral Skills to TA Total Scores	94
Table 4.30	Descriptive Statistics for Comprehensibility Scale Items in the Exploratory Dataset	95
Table 4.31	Descriptive Statistics for Oral Skills to TA Scale Items in the Exploratory Dataset	96
Table 4.32	Descriptive Statistics for Attitude Homophily Scale Items in the Exploratory Dataset	97
Table 4.33	Correlations between Attitude Homophily, Comprehensibility, and Oral Skills to TA Total Scores in the Exploratory Dataset	98

Table 4.34	Descriptive Statistics for Teacher Personality Scale Items in the Exploratory Dataset	99
Table 4.35	Descriptive Statistics for Listener-based Measures in the Exploratory Dataset	100
Table 4.36	Ordinal Transformation of ITA Familiarity	101
Table 4.37	Descriptive Statistics for Transformed ITA Familiarity Variable for the Exploratory Dataset	102
Table 4.38	Correlations among Listener Background Measures and Comprehensibility Total Scores	104
Table 4.39	Correlations between ITA Familiarity and Experience Measures, and Oral Skills to TA Total Scores	104
Table 4.40	Excerpted Results of the Lagrange Multiplier Test for Adding Parameters to the Three-factor Correlated Traits Model of Teaching Effectiveness	108
Table 4.41	Largest Absolute Standardized Residuals for the Three-Factor Correlated Traits Model of Components of Teaching Effectiveness for the Exploratory Dataset	109
Table 4.42	Enthusiasm Scale Items with Large Positive Standardized Residuals	109
Table 4.43	Fit Indices for the Revised Three-factor Correlated Traits Model for Components of Teaching Effectiveness	112
Table 4.44	Standardized Solutions for Structural Equations in the Preliminary Conceptual Model	119
Table 4.45	Fit Indices for the Preliminary Structural Model	120
Table 4.46	Excerpted Results of the Wald Test for Dropping Parameters for the Preliminary Model	121
Table 4.47	Excerpted Results of the Lagrange Multiplier Test for Adding Parameters to the Preliminary Model	122
Table 4.48	Characteristics of Undergraduates Who Participated in the Follow-up Interviews by Group Membership	132
Table 4.49	Fit Indices for the Unidimensional and Two-dimensional Models of Listener Perceptions of Speakers' Oral Language Use with the Exploratory Dataset	139
Table 4.50	Fit Indices for the Unidimensional Model of Comprehensibility after Removing Item TA2 with the Exploratory Dataset	140

Table 4.51	Standardized Solutions for Structural Equations in the Revised Conceptual Model (with Direct Effects only)	146
Table 4.52	Fit Indices for the Revised Structural Model	147
Table 4.53	Descriptive Statistics for TOP Oral Proficiency Measures in the Cross-validation Dataset	149
Table 4.54	Correlations between TOP Oral Proficiency Measures, and Comprehensibility and Teacher Personality Total Scores in the Cross-validation Dataset	150
Table 4.55	Descriptive Statistics for Teacher Personality Measures in the Cross-validation Dataset	151
Table 4.56	Descriptive Statistics for Comprehensibility as an ITA items in the Cross-validation Dataset	153
Table 4.57	Descriptive Statistics for Attitude Homophily Items in the Cross-validation Dataset	154
Table 4.58	Descriptive Statistics for Teacher Personality Items in the Cross-validation Dataset	155
Table 4.59	Descriptive Statistics for Listener-based Measures in the Cross-validation Dataset	156
Table 4.60	Descriptive Statistics for Transformed ITA Familiarity Variable in the Cross-validation Dataset	156
Table 4.61	Correlations among Listener Background Measures and Comprehensibility an ITA Total Scores for the Cross-validation Dataset	158
Table 4.62	Correlation between ITA Familiarity and Teacher Personality Total Scores for the Cross-validation Dataset	158
Table 4.63	Standardized Solutions for Structural Equations in the Revised Model with Cross-validation Dataset	165
Table 4.64	Fit Indices for the Revised Structural Model using the Cross-validation Dataset	166

ACKNOWLEDGMENTS

I am deeply indebted to faculty and staff within a number of departments at UCLA and several organizations outside the university. My development as a language assessment researcher has been indelibly shaped and influenced by my advisor and dissertation chair, Professor Lyle Bachman. His unparalleled understanding of conceptual issues in educational measurement and applied linguistics is complimented by an unwavering pragmatism, and he serves as a true inspiration for those who strive to design conceptually and methodologically sound research to address real-world problems. I will always be grateful for the remarkable good fortune I had to be afforded his guidance throughout my doctoral study.

I owe much of my growth and progress as a researcher to UCLA's Advanced Qualitative Methods (AQM) in Education program faculty and students. My mentor Professor Noreen Webb provided guidance and encouragement throughout my doctoral training, and insight that clearly derives from decades of productive research and mentoring. Professors Jose-Felipe Martinez, Peter Bentler, Li Cai, Steve Reise, and Mike Seltzer provided rigorous and thought-provoking instruction, and useful feedback on course projects. I am particularly grateful to Professor Martinez for his careful attention to methodological issues in generalizability theory and classroom assessment. In addition, part of this research is made possible by a pre-doctoral advanced quantitative methodology training grant (#R305B080016) awarded to UCLA by the Institute of Education Sciences of the US Department of Education.

This research would not have been possible without the support and encouragement of faculty and staff in UCLA's Office of Instructional Development. Directors Larry Loeher and Kumiko Haas demonstrated an interest in and commitment to research activities that helped provide the foundation and inspiration for this study, and I am grateful for their appreciation of

the need to justify the use of test scores with ongoing validity research. I would also like to thank Caitlin Cartmell for research support, as well as the raters and Questioners of the Test of Oral Proficiency (TOP) program who participated directly or indirectly in various facets of this study.

Finally, I would like to thank my colleagues and administrators at Educational Testing Service (ETS) for providing invaluable training, encouragement, and financial support. Dr. Xiaoming Xi provided me with a number of opportunities to contribute to language assessment research projects as a research assistant and summer intern at ETS, and her mentorship played a crucial role in my development as a researcher. I would also like to thank Mary Enright, Yasuyo Sawaki, Yeonsuk Cho, Brent Bridgeman, and Fred Cline for their support and mentorship. A key component of the conceptual model evaluated in this study would not have been possible to include without a grant provided by the TOEFL Grants and Awards Committee under the Small Grants for Doctoral Research in Second or Foreign Language Assessment program.

The views expressed in this paper are the author's alone and do not reflect the views/policies of the funding agencies or grantees.

VITA

MA Cognitive Psychology, University of Colorado/Boulder (2001)

BS Psychology, University of Illinois/Urbana-Champaign (1998)

Grants and Awards

Fellow, Advanced Quantitative Methods in Education Research, UCLA Graduate School of Education and Information Studies

September 2011 – June 2013

Small Grant for Doctoral Research in Second or Foreign Language Assessment, Educational Testing Service

December 2012

Selected Presentations

Schmidgall, J. E. (2013, May). *Evaluating score consistency within the framework of an argument for test use*. Paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Schmidgall, J. E. (2012, April). *The relationships between speaker proficiency variables and contextualized listeners perceptions of oral language use*. Paper presented at the 15th annual conference of the Southern California Association of Language Assessment Research, Los Angeles, CA.

Schmidgall, J E. (2012, April). *Using an ITA assessment to provide detailed feedback on performance: Implications for teachers, learners, and validity*. Paper presented at the 34th annual conference of the Language Testing Research Colloquium, Princeton, NJ.

Xi, X., Schmidgall, J.E., & Wang, Y. (2011, June). *User reactions to using SpeechRater for a practice test and validity implications*. Paper presented at the 33rd annual conference of the Language Testing Research Colloquium, Ann Arbor, MI.

Schmidgall, J.E., & Choi, I.K. (2011, May). *Frameworks for validity: A comparison of traditional and argument-based approaches for reviewing research*. Paper presented at the 14th annual conference of the Southern California Association of Language Assessment Research, Los Angeles, CA.

Schmidgall, J.E. (2011, March). *Confidence in the cut score: Reliability and conditional standard errors for a test of oral English*. Paper presented at the annual conference of the American Association of Applied Linguistics, Chicago, IL.

Schmidgall, J.E. (2010, May). *Developing and evaluating an annotation scheme for grammatical errors in speech*. Paper presented at the 13th annual conference of the Southern California Association of Language Assessment Research, Los Angeles, CA.

Schmidgall, J.E. (2007, October). *Communicative Language Teaching and the Confucian culture of learning: Why can't EFL students speak?* Paper presented at the 38th annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.

Chapter 1: Statement of the problem

1.1 Setting and Importance

As the characteristics of instructors and learners in North American academic settings have become increasingly diverse over the past two decades (Wells, 2007), proficiency in the language of instruction has become an increasingly important issue. In higher education, the increasing number of international faculty members and graduate students has led to institutional assessments of teaching candidates' oral proficiency as well as the need for students to understand different accents and English varieties to communicate in the classroom. In K-12 education, some policy makers have responded to this increased diversity by incorporating controversial evaluations of oral proficiency into teacher certification decisions (Hanna & Allen, 2012). Given the increasing relevance of oral proficiency to this domain and the danger of institutionalizing discriminatory policies by confusing oral proficiency with accent, it is critical for policy makers, researchers, and test developers to carefully and fully consider the construct of oral language proficiency.

Typically, developers of oral language assessments define the construct of oral proficiency in terms of the test-taker's ability to use oral language within a particular domain. Test developers operationalize the construct in rating scales that are in turn used by trained raters to provide scores that are interpreted as indicators of oral proficiency within the domain defined by the test tasks. Raters of oral assessments are trained to focus only on the aspects of performance identified in the rating scales, or construct-relevant indicators. Rater training often incorporates the use of benchmark responses in order to anchor judgment and provide feedback to promote consistent decision-making within and across raters, and minimize the leniency or severity of individual raters. In addition, raters are typically monitored in order to ensure that

ratings are not affected by irrelevant aspects of performance or rater background characteristics. From this perspective, the ideal rater (a) is internally consistent, (b) only attends to construct-relevant indicators, and (c) is not influenced by education, life experience, or personality characteristics that may favor some test-takers and disfavor others. This perspective drives research that examines the consistency and validity of rater scores, and provides important evidence to support test developer claims regarding the interpretation and use of test scores.

What may be lost in this focus on the idealized expert rater is the reality of the naïve listener or interlocutor. The naïve listener lacks the training of the expert rater[1], and research across disciplines suggests that listener perceptions of a speaker's oral proficiency may be influenced by a variety of speaker-, listener-, and context-related factors. Primary speaker-related factors include aspects of the speaker's proficiency in the target language such as pronunciation (Jenkins, 2002), fluency (Derwing, Munro, & Thomson, 2008), and grammatical accuracy (Munro & Derwing, 1995); previous exposure to the target language (Derwing, Munro, & Thomson, 2008); and willingness to communicate (Derwing, Munro, & Thomson, 2008). Listener-related factors include the listener's proficiency in the target language (Coetzee-Van Rooy, 2009), attitudes towards the speaker (Coetzee-Van Rooy, 2009), familiarity with the particular speaker (Brodkey, 1972; Gass & Varonis, 1984), familiarity with the speaker's native language or accent (Brodkey, 1972), familiarity with the speaker's topic (Gass & Varonis, 1984), and familiarity with non-native speakers or speech in general (Kennedy & Trofimovich, 2008). Finally, a variety of contextual factors including the norms of interaction or interpretation, speech form and content, and purpose of communication (Kachru, 2008) may play a role in the listener's perceptions of a speaker's oral proficiency.

[1] Naïve listeners may also be characterized as ordinary listeners within the target language use (TLU) domain as defined by the test developer. A more detailed description is provided in the definition of terms in section 1.4.

This shift in perspective – from expert rater evaluations of proficiency to key stakeholder perceptions – is important for a number of reasons. Since a fundamental goal of oral communication is for the speaker to be understood by listeners in the target language use (TLU) domain, stakeholders may desire to interpret scores on oral proficiency tests as indicators of how well the test-taker (speaker) would be understood by the target population of listeners or interlocutors. Thus, naïve listener perceptions may be considered a useful criterion measure for oral proficiency test scores (e.g., Bridgeman, Powers, Stone, & Mollaun, 2012). In addition, naïve listeners may be key stakeholders who need to be represented somehow in the assessment procedure. Standard-setting procedures may provide a link between test scores and key stakeholder perceptions, but when the target population of listeners is large and diverse it might be difficult to fully consider their perspective during the standard-setting process.

Finally, theory and research in language teaching, learning, and assessment has increasingly viewed naïve listeners as an important consideration for the development of speaking skills, or oral proficiency. Researchers interested in teaching and learning have argued that the goal of oral production should be successful communication outside of the classroom, not native-like production (Kennedy & Trofimovich, 2008). Thus, oral language skills are important to the extent that oral language use exceeds a threshold level at which listeners can understand the speaker, variously characterized as the intelligibility, comprehensibility, or interpretability[2] of speech from the listener's perspective (Coetzee-Van Rooy, 2009). Language assessment has largely followed suit: many assessments of oral proficiency utilize rating scales

[2] Intelligibility, comprehensibility, and interpretability may refer to different levels of listener comprehension or a listener's perceived ease of comprehension. A clear distinction between these terms and their definitions will be provided in section 2.1, below.

that focus on the expert listener's (rater's) ease of comprehension, or the comprehensibility[3] of speech, rather than the expected characteristics of the speaker's utterances. Given the assertions that (a) producing comprehensible speech is more important than producing native-like speech, and (b) comprehensibility is a relative standard (Pickering, 2006), it is important to understand the factors that may influence comprehensibility for relevant populations of listeners. Modeling listener perceptions as a function of speaker-, listener-, and context-related factors also recognizes a growing consensus that language use is co-constructed. In other words, "the responsibility for interpreting a verbal act is equally shared by speaker and hearer" (Kachru, 2008: p. 311).

1.2 Research Questions

Research across disciplines suggests that listener perceptions of a speaker's oral language use may be influenced by a variety of speaker, listener, and contextual factors. Primary speaker factors include aspects of the speaker's proficiency in the target language such as pronunciation and grammatical accuracy, and domain-related skills (e.g., teaching skills). Listener factors include proficiency in the target language, attitudes towards the speaker, familiarity with the speaker's native language or accent, familiarity with the speaker's topic, and familiarity with non-native speakers or speech in general. Finally, a variety of contextual factors may play a role including the norms of interaction or interpretation, speech form and content, and purpose of communication. Although previous research has identified many of these factors, none of the studies reviewed have attempted to integrate many of these factors into a larger conceptual model. Research that has examined the relationships between several of these factors has been

[3] See section 1.4.1 for a formal definition of comprehensibility, or section 2.1.4 for a full discussion of how this term has been used in research.

limited by small sample sizes, constrained or inauthentic speech samples, and homogenous groups of speakers or listeners.

The primary goal of this study is to address these inadequacies by examining the relationships among naïve listener perceptions of oral language use and speaker, listener, and contextual factors within a single model, using structural equation modeling (SEM) with a large sample of research participants.

The following research questions will be addressed:

1) What are the relationships between listener perceptions of oral language use, and speaker- and listener-related factors for a subpopulation of listeners from the TLU domain?
2) To what extent do speaker-related versus listener-related factors affect listener perceptions of comprehensibility?
3) To what extent do construct-relevant factors, i.e., pronunciation, lexical-grammar, rhetorical organization, question handling, versus construct-irrelevant factors, e.g., listener attitudes towards the speaker, affect listener perceptions of whether the speaker has the oral language skills necessary to TA?

1.3 Assumptions

While expert rater evaluations of speaker oral proficiency may be susceptible to the same types of biases that impact naïve listener perceptions of oral proficiency, previous research suggests that differences between first and second ratings for the oral proficiency assessment used in this study, the Test of Oral Proficiency or TOP (UCLA Office of Instructional Development, 2004), account for a relatively small proportion of the variance in test scores

(Schmidgall, 2011). Given the evidence of consistency of TOP ratings and the focus of this study, expert rater-related sources of bias or error will not be examined. Thus, speaker-related factors in the conceptual model will be based on expert rater judgment using established instruments and minimal expert rater-related bias will be assumed.

1.4 Definition of terms

1.4.1 Naïve listener

Naïve listeners will be defined as actual or authentic listeners within the target language use domain. In the case of this study, naïve listeners are undergraduate students who either have or will potentially enroll in courses taught by international teaching assistants (ITAs). Listeners are naïve in the sense that they have not been formally trained to evaluate the oral proficiency of ITAs using a particular rating scale or operationalization of the construct of oral proficiency. Naïve listeners may be expected to provide impressionistic judgments of oral proficiency that would not require linguistic expertise or training. It is important to note that naïve listeners may vary in the extent to which they are familiar with foreign languages, accents, and non-native speakers of English in general. Their key characteristics are (a) being members of the population of listeners corresponding to the TLU domain, and (b) lacking the formal training and certification of expert raters.

1.4.2 Comprehensibility

For the purpose of this study, comprehensibility will be broadly defined as a listener's perception of how easy or difficult a speaker is to understand. This will be operationalized as a listener's self-reported (a) effort required to understand a speaker, (b) degree to which the speaker's proficiency interfered with the listener's understanding, (c) confidence that the speaker

has been understood, and (d) ease or difficulty in understanding the speaker. A full review of this and related constructs is provided in section 2.1, below.

1.4.3 Oral language skills necessary to TA

Oral language proficiency is generally assessed in reference to a particular language use domain (Bachman & Palmer, 2010). In this study, the language use domain is defined as oral language use relevant to the teaching duties of teaching assistants (TAs). The construct of oral language use is conceptualized as consisting of four sub-constructs including pronunciation, lexical-grammar, rhetorical organization, and question handling. Language use tasks relevant to TA duties are varied, and range from informal conversations with students during office hours to formal and informal interactions with students during TA-led classroom sessions. Tasks that are representative of the latter are assumed to include (a) syllabus or assignment presentations, and (b) mini-lecture presentations. These two types of tasks and four sub-constructs are operationalized in the measurement design of the Test of Oral Proficiency, or TOP (UCLA Office of Instructional Development, 2004). A full review of this construct is provided in section 2.3, below.

1.5 Delimitations

There are several approaches to measuring speaker-related factors, each with its own strengths and weaknesses. In one approach, transcripts of speaker performances are analyzed to identify dozens of speaker-related factors based on brief segments (e.g., 20-30 seconds) of speech. The strength of this approach lies in the richness of the information that can be gleaned from the transcripts. A potential limitation is the relatively limited samples of speech that may be used due to the resources required to produce useful transcripts. Producing phonetic transcripts is a manual process that requires trained transcribers and is laborious even for brief

samples of speech. Given the length of speaker-listener interactions for the TOP task used in this study – typically, 5 to 10 minutes for TOP Task 3 – and the large number of speakers needed to evaluate the model, adopting this approach would require hundreds if not thousands of hours of transcription.

Another approach to measuring speaker-related factors is to use more global evaluations of speaker proficiency. In the case of the TOP, global evaluations are provided by trained raters using analytic rating scales for pronunciation, lexical-grammar, rhetorical organization, and question handling. This approach is much more efficient for a large sample of speakers since it utilizes pre-existing data: rater scores from previously administered TOP exams. Another strength of this approach is that it targets specific sub-constructs or speaker-related factors that are difficult to measure directly using more objective transcript- or corpus-based measures. Weaknesses of this approach include inconsistency in ratings and less specific information.

Given the evidence for the consistency of TOP ratings (see section 3.3.1.1 for additional discussion), the impracticality of producing transcripts for hundreds of speakers, and the benefits of employing established measures of relevant sub-constructs, more global evaluations of speaker proficiency will be employed in this study. As a result, generalizations made based on this study may be limited by the conceptual and psychometric qualities of TOP ratings.

While a large number of context-related factors may be expected to impact comprehensibility, these factors are largely consistent across the two scored TOP tasks and TOP administrations. The extent to which the proposed model generalizes to the TLU domain of the oral language skills necessary to TA depends in part upon the correspondence between the construct as defined and measured in TOP tasks and the construct as defined in the TLU domain. This correspondence will be briefly discussed in section 2.3.1 but will not be examined

systematically as part of this research study. The relationship between TOP tasks and the TLU domain may be examined systematically by conducting a needs analysis or domain analysis, but this is beyond the scope of this study. The results of this study are expected to contribute a model to describe listeners' perceptions of speakers' oral proficiency, but this model will reflect the constraints inherent in data collection and the particular language use context defined by TOP tasks.

1.6 Importance of the study

1.6.1 Theoretical importance

The primary goal of this study is to examine the relationships between listener perceptions of oral language use and speaker-, listener-, and context-related factors within a single conceptual model. In doing so, it operationalizes the view that oral communication is co-constructed by the listener and the speaker (e.g., Pickering, 2006). In such a model, individual listener perceptions of a speaker's oral proficiency may be expected to vary based on a number of factors. Although previous research has identified many of these factors (see Chapter 2), none of the studies reviewed have attempted to integrate many of these factors into a larger model. This study has thus been designed with the limitations of previous research in mind, including small sample sizes, constrained speech samples, and homogenous groups of speakers. The resulting model will enable a richer description of how listener perceptions are related to these factors, as well as listener judgments of a speaker's adequacy to perform in a particular TLU domain.

The proposed model can also be viewed as a more specific form of a general model of the relationships between teacher-based factors (in this case, international teaching assistants) and student-based factors (undergraduate students) in classroom communication, centered on student

perceptions of the teacher as a communicator. Thus, the model produced by this study may also be of interest to researchers and policy makers in education and communication studies.

1.6.2 Practical importance

Listener perceptions of oral proficiency and speaker-, listener-, and context-related factors in the model are referenced to a particular language use domain: oral language use by teaching assistants at a large North American research university. Since the listeners in this study will be a subgroup of listeners from the TLU domain, the relationship between construct-relevant speaker-related factors (e.g., Pronunciation) and listener perceptions of speech (e.g., Comprehensibility) may be regarded as evidence to strengthen or weaken the claim that test scores are meaningful indicators of oral proficiency (Bachman & Palmer, 2010). Within the context of the assessment of the oral proficiency of international teaching assistants (ITAs), undergraduate student judgments are often cited as a preferred criterion measure (e.g., Bridgeman, Powers, Stone, & Mollaun, 2012; Hardman, 2010; Hsieh, 2011; Isaacs, 2008; Plough, Briggs, & Van Bonn, 2010; Powers, Shedl, Wilson-Leung, & Butler, 1999; Yule & Hoffman, 1993).

Undergraduate students are key stakeholders of ITA assessments whose perspective needs to be accounted for in the assessment procedure (Hsieh, 2011). Undergraduates may participate in assessments as interlocutors, but generally are not considered qualified to work as raters. One way in which a community of undergraduate students may be brought into the decision-making process is during a standard setting process. Standard setting committees typically consist of a small group of experts or political appointees who represent key stakeholder groups. Given the size and diversity of the undergraduate population at a large research university, including undergraduates in the standard setting process is a challenge. The

model produced in this study will be based on a larger group of undergraduates than could be included in a standard setting study, and thus may be a useful reference during standard setting. A standard setting study for this test will be designed in the near future, and the results of this study are expected to have an impact on procedure for examining the relevance of proficiency thresholds.

This study is also designed to address a number of issues in education. International graduate students play an increasingly important role in higher education in the U.S., particularly in the sciences, where a large number of teaching assistants are necessary to support instruction. The conceptual model evaluated by this study examines the relationship between measures of ITAs' teaching skills, oral proficiency, and undergraduate perceptions while accounting for individual differences in undergraduates' attitudes and experiences. It is critical for administrators and teacher trainers to better understand the relative impact of these various factors as they have a direct impact on the quality of undergraduate education. Results could inform training programs for ITAs on campus, as well as outreach programs for undergraduate students designed to address attitudes towards international students in university communities that are becoming increasingly diverse.

More broadly, the role of a teacher's accent in perceived quality of instruction has been a high-profile political issue in both K-12 and higher education. The issue is often simplified by focusing entirely on a teacher's accent although research suggests that communication in the classroom is affected by a variety of factors, including many of those included in this study such as teaching effectiveness (e.g., teacher clarity, use of visuals) and student attitudes and familiarity. Thus, the conceptual model and quantitative methods employed in this study will produce findings that will contribute to the debate over the relative impact of a teacher's accent

in classroom communication. For example, if student attitudes and familiarity play as big a role in the conceptual model as some previous research suggests, it might be important to help enhance the resources students bring to oral communication via classes or workshops.

Chapter 2: Review of relevant literature

The purpose of this section is to review literature in order to clarify this study's approach to (1) conceptualizations of naïve listener perceptions of speech, (2) factors that may influence a listener's perceptions of speech, (3) the construct of the oral language skills necessary to TA, and (4) the construct of teaching effectiveness in higher education. Based on this review, and the assumptions and delimitations previously described, a model will be proposed specifying the relationships between listener perceptions of oral language use and the various factors that might impact them.

2.1 Conceptualizations of listener perceptions of oral proficiency

2.1.1 Overview

A listener's perception of oral communication has been alternately referred to as *intelligibility*, *comprehensibility*, *interpretability*, *accentedness,* and *ease of understanding*. While used interchangeably and inconsistently, three of these terms – intelligibility, comprehensibility, and interpretability – have often been related to one another in order to distinguish levels of listeners' understanding of speech. *Intelligibility* is typically viewed as word and utterance recognition (Coetzee-Van Rooy, 2009; Kennedy & Trofimovich, 2008) and measured as the number of words a listener can accurately reproduce (Gass & Varonis, 1984; Isaacs, 2008) or via impressionistic listener evaluations of accentedness (Anderson-Hsieh, Johnson, & Koehler, 1992). *Comprehensibility* has been alternately defined as understanding word and utterance meaning (Coetzee-Van Rooy, 2009; Smith & Nelson, 1985), or the degree of difficulty the listener has understanding speech (Derwing, Munro, & Thomson, 2008; Kennedy & Trofimovich, 2008). Researchers have often focused on comprehensibility in the latter sense, and measured it using impressionistic Likert-type scales where listeners indicate how easy or

difficult a speaker was to understand (Derwing & Munro, 2009; Kennedy & Trofimovich, 2008). *Interpretability* refers to the degree to which the speaker's intentions or meaning beyond the utterance level can be understood (Nelson, 2008; Smith & Christopher, 2001) and typically measured by listening comprehension questions (Coetzee-Van Rooy, 2009).

While these constructs suffer from a lack of specificity and are generally disconnected from cognitive models of speech perception, they embrace the notion that oral communication is not simply a matter of speaker skill or proficiency but constitutes an interaction between speaker and listener (Kachru, 2008). Research on intelligibility, accentedness, comprehensibility, and intelligibility will be briefly discussed in order to justify the definition of comprehensibility adopted by this study.

2.1.2 Intelligibility

Intelligibility is typically defined in terms of word and utterance recognition (Coetzee-Van Rooy, 2009; Kennedy & Trofimovich, 2008; Smith & Nelson, 1985), but has also been described as a multifaceted construct that includes recognition of an expression, understanding its literal meaning, and understanding its pragmatic function or meaning within the sociocultural context (Bangbose, 1998). It is usually measured by having listeners perform dictation exercises and determining the overall percentage of words correctly identified (Coetzee-Van Rooy, 2009; Derwing & Munro, 1997; Gass & Varonis, 1984; Munro, Derwing, & Morton, 2006), the percentage of thought groups correctly transcribed (Brodkey, 1972), or the percentage of key words correctly identified (Kennedy & Trofimovich, 2008).

Other researchers have used indirect or self-reported measures of whether the listener understood the literal meaning and propositional content of a segment of speech (Nelson, 2008). Indirect measures include 5- and 7-point Likert-type scales with extreme points marked as

"unintelligible speech" and "near native-like speech" (Anderson-Hsieh, Johnson, & Koehler, 1992; Fayer & Krasinski, 1987). Isaacs (2008) instructed listeners to estimate the overall percentage of words in a speech segment that they understood, and to identify features of speech that they believe inhibited their ability to understand.

Researchers have found that more direct measures of intelligibility or literal comprehension (i.e., transcription tasks) have been highly correlated with speaker- and context-related factors, as well as indirect measures of intelligibility (Smith & Rafiqzad, 1979). In a critical review of intelligibility studies that took a broad view of the term, Rajadurai (2007) noted that a speaker's pronunciation appeared to be a critical component. Other researchers have identified a relationship between intelligibility and various measures of suprasegmentals (Anderson-Hsieh & Koehler, 1998; Tajima, Port, & Dalby, 1997).

2.1.3 Accentedness

Accentedness has been defined as the degree to which a sample of speech differs from a specific variety (Derwing & Munro, 2009; Munro, Derwing, & Morton, 2006). It has been measured using impressionistic 7- or 9-point Likert-type scales that are labeled at extreme points (e.g., 1=No non-native accent, 9=Strong non-native accent), and more recently using computer-assisted acoustic analysis (Kang, 2008). Studies have found that measures of accentedness may be correlated with measures of comprehensibility (Varonis & Gass, 1982; Hsieh, 2011), grammatical accuracy (Munro & Derwing, 1995), phonemic production accuracy (Anderson-Hsieh, Johnson, & Koehler, 1992; Munro & Derwing, 1995; Rinly & Flege, 1998; Riney, Takada, & Ota, 2000), suprasegmentals (Anderson-Hsieh et al, 1992, Munro & Derwing, 1998, 2001; Trofimovich & Baker, 2006), musical ability (Isaacs, 2010), and oral proficiency in general (Hsieh, 2011).

2.1.4 Comprehensibility

Comprehensibility has been alternately defined as understanding word and utterance meaning (Coetzee-Van Rooy, 2009) or the degree of difficulty a listener has understanding an utterance (Derwing & Monro, 2009; Derwing, Munro, & Thomson, 2008; Isaacs, 2010; Kennedy & Trofimovich, 2008; Munro, Derwing, & Morton, 2006). Researchers who define comprehensibility in terms of understanding an utterance typically measure it using listening comprehension questions (Coetzee-Van Rooy, 2009; Bridgeman, Powers, Stone, & Mollaun, 2012). Some argue that there are meaningful differences between empirically verifiable measures of comprehension (e.g., comprehension questions) and perceived comprehensibility (Matsuura, Chiba, & Fujieda, 1999; Rubin, 1992) while others have reported high correlations between listening comprehension questions and indirect measures such as listeners' perceived effort to understand the speaker (Bridgeman, Powers, Stone, & Mollaun, 2012).

Comprehensibility is also defined as the ease or difficulty with which a listener understands an utterance, and measured using Likert-type scales with the extreme points labeled (Derwing & Munro, 2009; Derwing, Munro, & Thomson, 2008; Isaacs, 2008; Kennedy & Trofimovich, 2008). The extreme points are typically labeled "very easy to understand" and "extremely difficult to understand," respectively. Isaacs (2010) investigated the relationship between comprehensibility judgments made by raters and measures from four linguistic categories of speech: phonology, fluency, linguistic resources, and discourse. Based on the strength of the correlation between each individual measure and overall judgments of comprehensibility, as well as feedback from linguistic experts who annotated speaker transcripts, she identified five measures that were significant predictors of comprehensibility levels. An index of word stress errors (phonology) had the largest effect size, while word type frequency

(linguistic resources) and mean length of run (fluency) significantly distinguished lower and intermediate levels of comprehensibility. Measures of grammatical accuracy (linguistic resources) and story breadth (discourse) significantly distinguished intermediate and higher levels of comprehensibility. These findings are broadly consistent with other research that has found correlations between comprehensibility and fluency ratings (Derwing, Munro, & Thomson, 2008), phonological awareness (Venkatagiri & Levis, 2007), various measures of suprasegmentals (Munro & Derwing, 1998, 2001), grammatical accuracy and lexical variation (see Pickering, 2006), and discourse structure (Meierkord, 2004).

2.1.5 Interpretability

When comprehensibility is used to refer to a listener's understanding of word or utterance meaning, it may be contrasted with interpretability. Interpretability refers to perception and understanding of the speaker's intentions, or meaning beyond the word or utterance level (Coetzee-Van Rooy, 2009; Nelson, 2008; Smith & Christopher, 2001). In terms of Bachman's (1990) components of language competence, the listener would be utilizing illocutionary competence. Interpretability is thus at one end of a hypothesized continuum of understanding that moves from intelligibility (decoding, or formal recognition) to comprehensibility (word or utterance meaning) to interpretability (Smith, 1992). Interpretability is typically measured by using listening comprehension questions (Coetzee-Van Rooy, 2009).

2.1.6 Summary of conceptualizations of listener perceptions of proficiency

When researchers have looked at bivariate distributions of measures of these constructs, they have typically had strong correlations. Research suggests that comprehensibility is highly correlated with intelligibility (Issacs, 2008; Smith & Rafiqzad, 1979), accentedness (Varonis & Gass, 1982), and interpretability (Coetzee-Van Rooy, 2009). The distinction maintained by

those who advocate a continuum of understanding from intelligibility to interpretability has been supported anecdotally in small-scale studies (e.g., Coetzee-Van Rooy, 2009), but remains difficult to operationalize and measure with precision. In contrast, researchers who have defined comprehensibility based on listener perceptions (i.e., perceived ease or difficulty of understanding) have found that listeners find the concept intuitive and can use the scales with little training (Munro & Derwing, 1998).

In addition, research suggests that comprehensibility may be more important than accentedness in terms of communicative success (Derwing & Munro, 1997; Munro, 2008; Munro & Derwing, 1995). Kennedy and Trofimovich (2008) argue that listeners may find non-native characteristics in speech acceptable if the speaker is understandable. Thus, when successful communication is not defined in terms of native-like production, comprehensibility is a more relevant construct than accentedness.

Studies of these constructs have been criticized for the methodology employed and the misconceptions they may perpetuate about speech varieties (Rajadurai, 2007). Researchers who define comprehensibility in terms of word or utterance meaning to be measured by listening comprehension questions often use scripted or rehearsed monologues that may lack authenticity. In addition, listeners are usually asked to evaluate speech based on brief recorded segments. In this paradigm, listeners do not have the opportunity to interact with the speaker in any meaningful way. As a result, Smith and Nelson's (1985) claim that "intelligibility is not speaker- or listener-centred, but is interactional between speaker and listener" (p. 333) may be undermined by study design.

Misconceptions about speech varieties may be perpetuated in a number of ways by study design. When accentedness is a focus, the notion that all speech is accented may be lost. If

important listener-related factors are not measured appropriately, such as social-psychological attitudes (see section 2.2.3, below), research may exaggerate the claim that non-native speech lacks intelligibility (Rajadurai, 2007). Research has repeatedly found that listener perceptions of communicative success may have as much to do with their own attitudes as speaker oral proficiency (Lindemann, 2011; Pickering, 2006; Rubin, 1992).

With these concerns in mind, this study will adopt the definition of comprehensibility as a listener's perception of the ease or difficulty with which a speaker is understood. It will not include measures of accentedness, intelligibility, or interpretability. Since the communicative goal in the TLU domain of this study is to be understood by listeners – not to achieve native-like production – the construct of accentedness is not appropriate here.

2.2 Factors that impact the comprehensibility of speech

2.2.1 Overview

Researchers have found evidence that a number of speaker-, listener-, and context-related factors may impact the intelligibility, comprehensibility, or interpretability of speech. Primary speaker-related factors include measures of the speaker's proficiency in the target language such as pronunciation (Jenkins, 2002), fluency (Derwing, Munro, & Thomson, 2008), and grammatical accuracy (Munro & Derwing, 1995); previous exposure to the target language (Derwing, Munro, & Thomson, 2008); and willingness to communicate (Derwing, Munro, & Thomson, 2008). Listener-related factors include proficiency in the target language (Coetzee-Van Rooy, 2009); attitudes towards the speaker (Coetzee-Van Rooy, 2009); familiarity with the particular speaker (Brodkey, 1972; Gass & Varonis, 1984); familiarity with the speaker's native language or accent (Brodkey, 1972); familiarity with the speaker's topic (Gass & Varonis, 1984); and familiarity with non-native speakers or speech in general (Kennedy & Trofimovich, 2008).

Finally, a variety of contextual factors may play a role including the norms of interaction or interpretation, speech form and content, and purpose of communication (Kachru, 2008).

2.2.2 Speaker-related factors

The most commonly investigated speaker-related factors related to the comprehensibility of speech include pronunciation, grammatical or lexical proficiency, and discourse structure (Pickering, 2006). A speaker's pronunciation has been identified as the most frequent cause of communication breakdown (Berns, 2008; Pickering, 2006). Suprasegmental errors have been found to have a stronger impact on comprehensibility than segmental errors (Anderson-Hsieh, Johnson, & Koehler, 1992; Anderson-Hsieh & Koehler, 1988; Isaacs, 2010; Munro & Derwing, 1995). Measures of a speaker's fluency, including objective measures related to speech rate and impressionistic rating scales, also have found to be highly correlated with comprehensibility (Derwing, Munro, & Thomson, 2008; Issacs, 2010; Munro & Derwing, 1998).

Measures of target language and cultural familiarity have been shown to be related to comprehensibility. Derwing, Munro, and Thomson (2008) asked speakers to report their frequency of exposure to target language speech on a daily basis; frequency of exposure was related to speakers' perceived comprehensibility. In a series of studies, speakers' reported age of second language (L2) learning and length of residence in the target language environment predicted comprehensibility (Flege & Fletcher, 1992; Flege, Munro, & MacKay, 1995). Finally, a speaker's pragmatic competence and knowledge of the L2 culture may be related to comprehensibility (Pickering, 2006; Smith & Christopher, 2001).

2.2.3 Listener-related factors

Listener-related factors can generally be grouped into two primary categories: familiarity and attitudes (Pickering, 2006). Listeners can be expected to have varying degrees of familiarity

with (a) a particular speaker, (b) non-native speakers with the same native language (L1) as the particular speaker, (c) the phonology of the speaker's L1, (d) non-native speakers in general, and (e) the topic the speaker is discussing. Research has found a link between comprehensibility and all of these aspects of familiarity (Brodkey, 1972; Carey, Mannell, & Dunn, 2011; Gass & Varonis, 1974; Harding, 2008; Kennedy & Trofimovich, 2008; Pickering, 2006; Rubin, 1992). Familiarity with the phonology of the speaker's L1 may lead to an interlanguage phonology accommodation, even with trained raters. Carey, Mannell, and Dunn (2011) found that raters' familiarity with a speaker's accent was related to higher Pronunciation scores on IELTS oral proficiency interviews.

Listeners' expectations and attitudes towards the speaker appear to be related to perceived comprehensibility. In order to explain this phenomenon from a cognitive processing perspective, Levi-Ari (2010) proposed the *expectations-guided processing model*. Levi-Ari argued that native speakers expect non-native speech to be less reliable (i.e., contain pronunciation, grammatical, and lexical errors) and thus rely more on top-down processing when listening. As a result, native speakers may be expected to weight contextual information more heavily when listening to non-native speakers. Levi-Ari found that listeners had less detailed representations of the speech of non-native listeners based on their prior expectations. Other studies have also found that listeners who expect to understand a speaker are more likely to find the speaker comprehensible (Lindemann, 2002; Smith & Nelson, 1985).

When listeners have more positive attitudes towards speakers, perceived comprehensibility may increase. Rubin (1992) asked undergraduates to listen to segments of instructor speech and rate the speaker's acceptability as an instructor and their attitudes towards the speaker. Attitude homophily, a measure of listeners' perceived similarity to a speaker, was a

significant predictor of undergraduate judgments of the speaker's acceptability as an instructor. Coetzee-Van Rooy (2009) asked listeners to rate speakers on sixteen different attributes (e.g., friendly/unfriendly) and found that attitude towards the speaker was as correlated with comprehensibility as other perceptual measures of oral proficiency. Lindemann (2003) found that listeners rated non-native speakers significantly lower than native speakers for status-related characteristics (e.g., intelligent, successful) and has argued that "listener's assessments of the success of an interaction may have more to do with their own attitudes than speaker proficiency." (Lindemann, 2011: p. 228).

Cognitive factors may impact comprehensibility. Listeners with higher levels of cognitive fatigue may be less tolerant of speech errors and judge an interaction to be less successful (e.g., Field, 2003). Models of listening comprehension and speech perception emphasize the role of the listener's ability to efficiently allocate attentional resources and individual differences working memory capacity (e.g., Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Field, 2011; Londe, 2008; Rost, 2005). For example, Londe (2008) found that listening comprehension ability was predicted by listeners' working memory and short-term memory capacities. Levi-Ari (2010) concluded that working memory influences listeners' ability to adapt to non-native speakers – in particular, that higher working memory capacity leads to greater "good enough" processing, which is more tolerant of speech errors. Other cognitive abilities, such as musical ability, have been shown to impact listener ratings of speaker accentedness (Isaacs, 2010).

2.2.4 Context-related factors

Contextual or situational factors have been less studied but can be expected to impact comprehensibility (Derwing, Munro, & Thomson, 2008; Pickering, 2006; Nelson, 2008). While

often overlooked, these factors may be critical to understanding listeners' evaluations of non-native speakers, due to a greater reliance on top-down processing (Levi-Ari, 2010). Kachru (2008) lists a variety of contextual factors that may influence comprehensibility, including the participants, purpose of the interaction, norms of interaction, and environmental noise. Language assessment researchers have long sought to carefully define how facets of the environment, language input, and expected response may impact language use and judgments of proficiency (e.g., Bachman, 1990).

2.2.5 Summary

Although a wide variety of speaker-, listener-, and context-related factors may impact listener judgments of comprehensibility, major components can be identified. Primary speaker-related factors include aspects of pronunciation, particularly suprasegmentals, as well as grammatical accuracy, lexical accuracy, and discourse organization. Listener-related factors are generally related to various measures of familiarity and attitudes towards the speaker. Context-related factors may vary widely based on the language use task, but also need to be carefully considered when investigating listener comprehensibility.

Figure 2.1, below, provides a visual illustration of the relationship between comprehensibility and primary speaker-, listener-, and context-related factors.

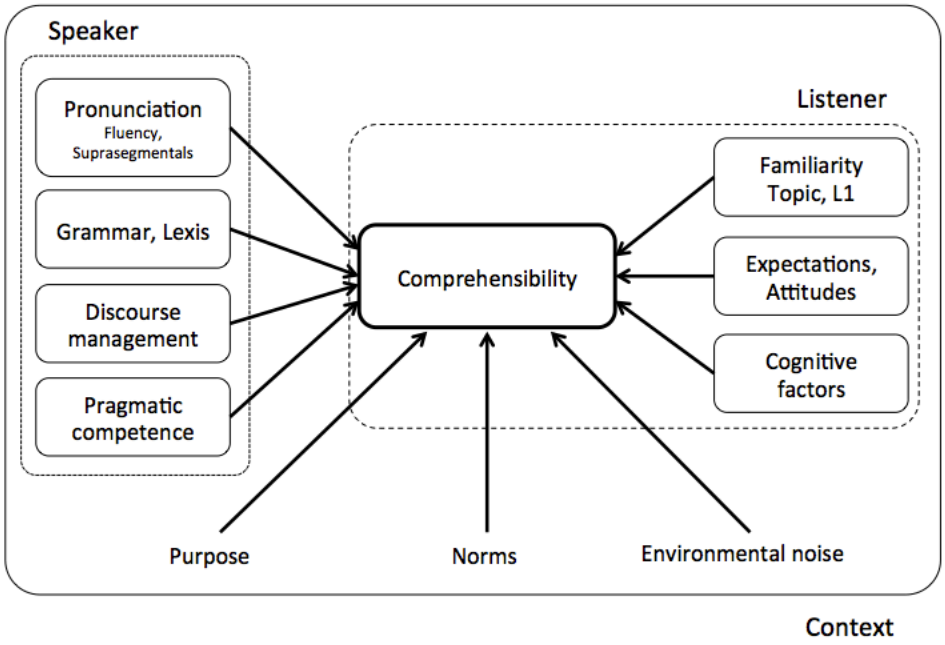

Figure 2.1. Primary speaker-, listener-, and context-related factors expected to influence comprehensibility.

2.3 Oral language skills necessary to TA

In one language use domain, the oral language skills necessary to perform as a graduate teaching assistant (TA), administrators use scores from assessments to make certification decisions. These assessments often require international graduate students who intend to TA (ITAs) to perform teaching tasks in a simulated classroom environment. Such ITA assessments may evaluate test-takers based on their oral skills, listening comprehension, and/or their interaction with students.

2.3.1 The construct of the oral language skills necessary to TA: ITA assessments

ITA assessments differ from each other in the extent to which they focus on evaluating oral skills, listening comprehension, and interaction or teaching skills, as well as how they operationalize the broader construct of the oral skills necessary to TA. An internet search of U.S. universities' approaches to ITA assessment was conducted in March 2012 and 29 ITA

assessment programs were identified. In general, universities conduct ITA assessments using locally developed assessments (e.g., the University of California, Los Angeles' *Test of Oral Proficiency,* or TOP) or one produced by an independent testing agency such as Educational Testing Service's (ETS) Speaking Proficiency English Assessment Kit (SPEAK). Of the 29 ITA assessment programs reviewed, 13 used locally developed assessments, 12 used the SPEAK exam, three used both, and one used the TOEFL iBT®, a large-scale assessment of academic English proficiency.

Although the SPEAK is no longer distributed or supported by ETS, it is still used as a measure of oral proficiency in many ITA assessment programs. The SPEAK was designed to be evaluated by two trained raters using a holistic rating scale containing five levels. Each level of the rating scale is characterized by four aspects of performance, including (1) the use of linguistic features, (2) the use of cohesive devices, (3) appropriateness of response, and (4) performance of task functions. Thus, the rating scale focuses on both the effectiveness of communication and task completion.

Since many of the SPEAK tasks do not relate to the domain of language use defined by TA duties or even a broader academic context, a number of ITA assessment programs have developed their own approach to ITA oral proficiency assessment. Unlike the SPEAK, tasks in these assessments correspond more closely to the target language use domain of TA language use, and may include an office hour simulation, a classroom role play, mini-lecture, and explaining an academic topic. Several of these assessments manage to include undergraduate students in the assessment procedure as interlocutors that interact with test-takers during relevant tasks.

In general, the construct of the oral language skills necessary to TA as defined by ITA assessments is dominated by components of oral skills (e.g., pronunciation), but may also include several other subconstructs such as aural skills and language use in interaction. A summary of the constructs evaluated by local assessments is provided in Table 2.1, below. Columns in the table indicate the constructs most commonly included in ITA assessments, and each row corresponds to a specific ITA assessment. Constructs included in a specific ITA assessment are indicated by a check mark.

Table 2.1

Summary of the Constructs Evaluated by Selected U.S. Universities' Local ITA Assessments

	Oral skills						Interaction			
ITA assessment (university or authors)	PR	FL	VC	GR	OR	OTH	QH	OTH	LIS	VAR
Test of Oral Proficiency (UCLA)	✓		✓	✓	✓		✓			
Oral Proficiency Test (Southern Illinois University)	✓	✓	✓	✓	✓				✓	
ITA test (Smith, Meyers, & Burkhalter, 1992)	✓	✓		✓	✓	✓	✓	✓	✓	✓
OECT (Iowa State University)	✓	✓	✓	✓	✓				✓	
TEPAIC (Indiana University)	✓	✓	✓	✓			✓		✓	
English Proficiency Interview (University of Illinois)	✓	✓	✓	✓	✓	✓	✓			
ITA test (University of Victoria)	✓		✓	✓	✓					✓
ITA test (Carnegie Mellon)	✓	✓		✓		✓			✓	
ITA test (Stanford University)		✓				✓		✓		
ITA test (University of Texas)	✓	✓		✓		✓				
Interactive Performance Test (University of Pennsylvania)		✓				✓				

Note. PR = Pronunciation; FL = Fluency; VC = Vocabulary; GR = Grammar; OR = Organization; OTH = Other; QH = Question handling; LIS = Listening; VAR = Various.

While local ITA assessments vary in terms of how they operationalize oral skills in their rating scales, most of these explicitly evaluate pronunciation, fluency, and grammar. Vocabulary

use, rhetorical organization, and comprehensibility are also frequently evaluated. Despite the relatively strong agreement across local assessments in terms of the construct of oral skills, the conceptual overlap between many of these categories (e.g., grammar and vocabulary) and lack of agreement regarding component definitions (e.g., fluency as a subcomponent of pronunciation or a separate component of oral proficiency) leads to differences in how subconstructs are operationalized. These differences may be further magnified by the use of holistic rubrics that use implicit weights, or analytic rubrics that use explicit weights to prioritize some components over others. For example, UCLA's Test of Oral Proficiency (TOP) uses an analytic scale in which grammar and vocabulary use are combined into one subscale, which is weighted less than the subscale for pronunciation. As a result, the construct of oral skills in the TOP is dominated by pronunciation.

Two other broad aspects of language use are often evaluated by local ITA assessments: aural skills, and language use in interaction. Some ITA assessments evaluate listening comprehension separately (e.g., Iowa State's Oral English Certification Test, or OECT), while others evaluate listening within the context of an interaction with an interlocutor (e.g., UCLA's TOP). The interaction component of an ITA assessment is often described as question handling (UCLA's TOP; Smith, Meyers, & Buckhalter's ITA test; University of Illinois' English Proficiency Interview, or EPI), and may include additional measures such as rapport and eye contact, audience awareness, and communication style.

Approximately half of the ITA assessments summarized in Table 2.1 also included an assessment component related to the interaction between the ITA and an interlocutor, typically an undergraduate. This component, labeled "Question handling" in the table, typically assesses the interaction between ITA and undergraduate that often takes the form of question-and-answer.

Successful interaction requires both speaking and listening skills on the part of the ITA, as well as interpersonal qualities that may be assessed explicitly as *rapport*, *eye contact*, or *audience awareness*.

2.3.2 Stakeholder perceptions of the oral skills necessary to TA

Plough, Briggs, and Van Bonn (2010) analyzed transcriptions of speaker performances, ratings, and rater comments from an ITA assessment in order to examine the features of language use that predicted certification decisions. The authors concluded that listening comprehension and pronunciation features predicted approval but this varied across disciplines. For speakers from literature, science, and arts, pronunciation largely predicted approval. For speakers from engineering, both pronunciation and interactional language use predicted approval.

Based on an empirical study comparing trained raters with untrained undergraduate ratings of speaker proficiency, Hseih (2011) identified six major conceptual categories that raters used to justify their ratings of oral proficiency: linguistic resources, phonology, fluency, content, global assessment, and non-linguistic factors. Hsieh concluded that undergraduates are more likely to make global assessments of speaker proficiency based on a speaker's perceived accent.

Given the prominence of undergraduate students as key stakeholders of ITA assessments, it is not surprising that several researchers have attempted to examine the relationship between undergraduate judgments of the oral skills necessary to TA and scores on ITA assessments. Rubin (1992) had 148 undergraduate students listen to three different instructors with varied accents and rate the speakers on their qualification to be a teacher, their accent, and perceived ethnicity. Rubin also collected listener-related background variables related to attitude and familiarity with non-native speakers. He found that listener ratings of teacher qualification were predicted by attitude homophily and perceived accent. While the ratings of teacher qualification

were not specific to oral skills, Rubin's findings suggest that listener attitudes may impact holistic judgments of competence as a TA.

Isaacs (2008) asked 18 undergraduate listeners to rate the speech of 8 ITAs based on their comprehensibility and whether they believed the speaker had the oral skills necessary to TA. She found that these two judgments were strongly correlated (r=0.78). She also asked listeners to justify their ratings using written comments. Listeners primarily referenced ITAs' speech rates and orating skills in these comments. Isaacs suggested that undergraduates' perceptions of whether speakers had the oral skills necessary to TA could be partially explained by comprehensibility, but other factors (i.e., those related to teaching skills) needed to be considered.

A potential model of primary speaker- and listener-related factors related to stakeholder perceptions of the oral skills necessary to TA is shown in Figure 2.2, below.

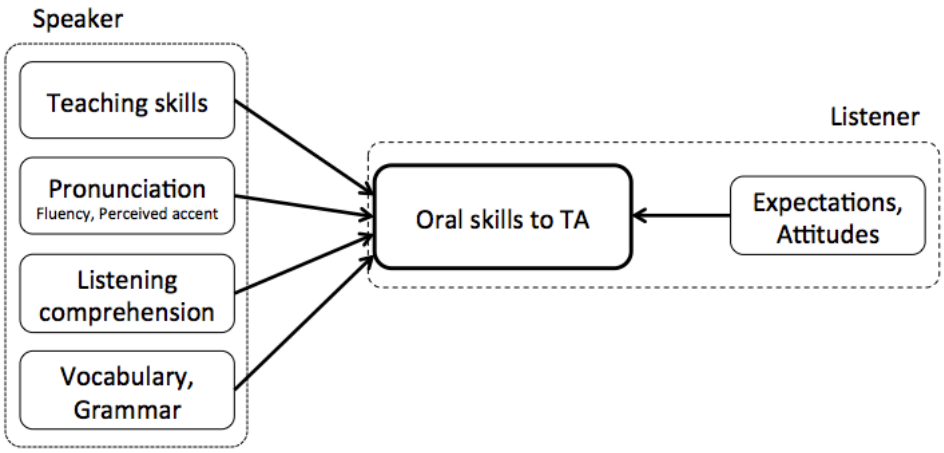

Figure 2.2. Primary speaker- and listener-related factors expected to influence naïve listener judgments regarding whether a speaker has the oral skills necessary to TA.

2.4 Teaching effectiveness

There is currently no consensus on how to define the construct of teaching skills in higher education, which has also been described as instructional quality (Junker, Weisberg, Matsumura,

Crosson, Kim-Wolfe, Levison, & Resnick, 2006), teaching competence (Catano & Harvey, 2011; Roelofs & Sanders, 2007), and *teaching effectiveness* (Heckert, Latier, Ringwald, & Silvey, 2006; Patrick & Smart, 1998; Polk, 2006; Rothstein & Mathis, 2013; Shaw, Young, Shaffer, & Mundfrom, 2003). It is generally recognized as a multidimensional construct with components that relate to teacher-centered behavior and skills, and student-centered interaction (Kang, 2008). Since the purpose of defining the construct is often to inform evaluation frameworks for teachers, the goal of teaching is usually specified as part of a scale or framework. This goal is typically framed in terms of instructional outcomes (Hosek, 2011), and may narrowly focus on student understanding or learning of content, or may include student outcomes such as affective learning (e.g., Houser & Frymier, 2009; McCroskey, 1994), learner empowerment (Mazer, 2012), or an abstract notion of course value (Heckert et al, 2006).

Researchers have proposed a variety of scales and conceptual frameworks to describe the overall construct that differ in terms of their components and dimensionality. In order to provide a coherent overview of the construct, similarities and differences between components across frameworks and scales will be explored. Next, a brief review of studies of the dimensionality of componential scales will be provided. Based on this overview and the context in which the construct is being applied, relevant components of teaching effectiveness will be identified and described in more detail in order to justify their specification in the conceptual model used in this study.

2.4.1 Components of teaching effectiveness

A review of 18 scales or evaluation frameworks for the construct of teaching effectiveness identified six major themes that captured most of the components within and across frameworks. These themes are shown in the top row of Table 2.2, below, and included (1)

Pedagogic/Organization skills, (2) *Communication* skills, (3) Teacher *enthusiasm*, (4) *Subject expertise*, (5) *Relevance and challenge*, and (6) *Respect/rapport* with students. Components associated with themes 1-4 were centered on teacher behavior while those associated with themes 5 and 6 were based on student interaction with the teacher and course content. A seventh general theme, *Other*, was included in Table 2.2 to capture components of frameworks not easily captured by one of the six primary themes. One additional column, *Global*, was inserted to identify when scales included items that provided a global assessment of the construct (e.g., "Everything considered, I would rate the instructor's effectiveness…").

Table 2.2

A Summary of Components of Scales or Frameworks for Teaching Effectiveness

		Teacher				Interaction		
Construct (reference)	Global	ORG	COM	EN	SUB	REL	RES	Other
Teaching effectiveness (Heckert et al, 2006)	✓	✓				✓	✓	✓
Teaching competence (Marsh & Roche, 1997)		✓	✓			✓	✓	✓
Teaching competence (Fulton, 1996)		✓	✓	✓	✓		✓	✓
Teaching competence (Cohen, 1981)		✓			✓	✓	✓	✓
Teaching competencies (Catano & Harvey, 2011)		✓	✓				✓	✓
Teaching effectiveness (Patrick & Smart, 1998)		✓	✓			✓	✓	
Instructional quality (Junker et al, 2006)		✓	✓					
Teaching competence (Roelofs & Sanders, 2007)		✓			✓		✓	✓
Lecturing performance (Ting, 2000)	✓	✓	✓	✓		✓		
Effective teaching (FFT: Kane et al, 2013)		✓	✓	✓			✓	✓
Effective teaching (CLASS: Kane et al, 2013)		✓	✓	✓	✓		✓	✓
Teacher ratings (Basow, 1990)	✓	✓	✓	✓	✓		✓	✓
Traits of effective teachers (Polk, 2006)		✓	✓	✓				✓
Teaching skills (Gibbs & Coffey, 2004)		✓		✓			✓	
Classroom performance (Guyton & Farokhi, 1987)		✓	✓	✓			✓	
Teaching effectiveness (Shaw et al, 2003)	✓	✓	✓	✓	✓	✓	✓	✓
Teaching skills (Smith et al, 1992)		✓	✓	✓		✓	✓	
Teaching skills (Farnsworth, 2004)	✓		✓	✓		✓	✓	

Note. ORG = Organization and pedagogy; COM = Communication; EN = Enthusiasm and attitude; SUB = Subject expertise; REL = Relevance and challenge; RES = Respect and rapport with student

Despite differences among scales and frameworks in terms of how each theme was defined and subsequently operationalized, a number of consistencies can be observed in Table 2.2. First, nearly all frameworks (17/18) included at least one component related to teacher-centered themes (themes 1-4), and one component related to interaction themes (themes 5-6). Among components related to teacher-centered themes, *Pedagogic/Organization* skills or *Communication* skills were included in all of the frameworks reviewed. As seen in Table 2.2, some frameworks included separate components for *Pedagogic/Organization* and *Communication* skills (5/18), while others combined them into a single component (9/18) such as *Organization/Clarity* in March & Roche's (1997) teaching competence scale. Teacher *enthusiasm* was also a frequently included component in the frameworks reviewed (12/18). Teachers' *subject expertise* or competence was also included in a minority of frameworks (5/18).

Two themes were identified that related to interaction: *Relevance and challenge* of the lecture or academic content, and *Respect/rapport*. Among these themes, a component related to *Respect/rapport* was included in most (15/18) of the frameworks reviewed. Components categorized under this theme focused on the degree to which a teacher cultivated respectful and nurturing relationships with students. The *relevance and challenge* theme appeared in half (9/18) of the frameworks, and components related to this theme focused more on the interaction between students and class content: its relevance, difficulty, and/or the degree to which it provided intellectual stimulation.

A majority (11/18) of the frameworks reviewed included at least one component not easily classified into one of the six themes just described and were categorized as *Other*. Components under this theme typically related to teachers' professional development, or use of assessments and grading.

Among the prevailing themes in this summary, most are relevant to teaching effectiveness in the context of the domain of oral language use by ITAs. The most frequently included components relating to teacher-centered themes – *Pedagogy/Organization*, *Communication*, and *Enthusiasm* – are relevant and could potentially be evaluated by a rater observing the ITA present a single mini-lecture. One of the concerns about the measurement of components related to subject matter expertise is the competence of the rater to evaluate it. Given the diverse range of topics presented in the domain of ITA mini-lectures, it would be extremely difficult to evaluate this component without a large number of content experts to serve as raters. The most common components related to interaction – those within the *Respect/rapport* theme – are relevant to this domain as well, due to the interaction between ITAs and students. The other theme related to interaction, *Relevance and challenge*, may also be relevant to this domain but difficult to assess as components such as academic rigor or appropriateness may require the judgment of content experts. Other components included in this theme such as difficulty and intellectual stimulation may be relative to the individual student. In addition, in this domain ITAs are instructed to present an introductory topic of limited complexity. Thus, content is relevant to the lecture task under the condition that it is not perceived as extremely complex by a student.

2.4.2 Dimensionality of teaching effectiveness scales

Despite the prevailing view that teaching effectiveness is multidimensional, researchers have disagreed about whether measures consisting of multiple components should report a global score and treat teaching effectiveness as a unidimensional construct, or report component scores separately. This disagreement may be more or less pronounced depending on the purpose for which scores on teaching effectiveness measures are used (Ryan & Harrison, 1995). If scores

are used for high-stakes decisions, such as personnel decisions, some researchers argue that global scores are more appropriate than component scores. Abrami and d'Apollonia (1991) argued that established rating forms might not be able to consistently reproduce multidimensional factor structures across the diverse subgroups that exist in a university setting. In addition, despite reasonable agreement on the themes that constitute the overall construct of teaching effectiveness, there is considerable variation in terms of which components constitute a theme and how components are subsequently operationalized. There is less controversy over the use of scores for low-stakes purposes, such as diagnostic feedback or theory development.

Studies of the dimensionality of rating forms using exploratory and confirmatory factor analysis often find evidence for multiple factors related to effective teaching that are consistent with the summary presented in the previous section. Erdle, Murray, and Rushton (1985) identified and coded teacher "predictive classroom behaviors" into 26 items and conducted exploratory factor analysis to examine the dimensionality of these items. Their analyses suggested a two-factor solution consisting of a charisma dimension (communicative and interpersonal) and organization dimension. Entwistle and Tait (1990) also found evidence for a two-factor solution characterized by teaching ability and openness to students. Swartz, White, and Stuck (1990) presented a two-factor solution consisting of clear instructional presentation and management of student behavior. These findings are consistent with the previous organization of teaching effectiveness components into two primary orientations: teacher-centered and interaction.

Shaw, Young, Schaffer, and Mundfrom (2003) found evidence of good model fit for both unidimensional and multidimensional models for the same multi-componential scale of teacher effectiveness. Shaw et al.'s scale included 11 items, each corresponding to a single component

of teaching effectiveness. In the unidimensional model, all of the items had high loadings (0.73 – 0.89) with the exception of the subject matter knowledge item, which had a moderately high loading (0.55). The researchers also produced a five-factor solution in which four items loaded on a *Respect/rapport* factor (Comfortable atmosphere, Respectful of students, Warmth and friendliness, Concern for learning), two items loaded on a *Relevance and challenge* factor (Increased interest, Increased understanding), two items loaded an a factor that combined *Communication* and *Enthusiasm* (Communication skills, Motivate and stimulate, Enthusiasm), one item composed a *Pedagogy/Organization* factor (Course organization), and one item a *Subject expertise* factor (Subject matter knowledge).

Since the purpose of using the construct of teaching effectiveness in this study is to examine how it might influence listener judgments of a speaker's oral language use in the presence of speaker oral language proficiency, a multi-dimensional or componential approach to the construct is preferred. Evidence suggests that the overall construct can be measured in terms of its components and that the components are measuring different aspect of teaching effectiveness. One of these components is communication skills. Since this component is related to speaker oral proficiency, using a multi-componential model that includes communication skills may help control for aspects of teaching effectiveness more directly related to oral proficiency. Finally, the psychometric quality and validity evidence for established measurement instruments can be deemed acceptable given the low-stakes nature of this research.

2.5 A preliminary conceptual model of listener perceptions of the oral language skills necessary to TA

UCLA's Test of Oral Proficiency (TOP) is an oral language assessment for international graduate students used to evaluate language use relevant to teaching assistant (TA) duties.

Trained raters evaluate test-takers during two scored tasks using an analytic rubric for pronunciation, lexical-grammar, rhetorical organization, and question handling. During the first scored task, test-takers present a generic syllabus to a mock class of two undergraduate students trained for this purpose. During the second task, test-takers give a brief prepared presentation on a basic topic in their academic field. The undergraduate students are a crucial component of this oral assessment in that (a) their interaction with the test-taker serves as the basis for the subscale *Question handling*, and (b) their presence provides a degree of authenticity that might otherwise undermine the validity of score interpretations if lacking. However, despite their role in the assessment, undergraduate perceptions of test-taker oral proficiency do not play a role the evaluation process. Given their role as primary stakeholders, it is important for the test administrator to understand the relationship between the TOP's measures of oral proficiency and undergraduate perceptions of test-takers' oral language use.

The characteristics of the TOP scoring and testing procedures are ideal for examining the complexity of interactions between speaker-, listener-, and context-related factors. Schmidgall (2012) proposed a model that specified the relationship between components of speaker oral proficiency as measured by the TOP (pronunciation, lexical-grammar, rhetorical organization, question handling) and listener perceptions of speaker language use (comprehensibility, oral skills to TA) that accounted for two listener-based factors: familiarity with the speaker's native language (L1), and familiarity with non-native speakers in general. After specifying and evaluating the initial model with an exploratory dataset using structural equation modeling, Schmidgall (2012) fit a revised model using a cross-validation dataset. This model provided a close fit to the data ($\chi^2(9) = 4.50$, $p = .88$; RMSEA = 0.00 (0.00, 0.05); CFI = 1.00) and is reproduced in Figure 2.3, below.

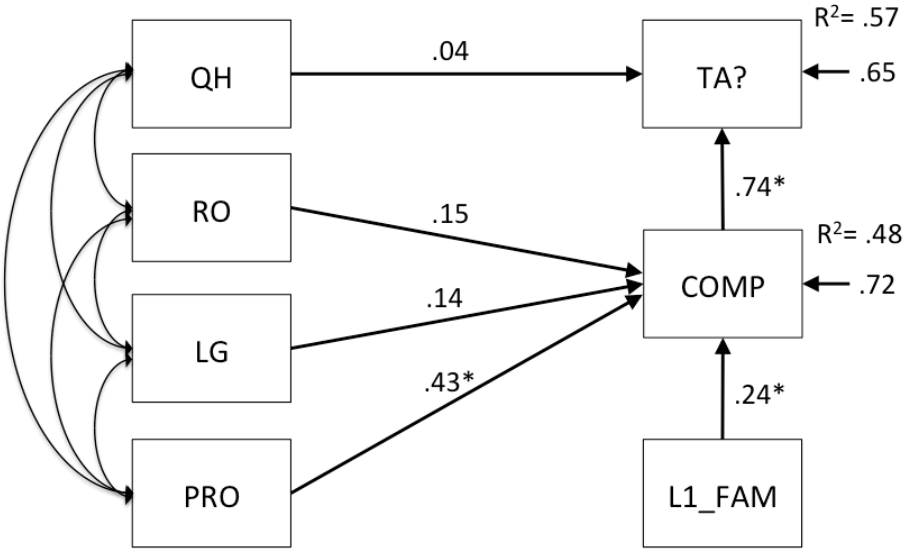

Figure 2.3. Final model of components of speaker oral proficiency and listener-related factors and perceptions of oral language use for the TOP (Schmidgall, 2012). TA? = TA acceptability; COMP = Comprehensibility; L1_FAM = Familiarity with speaker's L1; QH = Question handling; RO = Rhetorical organization; LG = Lexical-grammar; PRO = Pronunciation. * $p < .05$.

As shown in the model, listener ratings of speaker comprehensibility were significantly predicted by speaker pronunciation (0.43), while only weak positive relationships with a speaker's lexical-grammar (0.14) and rhetorical organization (0.15) were found. These results are consistent with the findings of previous research, that pronunciation has the greatest impact on comprehensibility. A listener-related factor, the listener's familiarity with the speaker's native language, was also a significant predictor of comprehensibility (0.24). While these four variables only explained 48% of the variance in comprehensibility ratings, the analysis helped clarify the relative impact of different aspects of language use and several listener-related factors on comprehensibility.

The relative impact of subcomponents of oral language proficiency and comprehensibility on listener perceptions of a speaker's proficiency within a particular context – in this case, oral language use for TA duties – was clarified. Comprehensibility was a strong predictor of perceived adequacy of oral language skills for TA duties, and the measured aspects of oral proficiency were relevant to this judgment to the extent that they predicted comprehensibility. The indirect effect of Pronunciation on perceived adequacy of oral language skills to TA was 0.32, Lexical-grammar was 0.10, and Rhetorical Organization was 0.11, larger than the direct effect of 0.04 of Question Handling on perceived adequacy to TA. Again, although a limited amount of the variance in listener ratings of perceived adequacy to TA was explained by the variables in the model (R^2=0.57), this study provides an initial indication of aspects of speaker proficiency and listener-related factors that influence these more global, context-related perceptions of language proficiency.

In order to try to account for a larger percentage of the variance in listener perceptions of speaker proficiency, additional factors that have been found to be related to comprehensibility or language use in context need to be measured and included in an expanded model. Some of these factors may be related to the speaker, such as domain-related skills (e.g., teaching effectiveness). Additional listener-related factors such as listener attitudes, and familiarity and interest in the speaker's topic may also help explain some of the variation in comprehensibility ratings. Finally, another limitation of the current model is the precision with which some of the variables have been measured. Comprehensibility remains an impressionistic construct that may be further stabilized by using multiple indicators in a measurement model.

Given the promise of this approach for clarifying the relationships between listener perceptions and speaker proficiencies within specific language use domains and the growing

recognition that oral language communication is co-constructed by the speaker and the listener, an integrated version of the models presented earlier in Figures 2.1 and 2.2 that incorporates results from the model relevant to language use domain defined by the TOP in Figure 2.3 is presented below in Figure 2.4. This preliminary conceptual model is centered around a listener's perception of a speaker's comprehensibility and indicates hypothesized relationships between measures of a speaker's oral proficiency, a key speaker-related factor (teaching effectiveness), and listener-related factors (attitudes, L1 familiarity, topic familiarity, cognitive fatigue. Context-related elements are missing from this model as they are fixed across speakers and questioners due to the characteristics of the TOP assessment procedure.

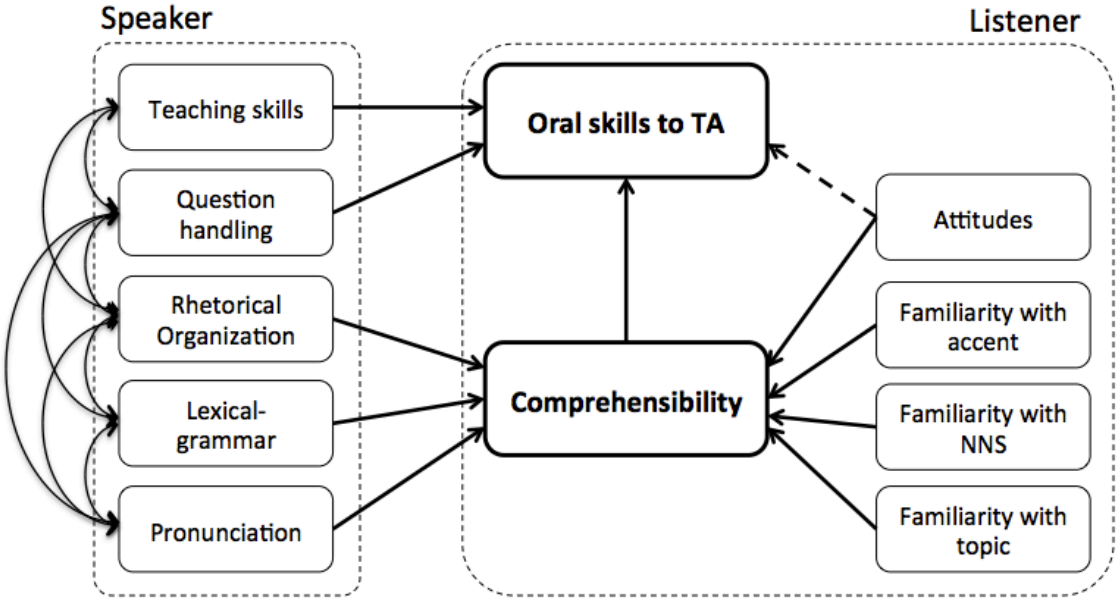

Figure 2.4. An integrated model of components of speaker oral proficiency, speaker-related factors, listener-related factors, and listener perceptions of oral proficiency for the TOP.

The speaker-related variables specified in the model and measured by the TOP (Teaching skills/effectiveness, Question handling, Rhetorical organization, Lexical-grammar,

Pronunciation) correspond well to those identified as important in this domain by prior research and similar assessment programs. The curved line with two arrows linking these variables indicates that they are expected to be correlated with one another. In particular, the four analytic subscales of the TOP are expected to correlate as they are designed to measure a single overall construct: the oral skills required to perform TA duties. The Rhetorical organization and Question handling subscales may also be expected to correlate with measures of teaching effectiveness (Farnsworth, 2004).

A listener's perception of the comprehensibility of a test-taker's speech is expected to be influenced by his or her attitude towards the speaker, familiarity with the speaker's accent, familiarity with non-native speakers of English in general, and familiarity with the topic presented by the speaker. The perceived comprehensibility of the speaker is in turn expected to influence the listener's perception of whether the test-taker has the oral language use skills to adequately perform TA duties for a typical class. This latter perception is expected to require a broader judgment on the part of the listener, in that he or she needs to consider additional aspects of language use beyond the comprehensibility of speech relative to a specific context or domain. Since the domain is TA duties for a typical undergraduate class, the listener is being asked to consider the perspective of other undergraduates as well. The model in Figure 2.4 implies that this judgment is not directly affected by the same listener-related factors that affected the intuition-based judgments of comprehensibility, although indirect effects may be present. The dashed line between the listener-based attitudes factor and the listener judgment of oral skills necessary to TA hypothesizes that a direct effect of attitudes on the oral skills judgment can be hypothesized based on previous research.

Chapter 3: Methodology

3.1 Description of the research approach used

This study utilized structural equation modeling (SEM) to evaluate a model that relates listener perceptions of comprehensibility and domain-appropriate language use, listener-related factors, and speaker-related factors. The study analysis comprised several pilot studies designed to evaluate the psychometric characteristics of new measurement instruments, followed by the analysis of a structural model in an exploratory phase and a cross-validation phase. The exploratory phase examined the statistical fit and conceptual coherence of a structural model derived from previous research using a dataset composed of relatively heterogeneous groups of speakers and listeners. The initial model was evaluated using exploratory statistical techniques, including specification search, within the SEM framework (Bentler, 2006). Based on statistical and conceptual considerations, a revised model was proposed. The cross-validation phase used the SEM framework to evaluate the statistical fit of the revised model using an additional data set.

3.2 Participants

3.2.1 Speakers (international graduate students)

Speakers were sampled from test-takers who took the TOP for the first time between September 2010 and April 2012. All of the test-takers were international graduate students who had taken the TOP for the purpose of certification for TA duties. Written informed consent to utilize videos of test performances for research purposes was obtained from test-takers in accordance with institutional review board (IRB) procedures. Prior to sampling, videos were screened by a researcher in order to remove any that contained (a) personally-identifying information such as the test-taker's name written on a whiteboard, (b) poor visual or audio quality, or (c) repeat test-takers.

A total of 500 videos of test-taker TOP mini-lecture performances were included in the main dataset, which was stratified based on score level. A stratified sample based on score level was used instead of a random sample because distributions of TOP scores typically have a negative skew. In addition, the grand mean of TOP score distributions tend to be relatively high. In order to ensure more variability with respect to speaker oral proficiency, four scores levels were identified that roughly corresponded to decision categories based on TOP scores: (1) Fail, or TOP scores less than 6.4; (2) Provisional Pass, or TOP scores between 6.4 and 7.0; (3) Pass, or TOP scores between 7.1 and 7.7; and (4) High Pass, or TOP scores higher than 7.7. The target sample distribution based on score levels 1-4 was 17%, 33%, 33%, and 17%, respectively. Due to a lack of test-takers at lower score levels, the actual stratified sample distribution based on score level was 11%, 29%, 34%, and 25%, respectively.

The native languages (L1s) of test-takers in the sample corresponded to TOP operational norms, and included Chinese/Mandarin (55%), Korean (10%), Hindi (5%), Spanish (5%), and 25 others. While test-takers belong to dozens of departments throughout the university, a majority of them were associated with departments within the School of Engineering (56%), with the rest coming from departments within physical sciences (15%), life sciences (11%), humanities (9%), social sciences (8%), and business (1%).

3.2.2 Listeners (undergraduate students)

Listeners were recruited from the undergraduate population at UCLA during the Fall 2012, Winter 2013, and Spring 2013 academic quarters. E-mail invitations to participate in the study were forwarded to students by departments and student groups, and flyers for the study were posted on campus by departments and researchers. Students were recruited across a variety of departments and class levels in order to ensure that the sample of listeners varied across key

background variables such as familiarity with the speaker's topic, familiarity with the speaker's accent, familiarity and experience with ITAs and non-native speakers of English in general, and attitude homophily. Given that the sample of speakers contained so many graduate students from the School of Engineering, an effort was made to target undergraduates in these programs by contacting professors in the school of engineering who permitted researchers to make an announcement about the study and distribute flyers before their classes. Amongst the students who participated in the study, 65% were recruited via departmental e-mails, 12% via flyers posted on campus, 12% through word-of-mouth, 5% via student group e-mails, and 5% via social media.

A total of 205 students participated in the study. A majority (64%) of the participants were female. The median age of participants was 20 years (M=20.31, SD=2.69). The sample was heterogeneous with respect to participants' class level: 29% were first year students, 19% were second year, 32% were third year, 19% were fourth year, and 2% were fifth year or more. The participants came from 47 different departments across campus, with the largest numbers coming from Economics (13%), Mathematics (10%), Computer Science (8%), and Engineering (8%). A majority (55%) of participants were self-identified native speakers of English. Amongst those who self-identified as non-native speakers of English, nearly all (98%) described their level of listening comprehension in English as "proficient" or "highly proficient."

Prior to participating in the study, students provided informed consent to use the data collected throughout the study for research purposes in accordance with IRB procedures. Listeners were compensated for their participation in the study by receiving a $10 gift certificate immediately upon completion. In addition to the information provided by the listener-related

measurement instruments below, listeners were asked to provide demographic data such as their age, major, class level, and gender.

3.2.3 Raters

Trained TOP raters participated in separate procedures to obtain ratings for speaker-related factors. Ratings of language skills were obtained from operational ratings of TOP tasks (see section 3.2.3.1, below). Ratings of teaching effectiveness were obtained via a separate research project that utilized a different group of trained raters (see section 3.2.3.2, below).

3.2.3.1 Raters of TOP pronunciation, lexical-grammar, rhetorical organization, and question handling

TOP rater scores from operational rating were obtained for each speaker included in this study. TOP raters are graduate students with a background in linguistics, language teaching, and/or oral language assessment. They are recruited from a variety of departments at the university but the raters included in this study primarily belonged to the departments of Applied Linguistics and Linguistics. All raters attended an annual training session where they were required to pass a certification test. Before each testing session, raters complete calibration activities. Ratings of TOP subscores in the TOP data set were obtained from 20 raters, corresponding to the test-takers included in the sample.

3.2.3.2 Raters of teaching effectiveness

Raters of teaching effectiveness were certified TOP raters who also had teacher training and teaching experience (1 year or more). TOP Raters with teaching experience were identified by the TOP coordinator and invited to participate in a research project. Those who agreed to participate were compensated at a rate of $20 per hour. In order to ensure that ratings of teaching effectiveness would not be contaminated by prior exposure to a particular speaker, none

of the raters who rated speakers' TOP language skills (section 3.2.3.1, above) were included in this group. All raters completed a brief training session that included four calibration ratings. Following training, raters were individually debriefed by the researcher in order to ensure that the rating procedure was clear. Ratings of teaching effectiveness were obtained from 8 raters.

3.3 Measurement instruments

3.3.1 Test of Oral Proficiency (TOP) tasks

The TOP includes three tasks: a self-introduction, syllabus or assignment presentation, and mini-lecture. When the test-taker arrives for his or her scheduled test session, he or she is given an overview of the administration and scoring procedure by a test coordinator. After receiving the overview, the test-taker is given a copy of the Task 2 syllabus they will present and approximately 10 minutes to review the syllabus and take notes. After reviewing the syllabus, the test-taker is escorted to a test room by a coordinator and the test session begins. Each test room includes two undergraduate questioners, who interact with test-takers throughout the exam, and two raters, who document test-taker language use and score performances using the analytic rating scales.

Once a test-taker has been introduced to a test room by a coordinator, a rater or questioner asks the test-taker to introduce himself or herself. This self-introduction is intended to put the test-taker at ease and allow him or her to briefly interact with questioners without being scored. Questioners are encouraged to engage in small talk with the speaker. Questioners may ask follow-up questions based on the test-taker's self-introduction (e.g., "Where in China are you from?") or simple questions such as "What do you do in your free time?" or "What do you think of Los Angeles?" After several minutes of conversation, one of the raters instructs the test-taker to move to the second task.

During the second task, the test-taker presents a syllabus or assignment and is questioned by the undergraduates and scored by the raters. Test-takers are permitted to refer to the syllabus form for the duration of the task, and to utilize the classroom white board at their discretion. Questioners choose approximately eight questions from a pre-determined list of 12 questions supplied by the test coordinators for the syllabus given to the test-taker. The questions may be related to the syllabus (e.g., "Is the final exam cumulative?") or general classroom administration questions (e.g., "Will you have a review session before the exam?"). After approximately five minutes of presentation and questions, the test-taker is instructed to move to the third task by one of the raters.

During the final task, the speaker presents a mini-lecture on a topic of his or her choice, and is questioned by the undergraduates and scored by the raters. Speakers are instructed to choose a basic topic related to their field of study prior to the testing session. Since the specific topic that the speaker will present is not known to test administrators in advance, undergraduates are trained to ask types of questions (e.g., clarifying questions) rather than specific rote questions. Test-takers are not permitted to use any notes or any visual aids for this task, but are again permitted to utilize the classroom white board. After approximately ten minutes of presentation and questions, the test-taker is thanked by one of the raters and leaves the testing room.

Due to the constraints of the data collection, only the third task was included in the main study. Thus, the undergraduate in the sample only observed the test-taker performing the mini-lecture task.

3.3.2 Measures of speaker-related factors

Speaker-related factors can be divided into two overall categories: language use and teaching skills. The four aspects of language use that were included in this study included

pronunciation, lexical-grammar, rhetorical organization, and question handling. In addition, several rating scales that measure different teaching skills were included to investigate whether listener perceptions of domain-appropriate oral language use were influenced by a speaker's expertise in the domain.

3.3.2.1 Pronunciation, lexical-grammar, rhetorical organization, question handling

The four aspects related to speaker oral language use were rated using the TOP's analytic rating scale, which consists of separate holistic subscales for each aspect of language use. *Pronunciation*, or phonetic and phonological competence, was scored based on the frequency and type of segmental and suprasegmental errors. *Lexical-grammar*, or Lexical and grammatical competence, was scored based on the frequency and type of grammatical and lexical errors. *Rhetorical organization* was scored based on the speaker's use of language to organize discourse (e.g., cohesive devices). *Question handling* was scored based on the speaker's responses to undergraduate questions – in particular, the clarity and comprehensiveness of the response. The rating scale for each sub-construct ranges from 1-4 (see Appendix A).

Each speaker was scored by two raters using the four subscales for the mini-lecture task. Thus, each speaker received four scores for each subscale that range from 2-8 when summed. One problem with summing or averaging scores across raters is that it ignores dependencies that might exist between this facet of the measurement design. Generalizability theory may be used to isolate and identify facets of the measurement procedure that contribute to variance in scores (Brennan, 2001). Schmidgall (2011) examined the dependability of these rating scales across six datasets using univariate and multivariate G-studies and found that a trivial amount of variance was associated with the rating main effect. Slightly larger percentages of overall variance (6-15%) were associated with the person by rating interaction, particularly for the Rhetorical

organization and Question handling subscales. Given the intended use of these subscores for this study, however, the amount of variance attributed to ratings was considered low enough to justify the use of a summed score for each sub-construct ranging from 2-8.

3.3.2.2 Teaching effectiveness component measures

Based on the research into teaching effectiveness in the context of a mini-lecture task in the ITA language use domain (see Section 2.4), three components of teaching skills were included in the model: *Organization/Clarity*, *Enthusiasm*, and *Respect/rapport* (see Appendix B). Due to the constraints of the rating procedure in this study, which required evaluation based on a single observation of a mini-lecture task, most of the established scales reviewed could not be used without some modification. The *Organization/Clarity* scale consisted of 7 items and was formed by slightly modifying the wording of items from Patrick and Smart's (1998) *Organization and presentation skills* scale, and Heckert et al.'s (2006) *Pedagogic skill* scale. The *Enthusiasm* scale included 8 items and primarily consisted of items adapted from Hosek's (2011) *Nonverbal immediacy* scale. The *Respect/rapport* scale consisted of 6 items was formed by selecting and modifying items from Hosek's (2011) *Confirmation* scale, Marsh and Roche's (1997) *Individual rapport scale*, and Patrick and Smart's (1998) *Respect* scale.

In addition to the multi-item componential measures, four holistic ratings were included: one corresponding to each measure (*Organization/Clarity*, *Enthusiasm*, *Respect/rapport*), and one overall judgment of teacher effectiveness (see Appendix C). Raters completed the multi-item componential measures first, followed by the holistic items.

The purpose of including the holistic items was twofold. First, each holistic item could be considered a criterion variable for its relevant multi-item measure in order to provide additional information with which to evaluate the validity of the measure. If holistic ratings for

each component correlated highly with individual items and/or item total scores from relevant multi-item scales, this may provide additional evidence of construct validity. In addition, performance-based ratings of components of teaching effectiveness may have a low level of consistency across raters (i.e., low interrater reliability). Although the measurement properties of the multi-item componential scales are more desirable for use in the context of this study (i.e., they facilitate latent variable modeling), if a corresponding holistic rating is determined to have a much higher level of interrater consistency then the holistic rating may replace the componential scale in the model.

Thus, although teaching effectiveness scales used in this study have been formulated and justified based on previous research, their psychometric qualities (internal consistency, interrater consistency) and validity (correlation with criterion variables) were examined during the exploratory phase of the analysis in order to determine whether holistic or multi-item componential measures are more appropriate to include in the conceptual model. Due to financial constraints, these scales were not able to be pilot tested using a separate dataset as with the listener-based measures.

3.3.3 Measures of comprehensibility and listener-related factors

Two judgments of listener perceptions of speaker oral language use were collected, including impressionistic judgments of comprehensibility and domain-related evaluations. Other listener-based measures relate to familiarity with the speaker's accent and lecture topic, and listener attitude towards the speaker.

3.3.3.1 Comprehensibility

Following the definition of comprehensibility provided in section 1.4, comprehensibility was measured by five items that require the listener to indicate the perceived ease or difficulty

with which they understood the speaker (see Appendix D). Many previous studies that have adopted this definition of comprehensibility have measured it using a single Likert-type item (e.g., Derwing & Munro, 2009; Isaacs, 2010; Kennedy & Trofimovich, 2008; Munro & Erwing, 1998). Given this study's use of the SEM framework in which comprehensibility is represented as a latent variable, multiple indicators of the construct were necessary.

While previous studies have not attempted to represent comprehensibility as a latent variable, several studies have asked listeners to report their perceived comprehension based on multiple items that may be useful to inform the development of such a scale. In a criterion-related validity study[4] of the Test of Spoken English (TSE®), Powers, Schedl, Wilson-Leung, and Butler (1999) asked undergraduate listeners to respond to five types of items using Likert-type rating scales which indicated listener perceptions of (a) the effort required to understand the speaker, (b) confidence that the speaker was understood, (c) the degree to which the speaker's proficiency interfered with comprehension, (d) the level of persuasiveness of the speaker, and (e) whether the speaker fulfilled the task required. The first three types of items (a, b, c) are closely related to the construct of comprehensibility and were moderately to highly correlated (0.47 – 0.79).

Bridgeman, Powers, Stone, and Mollaun (2012) conducted a similar study for the TOEFL iBT® wherein undergraduate listeners responded to four types of items indicating perceptions of (a) the effort required to understand the speaker, (b) confidence that the speaker was understood, (c) the degree to which the speaker's interference interfered with comprehension, and (d) whether the speaker fulfilled the task required. Bridgeman et al (2012) found that the first three types of items (a, b, c) were very highly correlated (0.95 – 0.98) and aggregated all of the

[4] A criterion-related validity study examines the degree of correspondence between two similar indicators of the same construct. This is typically performed by looking at the correlation between scores on a particular test and corresponding scores on a measure of a similar construct for the same group of test-takers.

perceptual measures into a single total score. This summed score correlated very highly with listeners' comprehension scores based on more traditional selected-response listening comprehension items.

The five items contained in the comprehensibility measure in Appendix D were adapted based on Powers et al (1999) and Bridgeman et al. Previous research has utilized 5-, 7-, and 9-point Likert scales to measure comprehensibility, but at least one systematic study of scale use has indicated that listeners have difficulty distinguishing much more than five levels of comprehensibility (Isaacs, 2010). The items proposed here utilize a 6-point Likert scale in order to eliminate the middle category and form forced-choice items. The extreme points on the scale were clearly labeled.

3.3.3.2 Oral proficiency to TA

Listeners responded to four items that indicated their perceptions of whether the speaker has the language skills necessary to perform important tasks within the TLU domain (see Appendix E). These items have been adapted from Clark and Swinton (1980), who conducted a criterion-related validity study of the TSE® using undergraduate listeners. Again, the items proposed here utilize a 6-point Likert scale in order to eliminate the middle category and form forced-choice items. The extreme points on the scale were clearly labeled.

3.3.3.3 Familiarity with speaker's accent and native language

Listeners indicated their familiarity with the speaker's accent and native language using a 5-point Likert-type scale (see Appendices F and G). As part of the background questionnaire, listeners indicated their familiarity with a variety of accents and languages based on (a) the native language backgrounds of the speakers they viewed, and (b) additional accents which served as filler items. The extreme points on the scale were clearly labeled.

3.3.3.4 Familiarity with non-native speakers of English

To indicate their familiarity with non-native speakers of English (NNS) in general, listeners responded to the following question: "During a typical week, how frequently do you interact with non-native speakers of English?" Listeners responded using a 5-point item that indicated frequency: "Not at all," "Not very frequently," "Somewhat frequently," "Frequently," and "Very frequently."

3.3.3.5 Familiarity with ITAs

Listeners were asked to indicate the number of discussion or lab sections they had previously had with an ITA, and to describe their overall experience with ITAs in classes using a 5-point Likert-type scale that was labeled "Negative" at one end and "Positive" at the other. Listeners were also asked to provide any comments about their experiences with ITAs in classes in an optional open-ended item. The purpose of the latter item was to potentially identify students that had extremely strong negative or positive previous experiences with ITAs that may have an impact on the validity of their other ratings.

3.3.3.6 Familiarity with, complexity of, and interest in speaker's topic

Listeners' familiarity with the topic presented by the speaker during their interaction was measured using a 6-point Likert-type scale. Familiarity with the speaker's topic was self-reported in response to the following question: "How familiar were you with the topic and information presented in Task 3?" The extreme points on the scale were clearly labeled (1=No prior knowledge; 6=very familiar).

Listeners' interest in the topic presented by the speaker was measured by an item using the same format, in response to the following question: "How interested were you in the lecture

topic?" The extreme points of the scale were clearly labeled (1=Not interested, 6=Very interested).

Listeners' perception of the complexity of the topic presented by the speaker was also measured using a 6-point Likert-type scale, in response to the question: "How complex was the lecture topic?" The extreme points of the scale were clearly labeled (1=Not complex, 6=Very complex).

3.3.3.7 Attitude homophily

Attitude homophily will be measured using semantic differential items from McCroskey, Richmond, and Daly's (1975) attitude homophily subscale of their perceived homophily measure (see Appendix H). The four items in the scale include brief descriptors on each extreme of a 6-point Likert-type scale (e.g., *Thinks like me*/*Doesn't think like me*, *Similar to me*/*Different from me*).

3.3.3.8 Teacher personality

Student perceptions of the relevant personality characteristics of the teacher were measured using semantic differential items adapted from Coetzee-Van Rooy (2009). As shown in Appendix I, the five items in the scale include brief descriptors on each extreme of a 6-point Likert-type scale (e.g., *Friendly*/*Unfriendly*, *Active*/*Passive*).

3.4 Data collection procedure

Data were collected in two phases: a series of small-scale pilot studies, and a larger-scale data collection. The purpose of the first phase was to examine the psychometric properties of the newly developed scales (comprehensibility, oral skills to TA, attitude homophily, teacher personality) using convenience samples. The goal of the second phase was to evaluate the proposed conceptual model with a more representative sample of listeners from the target

language use domain using the measures that had been developed during the first phase. A description of the sampling plans, procedures, and proposed analyses for each phase follows.

3.4.1 Small-scale pilot studies

The pilot study was conducted using two convenience samples: one from a TOP administration prior during September 2012, and another from November 2012. Prior to each academic year, TOP raters and undergraduate questioners are required to undergo re-training. Following these training sessions, approximately 150 test-takers complete the TOP exam across 3 to 4 days. During their training session, undergraduate questioners completed an expanded version of the familiarity with speaker's accent measure that included an exhaustive list of possible TOP test-taker native languages based on previous administrations. This was necessary as TOP questioners are randomly assigned to test-takers during test administration. Undergraduates also completed the measure of listeners' familiarity with non-native speakers of English. During the training session, undergraduates were familiarized with the various ratings scales used to evaluate each test-taker (speaker). Eleven undergraduate Questioners and 15 raters participated.

The procedure for collecting data for the pilot study during the Fall 2012 TOP administration was designed to have no impact on the assessment procedure. In operational testing, TOP raters and questioners are scheduled for 3- or 4-hour testing sessions (AM, PM) wherein one test-taker is scheduled for evaluation every half hour. Each individual testing session takes between 15 and 25 minutes, and time between individual testing sessions is used for filling out rating scales and taking breaks. After each test, the listeners complete (a) the comprehensibility scale, (b) the oral skills to TA scale, (c) the attitude homophily scale, (d) the teacher personality scale (November 2012 administration only), and (e) the topic interest,

familiarity, and complexity scales. During each operational testing session, two trained raters assigned scores using the TOP's analytic rating scale (see section 3.3.1.1).

During the September 2012 administration, a total of 11 Questioners participated and 149 test-takers were evaluated. Most test-takers were evaluated by two Questioners. For the purpose of this pilot study, Questioners were randomly assigned to a 'Questioner 1' (Q1, n=149) or 'Questioner 2' (Q2, n=146) dataset. Thus, the datasets largely contained the same speakers (test-takers) but different listeners (Questioners). The structure of this convenience sample contained some undesirable characteristics (e.g., overlap between speakers, multiple ratings by the same listeners, and variation in the frequency of listeners' ratings) but is arguably useful for the purpose of this pilot study because the samples of speakers used in the main data collection were drawn from test administrations. All analyses were conducted for both datasets in order to examine the consistency of results across a different group of listeners.

During the November 2012 TOP administration, a total of 12 TOP Questioners participated and 78 test-takers were evaluated. Most test-takers were evaluated by two Questioners. For the purpose of this pilot study, Questioners were randomly assigned to a 'Questioner 1' (Q1, n=74) or 'Questioner 2' (Q2, n=71) dataset. All analyses were conducted for both datasets in order to examine the consistency of results across a different group of listeners.

Data obtained from the two pilot studies were analyzed to examine whether every item in the scale was functioning as intended as well as to investigate the factor structure of scales. Thus, the scales were analyzed to investigate their internal consistency, dimensionality, interrater consistency, and relationship with important constructs.

3.4.2 Main data collection

The design of the larger-scale data collection differed from the pilot study in several important aspects. First, it was aimed at recruiting a larger and more diverse sample of undergraduate listeners than TOP administrations can provide. Second, listeners in the large-scale data collection observed speaker performances on video and provided their responses using a computer-based survey, while listeners in the pilot studies were undergraduate participants during TOP administrations. Finally, listeners in the large-scale data collection were exposed to a sample of speech from only one TOP task, the mini-lecture in TOP Task 3.

Each undergraduate survey lasted approximately 50 minutes and required participants to view and evaluate two randomly-assigned TOP Task 3 performances. A diagram of the procedure for collecting the data in the large-scale study is given in Figure 3.1 (below).

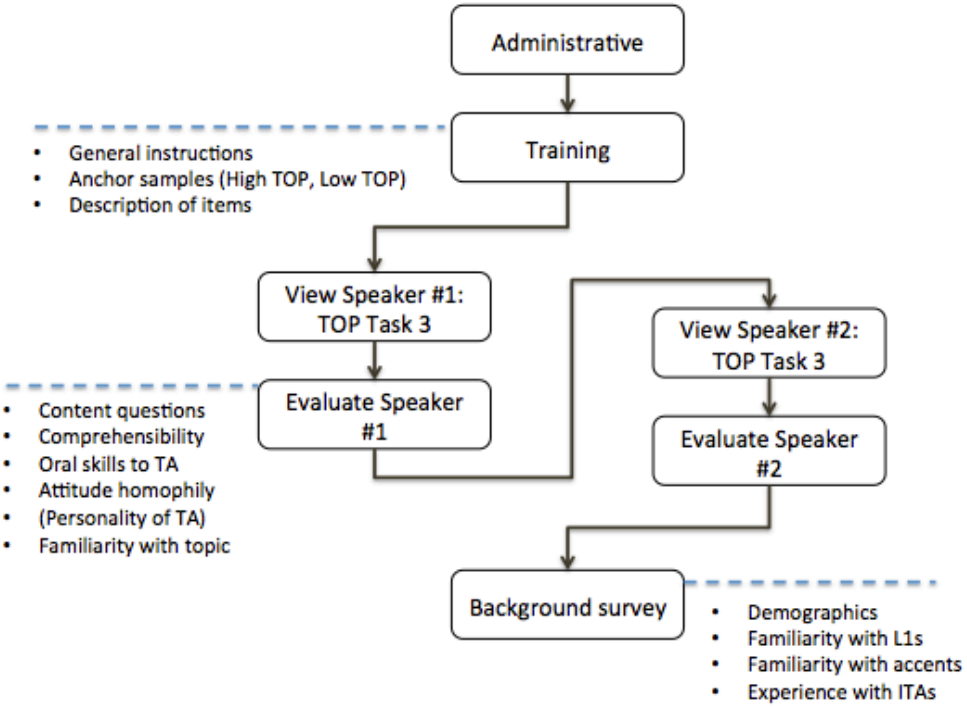

Figure 3.1. Procedure followed in the main study.

First, each participant read a brief description of the purpose of the study and completed an informed consent form. Participants then were given an ID and password that were used to log on to a website. Once the participant logged in, he or she read a brief description of the purpose of the study (see Appendix J). Next, the participant watched a brief video clip of a high-performing TOP test-taker (Sample video #1) and read a description identifying the speaker as easy to understand or comprehensible. The participant then watched different video clip of a low-performing TOP test-taker (Sample video #2) and read an accompanying description identifying the speaker as difficult to understand. After viewing the anchoring samples, the participant read a description of the types of questions he or she would be asked after viewing each video and was encouraged to provide careful and honest responses.

After completing instructions and training, participants viewed a 6-10 minute video clip (Speaker #1). Once the clip was viewed in its entirety, participants completed the content questions, comprehensibility scale, oral skills to TA scale, attitude homophily scale, teacher personality scale, and topic familiarity items. Next, the participant viewed another 6-10 minute video clip for another speaker (Speaker #2). The participant then completed the same measures described previously for the first video.

Once both videos had been viewed and evaluated, participants were given a background survey to complete that included demographic items (gender, age, academic department, class level, status as a native speaker of English), language familiarity questions (see Appendix D), accent familiarity questions (see Appendix E), and experience with ITAs and non-native speakers of English in general. Upon completion, participants were thanked and compensated with a $10 gift certificate.

An additional step was added to the data collection procedure for the main study to solicit participants to investigate undergraduates' decision-making processes during the rating process. As part of the background survey, participants indicated whether they would be interested in participating in an interview-based follow-up study that asked them more specific questions about how they evaluated ITAs' comprehensibility. Twenty-four participants who expressed interest and met selection criteria based on the exploratory analysis (see section 4.2.1.5) were invited for follow-up interviews. Among those invited, 17 completed the follow-up interviews (see section 4.2.1.6). The purpose of this procedure was to supplement the primarily quantitative results with more qualitative descriptions of participants' decision-making processes.

3.5 Data analysis for the main study

After the larger-scale data collection was completed, the dataset was segmented into two datasets of equal size. Each dataset contained the same group of listeners but a unique group of speakers. The first dataset was used to conduct exploratory data analysis based on the prosed model while the second dataset was used to cross-validate a revised model.

3.5.1 Exploratory data analysis

An exploratory data analysis was conducted using the framework of structural equation modeling (SEM) and the computer program *EQS 6.2* (Bentler, 2006) with the first dataset. First, the data was cleaned to ensure its validity. Patterns of responses to scale items were examined to identify possible cases or responses that may be considered errors or noise. Participants who self-identified as non-native speakers of English with low levels of listening comprehension were flagged for possible removal from the dataset. Participants who provided inconsistent or contradictory responses with respect to their ratings of familiarity with languages and self-

reported proficiency with various languages were also flagged for possible removal. If a participant indicated that he or she was already familiar with a speaker in one of the videos he or she watched, that case was flagged for removal to ensure that all participants listened to unfamiliar speakers.

Once the data were cleaned, descriptive statistics were provided for each scale, including means, standard deviations, estimates of skewness and kurtosis, and histograms. When appropriate, transformations were made to variables when large estimates of skewness or kurtosis indicated severe departures from normality. Researchers have found that when univariate non-normality is slight or moderate, transformation may not substantially reduce univariate skewness or kurtosis (Gao, Mokhtarian, & Johnston, 2008; Muthen & Kaplan, 1985). Muthen and Kaplan (1985) observed that when univariate skewness or kurtosis is smaller in absolute value than 1.0, distortions of ML chi-squares and standard error tend to be minimal. In addition, variable transformations may improve normality but complicate relationships within a hypothesized model in SEM (Gao, Mokhtarian, & Johnston, 2008). When variables were transformed, revised descriptive statistics were reported.

Mardia's coefficient was estimated in order to investigate multivariate normality. Potential outliers were identified using EQS. Any items flagged were examined to determine whether they should be excluded from the dataset. Harlow (1985) found that larger values of multivariate kurtosis (Mardia's coefficient > 7.98) were most likely to lead to biases in standard error estimates, as opposed to univariate skewness and kurtosis.

First, measurement models were specified and evaluated for all latent variables in the structural model. Since the Teaching effectiveness measures were not able to be evaluated in a pilot study, exploratory and cross-validation analyses of its dimensionality were conducted.

Although it was not included in the preliminary structural model, the measurement model for Teacher personality was also evaluated during this phase of the analysis.

Next, the structural model was specified and evaluated. All models were estimated using the Maximum Likelihood (ML) procedure in the computer program *EQS 6.2*. ML performs well for continuous data that is multivariate normally distributed, and can be applied with relatively small samples (Bentler, 2006).

The initial model proposed earlier in section 2.4 is reproduced as a structural model in Figure 3.2, below.

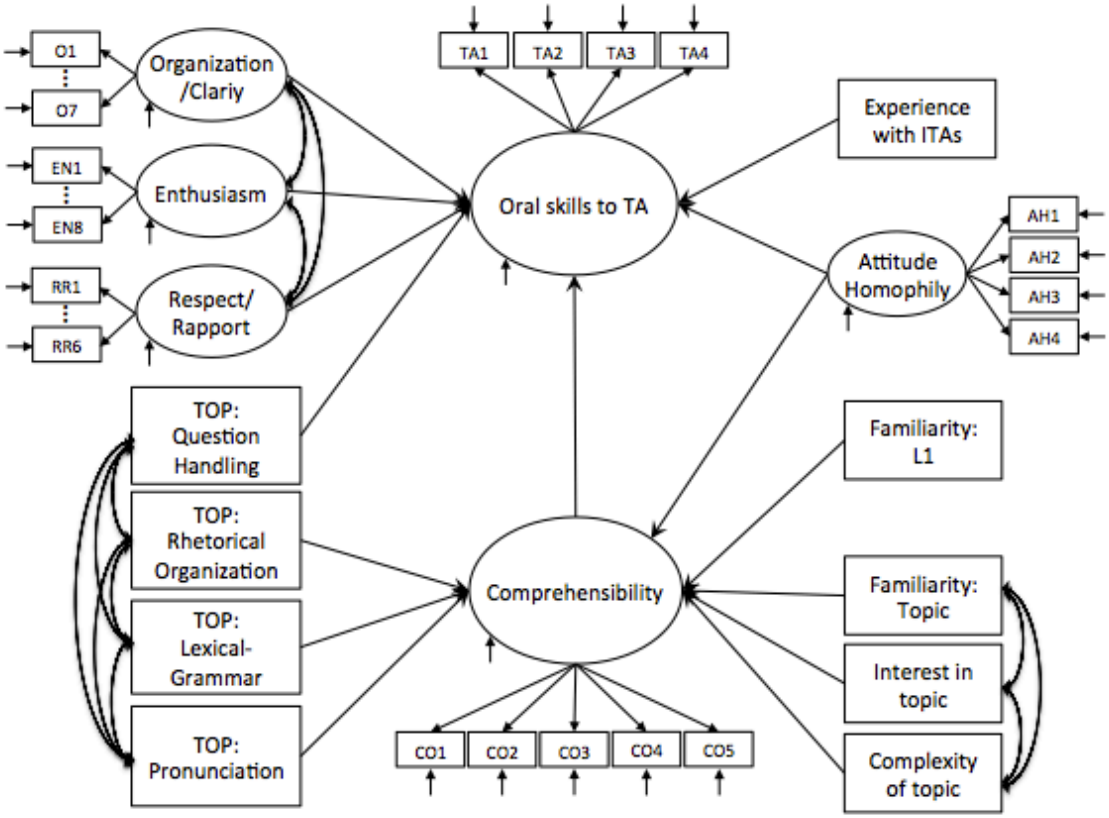

Figure 3.2. Preliminary SEM model specifying the relationships between listener perceptions of oral proficiency, and speaker- and listener-related factors.

The fit of the model was examined using the χ^2 statistic and several residual-based (RMSEA, SRMR) and comparative (CFI, NNFI) fit indices. In order to investigate possible revisions to the model based on statistical considerations, the Wald test and Lagrange multiplier test were employed using *EQS*. Output from the Wald test indicates whether removing paths between variables or factors in the model would explain more variance in the data, while output from the Lagrange multiplier test indicates whether adding paths would improve the fit of the model.

Follow-up interviews with a sample of listeners were conducted in order to further examine listener perceptions and decision-making processes during the survey. Questions for listeners who participated in the semi-structured follow-up interviews were designed to address concerns raised by the exploratory analysis.

Based on statistical fit and conceptual coherence, a revised model was proposed and compared with the initial model. Data collected from the semi-structured interviews with listeners were used in conjunction with expert recommendations derived from a review of relevant literature to justify any conceptual alterations to the initial model.

3.5.2 Cross-validation data analysis

The statistical fit of the revised model proposed at the conclusion of the exploratory analysis was examined using the second dataset. Data were cleaned and screened for outliers and the assumptions of univariate and multivariate normality were investigated as described above. The model was estimated using ML and model fit will be examined using the indices described above.

3.6 Software used for data analyses

Four computer software programs were used for the various phases of data analyses: Microsoft Excel (2011), *R* (R Development Core Team, 2008), Edu-G (Cardinet, Johnson, & Pini, 2010), and EQS 6.2 (Bentler, 2006). Microsoft Excel was used to input and organize data. *R* was used to produce descriptive statistics, investigate univariate normality, investigate bivariate relationships (including the estimation of intraclass correlation coefficients), investigate the internal consistency of scales (e.g., coefficient alpha), and transform variables. Edu-G was used to estimate variance components and generalizability coefficients for G-study designs. EQS 6.2 was used for CFA and SEM analyses.

Chapter 4: Results

4.1 Pilot studies

4.1.1 Pilot study 1

The internal consistency of each of three listener rating scales (Comprehensibility, Oral skills to TA, Attitude homophily) was investigated in order to evaluate their psychometric quality. In addition, correlations between scale total scores and TOP scaled scores were examined to determine if they were functioning as expected. Results suggest that each scale was measuring a unidimensional construct, exhibited satisfactory internal consistency, and correlated with criterion variables as expected.

4.1.1.1 Pilot study 1: Dataset

First, reverse-keyed items were transformed so all items within a scale were oriented in the same direction. Next, patterns of responses to items within a scale within pairs of Questioners for the same test-taker were examined. Extreme discrepancies were noted (e.g., Questioner 1 gave test-taker a '1' and Questioner 2 gave test-taker a '6') and edited when patterns of responses clearly indicated that a mistake had been made. For example, if Questioner 1's pattern of responses to the four Comprehensibility items was '1-6-6-6', and Questioner 2's pattern of responses was '6-5-6-6', the '1' response was changed to '6' as it was clearly logically inconsistent and occurred in a reverse-keyed item. Such alterations were only justified in extreme cases (i.e., '6-1-6-6', not '4-3-4-4'). For Comprehensibility items, this rarely occurred; for TA items, it occurred more frequently. This suggested that more Questioners were not reading items closely in the TA scale, as these mistakes occurred in reverse-keyed items.

For each Questioner, the percentage of adjacent or exact agreement for items within each scale was estimated in order to identify Questioners who may have used scales inappropriately.

For the Attitude homophily scale, two Questioners consistently treated all items in the scale as one item by scoring each item in the same way, regardless of whether it was reverse-keyed or not. Their response to this scale typically included a single response across all items (i.e., a single large circle that spanned items). These Questioners' responses to items in the Attitude homophily scale were removed from the data due to concerns about the validity of their data.

4.1.1.2 Pilot study 1: Comprehensibility scale

4.1.1.2.1 Descriptive statistics

Descriptive statistics for the four items in the Comprehensibility scale for each dataset are shown in Table 4.1, below.

Table 4.1

Descriptive Statistics for Items in the Comprehensibility Scale for Pilot Study 1

	Q1 dataset (n=149)				Q2 dataset (n=146)			
Item	*M* (*SD*)	range	skew	kurtosis	*M* (*SD*)	range	skew	kurtosis
CO1	4.44 (1.60)	1-6	-0.59	-0.99	4.35 (1.46)	1-6	-0.36	-1.10
CO2	4.87 (1.24)	1-6	-1.06	0.53	4.63 (1.32)	2-6	-0.53	-1.07
CO3	5.19 (1.14)	1-6	-1.58	2.17	4.83 (1.29)	1-6	-1.00	0.03
CO4	4.68 (1.27)	1-6	-0.70	-0.13	4.47 (1.34)	1-6	-0.54	-0.72

Items consistently exhibited a negative skew, but there were no gross violations of normality assumptions. One explanation for the relatively high Comprehensibility item means and negatively skewed distributions may be that the sample of speakers used in this pilot study were relatively proficient speakers, as indicated by their TOP scaled scores. Comprehensibility ratings have been shown to be positively correlated with oral proficiency scores (Schmidgall,

2012), and oral proficiency scores for this sample were relatively high, with a negatively skewed distribution.

4.1.1.2.2 Internal consistency

In order to examine the internal consistency of the scale, Cronbach's alpha was estimated for each scale and relevant item characteristics (intercorrelations, item-total correlations, alpha if deleted) are reported below.

The internal consistency of the scale as indicated by Cronbach's alpha was 0.952 for the Q1 dataset, and 0.936 for the Q2 dataset.

The correlations among scale items are presented in Table 4.2, below. Intercorrelations for the Q1 dataset are shown in the lower diagonal, and intercorrelations for the Q2 dataset are shown in the upper diagonal.

Table 4.2

Correlations among Items in the Comprehensibility Scale for Pilot Study 1

	CO1	CO2	CO3	CO4
CO1	1.00	0.85	0.71	0.81
CO2	0.84	1.00	0.76	0.79
CO3	0.81	0.85	1.00	0.81
CO4	0.89	0.88	0.84	1.00

Item intercorrelations appear to be slightly higher in the Q1 dataset. Overall, however, the magnitude of item intercorrelations was similar across datasets and item pairs.

Item statistics are shown in Table 4.3, below.

Table 4.3

Item-total Correlations for Items in the Comprehensibility Scale for Pilot Study 1

	Q1 dataset		Q2 dataset	
Item	Item-total correlation	Alpha if deleted	Item-total correlation	Alpha if deleted
CO1	0.893	0.945	0.851	0.917
CO2	0.898	0.933	0.870	0.909
CO3	0.870	0.944	0.808	0.929
CO4	0.924	0.925	0.870	0.909

Overall, the measure appears to have high internal consistency as indicated by strong item inter-correlations, and Cronbach's alpha. None of the items appear particularly problematic, as evidenced by strong item-total correlations and lowered estimates of alpha if an item is deleted.

4.1.1.2.3 Interrater consistency

Interrater consistency was estimated for individual items in the scale and the total score using the intraclass correlation coefficient (ICC). Based on previous research, moderate to strong intraclass correlations were expected (0.50 – 0.80). ICCs were estimated using the computer program *R* and are presented in Table 4.4, below.

Table 4.4

ICCs for Items in the Comprehensibility Scale for Pilot Study 1

Item CO1	Item CO2	Item CO3	Item CO4	Total score
0.50**	0.48**	0.30**	0.38**	0.49**

** $p < .01$

ICCs for individual items were slightly lower than expected (0.30-0.50). However, this is not necessarily a problematic finding. For this scale, raters are expected to vary in their response to the same speaker based on a number of different factors.

4.1.1.2.4 Correlations with criterion variables

Based on previous research, it was expected that comprehensibility total scores would have moderate or strong correlations with TOP scaled scores (oral proficiency) and oral skills to TA total scores, and moderate to weak correlations with attitude homophily total scores. Correlations between total scores on these scales for each dataset are shown in Table 4.5, below.

Table 4.5

Correlations between Comprehensibility Total Scores and Criterion Variables for Pilot Study 1

	Q1 dataset			Q2 dataset		
	TOP	Oral skills to TA	Attitude homophily	TOP	Oral skills to TA	Attitude homophily
Comprehensibility	0.55**	0.90**	0.44**	0.66**	0.87**	0.47**

** $p < .01$

As expected, there was a strong correlation between comprehensibility and oral skills to TA total scores (Q1 r=0.90; Q2 r=0.87). Both of these rating scales measure perceptions of aspects of oral proficiency, and thus were expected to be highly correlated. The correlations between comprehensibility total scores and oral proficiency scores (TOP) were slightly lower than expected (Q1 r=0.55; Q2 r=0.66). One possible explanation for this discrepancy may be the restricted range of TOP scores found in this sample, which may have attenuated the relationship. The moderately strong relationships observed between comprehensibility and attitude homophily total scores (Q1 r=0.44; Q2 r=0.47) were as expected.

4.1.1.3 Pilot study 1: Oral skills to TA scale

4.1.1.3.1 Descriptive statistics

Descriptive statistics of the four items in the Oral skills to TA scale for each dataset are shown in Table 4.6, below.

Table 4.6

Descriptive Statistics for Items in the Oral Skills to TA Scale for Pilot Study 1

	Q1 dataset (n=149)				Q2 dataset (n=146)			
Item	*M* (*SD*)	range	skew	kurtosis	*M* (*SD*)	range	skew	kurtosis
TA1	5.17 (1.18)	1-6	-1.56	2.03	4.83 (1.29)	2-6	-0.73	-0.80
TA2	4.80 (1.36)	1-6	-1.07	0.35	4.64 (1.38)	1-6	-0.62	-0.91
TA3	5.19 (1.11)	1-6	-1.66	2.84	5.01 (1.19)	1-6	-1.14	0.52
TA4	5.01 (1.21)	1-6	-1.11	0.47	4.80 (1.26)	2-6	-0.72	-0.71
TA5	4.91 (1.12)	1-6	-1.03	0.83	4.71 (1.27)	1-6	-0.76	-0.23

As in the Comprehensibility measure, items consistently exhibited a negative skew, but there were no gross violations of normality assumptions. Again, the characteristics of the sample of speakers (higher level of speaking proficiency as indicated by TOP scores) provide a plausible explanation for the negative skew.

4.1.1.3.2 Internal consistency

In order to examine the internal consistency of the scale, Cronbach's alpha was estimated for each scale and relevant item characteristics (intercorrelations, item-total correlations, alpha if deleted) are reported below.

The internal consistency of the scale as indicated by Cronbach's alpha was 0.931 for the Q1 dataset, and 0.929 for the Q2 dataset.

The correlations among scale items are presented in Table 4.7, below. Intercorrelations for the Q1 dataset are shown in the lower diagonal, and intercorrelations for the Q2 dataset are shown in the upper diagonal.

Table 4.7

Correlations among Items in the Oral Skills to TA Scale for Pilot Study 1

	TA1	TA2	TA3	TA4	TA5
TA1	1.00	0.83	0.60	0.75	0.77
TA2	0.82	1.00	0.57	0.76	0.84
TA3	0.65	0.61	1.00	0.64	0.68
TA4	0.80	0.84	0.62	1.00	0.80
TA5	0.73	0.74	0.75	0.78	1.00

Item intercorrelations were strong across most item pairs for both datasets. The correlations between Item 3 and other items were slightly lower in both datasets.

Item statistics are shown in Table 4.8, below.

Table 4.8

Item-total Correlations for Items in the Oral Skills to TA Scale for Pilot Study 1

	Q1 dataset		Q2 dataset	
Item	Item-total correlation	Alpha if deleted	Item-total correlation	Alpha if deleted
TA1	0.847	0.910	0.833	0.909
TA2	0.847	0.912	0.852	0.906
TA3	0.713	0.935	0.675	0.938
TA4	0.862	0.907	0.834	0.909
TA5	0.839	0.913	0.880	0.900

Overall, the measure appears to have high internal consistency as indicated by strong item inter-correlations, and Cronbach's alpha. In both datasets, Item 3 had a lower item-total correlation than other items, and the estimate of alpha would slightly improve if the item were to be deleted.

4.1.1.3.3 Interrater consistency

Interrater consistency was estimated for individual items in the scale and the total score using the intraclass correlation coefficient (ICC). Moderate correlations were expected, as students are expected to differ in their assessments of the oral language use and teaching skills of a speaker. ICCs were estimated using the computer program *R* and are presented in Table 4.9, below.

Table 4.9

ICCs for items in the Oral Skills to TA Scale for Pilot Study 1

Item TA1	Item TA2	Item TA3	Item TA4	Item TA5	Total score
0.30**	0.33**	0.42**	0.27**	0.51**	0.43**

** $p < .01$

Interrater correlations for individual items were slightly lower than expected (0.27-0.51). Again, this is not necessarily a problematic finding. For this scale, raters are expected to vary in their response to the same speaker based on a number of different factors.

4.1.1.3.4 Correlations with criterion variables

Based on previous research, it was expected that oral skills to TA total scores would have moderate or strong correlations with oral proficiency ratings (TOP) and comprehensibility total scores, and moderate to weak correlations with attitude homophily total scores. For each dataset, correlations between total scores on these scales are shown in the table below.

Table 4.10

Correlations between Oral Skills to TA Total Scores and Criterion Variables for Pilot Study 1

	Q1 dataset			Q2 dataset		
	TOP	COMP	Attitude homophily	TOP	COMP	Attitude homophily
Oral skills to TA	0.48**	0.90**	0.37**	0.65**	0.87**	0.38**

Note. COMP = Comprehensibility.

** $p < .01$

As observed earlier, there were a strong correlation between comprehensibility total scores and oral skills to TA ratings (Q1 r=0.90; Q2 r=0.87). The correlation between oral skills to TA total scores and oral proficiency ratings (TOP) was slightly lower than expected (Q1 r=0.48; Q2 r=0.65). Again, a possible explanation for this discrepancy may be the restricted range of TOP scores found in this sample. The moderately weak relationship observed between oral skills to TA and attitude homophily total scores (Q1 r=0.37; Q2 r=0.38) was as expected.

4.1.1.4 Pilot study 1: Attitude homophily scale

4.1.1.4.1 Descriptive statistics

Descriptive statistics for the four items in the Attitude homophily scale for each dataset are shown in Table 4.11, below.

Table 4.11

Descriptive Statistics for Items in the Attitude Homophily Scale for Pilot Study 1

	Q1 dataset (n=121)				Q2 dataset (n=140)			
Item	*M* (*SD*)	range	skew	kurtosis	*M* (*SD*)	range	skew	kurtosis
AH1	3.59 (1.07)	1-5	-0.63	-0.02	3.81 (1.14)	1-6	-0.47	-0.31
AH2	3.51 (1.10)	1-6	-0.16	0.05	3.68 (1.13)	1-6	-0.50	-0.15
AH3	3.31 (0.94)	1-6	0.48	0.61	3.46 (1.12)	1-6	-0.20	-0.40
AH4	3.42 (0.97)	1-6	0.00	0.08	3.69 (1.05)	1-6	-0.48	0.02

Based on estimates of skew and kurtosis, no gross violations of the assumption of normality were observed for either dataset.

4.1.1.4.2 Internal consistency

In order to examine the internal consistency of the scale, Cronbach's alpha was estimated for each scale and relevant item characteristics (intercorrelations, item-total correlations, alpha if deleted) are reported below.

The internal consistency of the scale as indicated by Cronbach's alpha was 0.880 for the Q1 dataset, and 0.847 for the Q2 dataset.

The correlations between scale items are presented in Table 4.12, below. Intercorrelations for the Q1 dataset are shown in the lower diagonal, and intercorrelations for the Q2 dataset are shown in the upper diagonal.

Table 4.12

Correlations among Items in the Attitude Homophily Scale for Pilot Study 1

	AH1	AH2	AH3	AH4
AH1	1.00	0.62	0.46	0.71
AH2	0.62	1.00	0.50	0.58
AH3	0.54	0.53	1.00	0.63
AH4	0.75	0.72	0.76	1.00

Item intercorrelations ranged from moderate to strong (0.46 – 0.76) across item pairs and datasets.

Item statistics are shown in Table 4.13, below.

Table 4.13

Item-total Correlations for Items in the Attitude Homophily Scale for Pilot Study 1

	Q1 dataset		Q2 dataset	
Item	Item-total correlation	Alpha if deleted	Item-total correlation	Alpha if deleted
AH1	0.723	0.854	0.704	0.797
AH2	0.704	0.863	0.662	0.815
AH3	0.674	0.871	0.606	0.838
AH4	0.880	0.794	0.772	0.770

Overall, the measure appears to have acceptably high internal consistency as indicated by moderate to strong item inter-correlations, and reasonably high estimates of Cronbach's alpha. None of the items appear particularly problematic, as evidenced by moderately strong item-total correlations and lowered estimates of alpha if an item is deleted.

4.1.1.4.3 Interrater consistency

Interrater consistency was estimated for individual items in the scale and the total score using the intraclass correlation coefficient (ICC). Weak or nonsignificant interrater correlations were expected for items in this scale. ICCs are presented in Table 4.14, below.

Table 4.14

ICCs for Items in the Attitude Homophily Scale for Pilot Study 1

Item AH1	Item AH2	Item AH3	Item AH4	Total score
0.16*	0.08	0.07	0.22**	0.16*

** $p < .05$*

*** $p < .01$*

ICCs for individual items were weak, as expected (0.07-0.22).

4.1.1.4.4 Correlations with criterion variables

Based on previous research, it was expected that attitude homophily total scores would have moderate or weak correlations with oral proficiency scores (TOP), comprehensibility total scores, and oral skills to TA total scores. Correlations between total scores on these scales are shown in Table 4.15, below.

Table 4.15

Correlations between Attitude Homophily Total Scores and Criterion Variables for Pilot Study 1

	Q1 dataset			Q2 dataset		
	TOP	COMP	Oral skills to TA	TOP	COMP	Oral skills to TA
Attitude homophily	0.39**	0.44**	0.37**	0.34**	0.47**	0.38**

Note. COMP = Comprehensibility.

*** $p < .01$*

As observed earlier, there were moderate correlations between attitude homophily and comprehensibility total scores (r=0.44) and oral skills to TA total scores (r=0.37). The correlation between attitude homophily total scores and oral proficiency ratings (TOP) was also moderate, as expected (r=0.39).

4.1.1.5 Pilot study 1: Discussion

Overall, all three scales exhibited acceptable psychometric characteristics, as indicated by estimates of internal consistency and correlations with criterion variables. One of the items in the Oral skills to TA scale (item 3) appeared to attenuate the estimate of internal consistency, but was retained as it was considered important to the construct definition. This item was distinct from others in the scale in that it focused more on the speaker's listening skills, while other items focused primarily on oral skills. Since the construct of Oral skills to TA emphasized the interactional nature of speaking in this domain, the item focused on the speaker's aural skills was retained in order to preserve this aspect of the construct.

4.1.2 Pilot study 2

A second pilot study was conducted with to examine how the scales functioned with another sample of speakers with the following modifications. Given the centrality of the construct of comprehensibility in the hypothesized conceptual model, an additional item was added to the Comprehensibility scale in an attempt to further increase its internal consistency. In addition, a scale designed to measure listener perceptions of speaker personality characteristics (Teacher personality scale) was piloted with the new sample of speakers.

The internal consistency of two scales (revised Comprehensibility, Teacher personality) was investigated in order to evaluate their psychometric quality. In addition, correlations between scales and important variables were examined to determine if they were functioning as

expected. Results suggest that each scale is measuring a unidimensional construct, exhibits satisfactory internal consistency, and correlates with criterion variables as expected.

4.1.2.1 Pilot study 2: Dataset

First, reverse-keyed items were transformed so all items within a scale were oriented in the same direction. Next, patterns of responses to items within a scale within pairs of Questioners for the same test-taker were examined. Extreme discrepancies were noted (e.g., Questioner 1 gave test-taker a '1' and Questioner 2 gave test-taker a '6') and transformed when patterns of responses suggested that a mistake was made. For each Questioner, the percentage of adjacent or exact agreement for items within each scale was estimated in order to identify Questioners who may have used scales inappropriately. For the Attitude homophily scale, two Questioners consistently treated all items in the scale as one item by scoring each item in the same way, regardless of whether it was reverse-keyed. Their response to this scale typically included a single response across all items (i.e., a single large circle that spanned items). These Questioners' responses to items in the Attitude homophily scale were removed from the data due to concerns about the validity of their data.

4.1.2.2 Pilot study 2: Revised Comprehensibility scale

4.1.2.2.1 Descriptive statistics

The revised Comprehensibility scale included the addition of one more item, Item 5. Descriptive statistics for the five items in the revised Comprehensibility scale for each dataset are shown in Table 4.16, below.

Table 4.16

Descriptive Statistics for Items in the Revised Comprehensibility Scale for Pilot Study 2

	Q1 dataset (n=74)				Q2 dataset (n=71)			
Item	*M* (*SD*)	range	skew	kurtosis	*M* (*SD*)	range	skew	kurtosis
CO1	4.04 (1.52)	1-6	-0.16	-1.29	4.27 (1.49)	1-6	-0.48	-0.82
CO2	4.11 (1.47)	1-6	-0.16	-1.16	4.49 (1.33)	1-6	-0.57	-0.50
CO3	4.24 (1.47)	1-6	-0.37	-1.20	4.68 (1.39)	1-6	-0.92	-0.13
CO4	4.14 (1.42)	1-6	-0.21	-1.15	4.22 (1.38)	1-6	-0.20	-1.02
CO5	4.20 (1.41)	1-6	-0.38	-1.06	4.54 (1.15)	2-6	-0.36	-0.84

Items consistently exhibited a negative skew, but there were no gross violations of normality assumptions. One explanation for the relatively high Comprehensibility item means and negatively skewed distributions may involve the sample of speakers used in this pilot study. Comprehensibility ratings have been shown to be positively correlated with oral proficiency scores (Schmidgall, 2012), and oral proficiency scores for this sample were relatively high, with a negatively skewed distribution.

4.1.2.2.2 Internal consistency

In order to examine the internal consistency of the scale, Cronbach's alpha was estimated for each scale and relevant item characteristics (intercorrelations, item-total correlations, alpha if deleted) are reported below.

The internal consistency of the scale as indicated by Cronbach's alpha was 0.973 for the Q1 dataset, and 0.963 for the Q2 dataset.

The correlations between scale items are presented in Table 4.17, below. Intercorrelations for the Q1 dataset are shown in the lower diagonal, and intercorrelations for the Q2 dataset are shown in the upper diagonal.

Table 4.17

Correlations among Items in the Revised Comprehensibility Scale for Pilot Study 2

	CO1	CO2	CO3	CO4	CO5
CO1	1.00	0.89	0.83	0.87	0.87
CO2	0.93	1.00	0.83	0.86	0.85
CO3	0.88	0.88	1.00	0.81	0.79
CO4	0.89	0.88	0.92	1.00	0.84
CO5	0.89	0.85	0.83	0.85	1.00

Item intercorrelations appear to be slightly higher in the Q1 dataset. This is expected given that the estimate of alpha was slightly higher for the Q1 dataset. However, all of the item intercorrelations are strong (0.71 – 0.89).

Item statistics are shown in Table 4.18, below.

Table 4.18

Item-total Correlations for Items in the Revised Comprehensibility Scale for Pilot Study 2

	Q1 dataset		Q2 dataset	
Item	Item-total correlation	Alpha if deleted	Item-total correlation	Alpha if deleted
CO1	0.946	0.963	0.925	0.950
CO2	0.929	0.966	0.918	0.951
CO3	0.918	0.967	0.862	0.960
CO4	0.930	0.966	0.901	0.954
CO5	0.889	0.972	0.893	0.957

Overall, the measure appears to have high internal consistency as indicated by strong item inter-correlations, and Cronbach's alpha. None of the items appear particularly problematic, as evidenced by strong item-total correlations and lowered estimates of alpha if an item is deleted.

4.1.2.2.3 Interrater consistency

Interrater consistency was estimated for individual items in the scale and the total score using the intraclass correlation coefficient (ICC). Based on previous research and Pilot study 1, moderate ICCs were expected (0.30 – 0.60). Correlations are presented in Table 4.19, below.

Table 4.19

ICCs for Items in the Revised Comprehensibility Scale for Pilot Study 2

Item CO1	Item CO2	Item CO3	Item CO4	Item CO5	Total score
0.43**	0.47**	0.34**	0.36**	0.40**	0.45**

** $p < .01$

ICCs for individual items were moderate, as expected (0.34-0.47).

4.1.2.2.4 Correlations with criterion variables

Based on previous research and Pilot study 1, it was expected that comprehensibility ratings would have moderate correlations with oral proficiency scores (TOP) and oral skills to TA total scores, and moderate to weak correlations with attitude homophily total scores. Correlations between total scores on these scales for each dataset are shown in Table 4.20, below.

Table 4.20

Correlations between Comprehensibility Total Scores and Criterion Variables for Pilot Study 2

	Q1 dataset			Q2 dataset		
	TOP	Oral skills to TA	Attitude homophily	TOP	Oral skills to TA	Attitude homophily
Comprehensibility	0.61**	0.94**	0.48**	0.43**	0.93**	0.29*

* $p < .05$

** $p < .01$

As expected, there was a strong correlation between comprehensibility and oral skills to TA total scores (Q1 r=0.94; Q2 r=0.93). Both of these rating scales measure perceptions of aspects of oral proficiency, and are expected to have strong correlations. The correlations between comprehensibility total scores and oral proficiency scores (TOP) were moderate, as expected (Q1 r=0.61; Q2 r=0.43). The moderate to weak relationships observed between comprehensibility and attitude homophily total scores (Q1 r=0.48; Q2 r=0.29) were as expected.

4.1.2.3 Pilot study 2: Teacher personality scale

4.1.2.3.1 Descriptive statistics

Descriptive statistics of the five items in the Teacher personality scale for each dataset are shown in Table 4.21, below.

Table 4.21

Descriptive Statistics for Items in the Teacher Personality Scale for Pilot Study 2

	Q1 dataset (n=74)				Q2 dataset (n=71)			
Item	*M* (*SD*)	range	skew	kurtosis	*M* (*SD*)	range	skew	kurtosis
P1	5.09 (1.09)	1-6	-1.26	1.59	5.12 (0.99)	2-6	-0.76	-0.27
P2	5.28 (1.08)	1-6	-1.80	3.32	5.23 (1.02)	1-6	-1.42	2.44
P3	4.70 (1.46)	1-6	-0.89	-0.35	4.66 (1.30)	1-6	-0.60	-0.55
P4	4.58 (1.40)	1-6	-0.68	-0.75	4.34 (1.32)	2-6	-0.19	-1.26
P5	4.79 (1.32)	1-6	-1.08	0.43	4.36 (1.24)	1-6	-0.65	0.33

As in the Comprehensibility measure, items consistently exhibited a negative skew. Item 2 had high negative skew and large positive kurtosis in both datasets, and could present problems for subsequent analyses.

4.1.2.3.2 Internal consistency

In order to examine the internal consistency of the scale, Cronbach's alpha was estimated for each scale and relevant item characteristics (intercorrelations, item-total correlations, alpha if deleted) are reported below.

The internal consistency of the scale as indicated by Cronbach's alpha was 0.887 for the Q1 dataset, and 0.841 for the Q2 dataset.

The correlations between scale items are presented in Table 4.22, below. Intercorrelations for the Q1 dataset are shown in the lower diagonal, and intercorrelations for the Q2 dataset are shown in the upper diagonal.

Table 4.22

Correlations among Items in the Teacher Personality Scale for Pilot Study 2

	P1	P2	P3	P4	P5
P1	1.00	0.48	0.45	0.49	0.28
P2	0.63	1.00	0.64	0.58	0.55
P3	0.61	0.69	1.00	0.58	0.68
P4	0.54	0.53	0.73	1.00	0.56
P5	0.41	0.69	0.71	0.62	1.00

Item intercorrelations ranged from weak (0.28) to moderately strong (0.73) across item pairs for both datasets. The correlations between Item 1 and other items were slightly lower in both datasets.

Item statistics are shown in Table 4.23, below.

Table 4.23

Item-total Correlations for Items in the Teacher Personality Scale for Pilot Study 2

	Q1 dataset		Q2 dataset	
Item	Item-total correlation	Alpha if deleted	Item-total correlation	Alpha if deleted
P1	0.631	0.884	0.474	0.850
P2	0.753	0.861	0.692	0.800
P3	0.837	0.836	0.744	0.780
P4	0.724	0.865	0.690	0.797
P5	0.725	0.864	0.651	0.808

Overall, the measure showed moderate to high internal consistency as indicated by item inter-correlations and estimates of Cronbach's alpha. For most items in both datasets, item-total correlations suggested that the estimate of alpha would decrease the item were to be deleted.

4.1.2.3.3 Interrater consistency

Interrater consistency was estimated for individual items in the scale and the total score using the intraclass correlation coefficient (ICC). Moderate correlations were expected, as students are expected to differ in their evaluations of teacher personality characteristics. ICCs are presented in Table 4.24, below.

Table 4.24

ICCs for Items in the Teacher Personality Scale for Pilot Study 2

Item P1: Friendly	Item P2: Knowledgeable	Item P3: Helpful	Item P4: Active	Item P5: Experienced	Total score
0.44**	0.18	0.28**	0.05	0.23*	0.28*

* $p < .05$

** $p < .01$

ICCs for individual items were much lower than expected (0.05-0.44). This might suggest that these items may be largely capturing an individual listener's perception of the speaker rather than a stable personality characteristic of the speaker. As indicated by ICCs, there was a moderate level of agreement between Questioners regarding whether a speaker was friendly or unfriendly (item P1, ICC=0.44), but no relationship between Questioner ratings of whether a speaker was active or passive (item P4, ICC=0.05).

4.1.2.3.4 Correlations with criterion variables

Based on previous research, it was expected that listener perceptions of a speaker's personality would have low or moderate correlations with oral proficiency ratings (TOP) and comprehensibility ratings, and moderate correlations with attitude homophily ratings. For each dataset, correlations between total scores on these scales are shown in Table 4.25, below.

Table 4.25

Correlations between Teacher Personality Total Scores and Criterion Variables for Pilot Study 2

	Q1 dataset				Q2 dataset			
	TOP	COMP	TA	AH	TOP	COMP	TA	AH
Personality	0.34	0.62	0.70	0.44	0.39	0.62	0.59	0.35

Note. COMP = Comprehensibility total score; TA = Oral skills to TA total score; AH = Attitude homophily total score.

The correlations between teacher personality total scores and TOP scores were moderately weak (Q1 r=0.34, Q2 r=0.39). There were moderately strong correlations between teacher personality and comprehensibility total scores (Q1 r=0.62; Q2 r=0.62), and between teacher personality and oral skills to TA total scores (Q1 r=0.70; Q2 r=0.59). The correlations between teacher personality ratings and attitude homophily ratings were moderately weak (Q1 r=0.44; Q2 r=0.35).

4.1.2.4 Pilot study 2: Discussion

The results from the second pilot study suggested that both the revised comprehensibility scale and the teacher personality scale would be useful to include in the main study. The additional item added to the comprehensibility scale helped increase the internal consistency of the scale while exhibiting desirable item characteristics. The teacher personality scale had comparatively lower internal consistency and even lower interrater consistency, but correlated with criterion variables in an interesting way. Its strongest argument for inclusion is based on the latter finding.

Across both datasets (Q1, Q2), the teacher personality total score correlated more strongly with the Comprehensibility (r=0.62, 0.62) and Oral skills to TA (r=0.70, 0.59) total scores, but only moderately with the TOP oral proficiency scores (r=0.34, 0.39). In other words, teacher personality total scores explained 35-49% of the variance in Comprehensibility and Oral

skills to TA total scores, but only 12-15% of the variance in TOP oral proficiency scores. While this analysis was not causal in nature, it suggested that while TOP, Comprehensibility, and Oral skills to TA scores all indicated judgments of oral proficiency or language use, the listener-based scales (Comprehensibility, Oral skills to TA) are more related to listener perceptions of teacher personality. This suggests that the teacher personality scale could be a valuable addition to the model.

Thus, the teacher personality scale was included in the main data collection but not specified in the initial model (see Figure 3). The construct of teacher personality is defined relative to the listener and not in objective sense, as evidenced by the relatively low interrater consistency for teacher personality scale items (see section 4.1.2.3.3). This distinction is made since teacher personality can be viewed as either a speaker-based factor or a listener-based factor. If a more stable measure of personality were used, teacher personality could be considered a speaker-based factor. Since teacher personality items had such low interrater consistency, judgments of teacher personality appeared to be more dependent on individual listener perceptions of the speaker's personality, or more akin to attitudes. This is consistent with some previous research that has described judgments of a speaker's positive or negative personality characteristics as "attitudes" (e.g., Coetzee-Van Rooy, 2009). In order to maintain a conceptual distinction between listeners' attitude homophily (perceived similarity of speaker to themselves) and listeners' perception of the speaker's personality and background characteristics relevant to teaching, the term "teacher personality" will be maintained. Possible uses of the scale will be examined during the exploratory phase of the study.

4.2 Main study

4.2.1 Exploratory phase

4.2.1.1 Dataset

Recall, from Chapter 3, that the exploratory data set consisted of listeners (n=205) watching a video of a speaker (n=205) perform TOP Task 3, a mini-lecture. Prior to analyses the data, this exploratory dataset (n=205), was cleaned to ensure valid responses. Descriptive statistics were produced and evaluated to investigate univariate normality, and variables were transformed when necessary. Two cases were removed during data cleaning, one variable was transformed, but no univariate outliers were detected. In total, 203 valid responses were retained to use in the subsequent analysis.

4.2.1.2 Data cleaning

The data were cleaned to ensure valid responses by examining participants' response patterns to listener-based scales, and flagging participants who (a) self-identified as non-native speakers of English with low levels of listening comprehension, (b) self-identified as non-native speakers of English but did not identify a language other than English in which they were proficient, or (c) indicated that they knew the speaker in the video. As a result of the validity check, 2 cases were removed from the exploratory dataset.

First, participants' response patterns to the four listener-based scales (comprehensibility, oral skills to TA, attitude homophily, teacher personality) were examined. As was done in the Pilot studies above, when a response pattern clearly showed an inconsistency that could be explained as a response error, a correction was made. For example, if a participant's response pattern for the comprehensibility scale was '6-6-1-6-6', and the '1' response to a reverse-keyed item, it was considered sufficient evidence for a response error and the '1' response was

corrected to '6' (i.e., the inverse response). Response errors were identified and corrected for the comprehensibility scale (5 cases in the exploratory sample, 9 cases in the cross-validation sample), oral skills to TA scale (4 and 10 cases, respectively), and attitude homophily scale (2 and 10 cases, respectively).

Three participants were flagged for removal based on the criteria identified above. Two participants who self-identified as non-native speakers of English with low levels of listening comprehension ("Low proficiency" or "Somewhat proficient") were removed in order to ensure that the samples of listeners were relatively homogenous with respect to their listening comprehension skill in English. One participant indicated that he or she knew one of the speakers observed; in order to ensure that none of the listeners in the sample were previously familiar with the speaker observed, this case was removed from the cross-validation sample.

Finally, responses that included non-numerical information (e.g., "10+", "around 15") were simplified by removing extraneous information (e.g., "10", "15").

4.2.1.3 Descriptive statistics

Descriptive statistics, outlier detection, and investigations of the assumption of univariate normality were performed for all variables included in the analysis. Given the large number of variables involved, groups of variables were examined in turn based on their designation as speaker-based or listener-based components in the conceptual model.

4.2.1.3.1 Speaker-based components

4.2.1.3.1.1 TOP oral proficiency measures

Descriptive statistics for the four TOP oral proficiency variables are shown in Table 4.26, below.

Table 4.26

Descriptive Statistics for TOP Oral Proficiency Measures in the Exploratory Dataset

Variable	*M* (*SD*)	range	skew	kurtosis
Pronunciation	5.19 (1.28)	2-8	0.00	-0.18
Lexical-Grammar	5.87 (1.10)	3-8	0.02	0.04
Rhetorical organization	6.13 (0.82)	4-8	0.18	1.14
Question handling	6.27 (0.93)	3-8	-0.10	0.91

Distributions for variables appeared to be approximately normal, as evidenced by the estimates of univariate skew and kurtosis in the table above. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on these analyses, it was determined that no variable transformations were necessary to maintain the assumption of normality.

Potential outliers in the bivariate distributions between TOP oral proficiency variables and other variables specified in the preliminary conceptual model (i.e., Comprehensibility, Oral skills to TA) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, Comprehensibility and Oral skills to TA total scores were used in this analysis instead of latent variable models. No outliers were detected.

Pearson correlations among the four TOP proficiency ratings and relevant variables in the preliminary conceptual model are shown in Table 4.27, below.

Table 4.27

Correlations between TOP Oral Proficiency Variables, Comprehensibility Total Scores, and Oral Skills to TA Total Scores

	Pronunciation	Lexical-Grammar	Rhetorical organization	Question handling	COMP total
Pronunciation	1.00				
Lexical-Grammar	0.60**	1.00			
Rhetorical organization	0.47**	0.51**	1.00		
Question handling	0.57**	0.46**	0.54**	1.00	
COMP total	0.33**	0.34**	0.24**	--	1.00
Oral skills to TA total	--	--	--	0.38**	0.88**

Note. COMP = Comprehensibility.

** $p < .01$

All of the correlations included in the table were statistically significant and ranged from relatively low to moderately strong positive relationships. The bivariate correlations among the TOP oral proficiency measures were moderately strong (0.46 – 0.60). The TOP oral proficiency measures that were hypothesized to predict comprehensibility ratings (pronunciation, lexical-grammar, rhetorical organization) had relatively low to moderate correlations with comprehensibility total scores (0.24 – 0.34). The TOP oral proficiency measure hypothesized to predict Oral skills to TA ratings (Question handling) had a moderate correlation with Oral skills to TA total scores (r=0.38).

4.2.1.3.1.2 Teaching effectiveness measures

Descriptive statistics for the four holistic measures of teaching effectiveness (Overall, Organization/Clarity, Enthusiasm, Respect/rapport) and items in the componential scales (Organization/Clarity, Enthusiasm, Respect/rapport) are shown in Table 4.28, below.

Table 4.28

Descriptive Statistics for Teacher Personality Measures in the Exploratory Dataset

Variable	*M* (*SD*)	range	skew	kurtosis
Teaching (Holistic)	3.32 (1.05)	1-5	-0.30	-0.52
Organization/Clarity (Holistic)	4.02 (1.02)	1-5	-1.04	-0.43
Enthusiasm (Holistic)	3.56 (1.16)	1-5	-0.38	-0.88
Rapport (Holistic)	3.95 (0.93)	1-5	-0.53	-0.29
Organization/Clarity scale items				
O1	4.08 (1.01)	1-5	-1.03	0.33
O2	3.88 (1.14)	1-5	-1.01	0.14
O3	4.08 (1.04)	1-5	-1.07	0.48
O4	4.09 (1.02)	1-5	-1.12	0.58
O5	4.35 (0.91)	1-5	-1.44	1.38
O6	3.81 (1.15)	1-5	-0.56	-0.94
O7	4.11 (0.99)	1-5	-1.20	0.93
Enthusiasm (Non-verbal immediacy) scale items				
EN1	3.37 (1.36)	1-5	-0.34	-1.19
EN2	4.00 (1.21)	1-5	-1.12	0.12
EN3	4.22 (1.10)	1-5	-1.41	0.90
EN4	3.74 (1.25)	1-5	-0.74	-0.64
EN5	3.04 (1.39)	1-5	0.02	-1.38
EN6	3.46 (1.27)	1-5	-0.42	-1.02
EN7	3.27 (1.42)	1-5	-0.32	-1.25
EN8	3.95 (1.21)	1-5	-0.97	-0.28
Respect/Rapport scale items				
RR1	3.83 (0.96)	1-5	-0.27	-0.69
RR2	4.23 (0.89)	1-5	-0.88	-0.02
RR3	4.56 (0.64)	2-5	-1.37	1.68
RR4	3.95 (0.89)	2-5	-0.35	-0.82
RR5	3.45 (1.02)	1-5	0.14	-1.00
RR6	4.50 (0.75)	2-5	-1.30	0.75

The distributions of all of the teaching effectiveness measures exhibited negative skewness, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for teaching effectiveness items.

Potential outliers in bivariate distributions between teaching effectiveness variables and other variables specified in the preliminary conceptual model (i.e., Oral skills to TA) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, total scores were used in this analysis instead of latent variable models for the following multi-item scales: Organization/Clarity, Enthusiasm, Respect/Rapport, and Oral skills to TA. No outliers were detected.

Pearson correlations for bivariate relationships specified in the conceptual model are shown in Table 4.29, below. Since the exploratory phase of analysis will further explore how components of teaching effectiveness may be used in the overall model, correlations between all measures have been reported.

Table 4.29

Correlations among Teaching Effectiveness Measures and Oral Skills to TA Total Scores

	ORG		EN		RR		Teaching
	HOL	TOT	HOL	TOT	HOL	TOT	HOL
ORG HOL	1.00						
ORG TOT	0.88**	1.00					
EN HOL	0.42**	0.45**	1.00				
EN TOT	0.39**	0.40**	0.69**	1.00			
RR HOL	0.43**	0.47**	0.58**	0.56**	1.00		
RR TOT	0.39**	0.42**	0.62**	0.60**	0.84**	1.00	
Teaching HOL	0.75**	0.70**	0.63**	0.61**	0.64**	0.63**	1.00
Oral skills to TA TOT	0.14	0.05	0.10	0.22**	0.09	0.09	0.21**

Note. ORG = Organization/Clarity; EN = Enthusiasm; RR = Respect/Rapport; HOL = Holistic score; TOT = Total score.

* $p < .05$

** $p < .01$

For the teaching effectiveness component measures, total scores tended to be strongly correlated with holistic ratings (r=0.69 – 0.88). The correlation between total score and holistic rating was higher for Organization/Clarity (0.88) and Respect/Rapport (r=0.84) and slightly lower for Enthusiasm (r=0.69).

Correlations between teaching effectiveness component measures were moderately strong (r=0.39-0.62). Correlations between Organization/Clarity and the other components of teaching effectiveness (Enthusiasm, Respect/Rapport) were slightly lower than the correlations between those components. In other words, Enthusiasm and Respect/Rapport measures had higher correlations with each other than with Organization/Clarity measures.

All of the teaching effectiveness component measures were strongly correlated with teaching holistic ratings (r=0.61 – 0.75). Organization/Clarity measures appeared to be the strongest predictors of teaching holistic ratings (r=0.70 – 0.75).

Correlations between teaching effectiveness measures and the Oral skills to TA total scores were non-significant or weak. Only two teaching effectiveness measures were significant predictors of Oral skills to TA total scores: Enthusiasm total scores (r=0.22) and Teaching holistic scores (r=0.21).

4.2.1.3.2 Listener-based components

4.2.1.3.2.1 Comprehensibility

Descriptive statistics for the five items in the Comprehensibility scale are shown in Table 4.30, below.

Table 4.30

Descriptive Statistics for Comprehensibility Scale Items in the Exploratory Dataset

Item	*M* (*SD*)	range	skew	kurtosis
CO1	3.34 (1.44)	1-6	0.22	-1.07
CO2	3.64 (1.50)	1-6	0.04	-1.17
CO3	3.96 (1.45)	1-6	-0.26	-1.04
CO4	3.57 (1.39)	1-6	-0.04	-1.01
CO5	3.72 (1.31)	1-6	0.04	-0.72

Distributions of variables in the Comprehensibility scale consistently exhibited negative kurtosis, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined

that no variable transformations were necessary to maintain the assumption of normality for comprehensibility items.

In the preliminary conceptual model, only one factor, Oral skills to TA, is predicted by Comprehensibility. For expediency, Comprehensibility and Oral skills to TA total scores were used to investigate potential outliers instead of latent variable models. No outliers were detected.

The correlation between Comprehensibility total scores and Oral skills to TA oral scores was very high (r=0.88).

4.2.1.3.2.2 Oral skills to TA

Descriptive statistics for the four items in the Oral skills to TA scale shown in Table 4.31, below.

Table 4.31

Descriptive Statistics for Oral Skills to TA Scale Items in the Exploratory Dataset

Item	*M* (*SD*)	range	skew	kurtosis
TA1	3.70 (1.53)	1-6	-0.01	-1.14
TA2	4.54 (1.50)	1-6	-0.79	-0.53
TA3	3.87 (1.50)	1-6	-0.14	-1.16
TA4	3.77 (1.54)	1-6	-0.25	-0.99

Distributions of items in the Oral skills to TA scale consistently exhibited negative skewness and kurtosis, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Histograms were produced for each variable in order to examine distributions and check for potential outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for Oral skills to TA items.

4.2.1.3.2.3 Attitude homophily

Descriptive statistics for the four items in the Attitude homophily scale are shown in Table 4.32, below.

Table 4.32

Descriptive Statistics for Attitude Homophily Scale Items in the Exploratory Dataset

Item	*M* (*SD*)	range	skew	kurtosis
AH1	3.37 (1.33)	1-6	-0.02	-0.79
AH2	2.89 (1.26)	1-6	0.32	-0.66
AH3	2.89 (1.38)	1-6	0.36	-0.88
AH4	2.74 (1.16)	1-6	0.33	-0.81

Distributions of items in the Attitude homophily scale consistently exhibited negative kurtosis, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for Attitude homophily items.

In the preliminary conceptual model, Attitude homophily is hypothesized to predict Comprehensibility and Oral skills to TA. For expediency, Comprehensibility total scores, Oral skills to TA total scores, and Attitude homophily total scores were used to investigate potential outliers instead of latent variable models. No outliers were detected.

Pearson correlations for these bivariate distributions are shown in Table 4.33, below.

Table 4.33

Correlations between Attitude Homophily, Comprehensibility, and Oral Skills to TA Total Scores in the Exploratory Dataset

	Attitude homophily total	Comprehensibility total	Oral skills to TA total
Attitude homophily total	1.00		
Comprehensibility total	0.41**	1.00	
Oral skills to TA total	0.42**	0.88**	1.00

* $p < .05$

** $p < .01$

Correlations between Attitude homophily and Comprehensibility total scores were moderately strong (r=0.41) and similar to those between Attitude homophily and Oral skills to TA total scores (r=0.42).

4.2.1.3.2.4 Teacher personality

Although a teacher personality factor was not specified in the preliminary conceptual model, descriptive statistics for the five items in the Teacher personality scale are presented in Table 4.34, below.

Table 4.34

Descriptive Statistics for Teacher Personality Scale Items in the Exploratory Dataset

Item (Descriptor)	*M* (*SD*)	range	skew	kurtosis
P1 (Friendly)	5.08 (0.98)	2-6	-1.06	0.69
P2 (Knowledgeable)	4.92 (1.08)	1-6	-1.29	1.73
P3 (Helpful)	4.29 (1.34)	1-6	-0.69	-0.28
P4 (Active)	4.10 (1.33)	1-6	-0.48	-0.50
P5 (Experienced)	3.69 (1.50)	1-6	-0.23	-1.06

Distributions consistently exhibited negative skew, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for teacher personality scale items.

4.2.1.3.2.5 Other listener-based measures

Descriptive statistics for the remaining listener-based measures in the preliminary conceptual model are shown in Table 4.35, below.

Table 4.35

Descriptive Statistics for Listener-based Measures in the Exploratory Dataset

Item	*M* (*SD*)	range	skew	kurtosis
Topic familiarity	2.86 (1.92)	1-6	0.46	-1.37
Topic interest	3.58 (1.59)	1-6	-0.17	-1.16
Topic complexity	2.28 (1.42)	1-6	0.94	-0.23
LFAM	2.19 (1.61)	1-5	0.87	-0.98
NNS familiarity	3.73 (1.06)	1-5	-0.27	-1.09
ITA familiarity	3.36 (3.56)	0-30	2.80	14.76
ITA experience	3.49 (1.11)	1-5	-0.42	-0.61

Note. LFAM = Familiarity with the speaker's native language; NNS familiarity = Familiarity with non-native speakers of English by frequency of interaction.

Distributions consistently exhibited negative kurtosis for most variables in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary for most items to maintain the assumption of normality.

The ITA familiarity variable exhibited unacceptably high levels of univariate skewness (~2.8) and kurtosis (~14.7). After examining the histogram of this variable for each dataset, it was determined that the variable could be transformed from an interval scale (based on absolute frequency) to an ordinal scale (based on relative frequency) using the ordered categories in Table 4.36, below.

Table 4.36

Ordinal Transformation of ITA Familiarity

Range of original variable	New value	Interpretation
0	1	No prior experience with ITAs
1-2	2	Little prior experience with ITAs
3-5	3	Some prior experience with ITAs
6+	4	Substantial prior experience with ITAs

The original variable was expressed on an interval scale and indicated the self-reported number of courses a listener had previously had with ITAs. The transformed variable was expressed on an ordinal scale and indicated the relative frequency of experience the listener self-reported in regards to the number of courses previously taken with ITAs. The transformed scale more closely approximated a normal distribution but could also be justified conceptually. This variable contained a number of responses that needed to be simplified during the data cleaning phase due to the inclusion of non-numerical information (e.g., "10+") which indicated that some participants had difficulty providing exact estimates.

Histograms for the original and transformed values of ITA familiarity for the exploratory dataset are provided in Figure 4.1, below.

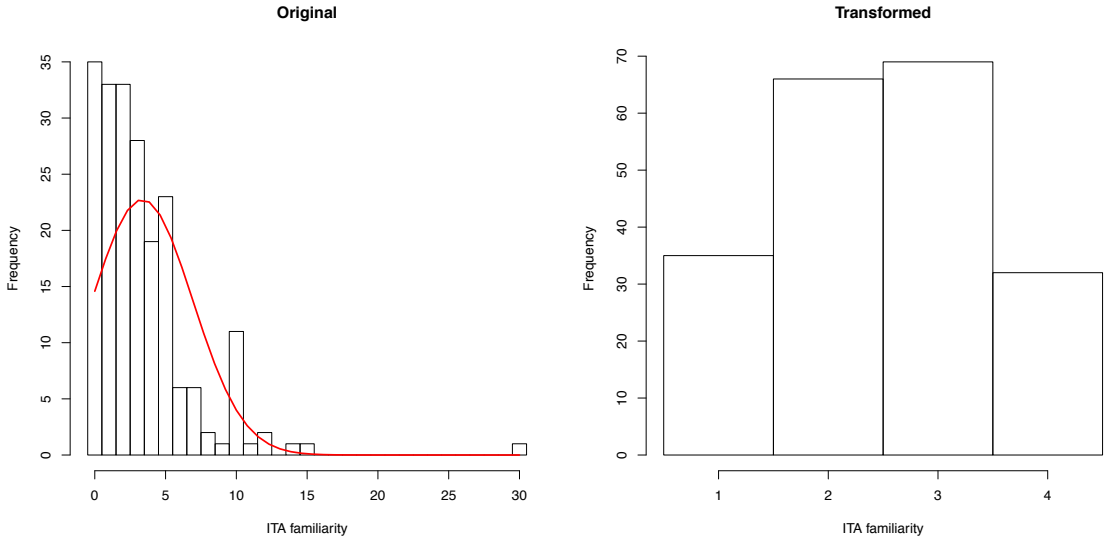

Figure 4.1. Histograms for the original and transformed values of ITA familiarity for the exploratory dataset.

As seen in Figure 4.1, the transformed distribution much more closely resembles a normal distribution while retaining its interpretability. Descriptive statistics for the transformed variable are provided in Table 4.37, below.

Table 4.37

Descriptive Statistics for Transformed ITA Familiarity Variable for the Exploratory Dataset

Item	*M* (*SD*)	range	skew	kurtosis
ITA familiarity	2.49 (0.96)	1-4	-0.01	-0.97

The distributions exhibited negative kurtosis, as evidenced by the estimates in the table above and the histogram provided in Figure 4 above. However, the relative size of skewness and kurtosis estimates for the transformed variable did not suggest severe departures from normality. Based on this analysis, it was determined that the transformed Topic familiarity variable should be retained.

Potential outliers in bivariate distributions between listener background variables and other variables specified in the preliminary conceptual model (i.e., Comprehensibility, Oral skills to TA) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, Comprehensibility and Oral skills to TA total scores were used in this analysis instead of latent variable models. No outliers were detected.

Pearson correlations for bivariate distributions specified in the conceptual model are shown in the tables below. Table 4.38 shows correlations between relevant listener background variables and Comprehensibility total scores. Table 4.39 shows correlations between relevant listener background variables and Oral skills to TA total scores.

Table 4.38

Correlations among Listener Background Measures and Comprehensibility Total Scores

	Topic familiarity	Topic interest	Topic complexity	LFAM	NNS familiarity	COMP total
Topic familiarity	1.00					
Topic interest	0.31**	1.00				
Topic complexity	-0.36**	-0.30**	1.00			
LFAM	--	--	--	1.00		
NNS familiarity	--	--	--	--	1.00	
COMP total	0.10	0.35**	-0.26**	0.15*	0.10	1.00

Note. LFAM = Familiarity with the native language (L1) of the speaker; NNS familiarity = Familiarity with non-native speakers of English by frequency of interaction.

* $p < .05$

** $p < .01$

Table 4.39

Correlations between ITA Familiarity and Experience Measures, and Oral Skills to TA Total Scores

	ITA familiarity	ITA experience	Oral skills to TA total
ITA familiarity	1.00		
ITA experience	--	1.00	
Oral skills to TA total	0.19**	0.24**	1.00

* $p < .05$

** $p < .01$

Correlations between listener background variables and Comprehensibility total scores were generally non-significant or moderately weak. Topic interest (r=0.35) and topic complexity (r=-0.26) had the strongest correlations with Comprehensibility total scores among listener background variables. Listener variables related to the topic (topic familiarity, interest, and complexity) had moderately weak bivariate correlations. Topic complexity has a negative correlation with other measures since it has an inverse relationship with most measures. Although the correlation was weak, as perceived topic complexity increased, Comprehensibility total scores decreased (r=-0.26).

Both listener background variables hypothesized to predict Oral skills to TA in the conceptual model were weakly correlated with Oral skills to TA total scores.

4.2.1.4 Measurement models

For each of the measurement models specified in the preliminary structural model (Teaching effectiveness components, Comprehensibility, Oral skills to TA, Attitude homophily) as shown previously in Figure 3.2, and for Teacher personality, confirmatory factor analysis (CFA) was performed as previously described in section 3.5.2.

4.2.1.4.1 Teaching effectiveness components

The teaching effectiveness scale was composed of three sub-scales (Organization/Clarity, Enthusiasm/Non-verbal immediacy, Respect/Rapport) that were hypothesized to measure distinct aspects of the construct of teaching effectiveness. Since the measure of teaching effectiveness used in this study consisted of items adapted from previous scales, an analysis of its dimensionality was conducted in order to determine the most appropriate measurement model for components of teaching effectiveness.

First, CFA was conducted using the exploratory dataset in order to examine parameter estimates and model fit indices. In order to investigate possible revisions to the model based on statistical considerations, the Wald test and Lagrange multiplier test were utilized. Based on statistical and conceptual considerations, a REVISED model was fit using the cross-validation dataset.

4.2.1.4.1.1 Teaching effectiveness components: Exploratory data analysis

A correlated traits three-factor model for the teaching effectiveness scale was fit to the exploratory dataset. Parameter estimates for the model are shown in Figure 4.2, below. All factor loadings and covariances were statistically significant.

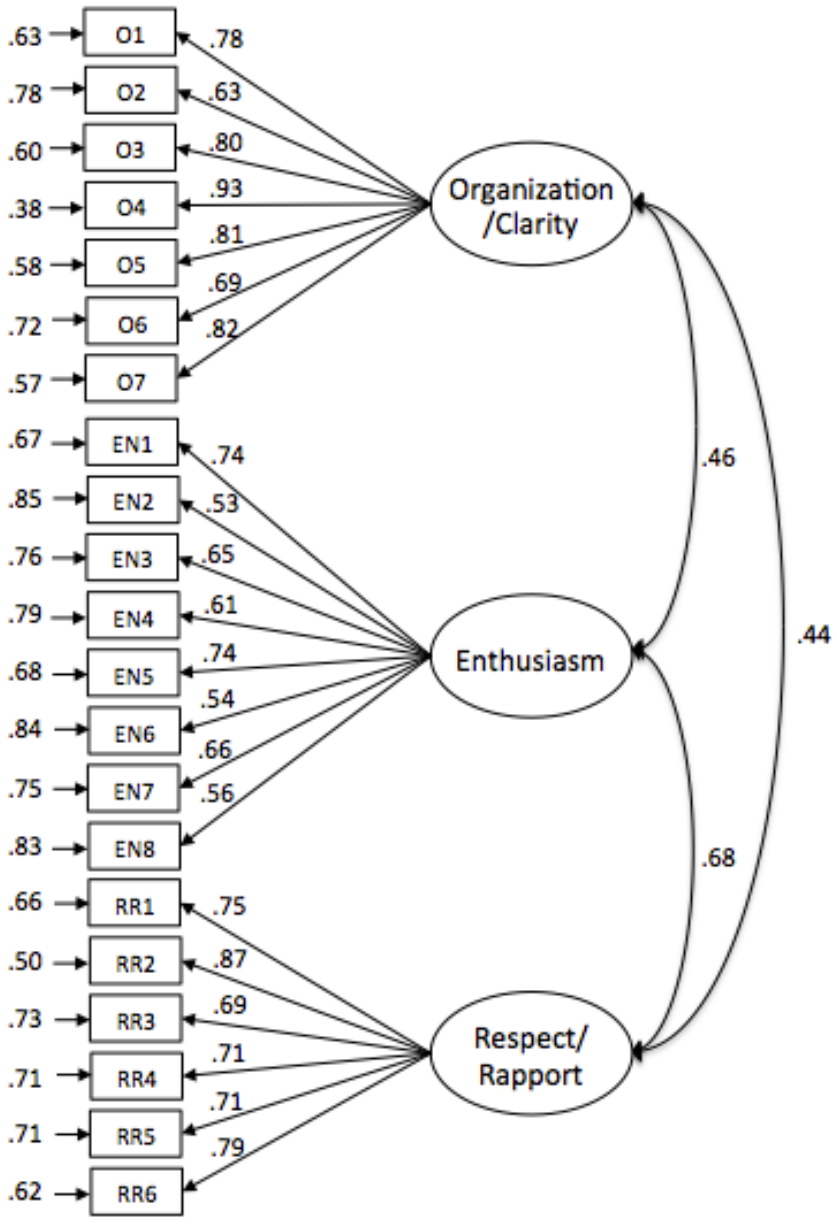

Figure 4.2. Measurement model for the three-factor correlated traits model of components of teaching effectiveness for the exploratory dataset.

Standardized factor loadings for each of the three factors were generally high (0.53 – 0.93). Correlations between factors ranged from moderate to strong (0.44 – 0.68). Enthusiasm and Respect/Rapport had a strong correlation (0.68), while correlations between these factors and Organization/Clarity were moderately strong (0.46 and 0.44, respectively).

The chi-square statistic suggested that the null hypothesis that the model fit the data closely should be rejected, χ^2 (186) = 734.77, $p < 0.01$. In addition, most of the incremental- and residual-based fit indices suggested poor model fit (CFI = 0.787; NNFI = 0.759; SRMR = 0.092; RMSEA = 0.122).

The Wald test was performed in order to identify parameters that might be dropped in order to improve the statistical fit of the model. None of the parameters in the model were identified using this test.

The Lagrange multiplier (LM) test was performed in order to identify parameters that could be added in order to improve the statistical fit of the model. Table 4.40 below summarizes results of this test.

Table 4.40

Excerpted Results of the Lagrange Multiplier Test for Adding Parameters to the Three-factor Correlated Traits Model of Teaching Effectiveness

		Cumulative multivariate statistics			Univariate increment	
Step	Parameter to add	χ^2	*df*	*p*	χ^2	*p*
1	EN7 <-> Respect/Rapport	21.646	1	0.00	21.646	0.00
2	EN4 <-> Organization/Clarity	38.092	2	0.00	16.446	0.00
3	EN3 <-> Respect/Rapport	50.177	3	0.00	12.085	0.00
4	O7 <-> Respect/Rapport	60.847	4	0.00	10.670	0.00
5	EN7 <-> Organization/Clarity	69.070	5	0.00	8.223	0.00

The table above identifies parameters that could be added to the model to improve model fit, as measured by the χ^2 estimate. For example, if a path were added between Respect/Rapport and EN7, the chi-square (χ^2) statistic could be expected to decrease by 21.646, leading to an incremental improvement in model fit. Most of the recommendations related to Enthusiasm

scale items, which suggests that items in this scale might have some conceptual overlap with items in other scales.

The largest absolute standardized residuals are shown in Table 4.41, below.

Table 4.41

Largest Absolute Standardized Residuals for the Three-Factor Correlated Traits Model of Components of Teaching Effectiveness for the Exploratory Dataset

Parameter	Estimate
EN8, EN3	0.449
EN8, EN2	0.303
EN4, O5	0.235
EN2, EN3	0.224
R2, EN7	0.209
EN2, O2	-0.200

A large positive (or negative) standardized residual indicates that residual covariance between variables in the data is much larger (or smaller) than expected based on the model. Three of the four largest standardized residuals were between three items in the Enthusiasm scale: EN2, EN3, and EN8. This suggests that there is additional positive covariance between these items in the data that is not accounted for in the model. These items are reproduced in Table 4.42, below.

Table 4.42

Enthusiasm Scale Items with Large Positive Standardized Residuals

Item	Text of item
EN2	The TA often turned his/her back on students.
EN3	The TA avoided eye contact while talking to students.
EN8	The TA maintained eye contact while talking to students.

The items in Table 4.41 all relate to a particular indicator of Enthusiasm (or Nonverbal immediacy): eye contact. Other items in the scale (see Appendix B) relate to the speaker's tone of voice ("dull voice", "variety of vocal expressions"), physical expressiveness ("gestures", "facial expressions"), and confidence ("very nervous").

One method that could account for residual covariance between the Enthusiasm items with very similar item content would be to allow correlated error terms for these items. An error term accounts for the unique variance in an item not explained by the latent factor (its communality), and error terms are generally not permitted to be correlated due to an underlying assumption regarding the independence of errors. In this case, correlated errors for these items may be justified by pointing to the conceptual overlap between them (eye contact) beyond that explained by the latent factor (Enthusiasm).

Another approach would be to drop one of the items in the Enthusiasm scale (E8). This item is largely redundant with another item in the scale (E3), and helps account for two of the largest estimates in the matrix of standardized residuals (see Table 4.41, above).

Given the questionable practice of correlated error terms, the redundancy of item E8 conceptually, and item E8's contribution to large estimates of standardized residuals, it was determined that this item should be dropped from the scale. Although this might only lead to a small improvement in model fit, it is more defensible conceptually than allowing error terms to be correlated.

A revised three-factor correlated traits model of components of teaching effectiveness with one Enthusiasm scale item dropped was fit to the exploratory dataset. Parameter estimates for the model are shown in Figure 4.3, below. All factor loadings and covariances were statistically significant.

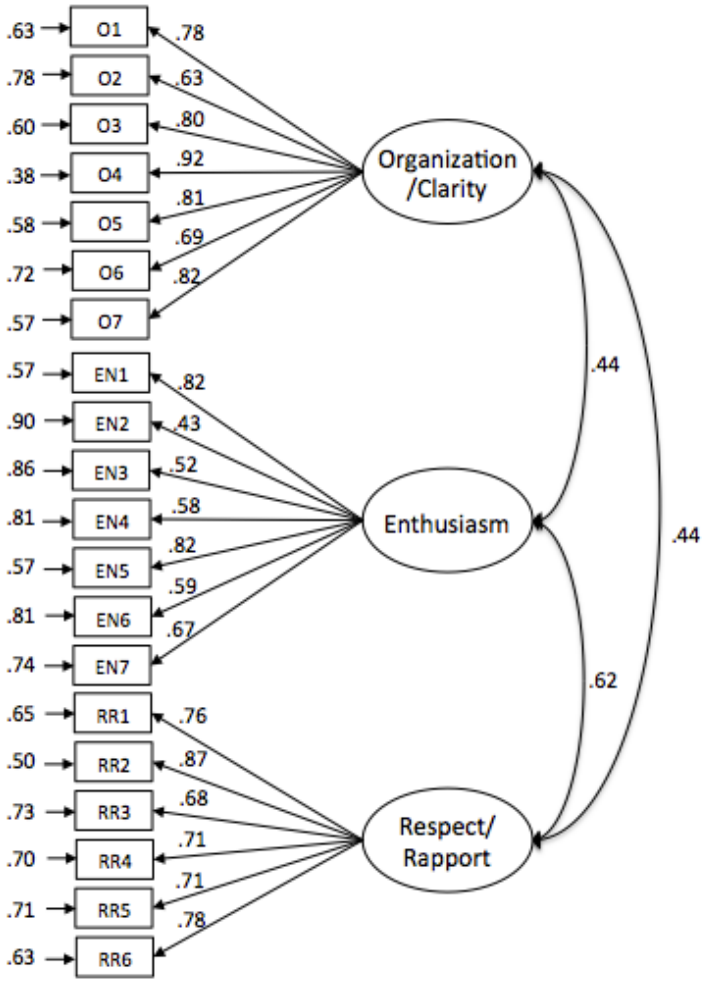

Figure 4.3. Measurement model for the revised three-factor correlated traits model of components of teaching effectiveness for the exploratory dataset.

Standardized factor loadings for each of the three factors were generally high (0.43 – 0.92). Correlations between factors ranged from moderate to strong (0.44 – 0.62). Enthusiasm and Respect/Rapport had a strong correlation (0.62), while correlations between these factors and Organization/Clarity were moderately strong (0.44).

The chi-square statistic suggested that the null hypothesis that the model fit the data closely should be rejected, χ^2 (167) = 532.87, $p < 0.01$. Most of the incremental- and residual-based fit indices still suggested relatively poor model fit (CFI = 0.844; NNFI = 0.822; SRMR =

0.090; RMSEA = 0.105), although model fit appeared to incrementally improve with the deletion of item EN8.

4.2.1.4.1.2 Teaching effectiveness components: Cross-validation data analysis

In order to investigate whether the model fit an additional dataset, the four-factor correlated traits model was fit to the teaching effectiveness scale items in the cross-validation dataset. Estimates of model fit for the exploratory and cross-validation datasets are provided in Table 4.43, below.

Table 4.43

Fit Indices for the Revised Three-factor Correlated Traits Model for Components of Teaching Effectiveness

Dataset	χ^2	*df*	RMSEA (ε^)	RMSEA 90% *CI*	CFI	NNFI	SRMR
Exploratory	532.87**	167	0.105	[0.095, 0.115]	0.844	0.822	0.090
Cross-validation	557.09**	167	0.110	[0.100, 0.120]	0.802	0.774	0.097

** $p < .01$

As seen in Table 3, model fit indices were comparable between the exploratory and cross-validation datasets. Estimates in the cross-validation dataset were slightly lower (for CFI, NNFI) or higher (for RMSEA, SRMR, χ^2), in both cases suggesting that model fit may be slightly worse in the cross-validation dataset. Overall, however, evidence of fit for the three-factor correlated traits model was considered adequate to utilize this structure for the measurement model for teaching effectiveness.

4.2.1.4.2 Comprehensibility

The measurement model for Comprehensibility is reproduced with parameter estimates in Figure 4.4, below.

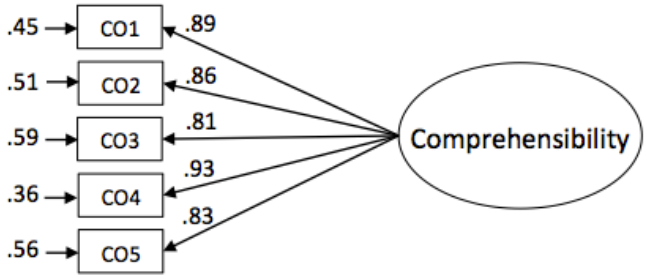

Figure 4.4. Measurement model for Comprehensibility for the exploratory dataset.

Model fit indices generally provided evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (5) = 33.46, p < 0.01. The RMSEA point estimate (RMSEA = 0.168) also suggested that the model did not provide a good fit. However, the 90% confidence interval around the point estimate was large (0.116, 0.223), which suggests that the point estimate should be interpreted with caution given its imprecision. Another residual-based fit index, the standardized root mean-square residual (SRMR), suggested adequate fit (SRMR = 0.03). Several incremental fit indices also suggested the model provided acceptable fit (CFI = 0.97; NNFI = 0.94).

4.2.1.4.3 Oral skills to TA

The measurement model for Oral skills to TA is reproduced with parameter estimates in Figure 4.5, below.

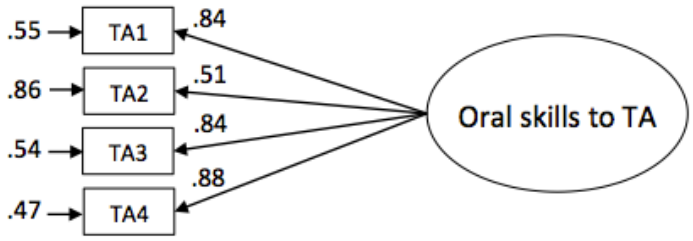

Figure 4.5. Measurement model for Oral skills to TA for the exploratory dataset.

Model fit indices generally provided evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (2) = 17.68, p < 0.01. The RMSEA point estimate (RMSEA = 0.197) also suggested that the model did not provide a good fit. Again, the 90% confidence interval around the point estimate was large (0.119, 0.285), which suggests that the point estimate should be interpreted with caution given its imprecision. The SRMR suggested adequate fit (SRMR = 0.04). Several incremental fit indices also suggested the model provided acceptable fit (CFI = 0.96; NNFI = 0.89).

4.2.1.4.4 Attitude homophily

The measurement model for Attitude homophily is reproduced with parameter estimates in Figure 4.6, below.

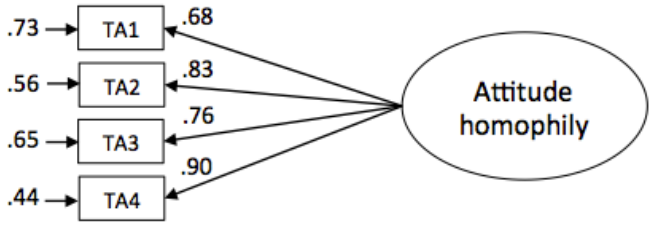

Figure 4.6. Measurement model for Attitude homophily for the exploratory dataset.

Model fit indices generally provided evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (2) = 11.58, p < 0.01. The RMSEA point estimate (RMSEA = 0.154) also suggested that the model did not provide a good fit. Again, the 90% confidence interval around the point estimate was large (0.076, 0.244), which suggests that the point estimate should be interpreted with caution given its imprecision. The SRMR suggested adequate fit, (SRMR = 0.03). Several incremental fit indices also suggested the model provided close fit (CFI = 0.98; NNFI = 0.93).

4.2.1.4.5 Teacher personality

The measurement model for Teacher personality is reproduced with parameter estimates in Figure 4.7, below.

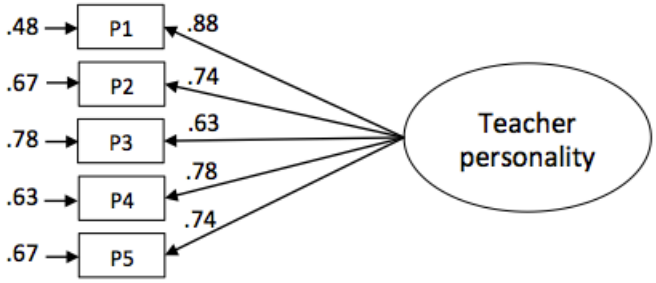

Figure 4.7. Measurement model for Teacher personality for the exploratory dataset.

Model fit indices generally provided evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (5) = 33.54, $p < 0.01$. The RMSEA point estimate (RMSEA = 0.168) also suggested that the model did not provide a good fit. Again, the 90% confidence interval around the point estimate was large (0.117, 0.224), which suggests that the point estimate should be interpreted with caution given its imprecision. The SRMR suggested adequate fit (SRMR = 0.06). The CFI estimate suggested the model provided adequate fit (CFI = 0.90), while the NNFI did not show evidence of good fit (NNFI = 0.79).

4.2.1.5 Structural model

The preliminary conceptual model presented in Chapter 1 above was updated based on the revision of the teaching effectiveness measurement model in section 4.2.1.4.1, and is presented in Figure 4.8, below.

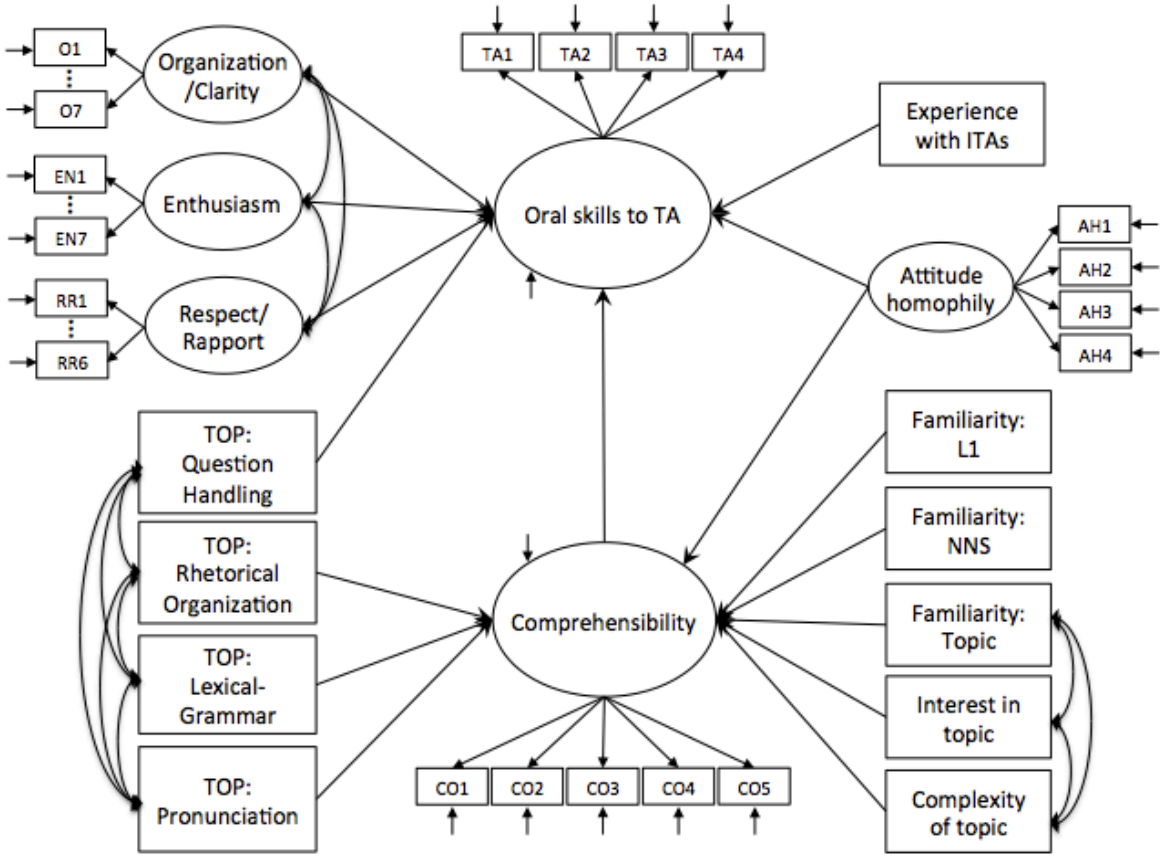

Figure 4.8. Revised preliminary SEM model specifying the relationships between listener perceptions of oral proficiency, and speaker- and listener-related factors.

In this revised model, a speaker's comprehensibility to the listener is hypothesized to be influenced by the speaker's oral language proficiency (TOP: Pronunciation, Lexical-Grammar, Rhetorical organization), the listener's attitude towards the speaker (Attitude homophily), the listener's familiarity with the speaker's native language and non-native speakers of English in general, and the listener's familiarity, interest, and perceived complexity of the topic. A listener's perception of the speaker's oral skills to TA is influenced by three components of speaker's teaching effectiveness (Organization/Clarity, Enthusiasm, Respect/Rapport), comprehensibility, the listener's attitude towards the speaker, and the listener's previous experience with ITAs.

4.2.1.5.1 Parameter estimates

An initial analysis identified one outlier that contributed to multivariate kurtosis. After removing this outlier, the estimate of multivariate kurtosis (Mardia's coefficient) was reduced from 3.64 to 2.61, below the threshold that may indicate departure from multivariate normality (Bentler, 2006). This reduced the number of cases in the exploratory dataset used to estimate the preliminary structural model to 202. In addition, listwise deletion was used to remove 8 cases that had missing data, further reducing the size of the dataset to 194 cases. Parameter estimates for the structural model with measurement models removed are shown in Figure 4.9, below.

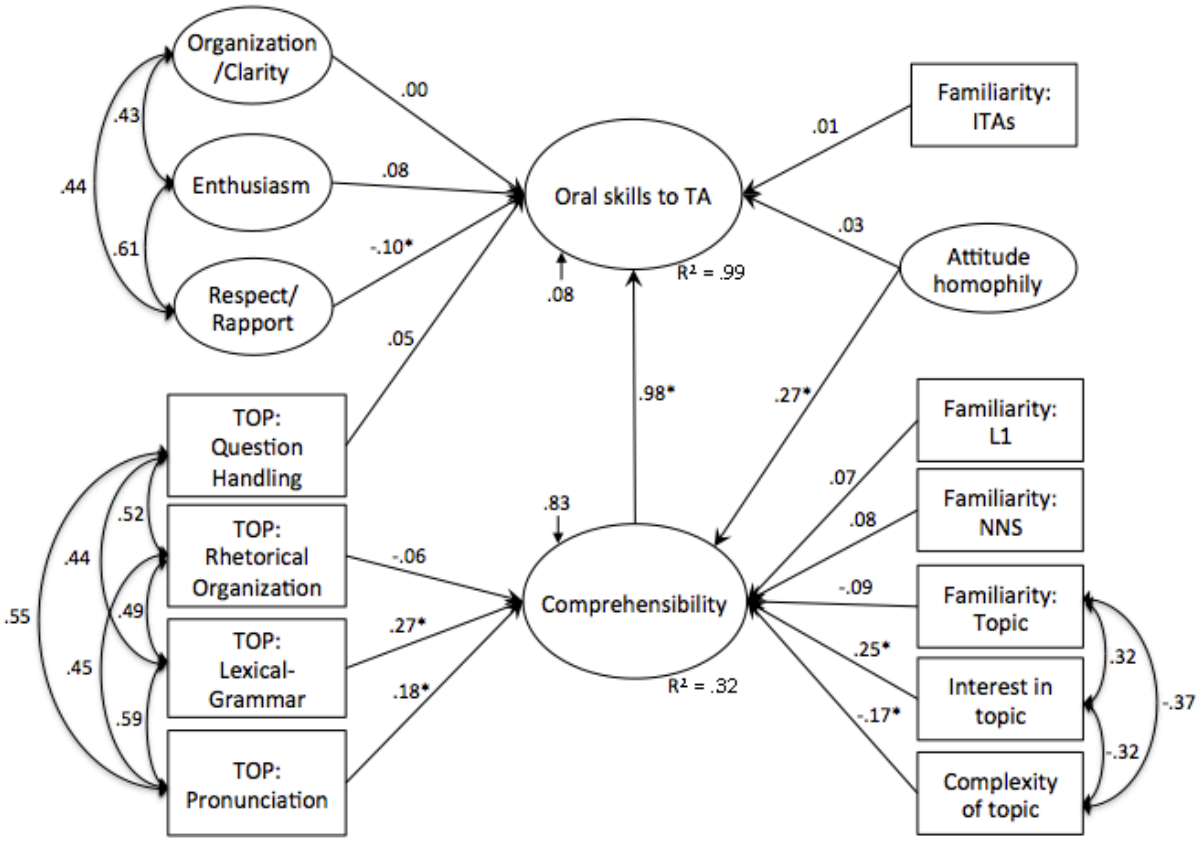

Figure 4.9. Parameter estimates for the preliminary structural model. * $p < .05$.

In this model, Comprehensibility is significantly predicted by two TOP speaker variables – a speaker's pronunciation (TOP: Pronunciation, .18) and lexical-grammar (TOP: Lexical-Grammar, .27) – and three listener measures: the listener's attitude towards the speaker (Attitude homophily, .27), interest in the topic (.25), and negatively by the perceived complexity of the topic (-.17). Listener judgment of the speaker's oral skills to TA is almost entirely predicted by the speaker's comprehensibility (.98), but also negatively by the speaker's rapport with students (-.10). This latter result is surprising and not immediately interpretable, although the size of the effect is quite small.

Most of the speaker- and listener-based factors that predict comprehensibility are also indirect predictors of oral skills to TA, as shown by a decomposition of the direct and indirect effects in the standardized solutions for the structural equations are presented in Table 4.44, below.

Table 4.44

Standardized Solutions for Structural Equations in the Preliminary Conceptual Model

Dependent factor		Independent variables and factors	R^2
Comprehensibility	=	0.18*Pronunciation + 0.27*Lexical-Grammar – 0.06*Rhetorical organization – 0.09*Topic familiarity + 0.25*Topic interest – 0.17*Topic complexity + 0.07*LFAM + 0.08*NNS familiarity + 0.27*Attitude homophily + 0.83 D1 (disturbance)	0.32
Oral skills to TA	=	0.98*Comprehensibility + 0.05*Question handling + 0.01*ITA familiarity + 0.03*Attitude homophily + 0.00*Organization/Clarity + 0.08*Enthusiasm – 0.10*Respect/Rapport + 0.08 D2 (disturbance)	0.99

Table 4.44 summarizes the parameter estimates for each independent variable or factor in the two structural equations specified in the model. Approximately 32% of the variation in Comprehensibility is explained by ten predictors in the model, five of which are statistically

significant predictors. The largest predictors of Comprehensibility are Attitude homophily (0.27) and Lexical-Grammar (0.27), followed by Topic interest (0.25), Pronunciation (0.18), and Topic complexity (-0.17).

Nearly all of the variance in Oral skills to TA is explained by the eight predictors in the model. Two of these predictors are statistically significant, but Comprehensibility accounts for almost all of the variance. The other predicator, Respect/Rapport, has a relatively small parameter estimate and is not interpretable; conceptually, it does not make sense that this aspect of teaching effectiveness would have a negative impact on Oral Skills to TA.

4.2.1.5.2 Model fit

Fit indices are shown in Table 4.45, below.

Table 4.45

Fit Indices for the Preliminary Structural Model

χ^2	*df*	RMSEA (ε^)	RMSEA 90% *CI*	CFI	NNFI	SRMR
1634.94**	842	0.070	[0.065, 0.075]	0.842	0.830	0.105

** $p < .01$

Model fit indices provided some evidence of adequate fit, but generally poor model fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (797) = 1398.24, $p < 0.01$. The RMSEA point estimate (RMSEA = 0.070) suggested that the model provided adequate fit, and the upper limit of the 90% confidence interval for the point estimate was below the threshold generally used to indicate adequate fit (0.080). The SRMR, CFI, and NNFI indices suggested poor model fit (SRMR = 0.105; CFI = 0.842; NNFI = 0.830).

The matrix of standardized residuals was examined to identify parameter estimates with large residuals. There were several parameters with very large (greater than 0.30) standardized residuals, and these were either related to an Attitude homophily item (AH1) or the listener's interest in the topic. Overall, the distribution of standardized residuals had a desirable shape: leptokurtic, centered at zero.

4.2.1.5.3 Specification search

In order to improve the statistical fit of the model, the Wald test was performed in order to identify parameters that might be dropped. Table 4.46 below summarizes results of this test.

Table 4.46

Excerpted Results of the Wald Test for Dropping Parameters for the Preliminary Model

		Cumulative multivariate statistics			Univariate increment	
Step	Parameter to remove	χ^2	*df*	*p*	χ^2	*p*
1	Oral skills to TA <- Clarity/Organization	0.001	1	0.98	0.001	0.98
2	D2 (*Oral skills to TA* disturbance term)	0.084	2	0.96	0.083	0.77
3	Oral skills to TA <- Experience with ITAs	0.268	3	0.97	0.184	0.67
4	Oral skills to TA <- Attitude homophily	0.923	4	0.92	0.655	0.42
5	Comprehensibility <- Rhetorical organization	1.637	5	0.90	0.714	0.40

The table above identifies parameters in the model that contributed very little to model fit, as measured by the χ^2 estimate. Most recommendations made by the Wald test as shown in Table 4.46 are related the parameter recommended to be removed in step two: the disturbance term for the Oral skills to TA structural equation. Disturbance for this structural equation is extremely small because most of the variation in Oral skills to TA is explained by Comprehensibility.

The Lagrange multiplier (LM) test was performed in order to identify parameters that could be added in order to improve the statistical fit of the model. Table 4.47 below summarizes results of this test.

Table 4.47

Excerpted Results of the Lagrange Multiplier Test for Adding Parameters to the Preliminary Model

		Cumulative multivariate statistics			Univariate increment	
Step	Parameter to add	χ^2	*df*	*p*	χ^2	*p*
1	EN7 (facial expressions) <-> Respect/Rapport	24.710	1	0.00	24.710	0.00
2	EN3 (eye contact) <-> Respect/Rapport	49.055	2	0.00	24.345	0.00
3	Attitude homophily <-> Topic interest	71.438	3	0.00	22.383	0.00
4	EN4 (nervous) <-> Organization/Clarity	91.165	4	0.00	19.727	0.00
5	C3 (certainty of understanding) <-> Topic complexity	106.627	5	0.00	15.462	0.00
6	TA1 (lecturing) <-> Attitude homophily	119.477	6	0.00	12.849	0.00
7	TA1 (lecturing) <-> Topic interest	134.268	7	0.00	14.791	0.00
8	Familiarity with ITAs <-> Familiarity with NNS	145.030	8	0.00	10.762	0.00

The table above identifies parameters that could be added to the model to improve model fit, as measured by the χ^2 estimate. For example, if a path were added between Attitude homophily and Topic interest, the chi-square (χ^2) statistic could be expected to decrease by 22.383, leading to an incremental improvement in model fit. Several of the recommendations related to Attitude homophily or Topic interest, which suggests that their roles in the model

might need to be re-examined. In addition, model fit might improve by specifying covariance between Familiarity with NNS and Familiarity with ITAs, as it is reasonable to expect that these variables would be correlated.

4.2.1.5.4 Recommendations for a revised model

Although the preliminary model appeared to provide an adequate or marginal fit to the data, there were several concerns that might be addressed by revising the model.

4.2.1.5.4.1 Comprehensibility and Oral skills to TA: One construct or two?

In this model, listener judgments of speakers' Oral skills to TA are almost entirely predicted by Comprehensibility: the parameter estimate for the path between Comprehensibility and Oral skills to TA was 0.98, and the Wald test suggested dropping the disturbance term for the Oral skills to TA structural equation since 99.9% of the variance in Oral skills to TA is explained by the model. As a result, Comprehensibility is essentially identified with Oral skills to TA. Factors that predict Comprehensibility as direct effects also predict Oral skills to TA as indirect effects, but at virtually the same magnitude.

There were several possibilities for dealing with this problem. One possibility could be to merge Comprehensibility and Oral skills to TA into a single factor. This would eliminate the conceptual distinction made earlier between naïve listener perceptions of a speaker's comprehensibility and perceptions of the speaker's competence using language to accomplish tasks in the domain (TA-led classroom). This approach could be justified in part by examining the statistical fit of a unidimensional model consisting of both Comprehensibility and Oral skills to TA items.

An alternate approach to deal with the same problem would be either drop one of the scales from the model, or to evaluate each of the two structural equations in the preliminary

model separately. If Comprehensibility were removed as a predictor of Oral skills to TA, the link between the two structural equations in the model would be severed and each structural equation could be examined separately. This approach would retain the conceptual distinction between the two constructs maintained in the current model.

An overriding concern was the degree to which a revised model would help address the research questions posed by this study. The three research questions posed in section 1.2 are reproduced below:

1) What are the relationships between listener perceptions of oral language use, and speaker- and listener-related factors for a subpopulation of listeners from the TLU domain?
2) To what extent do speaker-related versus listener-related factors affect listener perceptions of comprehensibility?
3) To what extent do construct-relevant factors, i.e., pronunciation, lexical-grammar, rhetorical organization, question handling, versus construct-irrelevant factors, e.g., listener attitudes towards the speaker, affect listener perceptions of whether the speaker has the oral language skills necessary to TA?

The first question could be addressed using either of the approaches suggested. The second question refers specifically to the construct of comprehensibility, and the extent to which it is affected by speaker- or listener-based factors. The third question refers specifically to the construct of Oral skills to TA, and the extent to which it is affected by construct-relevant versus construct-irrelevant factors.

If the preliminary model were to be separated into two models – one for Comprehensibility, and one for Oral skills to TA – all of the research questions could be

addressed without needing to lose the conceptual distinction between the two constructs. If the preliminary model was to be preserved as a single model, the constructs would need to be merged, with the more general (Comprehensibility) subsuming the more specific (Oral skills to TA). The merged construct could be interpreted as Comprehensibility defined within a particular domain: the domain of ITA language use. This context is implied in the original conceptual model of Comprehensibility presented earlier in Figure 1, but would need to be made explicit in the revised model. Thus, comprehensibility in this domain could also be interpreted as a listener's perceptions of whether the speaker has the oral skills necessary to TA.

4.2.1.5.4.2 Pronunciation as a predictor of Comprehensibility

A second issue in the preliminary model was the surprisingly low effect of TOP Pronunciation on Comprehensibility. In a study that informed the development of the preliminary conceptual model, TOP Pronunciation was the largest predictor of Comprehensibility (0.43). In the preliminary model, pronunciation was a statistically significant predictor, but it had a relatively small effect (0.18).

The path model presented earlier (see Figure 2.3) was fit to the exploratory dataset in order to examine differences in parameter estimates for the same model between the dataset described by Schmidgall (2012) and the exploratory dataset. The model was estimated using ML estimation in EQS 6.2, and the path model is shown in Figure 4.10, below.

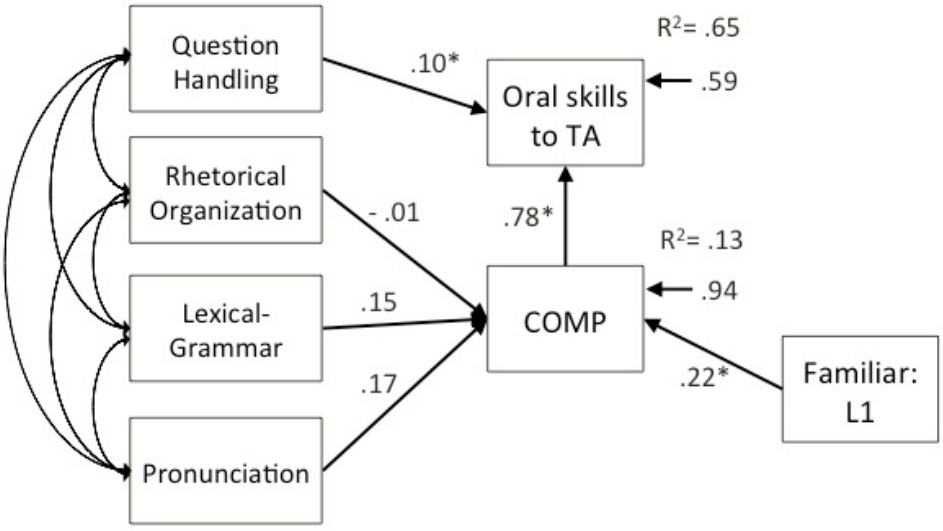

Figure 4.10. A path model of the relationships between speaker oral proficiency variables and listener perceptions of the speaker (Main study, exploratory sample). COMP = Comprehensibility; Familiar: L1 = Listener's familiarity with the speaker's native language. * $p < .05$.

The primary difference between the two models is the effect of Pronunciation on Comprehensibility. Pronunciation was a significant predictor of Comprehensibility in the model presented earlier in Figure 2.3 (0.43), but not in the model above based on the exploratory dataset (0.17). Rhetorical Organization had a larger effect on Comprehensibility in the original model (0.15), although it was statistically non-significant as a predictor of Comprehensibility in both models. Finally, Question Handling was a statistically significant predictor of Oral skills to TA in the model above, but not in the original.

Several relationships between predictors and outcome variables were consistent across the models. Oral skills to TA was largely predicted by Comprehensibility at a similar magnitude between the two models (0.74, 0.78). The listener's familiarity with the speaker's native language was a statistically significant predictor of Comprehensibility in both models at a similar magnitude (0.24, 0.22). Finally, the magnitude of Lexical-Grammar as a predictor of

Comprehensibility was similar across both models (0.14, 0.15), although it was statistically insignificant in both.

In order to further relate this earlier model to the preliminary model evaluated in the exploratory phase, the model was re-evaluated using the exploratory dataset with Comprehensibility and Oral skills to TA specified as latent variables rather than observed variables. This model is shown in Figure 4.11, below.

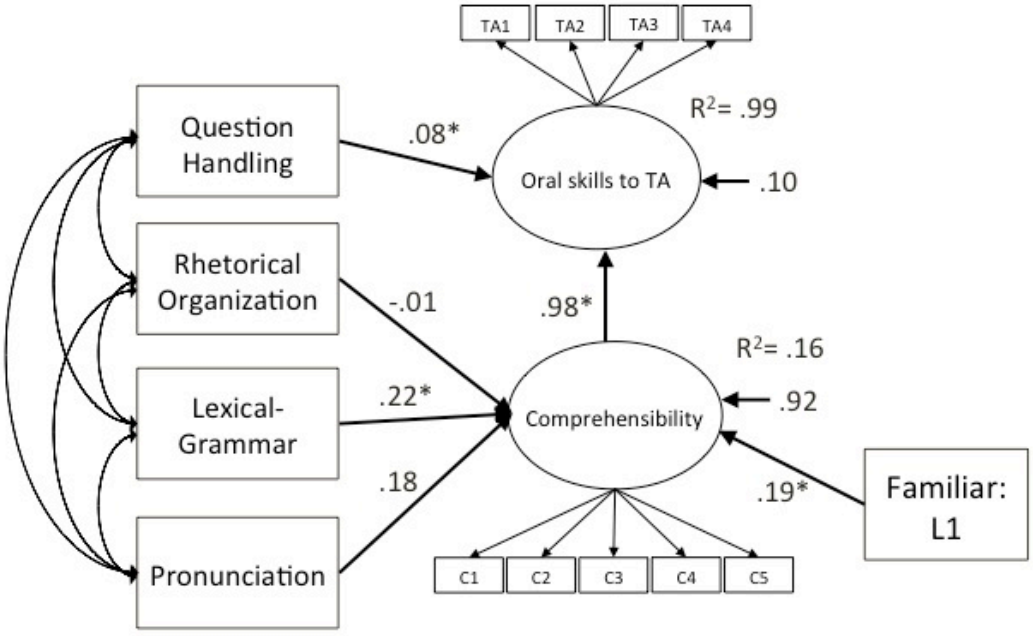

Figure 4.11. A structural model of the relationships between speaker oral proficiency variables and listener perceptions of the speaker (Main study, exploratory sample). * $p < .05$.

The magnitudes of the parameter estimates were similar to those in the previous model (see Figure 4.10). The direct effect of Lexical-Grammar on Comprehensibility is slightly higher in the model above (0.23 vs. 0.15), and is statistically significant. In addition, it is clear that when Comprehensibility and Oral skills to TA are specified as latent variables instead of

observed variables, the high correlation previously observed between the two (0.78) increases to virtual identity (0.98).

Thus, the difference in the effect of TOP Pronunciation on Comprehensibility across the two models does not necessarily appear to be related to changes to the conceptual model. There are important differences between the samples used in Schmidgall (2012) and the exploratory dataset, however. Key differences include (1) the number and diversity of listeners, (2) the duration of the listener's exposure to each individual speaker, (3) the nature of the listener's exposure to each individual speaker, and (4) the listener's experience with the Comprehensibility and Oral Skills to TA rating scales. The dataset used in Schmidgall (2012) contained less than 20 unique listeners, while the exploratory dataset contained a unique listener for each speaker – over 200 unique listeners. Listeners in the dataset used in Schmidgall (2012) listened to each speaker perform three tasks (self-introduction, syllabus presentation, mini-lecture) before rating, while those in the exploratory dataset only listened to each speaker perform one of the three tasks (mini-lecture). The nature of the interaction between speaker and listener varied between datasets. In the earlier dataset, listeners were directly engaged with speakers. Listeners in the exploratory dataset observed a video recording of that interaction and thus were not active participants. Finally, listeners in the earlier sample interacted with a large number of (ITA) speakers and thus were very familiar with both the speaking tasks and rating scales. Listeners in the second sample had very limited exposure to the speaking tasks and rating scales.

It is not entirely clear how the differences between the samples may explain the differences and similarities between the models. One fundamental difference is that ratings of Comprehensibility and Oral Skills to TA were based on speaker performance across three tasks in the first model, and only one task in the second: the mini-lecture. The content of the mini-

lecture task is chosen by the speaker, but is an academic topic that may be unfamiliar to the listener. Research suggests that topic familiarity may facilitate comprehensibility; topic familiarity or the lack thereof may be expected to have a larger impact on Comprehensibility and Oral Skills to TA ratings in the second model given that these ratings are entirely based on a task in which the speaker discusses a specific academic topic. If the listeners in the second sample are familiar with the topic, then they may have an easier time understanding the speaker regardless of the speaker's oral proficiency. In other words, when listener familiarity with the topic is high, there may be less of a predictive relationship between TOP Pronunciation and Comprehensibility. Conversely, TOP Pronunciation may provide a stronger prediction of Comprehensibility when the listener's familiarity with the topic is low.

In order to further investigate whether group differences (e.g., high or low in Topic familiarity) may have impacted the effect of TOP Pronunciation on Comprehensibility, grouping variables were created to identify listeners "high" or "low" on each listener-based factor. Next, correlations between TOP Pronunciation and Comprehensibility total scores were estimated based on group membership (High, Low) for the exploratory dataset.

Based on the results of this analysis, several variables were identified for which group membership led to a relatively higher or lower correlation between TOP Pronunciation and Comprehensibility. One such variable was a composite of Familiarity with speaker accent and Familiarity with speaker's native language (AFAM + LFAM). When a listener was high with respect to AFAM+LFAM, the correlation between TOP Pronunciation and Comprehensibility was low ($r=0.16$). Conversely, when a listener was low with respect to AFAM+LFAM, the correlation between TOP Pronunciation and Comprehensibility was moderately strong ($r=0.41$). This analysis was replicated using the cross-validation dataset and a similar relationship was

observed (High: r=0.12; Low: r=0.48). This implies that familiarity with the speaker's native language and accent may impact the relationship between TOP Pronunciation and Comprehensibility; in other words, when familiarity with the speaker's language and accent is low, TOP Pronunciation has a larger effect on Comprehensibility. Thus, familiarity with the speaker's native language and accent may diminish the effect of TOP Pronunciation on Comprehensibility.

4.2.1.6 Undergraduate interviews

Follow-up interviews were conducted with a subsample of the undergraduates who participated in the main study in order to investigate the following research questions, motivated by previous discussion:

1) Do naïve listeners (i.e., undergraduates who participated in the main study) consider different aspects of performance when rating a speaker's Comprehensibility and Oral skills to TA?
2) Do naïve listeners who are more familiar with the topic, and the speaker's native language and accent find the speaker more comprehensible?

4.2.1.6.1 Data and procedure

In order to address these questions in a qualitative fashion, undergraduates were selected to participate in interviews based on (1) High or Low familiarity with the speaker's native language and accent (AFAM+LFAM), and (2) High or Low familiarity with the speaker's lecture topic.

First, two speakers were identified to inform the selection of undergraduates. The first speaker (Speaker 1) was a native speaker of Korean with relatively low TOP Pronunciation and Lexical-Grammar scores, lecturing on a topic in Computer Science. He received relatively high

ratings for teaching effectiveness. The second speaker (Speaker 2) was a native speaker of Mandarin Chinese with slightly higher TOP Pronunciation and Lexical-Grammar scores, lecturing on a topic in Electrical Engineering. He received slightly lower scores for teaching effectiveness.

Based on group membership (AFAM+LFAM, Topic familiarity) relative to the speaker, undergraduates who previously participated in the main survey were invited to participate in follow-up interviews. Twenty-four undergraduates were invited to participate: twelve for each speaker. The twelve undergraduates invited for each speaker were selected based on their membership in one of the following four groups, based on their earlier response:

- High AFAM+LFAM, High Topic familiarity
- High AFAM+LFAM, Low Topic familiarity
- Low AFAM+LFAM, High Topic familiarity
- Low AFAM+LFAM, Low Topic familiarity

Among the participants who qualified based on these categories relative to each speaker, three were randomly selected for each group and speaker. In total, 17 participants responded to invitations to participate in follow-up interviews. Information about participants is summarized in Table 4.48, below.

Table 4.48

Characteristics of Undergraduates Who Participated in the Follow-up Interviews by Group Membership

Group	*n*	LFAM+ AFAM median	Topic familiarity median	Topic complexity median
Speaker 1: Korean L1				
High AFAM+LFAM, High Topic familiarity	2	9	2	4
High AFAM+LFAM, Low Topic familiarity	2	10	1	3.5
Low AFAM+LFAM, High Topic familiarity	3	2	3	2
Low AFAM+LFAM, Low Topic familiarity	3	2	1	5
Speaker 2: Mandarin Chinese L1				
High AFAM+LFAM, High Topic familiarity	3	9	6	1
High AFAM+LFAM, Low Topic familiarity	2	9.5	3.5	1.5
Low AFAM+LFAM, High Topic familiarity	1	3	6	1
Low AFAM+LFAM, Low Topic familiarity	1	4	3	2

The table above shows the median score for each category for each speaker. While the difference between language familiarity (LFAM+AFAM) group medians between speakers was similar, group differences between Topic familiarity medians between speakers were slightly different. Speaker 1 presented a relatively advanced topic, so even undergraduates from the same department as the speaker may have been unfamiliar with the topic. Conversely, Speaker 2 presented an introductory topic, so undergraduates from unrelated departments still reported some familiarity with the topic. This disparity is reflected in the Topic complexity group medians, which were generally higher for Speaker 1. Due to the small sample size, statistical comparisons between groups were not performed.

The follow-up interviews were conducted using the following procedure. First, the purpose of the study was described to participants, who signed informed consent forms in

accordance with IRB regulations. Next, participants viewed the instructional and training materials for viewing speaker performances as described in section 3.4.2. Instead of viewing randomly selected videos of two speakers as in the main study, participants viewed a video of one of the two speakers described above. After viewing the video and completing the listener rating scales (see section 3.4.2), each participant was asked to engage in a 10-minute semi-structured interview. Interviews were audio-recorded and transcribed with participants' consent. The semi-structured interview covered the following questions:

1) Overall, were there things that made it harder or easier for you to understand the TA? (factors influencing Comprehensibility)
2) Did it seem like the TA had a strong foreign accent? Was it an accent that you recognized, or were familiar with? (familiarity with, and impact of speaker's accent)
3) Based on the topic, is this is a class you have taken, or think you might take? (familiarity with, and relevance of speaker's topic)
4) Overall, what were the things that affected your ratings of Oral skills to TA? Were they the same or different from those described in (1)? (factors influencing Oral skills to TA)

When the interview began, the researcher focused the participant's attention on the Comprehensibility ratings she or he had previously completed. The first question above was thus clearly intended to identify aspects of the speaker's performance (or the listener's background) that may have influenced Comprehensibility. The second and third questions explored the importance of familiarity with the speaker's accent and topic, respectively. Prior to asking the fourth question, the researcher focused the participant's attention on the Oral skills to TA ratings she or he had previously completed. This question was intended to explore

differences or similarities between how participants responded to Comprehensibility and Oral skills to TA items. After the interview was completed, participants were compensated with two $10 gift certificates.

4.2.1.6.2 Results

4.2.1.6.2.1 Interview research question 1: Do naïve listeners consider different aspects of performance when rating Comprehensibility versus Oral skills to TA?

When asked directly, most participants (13 of 17) indicated that they considered the same aspects of performance when rating the two constructs. Several of the students suggested that there were some differences between how they rated the two constructs. For example, one of the participants said that he considered potential students' ability to adjust to the speaker's accent, as well as aspects of teaching effectiveness:

> I would probably say it was the same things, but also because considering [the speaker's] ability to teach over a longer span of time, I feel…his students would also become accustomed to the way he speaks as well so I feel like that is also something considered. I definitely weighted his actual skills as opposed to I guess his oral English proficiency more when I was considering how he would do teaching.
>
> -- Interview #8 (Low accent familiarity/High topic familiarity)

More indirectly, when asked to describe the aspects of performance that influenced their ratings for each scale independently, participants generally described the same or similar aspects for both Comprehensibility and Oral skills to TA. For Speaker 1, who received lower TOP Pronunciation and Lexical-Grammar scores but higher Teaching effectiveness scores, participants indicated that Comprehensibility was negatively influenced by the speaker's accent or pronunciation (6 of 10), grammar or vocabulary (4 of 10), and lack of familiarity with the topic (3 of 10). As one participant observed,

> I think the pronunciation was the huge part…sometimes when he was asking the questions to the audience or the students, he had to repeat several times because the other

> end couldn't understand what he was saying. So I think the pronunciation could probably affect negatively.
>
> -- Interview #5 (High accent familiarity/High topic familiarity)

Participants indicated that aspects of Teacher personality had a positive impact on Comprehensibility, including the speaker's friendliness (4 of 10), and activeness or enthusiasm (4 of 10). Participants linked these aspects of directly to Comprehensibility:

> Yeah, yah his personality was really...he was active so I mean like he pretty much answered the questions right away and he was really active I guess so he seemed to really be into teaching. Probably because if he's not proactive that might make the topic even more harder I guess to grasp.
>
> -- Interview #5 (High accent familiarity/High topic familiarity)

For eight of the ten participants who viewed Speaker 1's mini-lecture, aspects of Teaching effectiveness were cited as having a positive impact on Comprehensibility. Most participants (7 of 10) mentioned the speaker's organization and clarity, specifically his use of relevant examples:

> I know I thought he was a good TA in general because of the examples he gave, so I think that may have influenced how easy I thought it was to understand him and like the topic he was teaching. Because with his analogies or the examples that he asked the students, I thought that helped me a lot in understanding in general, not just understanding through his accent. So I think as a TA, he was a good TA. His organization made up for it.
>
> -- Interview #2 (Low accent familiarity/Low topic familiarity)

Comprehensibility was also influenced by the speaker's use of the white board. One participant, who rated the speaker very high in Comprehensibility (5.6/6), said the following:

> I think like it was he was pretty easy to understand because he wrote a lot of what he was talking about on the board…If he didn't write anything on the board, regardless of his accent, or anything, I would have not understood a big chunk of what he said. It would make a big pretty big difference because I tend to like learning looking at whatever the material.
>
> -- Interview #3 (High accent familiarity/High topic familiarity)

Participants believed that some aspects of their background, including familiarity with the speaker's topic, accent, and prior experience with ITAs potentially had positive or negative

effects on comprehensibility. One participant cited her familiarity with the speaker's accent as facilitating:

> I was familiar with his accent so it was a little easier to understand...I highly recognize the accent 'cause I think I'm the same ethnicity 'cause I'm Korean so it sounded like someone I would talk to normally but not in English. I think it made it a lot easier for me to understand as opposed to maybe a white or Hispanic person.
>
> -- Interview #9 (High accent familiarity/Low topic familiarity)

Another participant, who already had some familiarity with the topic, believed that more familiarity with it would have been more useful:

> I feel like yes, if I would have been more familiar with the topic it would affected my rating…I definitely think that knowing the topic, or like the terms used to describe and discuss the topic would have affected my comprehension of it as well.
>
> -- Interview #8 (Low accent familiarity/High topic familiarity)

Speaker 2 received slightly higher TOP Pronunciation and Lexical-Grammar scores than Speaker 1, but lower Teaching effectiveness scores. Participants who viewed Speaker 2's lecture cited accent or pronunciation (3 of 7), grammar or vocabulary (2 of 7), or fluency (1 of 7) as aspects of the speaker's oral proficiency that negatively impacted comprehensibility. For example:

> It was pretty much for this TA his accent, that he had a good command of English, that words weren't missing from the sentences or anything so you could get his ideas but certain words would be a little jumbled because he would pronounce them with an accent…Certain pronunciations would come out incorrect and you would have to be like I think this was this word.
>
> -- Interview #15 (Low accent familiarity/Low topic familiarity)

Participants also cited aspects of the speaker's performance related to Teacher personality and Teaching effectiveness. In general, participants suggested that the speaker's lack of enthusiasm had a negative impact on comprehensibility, but his knowledge of the material had a facilitating effect. One participant suggested that the speaker's enthusiasm or non-verbal communication could compensate in some way for his accent:

> I mean he was comprehensible but you know, with the accent also he doesn't seem very enthusiastic and almost a little bit monotone, like he's very knowledgeable and he knows what he's talking about, but it's almost to the point where he's just reciting everything that's in his head...It seemed like he was just trying to recite what was in his head and he needs to understand that we're students and we're just learning this. Like I said, just a little more engagement.
>
> -- Interview #12 (High accent familiarity/Low topic familiarity)

Finally, participants suggested that Speaker 2's comprehensibility benefitted from their familiarity with the topic, his accent, or from comparisons they made between him and other ITAs. As one participant observed:

> I'm taking Chinese and it seemed like a Chinese accent so it's easier for me. I recognize the phonology basically.
>
> -- Interview #11 (High accent familiarity/Low topic familiarity)

4.2.1.6.2.2 Interview research question 2: Do naïve listeners who are more familiar with the topic, and the speaker's native language and accent find the speaker more comprehensible?

Due to the size of the sample, it is not possible to test the significance of group comparisons. Participants who were familiar with the speaker's topic, and native language and accent typically provided high comprehensibility ratings. Participants with some or no familiarity of topic or accent provided a range of ratings, some of which were very low.

In general, participants who were unfamiliar with the speaker's native language and accent mentioned the speaker's pronunciation as an important factor that influenced comprehensibility (Speaker 1: 5 of 6; Speaker 2: 1 of 2). In contrast, participants who were familiar with the speaker's native language and accent infrequently mentioned pronunciation as a factor (Speaker 1: 1 of 4; Speaker 2: 2 of 5).

4.2.1.6.2.3 Summary of results

Overall, the interviews with these participants suggested that undergraduates did not make a distinction between the constructs of Comprehensibility and Oral skills to TA.

Undergraduates generally considered similar aspects of a speaker's performance when rating both scales. Aspects related to Teaching effectiveness and Teacher personality were frequently described as factors influencing Comprehensibility ratings, emphasizing the exchangeable nature of the constructs.

In addition, Comprehensibility ratings by subgroup(s) generally supported the hypothesis that a speaker's pronunciation was a more important factor when the listener's familiarity with the speaker's native language and accent was low, and comprehensibility was generally high when the listeners were already familiar with the speaker's topic, and native language and accent.

4.2.1.7 Revised conceptual model

Based on the analysis of the preliminary model and the follow-up interviews, several alterations to the model were suggested. The justification for each alteration is discussed in turn.

4.2.1.7.1 Revised Comprehensibility construct

First, the constructs of Comprehensibility and Oral skills to TA were merged into a single construct that indicated Comprehensibility as an ITA, or Comprehensibility within the ITA language use domain. This construct was hypothesized to be defined by items in both the Comprehensibility and Oral skills to TA scales.

A series of CFAs was conducted in order to investigate the psychometric qualities of a revised Comprehensibility construct defined by both Comprehensibility and Oral skills to TA scale items. First, a unidimensional model was estimated using the exploratory dataset in order to examine model fit. Next, a two-dimensional model specifying Comprehensibility and Oral skills to TA factors for designated scale items was fit using the exploratory dataset. Fit indices for each model are shown in Table 4.49, below.

Table 4.49

Fit Indices for the Unidimensional and Two-dimensional Models of Listener Perceptions of Speakers' Oral Language Use with the Exploratory Dataset

Model	χ^2	*df*	RMSEA (ε^)	RMSEA 90% CI	CFI	NNFI	SRMR
Unidimensional (Comprehensibility as an ITA)	140.73**	27	0.145	[0.121, 0.168]	0.934	0.912	0.043
Two dimensional (Comprehensibility, Oral skills to TA)	140.70**	26	0.148	[0.124, 0.172]	0.933	0.908	0.043

** $p < .01$

Based on the fit indices reported in Table 4.49, the two-dimensional model did not appear provide better fit to the data than the unidimensional model. The two dimensional model's use of an additional parameter (and thus, one less degrees-of-freedom) leads to a marginal reduction of the chi-square statistic (0.03), and no improvement or slightly degraded fit for the other indices.

Overall, model fit appeared to be adequate. Although the chi-square statistic and RMSEA point estimate were above the thresholds expected to indicate good fit, the CFI, NNFI, and SRMR estimates suggested good fit. Standardized factor loadings were high for all items (0.79 – 0.91) except one, an item in the Oral skills to TA scale relevant to the speaker's listening comprehension skills (TA3; see an earlier discussion of this item in section 4.1.1.5, in which it was labeled TA2).

Based on the item's comparatively low factor loading (0.45) and its conceptual distinctiveness, an additional CFA was conducted after removing it from the revised Comprehensibility scale. Again, a unidimensional model was specified using the exploratory

dataset in order to examine whether the removal of the item improved the fit of the model. Fit indices for this model are provided in the table below.

Table 4.50

Fit Indices for the Unidimensional Model of Comprehensibility after Removing Item TA2 with the Exploratory Dataset

χ^2	df	RMSEA (ε^)	RMSEA 90% CI	CFI	NNFI	SRMR
97.169**	20	0.139	[0.111, 0.166]	0.953	0.934	0.029

** $p < .01$

As seen in Table 4.50, the deletion of the item resulted in slightly improved measures of model fit. Based on its psychometric properties and superior conceptual coherence, the revised model of Comprehensibility that incorporated Oral skills to TA items TA1, TA3, and TA4 was retained for subsequent analyses.

The measurement model for the revised Comprehensibility construct is shown in Figure 4.12, below.

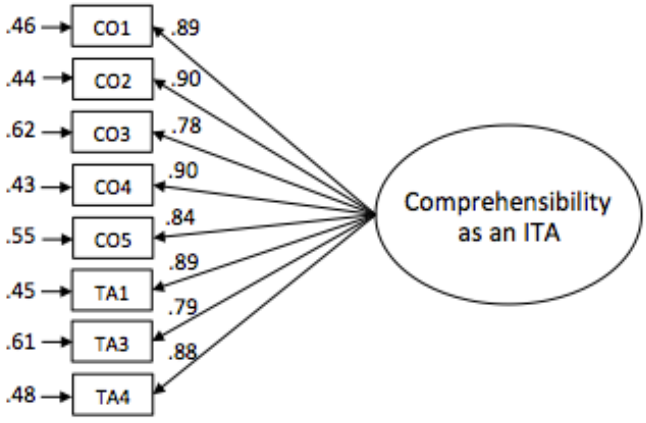

Figure 4.12. Measurement model for the construct of Comprehensibility as an ITA for the exploratory dataset.

4.2.1.7.2 Addition of Teacher personality measure to the conceptual model

Speaker enthusiasm, activeness, friendliness, knowledge, experience, and helpfulness were frequently mentioned by listeners in the follow-up interviews as important features of speakers that influenced Comprehensibility as an ITA. The Teacher personality measure described earlier (see sections 3.3.3.8, 4.1.2.3, and 4.2.1.4.5) but not included in the preliminary conceptual model may be introduced into a revised model to explain more of the variance in Comprehensibility as an ITA.

First, the specification of Teacher personality in the model must be justified conceptually and statistically. The construct of Teacher personality used in this study was defined as non-linguistic characteristics of the speaker expected to facilitate or inhibit classroom interaction between teachers and students, and included the following descriptive adjectives: *friendly*, *knowledgeable*, *helpful*, *active*, *experienced*. The scale is described as Teacher personality as it is a composite measure related to the speaker's personality and teaching skills. Speakers who score high on this measure are expected to be perceived as possessing positive personality characteristics with respect to the tasks performed in the domain (i.e., teaching). Prior analyses suggest that this scale is unidimensional (see section 4.2.1.4.5), and that Teacher personality ratings have relatively low interrater consistency (see section 4.1.2.3). Teacher personality total scores had a strong correlation with Comprehensibility (r=0.62) and Oral skills to TA (r=0.70) total scores in the exploratory dataset, and a moderate correlation with Attitude homophily total scores (r=0.44). Thus, based on the follow-up interviews and previous statistical analyses, Teacher personality is expected to have a direct effect on Comprehensibility.

The addition of Teacher personality to the revised conceptual model has implications for how other factors may be specified in the model. In the preliminary model, Attitude homophily

was hypothesized to have a direct effect on both Comprehensibility and Oral skills to TA. Conceptually, though, the relationship between Attitude homophily, Teacher personality, and Comprehensibility as an ITA needs to be reconciled.

In the current study, listeners' ratings of Teacher personality may be influenced by genuine characteristics of the speaker, and by listener preconceptions, expectations, or biases. Based on the low interrater consistency of the Teacher personality measure (see section 4.1.2.3), it appears that listener judgments of Teacher personality in this study may be more influenced by their expectations and biases than stable speaker personality characteristics. Attitude homophily may be considered a reflection of listener expectations, in that the listener provides an impressionistic measure of how similar or different the speaker is from themselves, a speculative judgment. When Attitude homophily scores are low, it implies that listeners believe that the speaker differs from them in fundamental ways. These impressionistic judgments may be expected to interact with relevant aspects of the teacher's performance to influence judgments of Teacher personality.

Several other listener-based factors may be expected to Teacher personality ratings. In the follow-up interviews, participants often compared the speaker they viewed to ITAs they have had in the past. Participants who had previously taken courses with ITAs described a process of "TA shopping", in which TA sections were chosen based on a comparative evaluation of TA English proficiency, teaching effectiveness, and personality characteristics. One participant was asked if she would stay enrolled in a course with the TA she had just viewed, whom she described as having moderate comprehensibility. She responded:

> Fifty-fifty. I would shop around to see who the other TAs were, but as far as this TA goes, I've seen and had to deal with worse [referring to his pronunciation] so this would be a better alternative to what I know is out there.
>
> -- Interview 15 (Low accent familiarity/Low topic familiarity)

Other participants in the interviews offered similar comments when describing aspects of Teacher personality, implying that their familiarity with ITAs in general may influence their judgments.

In a revised model, components of teaching effectiveness may have a more direct effect on Teacher personality judgments than Comprehensibility judgments. Given how the components of Teaching effectiveness have been defined in this study, this conceptual link is justified. These components – organization and clarity, non-verbal communication, teacher presence, and respect/rapport – would be expected to have a direct impact on the construct Teacher personality, which is defined the relevant characteristics *friendly*, *knowledgeable*, *helpful*, *active*, *experienced*.

Finally, one aspect of the oral skills to TA may be expected to have a more direct relationship with Teacher personality than Comprehensibility. TOP Question handling is defined by the speaker's interaction with TOP Questioners, the undergraduates trained to participate in the assessment procedure. This construct is thus based on the interaction between teacher and student (speaker and listener), and includes an evaluation of the speaker's receptive skills (listening comprehension) as well as production skills (speaking). Teacher personality judgments derive from an observation of this interaction, and are expected to be particularly influenced by it.

4.2.1.7.3 Relationship between familiarity with ITAs and familiarity with NNS

Based on the exploratory analysis, the fit of the model would improve if a relationship between familiarity with ITAs and familiarity with non-native speakers of English (NNS) were specified. This makes sense conceptually, as one would expect listeners who are more familiar

with ITAs to be more familiar with NNS in general, since interactions with ITAs are also interactions with NNS.

4.2.1.7.4 Revised conceptual model

Based on the justifications presented above, revisions were made to the preliminary model presented earlier to produce the revised conceptual model in Figure 4.13, below.

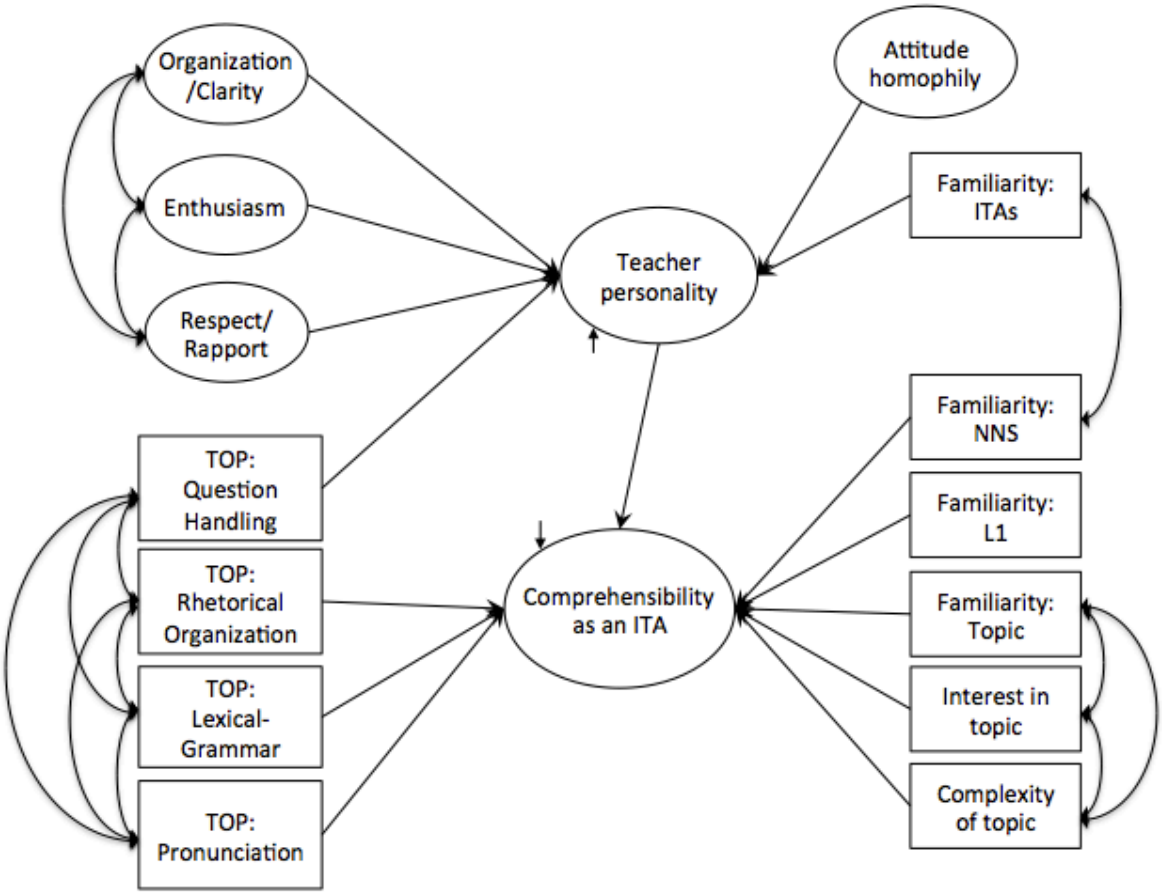

Figure 4.13. Revised conceptual model specifying the relationships between Comprehensibility as an ITA, Teacher personality, and speaker- and listener-related factors.

In the revised model, listener perceptions of Teacher personality are hypothesized to be influenced by the speaker's teaching effectiveness (Organization/Clarity, Enthusiasm,

Respect/Rapport), TOP Question handling, the listener's attitude towards the speaker (Attitude homophily), and the listener's familiarity with ITAs.

Comprehensibility as an ITA is in turn predicted by (a) a speaker's pronunciation, lexical-grammar, and rhetorical organization; (b) listener perceptions of Teacher personality; (c) listener familiarity with the speaker's native language, and with non-native speakers of English in general; and (d) listener familiarity with, interest in, and perceived complexity of the topic.

Parameter estimates for the revised structural model with the measurement models removed are shown in Figure 4.14, below.

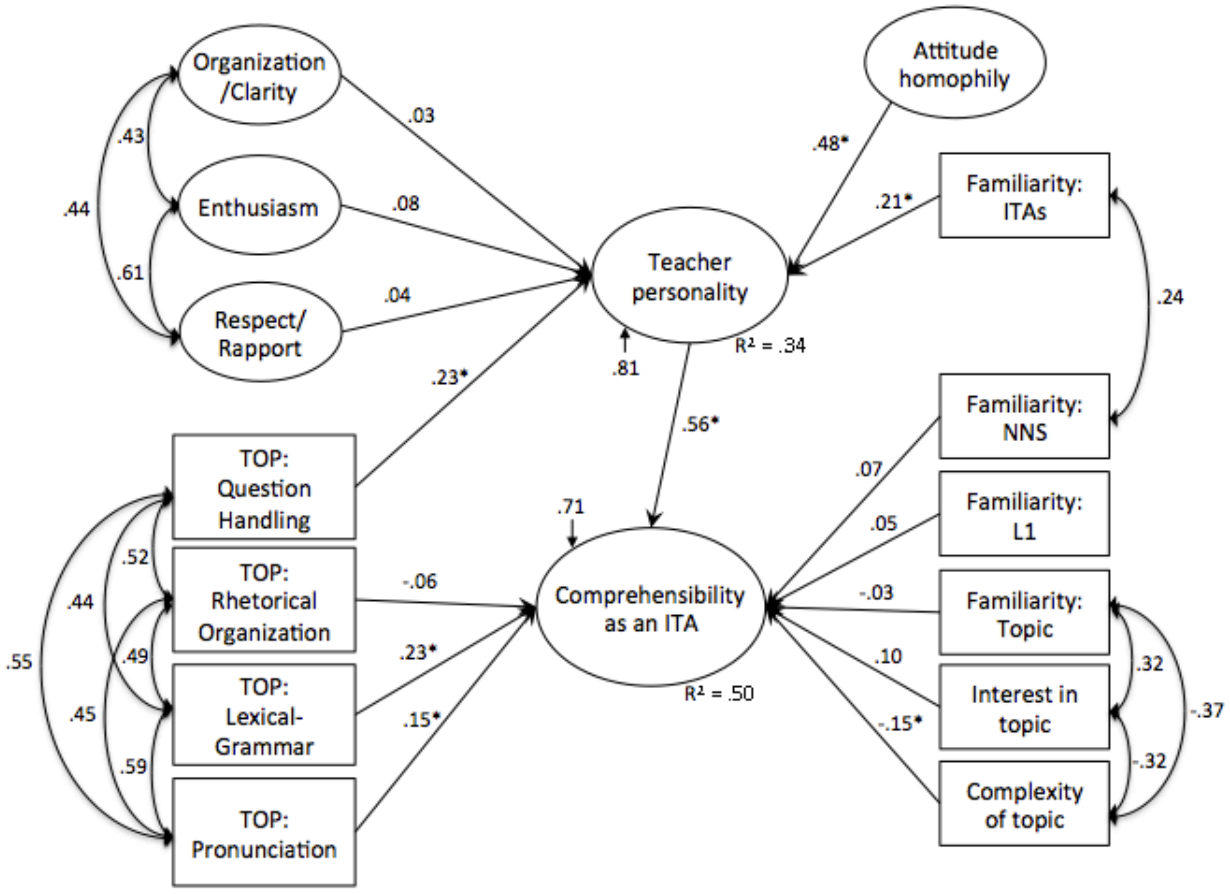

Figure 4.14. Parameter estimates for the revised structural model. * = $p < .05$.

In the model, Comprehensibility as an ITA s significantly predicted by a speaker's pronunciation (TOP: Pronunciation, .15), lexical-grammar (TOP: Lexical-Grammar, .23), listener judgments of Teacher personality (.56), and listeners' perceived complexity of the topic (-.15). Listener judgments of Teacher personality are predicted by a speaker's question handling (TOP: Question handling, .23), listeners' attitude homophily (.48), and listeners' familiarity with ITAs (.21). Standardized solutions for the structural equations are presented in Table 4.51, below.

Table 4.51

Standardized Solutions for Structural Equations in the Revised Conceptual Model (with Direct Effects only)

Dependent factor		Independent variables and factors	R^2
Comprehensibility as an ITA	=	**0.15*Pronunciation** + **0.23*Lexical-Grammar** – 0.06*Rhetorical organization + **0.56*Teacher personality** – 0.03*Topic familiarity + 0.10*Topic interest – **0.15*Topic complexity** + 0.05*LFAM + 0.07*NNS familiarity + 0.71 D1 (disturbance)	0.50
Teacher personality	=	**0.48*Attitude homophily** + **0.23*Question handling** + **0.21*ITA familiarity** + 0.03*Organization/Clarity + 0.08*Enthusiasm + 0.04*Respect/Rapport + 0.81 D2 (disturbance)	0.34

Note. Statistically significant ($p < .05$) parameters are shown in bold type.

Table 4.51 summarizes the parameter estimates for each independent variable or factor in the two structural equations specified in the model. Approximately 50% of the variation in Comprehensibility as an ITA is explained by nine predictors in the model, of which four are statistically significant. The largest predictor of Comprehensibility as an ITA is Teacher personality (0.56).

Approximately 34% of the variance in Teacher personality is explained by seven predictors in the model. Three of these predictors are statistically significant, including Attitude homophily (.48), Question handling (.23), and ITA familiarity (.21).

Fit indices for the revised model are shown in Table 4.52, below.

Table 4.52

Fit Indices for the Revised Structural Model

χ^2	*df*	RMSEA (ε^)	RMSEA 90% *CI*	CFI	NNFI	SRMR
1886.17**	1014	0.067	[0.062, 0.071]	0.838	0.828	0.101

** $p < .01$

Model fit indices provided some evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (1014) = 1886.17, p < 0.01. The RMSEA point estimate (RMSEA = 0.067) suggested that the model provided adequate fit, and the upper limit of the 90% confidence interval for the point estimate was below the threshold generally used to indicate adequate fit. The SRMR, CFI, and NNFI indices suggested poor fit (SRMR = 0.101; CFI = 0.838; NNFI = 0.828).

The matrix of standardized residuals was examined to identify parameter estimates with large residuals. There were several parameters with very large (greater than 0.30) standardized residuals. They primarily involved covariation between the listener's interest in the topic, and items related to Attitude homophily and Teacher personality. Based on follow-up interviews with participants, it was not clear whether interest in the topic influenced listener Attitude homophily or Teacher personality, or an inverse relationship held. Thus, no further changes to the model were recommended in order to account for this covariance.

4.2.2 Cross-validation phase

4.2.2.1 Dataset

The cross-validation dataset (n=205) was cleaned to ensure valid responses, descriptive statistics were produced and evaluated to investigate univariate normality, and variables were transformed when necessary. Three cases were removed during data cleaning, one variable was transformed, but no univariate outliers were detected. In total, 203 valid responses were retained to use in the subsequent analysis.

4.2.2.2 Data cleaning

Data was cleaned to ensure valid responses by examining participants' response patterns to listener-based scales, and flagging participants who (a) self-identified as non-native speakers of English with low levels of listening comprehension, (b) self-identified as non-native speakers of English but did not identify a language other than English in which they were proficient, (c) indicated that he or she knew the speaker in the video. This procedure was identical to the one described in section 4.2.1.2, for the exploratory dataset. As a result of the validity check, 3 cases were removed from the cross-validation sample (n=202).

4.2.2.3 Descriptive statistics

Descriptive statistics, outlier detection, and investigations of the assumption of univariate normality were performed for all variables included in the analysis. As with the exploratory dataset, groups of variables were examined in turn based on their designation as speaker-based or listener-based components in the conceptual model.

4.2.2.3.1 Speaker-based components

4.2.2.3.1.1 TOP oral proficiency measures

Descriptive statistics for the four TOP oral proficiency measures are shown in Table 4.53, below.

Table 4.53

Descriptive Statistics for TOP Oral Proficiency Measures in the Cross-validation Dataset

Variable	*M* (*SD*)	range	skew	kurtosis
Pronunciation	5.31 (1.25)	2-8	0.31	-0.42
Lexical-Grammar	5.87 (1.17)	4-8	0.18	-0.45
Rhetorical organization	6.24 (0.94)	4-8	0.25	0.16
Question handling	6.35 (0.97)	3-8	0.08	0.34

Distributions appeared to be approximately normal, as evidenced by the estimates of univariate skew and kurtosis in the table above. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality.

Potential outliers in bivariate distributions between TOP oral proficiency variables and other variables specified in the revised conceptual model (i.e., Comprehensibility as an ITA, Teacher personality) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, Comprehensibility as an ITA and Teacher personality total scores were used in this analysis instead of latent variable models. No outliers were detected.

Pearson correlations for bivariate distributions specified in the conceptual model are shown in Table 4.54, below.

Table 4.54

Correlations between TOP Oral Proficiency Measures, and Comprehensibility and Teacher Personality Total Scores in the Cross-validation Dataset

	PRO	LG	RO	QH	COMP total	Teacher personality total
PRO	1.00					
LG	0.57**	1.00				
RO	0.51**	0.39**	1.00			
QH	0.58**	0.58**	0.53**	1.00		
COMP total	0.36**	0.20**	0.24**	--	1.00	
Teacher personality total	--	--	--	0.31**	0.53**	1.00

Note. PRO = Pronunciation; LG = Lexical/Grammar; RO = Rhetorical organization; QH = Question handling; COMP = Comprehensibility.

** $p < .01$

All of the correlations included in the table were statistically significant and ranged from weak to moderately strong positive relationships. TOP oral proficiency measures had moderately strong bivariate correlations (0.39 – 0.58). The TOP oral proficiency measures hypothesized to predict comprehensibility ratings (pronunciation, lexical-grammar, rhetorical organization) had relatively weak to moderate correlations with comprehensibility total scores (0.20 – 0.36). The TOP oral proficiency measure hypothesized to predict Teacher personality ratings (question handling) had a moderately weak correlation with Teacher personality total scores (r=0.31).

4.2.2.3.1.2 Teaching effectiveness measures

Descriptive statistics for the items in the componential scales (Organization/Clarity, Enthusiasm, Respect/rapport) are shown in Table 4.55, below.

Table 4.55

Descriptive Statistics for Teacher Personality Measures in the Cross-validation Dataset

Variable	*M* (*SD*)	range	skew	kurtosis
Organization/Clarity scale items				
O1	3.98 (0.99)	1-5	-1.07	0.81
O2	3.71 (1.26)	1-5	-0.86	-0.36
O3	3.99 (0.79)	2-5	-0.58	0.09
O4	3.91 (1.01)	1-5	-0.70	-0.46
O5	4.18 (0.82)	2-5	-0.93	0.50
O6	3.78 (0.97)	1-5	-0.60	-0.18
O7	3.93 (1.05)	1-5	-0.85	-0.17
Enthusiasm (Non-verbal immediacy) items				
EN1	3.08 (1.35)	1-5	-0.08	-1.33
EN2	3.82 (1.12)	1-5	-0.88	-0.16
EN3	3.89 (1.16)	1-5	-0.96	-0.20
EN4	3.51 (1.26)	1-5	-0.36	-1.13
EN5	2.86 (1.36)	1-5	0.07	-1.38
EN6	3.52 (1.18)	1-5	-0.80	-0.31
EN7	3.05 (1.36)	1-5	-0.07	-1.29
Respect/Rapport scale items				
RR1	3.58 (0.95)	1-5	-0.06	-0.50
RR2	3.85 (1.00)	1-5	-0.44	-0.65
RR3	4.41 (0.74)	2-5	-1.32	1.82
RR4	3.88 (0.78)	2-5	-0.04	-0.87
RR5	3.23 (1.01)	1-5	0.02	-0.43
RR6	4.20 (0.90)	2-5	-0.80	-0.42

Distributions consistently exhibited negative skew, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for teaching effectiveness items.

Potential outliers in bivariate distributions between teaching effectiveness variables and other variables specified in the revised conceptual model (i.e., Teacher personality) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, total scores were used in this analysis instead of latent variable models for the following multi-item scales: Organization/Clarity, Enthusiasm, and Respect/Rapport. No outliers were found.

Pearson correlations for bivariate distributions specified in the conceptual model were estimated. Correlations between teaching effectiveness component measures were moderately strong (r=0.27-0.53). As expected, correlations between Organization/Clarity and the other components of teaching effectiveness (Enthusiasm, Respect/Rapport) were slightly lower than the correlations among those components. In other words, Enthusiasm and Respect/Rapport measures had higher correlations with each other than with Organization/Clarity measures.

Correlations between teaching effectiveness measures and Teacher personality total scores were moderately weak. All of the teaching effectiveness measures were significant predictors of Teacher personality total scores, with the largest predictors being Respect/Rapport (r=0.31), and Organization/Clarity (r=0.22).

4.2.2.3.2 Listener-based components

4.2.2.3.2.1 Comprehensibility as an ITA

Descriptive statistics for the eight items in the Comprehensibility as an ITA scale are shown in Table 4.56, below.

Table 4.56

Descriptive Statistics for Comprehensibility as an ITA items in the Cross-validation Dataset

Item	*M* (*SD*)	range	skew	kurtosis
CO1	3.37 (1.73)	1-6	0.09	-1.36
CO2	3.75 (1.58)	1-6	-0.16	-1.28
CO3	3.90 (1.61)	1-6	-0.36	-1.14
CO4	3.52 (1.62)	1-6	0.00	-1.29
CO5	3.81 (1.48)	1-6	-0.17	-1.12
TA1	3.77 (1.59)	1-6	-0.22	-1.20
TA3	3.84 (1.63)	1-6	-0.15	-1.30
TA4	3.80 (1.68)	1-6	-0.30	-1.21

Variable distributions consistently exhibited negative kurtosis, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for comprehensibility items.

4.2.2.3.2.2 Attitude homophily

Descriptive statistics for the four items in the Attitude homophily scale are shown in Table 4.57, below.

Table 4.57

Descriptive Statistics for Attitude Homophily Items in the Cross-validation Dataset

Item	*M* (*SD*)	range	skew	kurtosis
AH1	3.11 (1.45)	1-6	0.04	-1.10
AH2	2.94 (1.44)	1-6	0.28	-1.02
AH3	2.97 (1.37)	1-6	0.17	-1.00
AH4	2.78 (1.26)	1-6	0.26	-0.76

Variable distributions consistently exhibited negative kurtosis, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for Attitude homophily items.

In the revised conceptual model, Attitude homophily is hypothesized to predict Teacher personality. For expediency, Teacher personality and Attitude homophily total scores were used to investigate potential outliers instead of latent variable models. No outliers were detected. The Pearson correlation for this bivariate distribution was moderately strong (r=0.54).

4.2.2.3.2.3 Teacher personality

Descriptive statistics for the five items in the Teacher personality scale are shown in Table 4.58, below.

Table 4.58

Descriptive Statistics for Teacher Personality Items in the Cross-validation Dataset

Item	*M* (*SD*)	range	skew	kurtosis
P1	4.91 (1.10)	2-6	-0.91	0.11
P2	4.79 (1.24)	1-6	-1.07	0.56
P3	4.16 (1.43)	1-6	-0.46	-0.73
P4	4.21 (1.43)	1-6	-0.57	-0.60
P5	3.88 (1.51)	1-6	-0.33	-0.97

Variable distributions consistently exhibited negative skewness, as evidenced by the estimates in the table above. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary to maintain the assumption of normality for teacher personality items.

4.2.2.3.2.4 Other listener-based measures

Descriptive statistics for additional listener-based measures in the revised conceptual model are shown in Table 4.59, below.

Table 4.59

Descriptive Statistics for Listener-based Measures in the Cross-validation Dataset

Item	*M* (*SD*)	range	skew	kurtosis
Topic familiarity	3.01 (1.90)	1-6	0.33	-1.44
Topic interest	3.49 (1.68)	1-6	-0.10	-1.26
Topic complexity	2.73 (1.57)	1-6	0.55	-0.89
LFAM	2.19 (1.63)	1-5	0.89	-0.95
NNS familiarity	3.72 (1.06)	1-5	-0.26	-1.09
ITA familiarity	3.36 (3.57)	0-30	2.79	14.68

Note. LFAM = Familiarity with the native language (L1) of the speaker; NNS familiarity = Familiarity with non-native speakers of English by frequency of interaction.

For most variables in the table, distributions in both samples consistently exhibited negative kurtosis. However, the relative size of skewness and kurtosis estimates did not suggest severe departures from normality. Visual examinations of histograms for each variable also indicated that the distributions were approximately normal and that there were no outliers. Based on this analysis, it was determined that no variable transformations were necessary for most items to maintain the assumption of normality.

The ITA familiarity variable exhibited unacceptably high absolute levels of univariate skewness (~2.8) and kurtosis (~14.7). This variable was transformed into an ordinal variable using the procedure described in section 4.2.1.3.2.5. Descriptive statistics for the transformed variable are provided in Table 4.60, below.

Table 4.60

Descriptive Statistics for Transformed ITA Familiarity Variable in the Cross-validation Dataset

Item	*M* (*SD*)	range	skew	kurtosis
ITA familiarity	2.50 (0.96)	1-4	-0.02	-0.97

Variable distributions in both samples exhibited negative kurtosis, as evidenced by the estimates in the table above and the histogram provided in Figure 4 above. However, the relative size of skewness and kurtosis estimates for the transformed variable did not suggest severe departures from normality. Based on this analysis, it was determined that the transformed Topic familiarity variable should be retained.

Potential outliers in bivariate distributions between listener background variables and other variables specified in the revised conceptual model (i.e., Comprehensibility as an ITA, Teacher personality) were investigated by (1) examining scatterplots for all bivariate distributions, (2) estimating Bonferroni p-values for extreme observations, and (3) identifying influential observations using Cook's distance. For expediency, Comprehensibility as an ITA and Teacher personality total scores were used in this analysis instead of latent variable models. No outliers were detected.

Pearson correlations for bivariate distributions specified in the revised conceptual model are shown in the tables below. Table 4.61 shows correlations between relevant listener background variables and Comprehensibility as an ITA total scores. Table 4.62 shows the correlation between listener Familiarity with ITAs and Teacher personality total scores.

Table 4.61

Correlations among Listener Background Measures and Comprehensibility an ITA Total Scores for the Cross-validation Dataset

	Topic familiarity	Topic interest	Topic complexity	LFAM	NNS familiarity	COMP total
Topic familiarity	1.00					
Topic interest	0.25**	1.00				
Topic complexity	-0.39**	-0.28**	1.00			
LFAM	--	--	--	1.00		
NNS familiarity	--	--	--	--	1.00	
COMP total	0.15*	0.41**	-0.22**	0.07	0.00	1.00

Note. LFAM = Familiarity with the native language (L1) of the speaker; NNS familiarity = Familiarity with non-native speakers of English by frequency of interaction.

* $p < .05$

** $p < .01$

Table 4.62

Correlation between ITA Familiarity and Teacher Personality Total Scores for the Cross-validation Dataset

	ITA familiarity	Teacher personality
ITA familiarity	1.00	
Teacher personality	-0.08	1.00

Correlations between listener background variables and Comprehensibility as an ITA total scores were generally non-significant or moderately weak. Topic interest (r=0.41) and topic complexity (r=-0.22) had the strongest correlations with Comprehensibility as an ITA total scores among listener background variables. Listener variables related to the topic (topic familiarity, interest, and complexity) had moderately weak bivariate correlations.

4.2.2.4 Multivariate normality

The assumption of multivariate normality was investigated by estimating Mardia's coefficient for the revised structural model. Initially, a large estimate was obtained (Mardia's coefficient = 7.14), which indicated a high level of multivariate kurtosis. Six cases in the cross-validation dataset were found to disproportionately contribute to the estimate of multivariate kurtosis, and were removed from the dataset. The resulting estimate of multivariate kurtosis was closer to the threshold considered acceptable (Mardia's coefficient = 3.69).

4.2.2.5 Measurement models

For each of the measurement models specified in the revised structural model (Teaching effectiveness components, Comprehensibility, Teacher personality, Attitude homophily), confirmatory factor analysis (CFA) was performed specifying a single underlying factor.

4.2.2.5.1 Teaching effectiveness components

The measurement model for the components of Teaching effectiveness is reproduced with parameter estimates in Figure 4.14, below.

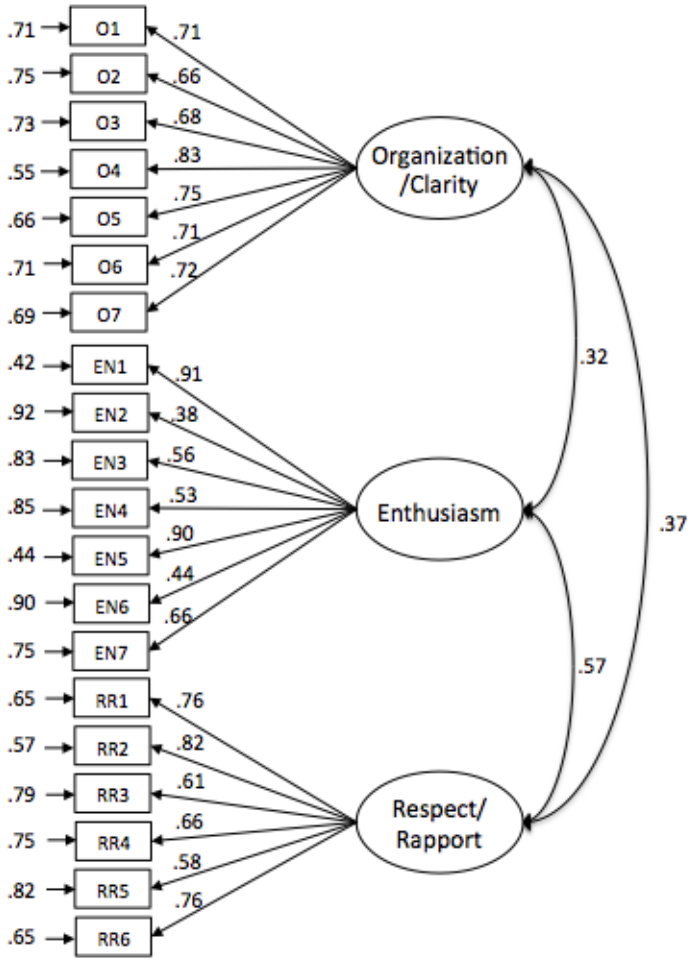

Figure 4.14. Measurement model for Teaching effectiveness components for the cross-validation dataset.

Model fit indices generally indicated poor fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (167) = 557.09, $p < 0.01$. The RMSEA point estimate (RMSEA = 0.110) was higher than the threshold typically used to indicate adequate model fit. Another residual-based fit index, the SRMR, suggested marginal fit (SRMR = 0.097). Several incremental fit indices suggested the model provided poor fit (CFI = 0.802; NNFI = 0.774).

4.2.2.5.2 Comprehensibility as an ITA

The measurement model for Comprehensibility as an ITA is reproduced with parameter estimates in Figure 4.15, below.

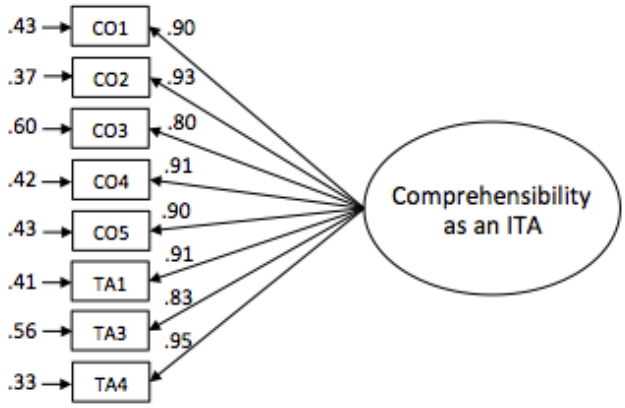

Figure 4.15. Measurement model for Comprehensibility as an ITA for the cross-validation dataset.

Model fit indices generally provided evidence of good fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (20) = 104.45, $p < 0.01$. The RMSEA point estimate (RMSEA = 0.150) also suggested that the model did not provide a good fit to the data. The standardized root mean-square residual (SRMR), suggested good fit, SRMR=0.029. Several incremental fit indices also suggested the model provided good fit (CFI = 0.955; NNFI = 0.938).

4.2.2.5.3 Attitude homophily

The measurement model for Attitude homophily is reproduced with parameter estimates in Figure 4.16, below.

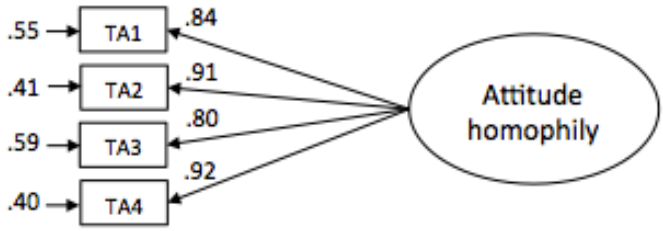

Figure 4.16. Measurement model for Attitude homophily for the cross-validation dataset.

Model fit indices generally provided evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (2) = 28.89, p < 0.01. The RMSEA point estimate (RMSEA = 0.263) also suggested that the model did not provide a good fit. However, the 90% confidence interval around the point estimate was large (0.183, 0.350), which suggests that the point estimate should be interpreted with caution given its imprecision. The SRMR suggested good fit, SRMR=0.032. The incremental fit indices provided mixed information, with the CFI indicating good fit (CFI = 0.957) and the NNFI indicating marginal fit (NNFI = 0.872).

4.2.2.5.4 Teacher personality

The measurement model for Teacher personality is reproduced with parameter estimates in Figure 4.17, below.

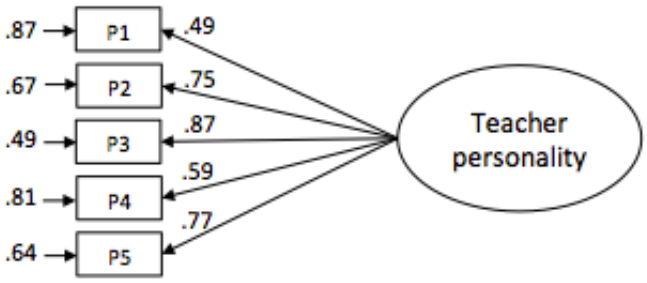

Figure 4.17. Measurement model for Teacher personality for the cross-validation dataset.

Model fit indices generally provided evidence of good fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (5) = 17.52, p < 0.01. The RMSEA point estimate (RMSEA = 0.113) also suggested that the model did not provide a good fit. Again, the 90% confidence interval around the point estimate was large (0.058, 0.173), which suggests that the point estimate should be interpreted with caution given its imprecision. The SRMR suggested adequate fit, SRMR=0.042. Both the CFI and NNFI estimates indicated good fit (CFI = 0.965, NNFI = 0.930).

4.2.2.6 Structural model

4.2.2.6.1 Parameter estimates

The number of cases in the cross-validation dataset used to estimate the revised structural model was reduced to 196 after removing the six cases contributing to multivariate kurtosis as described in section 4.2.3.4. In addition, listwise deletion was used to remove 6 cases that had missing data, further reducing the size of the dataset to 190 cases. Parameter estimates for the structural model with measurement models removed are shown in Figure 4.18, below.

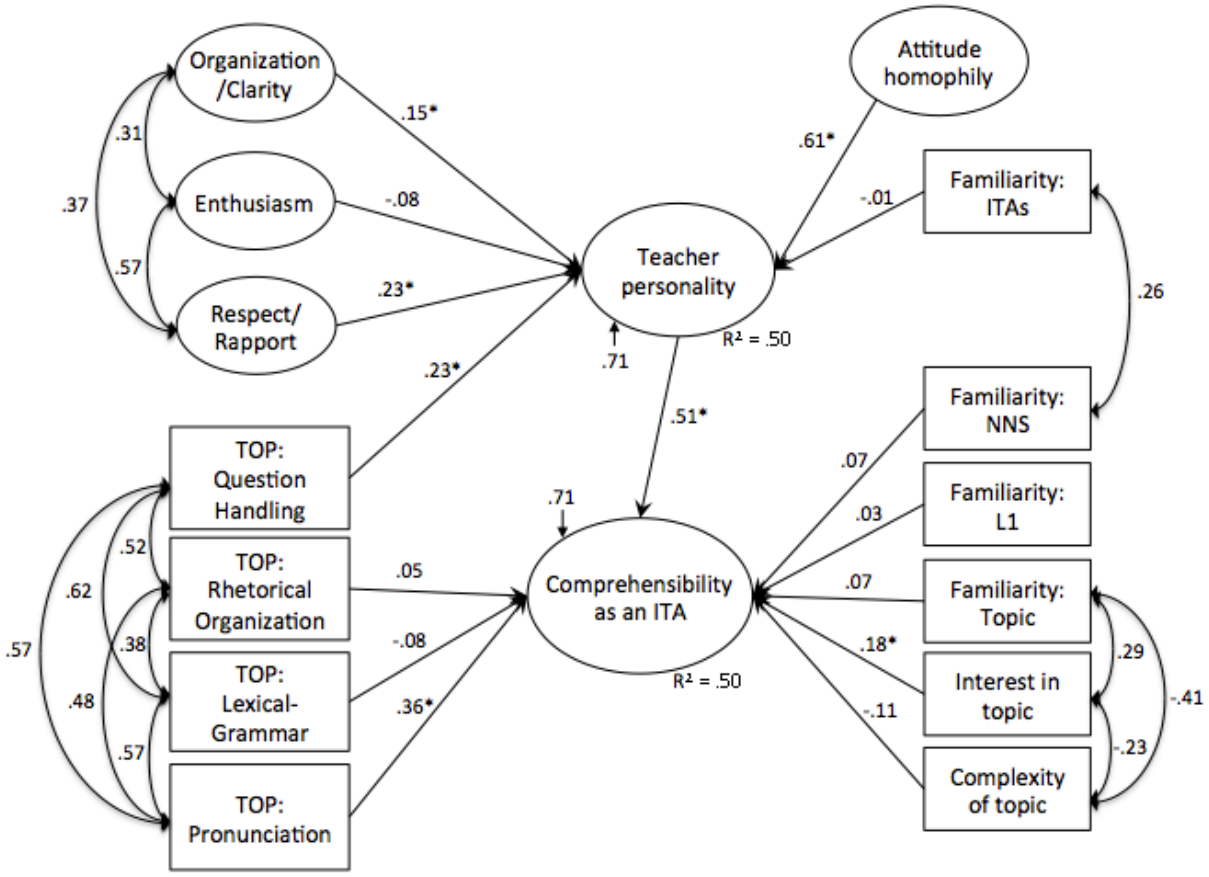

Figure 4.18. Parameter estimates for the revised structural model with cross-validation dataset. * $p < .05$.

In this model, Comprehensibility as an ITA is significantly predicted by a speaker's pronunciation (TOP: Pronunciation, .36), listener perceptions of Teacher personality (.51), and listener interest in the topic (.18). Listener perceptions of Teacher personality are largely predicted by Attitude homophily (.61), but also by the speaker's question handling (TOP: Question handling, .23), organization and clarity (.15), and respect/rapport with students (.23). Standardized solutions for the structural equations are presented in Table 4.63, below.

Table 4.63

Standardized Solutions for Structural Equations in the Revised Model with Cross-validation Dataset

Dependent factor		Independent variables and factors	R^2
Comprehensibility as an ITA	=	**0.36*Pronunciation** – 0.08*Lexical-Grammar + 0.05*Rhetorical organization + **0.51*Teacher personality** + 0.07*Topic familiarity + **0.18*Topic interest** – 0.11*Topic complexity + 0.07*LFAM – 0.03*NNS familiarity + 0.71 D1 (disturbance)	0.50
Teacher personality	=	**0.61*Attitude homophily** + **0.23*Question handling** – 0.01*ITA familiarity + **0.15*Organization/Clarity** – 0.08*Enthusiasm + **0.23*Respect/Rapport** + 0.71 D2 (disturbance)	0.50

Note. Statistically significant ($p < .05$) parameters are shown in bold type.

As can be seen in Table 4.63, approximately 50% of the variation in Comprehensibility as an ITA is explained by the nine predictors in the model, of which three are statistically significant. The largest predictors of Comprehensibility are Teacher personality (0.51) and Pronunciation (0.36).

Approximately 50% of the variation in Teacher personality is explained by the seven predictors in the model, of which four are statistically significant. The largest predictor of Teacher personality is Attitude homophily (0.61).

4.2.2.6.2 Model fit

Fit indices are shown in Table 4.64, below.

Table 4.64

Fit Indices for the Revised Structural Model using the Cross-validation Dataset

χ^2	df	RMSEA (ε^)	RMSEA 90% *CI*	CFI	NNFI	SRMR
1798.21**	1016	0.064	[0.059, 0.068]	0.861	0.852	0.098

** $p < .01$

Model fit indices provided some evidence of adequate fit. The chi-square statistic suggested that the null hypothesis that the model provided perfect fit to the data should be rejected, χ^2 (1016) = 1798.21, $p < 0.01$. The RMSEA point estimate (RMSEA = 0.064) suggested that the model provided adequate fit, and the upper limit of the 90% confidence interval for the point estimate was below the threshold generally used to indicate adequate fit. The SRMR, CFI, and NNFI indices suggested marginal fit (SRMR = 0.098; CFI = 0.861; NNFI = 0.852).

The matrix of standardized residuals was examined to identify parameter estimates with large residuals. There were several parameters with very large (greater than 0.30) standardized residuals, and they generally indicated unexpected positive covariance with Teacher personality or Attitude homophily items, and the listener's interest in the topic. Overall, the distribution of standardized residuals had a desirable shape: leptokurtic, centered at zero.

4.3 Summary of results for research questions

For several reasons, two of the primary constructs included in the preliminary model and research questions were redefined as a single construct. When Comprehensibility and Oral skills to TA were measured as latent variables, statistical analyses suggested that the combined scale items were indicators of a single underlying factor: Comprehensibility as an ITA. Follow-up interviews with listeners corroborated these findings by suggesting that listener judgments of Comprehensibility and Oral skills to TA were influenced by similar aspects of a speaker's performance and listener background characteristics. As a result, research questions were slightly modified to account for this change.

4.3.1 Research question 1

What are the relationships between listener perceptions of oral language use, and speaker- and listener-related factors for a subpopulation of listeners from the TLU domain?

The revised model with the cross-validation dataset and corresponding structural equations describe these relationships (see Figure 23 and Table 13, in section 4.2.3.6.1). Listener perceptions of oral language use, or Comprehensibility as an ITA, are largely predicted by listener impressions of Teacher personality. Comprehensibility as an ITA is also predicted by a speaker's pronunciation, and to a lesser extent, the listener's interest in the topic.

Teacher personality impressions, in turn, are largely predicted by attitude homophily. Several speaker-based factors also predict Teacher personality, including the speaker's question handling, organization and clarity, and respect/rapport with students.

4.3.2 Research question 2

To what extent do speaker-related versus listener-related factors affect listener perceptions of comprehensibility as an ITA?

Both speaker-related and listener-related factors affected Comprehensibility as an ITA. However, listener-related factors appeared to play a slightly bigger role. Based on the revised model with the cross-validation dataset, Comprehensibility is largely predicted by Teacher personality. Teacher personality itself is a composite of listener impressions related to the speaker's personality and experience related to teaching. Teacher personality impressions are largely predicted by listeners' attitude homophily judgments.

In addition to Teacher personality, a speaker's pronunciation was a significant predictor of Comprehensibility as an ITA. This result was consistent with findings from the initial pilot study and previous research. In addition, listener interest in the topic was a significant predictor of comprehensibility.

Several speaker- and listener-related factors had statistically significant indirect effects on comprehensibility. Speaker-related factors with very small but significant indirect effects included organization and clarity, respect/rapport with students, and question handling. Attitude homophily, a listener-related factor, had a slightly larger indirect effect (0.31).

4.3.3 Research question 3

To what extent do construct-relevant factors, i.e., pronunciation, lexical-grammar, rhetorical organization, question handling, versus construct-irrelevant factors, e.g., listener attitudes towards the speaker, affect comprehensibility as an ITA?

The construct-relevant factors that were statistically significant predictors of comprehensibility included speaker pronunciation (0.36, direct effect) and question handling (0.11, indirect effect) accounted for approximately 15% of the variance in Comprehensibility as an ITA. In contrast, construct irrelevant factors—Teacher personality (0.51, direct effect); listener-related factors such as attitude homophily (0.31, indirect effect) and topic interest (0.18,

direct effect); and speaker-related factors such as respect/rapport (0.11, indirect effect) and organization/clarity (0.08, indirect effect)—accounted for approximately 35% of the variance in Comprehensibility as an ITA. Thus, most of the statistically significant predictors of Comprehensibility as an ITA in the revised model were not relevant to the construct of Oral skills to TA as defined by the TOP.

Chapter 5: Discussion

This study specified and cross-validated a conceptual model centered on a speaker's comprehensibility to naïve listeners within a particular academic language use domain. As such, it facilitated a description and comparison of the various speaker- and listener-related factors that influenced comprehensibility in the domain of TA language use. As a model of interaction between speakers and listeners, it supports the interactional perspective on oral language and has implications for language assessment (see section 5.2.1) and language learning and teaching (see section 5.2.2). As a model of interaction between teachers and students, it implies that the construct of oral proficiency in teacher evaluation and certification needs to be more carefully examined to account for the perspective of naïve listeners (see section 5.2.3). After summarizing the key findings from this study, implications for language assessment, language learning, and education policy will be examined; methodological and conceptual constraints will be further delineated; and follow-up research activities will be explored. This discussion will end by emphasizing an important consideration for oral proficiency assessments in this domain: providing sufficient information about "real world" constructs to make appropriate decisions.

5.1 Key findings

The exploratory phase of the analysis found support for a revised conceptual model in which a listener's perception of a speaker's comprehensibility as an ITA was predicted by speaker-related factors (oral proficiency, teaching effectiveness), non-linguistic listener perceptions of the speaker (teacher personality), and listener-related factors (attitude homophily; topic familiarity, interest, and complexity; familiarity with the speaker and non-native speakers of English in general). A structural model that operationalized this revised conceptual model was estimated and then cross-validated. This model predicted more of the variance in a listener's

perception of a speaker's comprehensibility compared to the preliminary model (50% vs. 32%), and was more coherent conceptually.

This revised model differed from the preliminary model in several key aspects. First, the two constructs related to a naïve listener's perception of the speaker's oral language use – Comprehensibility and Oral skills to TA – were combined into a single construct, Comprehensibility as an ITA. This construct may be described as a listener's perception of a speaker's *language for specific purposes ability* (Douglas, 2000), which "results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics" (p. 40). This view of language ability emphasizes (a) the importance of carefully considering the context in which language is used, (b) strategic competence as an important component of language ability, and (c) constraining inferences about language ability to a particular discourse domain.

In the preliminary model, Comprehensibility was viewed as a relatively context-independent judgment of the ease or difficulty with which the listener understood the speaker, and Oral skills to TA was interpreted as the listener's impression of the speaker's oral language ability in the domain. The preliminary model hypothesized that the listener's more general impression of the speaker's comprehensibility could be isolated from domain-specific impressions of the speaker's functional competence, or Oral skills to TA.

However, the results of the study suggest that naïve listeners did not evaluate comprehensibility independent of the context of language use. In the revised model, the construct of Comprehensibility as an ITA reflects a listener's perception of the speaker's comprehensibility in relation to the speaker's functional competence in the domain of TA

language use. As such, it is influenced by linguistic and non-linguistic aspects of interaction and constrained to the domain in which interaction occurs. This subtle but important redefinition of the construct has theoretical and practical implications for oral assessment, discussed in section 5.2.1, below.

A second aspect in which the revised model differed from the preliminary model was the inclusion of the Teacher personality factor. This factor was characterized as a listener's perception of speaker attributes related to their role as a teacher, or TA. Speakers high in this factor were viewed as relatively more friendly, active, helpful, knowledgeable, and experienced. In the revised model, listener perceptions of Teacher personality predicted 26% of the variance in Comprehensibility as an ITA. In contrast, a speaker's pronunciation (TOP pronunciation) only predicted 13% of the variance in Comprehensibility as an ITA. This finding may be useful to researchers and practitioners interested in the "real-world" construct of Oral skills to TA, and has practical implications for oral proficiency assessment, language teaching, and educational policy, discussed in section 5.2 below.

This model provides further support for an interactional perspective on oral language use, and demonstrates the importance of the listener's perspective in the co-construction of meaning. While Comprehensibility as an ITA was predicted by both speaker- and listener-related factors, the listener factors were the strongest predictors, accounting for about 35% (vs. 15%) of the variation in Comprehensibility as an ITA. The quality of some of the measures of speaker-related factors may have attenuated the relationship between speaker-related factors and Comprehensibility as an ITA (see section 5.3.1), but the relative importance of listener-related factors suggests that the balance of responsibility for co-constructing meaning may be pushed to the listener's side in this domain. The implications of this finding are in the next section.

5.2 Implications of this study

The key findings of this study have both theoretical and practical implications for oral proficiency assessment, language teaching, and educational policy.

5.2.1 Implications for the assessment of oral proficiency

In this study, competence in the language use domain – the oral language skills necessary to TA – was operationalized by the TOP as phonological competence (TOP Pronunciation), lexical-grammatical competence (TOP Lexical/Grammar), rhetorical organization, and question handling. In the design and development of the TOP, non-linguistic features of a speaker's performance such as teaching effectiveness or aspects of the speaker's personality were considered not to be construct-relevant. This was a policy decision that considered institutional values – in this case, fairness. Since native speakers of English were not required to pass a pre-service certification test, it was determined that it would be unfair to evaluate non-linguistic features of an ITA candidate's performance.

However, many researchers and practitioners have observed that performing successfully as an ITA in the "real world" of English-medium university classes may require more than just proficiency in speaking English. Possible speaker-related factors include personality (Bailey, 1983), strategic competence (Douglas & Myers, 1989; Hoekje & Williams, 1992), intercultural competence (Deardorff, 2008), subject knowledge (Kavas & Kavas, 2008), teacher clarity (Chesebro, 2003), and teacher credibility (Chesebro, 2003; Li, Mazer, & Ju, 2011). Although many of these factors can be related to teaching effectiveness, many researchers and policy makers have regarded these factors as something not to be tested as part of the ITA certification, largely because of fairness issues (e.g., Farnsworth, 2004). Researchers studying language discourse in interaction or listener judgments as criterion measures have discussed the

importance of listener-related factors such as students' experience with ITAs, expectations, attitudes, and intercultural competence (Kavas & Kavas, 2008; LaRocco, 2011; Plakans, 1997).

This study provides insight into the factors that may characterize the "real-world" construct in this domain. From the perspective of naïve listeners, a speaker's comprehensibility as an ITA is influenced by linguistic and non-linguistic factors, and speaker- and listener-based factors. Based on the revised conceptual model, in this "real-world" construct the speaker's teaching skills matter; features of the speaker's personality matter; and who the listeners are matters. Thus, many of the speaker- and listener-based factors included in this study were relevant to the "real-world" construct of Comprehensibility as an ITA. The construct that informs the TOP (and many other such oral assessments) may underrepresent the "real-world" construct, although this may be the result of a deliberate design decision on the part of the test developer.

Assuming that the TOP measures the construct of oral skills to TA as it has been defined, a larger concern for test administrators and stakeholders might be the sufficiency of TOP scores for making decisions. In Bachman and Palmer's (2010) Assessment Use Argument (AUA), an argument-based approach to validity, an important claim regarding score interpretations is that these interpretations provide information that is sufficient to support decision-making. If the construct articulated by the TOP's design statement underrepresents the "real-world" construct, it follows that TOP score interpretations may provide insufficient information for decision-making.

Setting aside the question of whether the TOP *should* attempt to represent a "real world" construct in which features of the interaction outside of the speaker's control influence its outcome, the extent to which the TOP underrepresents this construct would have consequences. One such consequence may be differences in how decision errors are viewed: by naïve listeners

in the target language use domain, or by experts who have defined a narrower construct with institutional values in mind. ITAs who are certified as possessing the oral skills necessary to TA based on TOP score interpretations will not necessarily be comprehensible to all naïve listeners, which may diminish the credibility of ITA certification to naïve listeners and other stakeholder groups.

Thus, for the TOP, and for assessment practice in general, a crucial consideration is the extent to which the information provided by the test is sufficient for making correct decisions. Choi and Schmidgall (2011) reviewed ten years of language assessment research publications, and found that only 1% of validation studies addressed the issue of sufficiency. In the same study, the researchers found that approximately 40% of the research publications surveyed investigated research questions related to the meaningfulness of score interpretations – or, the extent to which scores can be interpreted as indicators of the test construct.

The results of the present study clearly suggest that regardless of whether scores may be interpreted as indicators of the construct as defined by the test, the question of the sufficiency of score interpretations for making decisions is an important and distinctly different consideration for research. The need to consider this quality of score interpretations extends to language assessment in general, and is comparatively under-represented in research publications. A practical follow-up study for the TOP to address this concern is explored in section 5.4.4, below.

5.2.2 Implications for language teaching

As the prevailing paradigm in language teaching has shifted to communicative language teaching and task-based language teaching, instructors and researchers have become increasingly sensitive to the importance of interaction for language learners. For instructors who help

learners develop speaking and listening skills with a focus on comprehensibility, this awareness is particularly acute.

In a review of factors that have been found to influence intelligibility (comprehensibility), Munro (2013) delineates the aspects of an interaction that are within a speaker's control, and those within a listener's control. Munro emphasizes that speakers may make adjustments to benefit one group of listeners, but these adjustments might not necessarily benefit another group; a great deal of listener variability may result from listeners' language experience, attitudes, and expectations. For instructors, this highlights the need to prepare students to negotiate interactions that may vary not only by the nature of the task, but also by the characteristics of the listener or interlocutor.

The results of this study support an orientation that recognizes the importance of interaction for language learners in this domain. A language instructor or learner who is focused only on core elements of oral proficiency (e.g., pronunciation, lexical-grammar, etc.) without using them in interaction may not be able to develop and enhance (1) other speaker-based skills and attributes valued by naïve listeners in this domain, and (2) their understanding of listener attitudes, expectations, and experiences. While these components may not be construct-relevant to an oral proficiency test for various reasons, they may help facilitate communication in the TLU domain, which is commonly the goal of language learning and instruction.

5.2.3 Implications for educational policy

A controversial policy decision that followed the No Child Left Behind Act of 2001 (NCLB) was the Arizona Department of Education's (ADE) decision to implement oral language tests for a subpopulation of Arizona public school teachers. More specifically, Structured English Immersion (SEI) teachers of English language learners (ELLs) in Arizona are required to

pass a performance assessment that evaluated the teacher's pronunciation and grammar (Zehr, 2010). SEI teachers that are judged to speak with a heavy accent or ungrammatical speech may face consequences that include removal from the classroom (Jordan, 2010).

Critics have argued that the ADE's oral assessment essentially targets accent and uses this policy to justify discriminatory practice (Hanna & Allen, 2012). In their critique of the ADE's assessment, Hanna and Allen argue that the rubric used by raters to evaluate test-takers only uses two criteria to evaluate pronunciation: incomprehensibility and impeding communication. The researchers argue that these constructs are inherently subjective, are essentially indicators of comprehensibility, and relative to the rater (listener). They assert that the consequences of decisions based on these test scores may be negative for students, teachers, and the educational system in general, by depriving students of the linguistic diversity of the English language that might be encountered in real-world settings. Students may also be deprived of instructors who have unique insight into students' educational experiences. Instructors who share a cultural background with students may also serve as role models as functionally proficient English users.

While the instructional and language use domain differs in this study, its findings may be relevant on a number of counts. First, by operationalizing NCLB's mandate to ensure that teachers have fluency in the English oral communications skills "pronunciation" and "grammar", the ADE may be implementing an assessment that has substantially underrepresented the construct. Given that the "real-world" construct is oral communication skills in the classroom, it is of considerable concern that the assessment does not evaluate interactional aspects of language use or consider the perspective of naïve listeners. Second, if the assessment of pronunciation is essentially operationalized as an assessment of comprehensibility using expert raters, then the

validity of score interpretations may also be undermined. As seen in this study, comprehensibility may be informed by a variety of factors, some non-linguistic; the relative importance of these factors may depend on characteristics of the population of naïve listeners. In this case, naïve listeners are ELL students of SEI instructors in Arizona. Assuming that the purpose of the ADE's test is to ensure that teachers are comprehensible to students – and not to discriminate against teachers based on accent as a matter of policy – it may be important to understand how the construct of comprehensibility as an SEI teacher, a "real world" construct, functions in that domain.

5.3 Limitations of the study

While limitations of this study were delineated at the outset (see section 1.5), several additional limitations arose during the progression of the study.

5.3.1 Methodological limitations: Validity and consistency of measures

One of the practical limitations imposed by the design of the study was the use of three distinct groups of raters: oral proficiency raters, teaching effectiveness raters, and listeners. A central assumption built into the model was that expert rater-related sources of bias or error would be minimal (see section 1.3). The usefulness of expert rater-based measures was constrained by the extent that raters were consistent and did not vary in terms of their severity.

This assumption was particularly problematic for the teaching effectiveness measures. These measures needed to be developed and were adapted based on several composite measures for use in this study. Due to practical constraints, raters received a limited amount of training (less than two hours). As a result, the g-coefficients of individual scale items and composite scales were low, ranging from 0.32 – 0.59. This suggests that ratings for these measures, while internally consistent, were influenced to a large degree by construct-irrelevant factors (e.g.,

differences between raters). The relative percentage of construct-irrelevant variance in expert rater scores – particularly in teaching effectiveness ratings – helps partially explain their limited value as predictors in the model.

Another possible explanation for the limited predictive value of the speaker-based oral proficiency measures (TOP scores) in the model was the characteristics of distributions of scores. TOP scores (TOP pronunciation, lexical-grammar, rhetorical organization, question handling) had a possible range of seven points, but for most scores were largely distributed between two to three points on the scale. In addition, scores had moderately high intercorrelations. As a result, the information provided by TOP oral proficiency scores is somewhat limited.

5.3.2 Conceptual limitations: Generalizing the conceptual model to other domains

One of the features of the conceptual model is that it accounts for speaker- and listener-related predictors of comprehensibility within a particular interactional domain, TA oral language use. Since aspects of the context may influence comprehensibility (e.g., purpose and norms of interaction; see section 2.2.4), the context was fixed and essentially controlled by the conditions of the TOP procedure. The rationale for merging the comprehensibility and oral skills to TA constructs in the conceptual model further emphasized the domain of TA oral language use.

While this domain is an important and growing part of the larger domain of academic language use at North American universities, researchers and administrators may be interested in how this model may be relevant to other domains. For example, the model in this study may be considered a more constrained version of a model of teacher-student interaction that is centered on student perceptions of the teacher's comprehensibility. In considering the extent to which the findings may generalize to other domains of teacher-student interaction, it is important to

consider differences between features of the domains, in terms of (a) characteristics of the teachers, (b) characteristics of the students, and (c) characteristics of the classroom context. In the context of this study, teachers were international graduate students from a variety of disciplines who varied in the degree to which they were familiar with local classroom norms; teacher-related factors consisted of teaching effectiveness and oral proficiency, operationalized relative to the domain. Students were undergraduates from diverse backgrounds who generally interacted with non-native speakers of English relatively frequently, but were not expected to have background knowledge of the topic being discussed. In the classroom, teachers (ITAs) presented brief academic lectures that varied in content and complexity and were expected to answer student questions, but not to fill the role of the primary instructor of the course. Depending on the degree to which another instructional domain differs in terms of these characteristics and operationalizes relevant constructs, this conceptual model may fail to generalize.

5.4 Future directions

A number of issues arose that were not fully addressed in the analysis due to the purpose and limitations of this study, but could be examined using additional data analyses or follow-up studies. This discussion will conclude by suggesting additional research.

5.4.1 Exploring the conditions under which oral proficiency measures better predict comprehensibility

One of the surprising results of this study was the relatively weak relationship between a speaker's oral proficiency (pronunciation, lexical-grammar) and comprehensibility. Although characteristics of the measurement design and instruments may have played a role in reducing these relationships, the results of the exploratory analysis suggested that there might be

conditions under which a speaker's pronunciation may better predict comprehensibility. An analysis of subgroups in the datasets found that the bivariate correlation between pronunciation and comprehensibility was much larger when the listener was unfamiliar with the speaker's native language and accent. This finding was generally supported by the follow-up interviews, in which participants speculated that being unfamiliar with the speaker's accent (and the topic) made a speaker's pronunciation much more important. In terms of the model, this suggests that the degree of the listener's familiarity with the speaker's native language and accent might function more as a grouping variable. Based on this grouping variable or subgroup, the importance of various predictors of comprehensibility might be expected to vary. Unfortunately, sample sizes prevented an exploration of this hypothesis using multi-group SEM, but this appears to be a plausible hypothesis that could be explored in the future.

5.4.2 Using more objective measures of speech in the model

One of the constraints of the current model is the limited information provided by the speaker-based oral proficiency measures: there is limited variation in scores, and inconsistency in ratings diminishes their usefulness. As discussed earlier in section 1.3, information from phonetic transcriptions of speaker performances could provide more descriptive and precise characteristics of speech. For this study, the resources required to produce and evaluate such transcriptions were not available, but a future study may be able to utilize this information. More precise information about aspects of speech, even summative indices such as speech rate, have been found to predict comprehensibility judgments (Munro & Derwing, 2001) and may prove to be a useful addition to the conceptual model. Given the continuing development and increasing proliferation of computer-based tools to analyze speech, this approach may become more feasible in the near future.

5.4.3 Availability of visual information, length of interaction, listener accommodation, and comprehensibility

One difference between the design of this study and many previous studies of comprehensibility was the nature of the sample of speech. In this study, listeners watched extended videos of speaker performances in which speakers provided a stream of non-verbal visual information – including gestures, facial expressions, and writing on a whiteboard – that may have impacted communication. The conceptual model and follow-up interviews suggested that listener judgments of comprehensibility were impacted by verbal and non-verbal information. In most previous studies, listeners monitored a brief audio recording of a speaker, and thus were only given access to verbal information.

In this study, listeners not only had access to visual and audio channels, they had access to much more of it than in previous studies. Speaker performances in this study ranged from six to eleven minutes in length, averaging around eight minutes. This allowed listeners a period of time to accommodate to features of the speaker (e.g., pronunciation), or to compensate for their lack of familiarity with a complex topic (e.g., by taking notes). In fact, many participants in the follow-up interviews suggested that they were able to adjust to features of the speaker fairly quickly, which may have resulted in higher comprehensibility scores:

Participant 6:	At first it was kind of hard, you have to get used to the accent the first couple minutes.
Researcher:	So, if you just watched the first couple of minutes, do you think you might have a different judgment?
Participant 6:	Yeah, definitely.

Participant 2:	Initially it was really hard. I had no idea what he was talking about.
Researcher:	If you had watched the first minute or two minutes, would your scores have changed?
Participant 2:	Yeah, probably.

Researcher: Did it get easier to understand the TA as time went by?

Participant 16: Uh…yeah like after the first minute or so you understand how he's going to speak and how you should interpret it, his syntax, his way of speaking, you can understand.

Thus, the length of the speech sample and quality of visual information may be important to consider when evaluating judgments of comprehensibility. Both of these features of the speech sample increase the amount of information available to the listener, and may hypothetically increase the variability of judgments of comprehensibility by allowing the listener more time and information to use to adjust to or accommodate features of the speaker or their own background.

5.4.4 The impact of the construct underrepresentation on decision errors for the TOP

As discussed in section 5.2.1, one of the consequences of the apparent misalignment between the TOP's and undergraduates' evaluations of oral skills to TA might be different views of what constitutes decision error. Decision error – or the rate of false positives and false negatives for each TOP decision category – could be investigated by comparing decisions based only on TOP scores (Pass, Provisional Pass, or Fail) with decision based only on scores for Comprehensibility as an ITA for each test-taker. Ultimately, construct underrepresentation might only be a concern to TOP administrators to the extent that it impacts decision errors in this manner.

5.5 Concluding comments

This study suggests that the field of language assessment needs to conduct research into the role of listener perceptions in oral communication, and explore ways to address or incorporate this into oral assessment. This is a challenging issue to consider, given the limitations of most assessment procedures and the broad and multifaceted nature of "real-world"

constructs. Listener perceptions in the real world may be influenced by listeners' experiences and biases over which the speaker has little or no control. Introducing these features of listeners into the assessment procedure has typically been avoided out of concern for introducing bias or error into the procedure. But as seen in this study, if speakers are expected to interact with naïve listeners in the domain of language use, then ignoring factors that influence listener perceptions may lead to an underrepresentation of the construct, a potential threat to the validity of score interpretations that has been comparatively under-researched. When the construct is clearly underrepresented, the sufficiency of score interpretations for making decisions may be called into question. The degree to which score interpretations are sufficient to make decisions may impact classification decision errors, undermining the usefulness of the test.

Appendix A – TOP Pronunciation, Lexical-grammar, Rhetorical organization, and Question handling rating scales

Score Categories	4	3	2	1
Phonetic & Phonological Competence	Accent not distracting. Pronunciation does not impede communication. Near-native pronunciation.	Accent slightly distracting. Pronunciation rarely or slightly impedes communication. Possible pronunciation errors, but not distracting.	Accent somewhat distracting. Pronunciation somewhat impedes communication. Persistent and frequent errors in pronunciation that distract listeners.	Pronunciation severely impedes communication.
Lexical / Grammatical Competence	Near-native word choice and/or oral grammar. If errors occur they are not very noticeable. Errors do not impede communication.	Some errors but rarely major. Somewhat distracting word choice and/or oral grammar. Appropriate use / range of vocabulary and grammar structure for situation, and errors slightly impede communication.	Grammar errors common in more complex constructions. Some errors in simple constructions. Major and severely distracting word choice and/or oral grammar. Lexical errors somewhat impede communication.	Major and severely distracting word choice and/or oral grammar. Lack of grammar/ lexis severely impedes communication. May be satisfactory for very simple communication.
Rhetorical Organization	Excellent overall organization and use of transitions between sentences and topics; effective use of rhetorical questions. Successful macro and micro rhetorical organization. Clearly organized discourse positively contributes to communication.	Good overall organization and use of transitions between ideas/ sentences. Discourse is appropriately organized and structured. Ideas are logically connected to one another with appropriate cohesive devices. Organization does not significantly impede communication.	Minimal overall organization and/or incorrect use of transitions between ideas/ sentences. Discourse not well organized and difficulty articulating main topics and/ or subtopics. Errors in use of cohesive devices and organization of ideas somewhat impede communication.	No overall organization and/ or ineffective use of transitions between ideas/ sentences. Discourse is generally not organized or structured. Errors in use of cohesive devices and lack of organization of ideas severely impede communication.
Question Handling	Provides substantial and comprehensive responses when appropriate. Clearly restates questions to demonstrate understanding. Provides appropriate answers and develops responses that connect answers to the presentation, and does not diverge from the presentation. Uses appropriate methods to ensure that responses are understood and thus that questions are adequately answered.	Responds appropriately to questions. May ask for clarification. Consistently shows evidence of question comprehension.	Sometimes does not respond appropriately to questions, showing evidence of insufficient question comprehension. Often asks for clarification, even for fairly simple questions. Attempts to answer questions but may provide inappropriate or incomplete responses. Hedges questions.	Unable to understand questions and/ or to answer questions. Often responds inappropriately, offering answers that are not relevant to question. Needs clarification very often, even for basic things.

Appendix B – Teaching skills rating scales

Adapted from: Hosek, 2011; Patrick & Smart, 1998

1=Strongly disagree, 2=slightly disagree, 3=Neither agree nor disagree, 4=Slightly agree, 5=Strongly agree

Respect/rapport:

RR1. The TA made students feel welcome to ask questions.	SD	1	2	3	4	5	SA
RR2. The TA was friendly towards individual students.	SD	1	2	3	4	5	SA
RR3. The TA listened attentively when students asked questions.	SD	1	2	3	4	5	SA
RR4. The TA was concerned with students' understanding.	SD	1	2	3	4	5	SA
RR5. The TA encouraged students to participate.	SD	1	2	3	4	5	SA
RR6. The TA treated students with respect.	SD	1	2	3	4	5	SA

Organization/Clarity:

O1. The TA used clear and relevant examples.	SD	1	2	3	4	5	SA
O2. The TA used the whiteboard well.	SD	1	2	3	4	5	SA
O3. The TA made the goals of the lesson clear.	SD	1	2	3	4	5	SA
O4. The TA structured the material well.	SD	1	2	3	4	5	SA
O5. The TA was well prepared.	SD	1	2	3	4	5	SA
O6. The TA summarized major points.	SD	1	2	3	4	5	SA
O7. The TA clearly defined major concepts.	SD	1	2	3	4	5	SA

Presence and Enthusiasm (Non-verbal immediacy)

EN1. The TA used a monotone or dull voice when talking. *(R)*	SD	1	2	3	4	5	SA
EN2. The TA often turned his/her back on students. *(R)*	SD	1	2	3	4	5	SA
EN3. The TA avoided eye contact while talking to students. *(R)*	SD	1	2	3	4	5	SA
EN4. The TA appeared to be very nervous. *(R)*	SD	1	2	3	4	5	SA
EN5. The TA used a variety of vocal expressions when he/she talked.	SD	1	2	3	4	5	SA
EN6. The TA gestured when he/she talked.	SD	1	2	3	4	5	SA
EN7. The TA had bland facial expressions when he/she talked. *(R)*	SD	1	2	3	4	5	SA
EN8. The TA maintained eye contact while talking to students.	SD	1	2	3	4	5	SA

(R) denotes reverse-keyed items

Appendix C – Teaching effectiveness holistic rating items

1=Strongly disagree, 2=slightly disagree, 3=Neither agree nor disagree, 4=Slightly agree, 5=Strongly agree

HOLISTIC

RR. The TA created a comfortable learning atmosphere.	SD	1	2	3	4	5	SA
ORG. The TA's lecture and discussion was clear and well organized.	SD	1	2	3	4	5	SA
ENTH. The TA was enthusiastic about teaching the class.	SD	1	2	3	4	5	SA

OVERALL. How would you rate the TA's in general (all-around) teaching effectiveness? (select one)

_____ An outstanding and stimulating instructor

_____ A very good instructor

_____ A good instructor

_____ An adequate, but not stimulating instructor

_____ A poor and inadequate instructor

Appendix D – Comprehensibility rating scale

Adapted from Powers, Schedl, Wilson-Leung, & Butler (1999), and Bridgeman, Powers, Stone, & Mollaun (2012).

Please answer the following general questions about the TA's comprehensibility.

CO1. *As a listener, how much effort was required to understand the TA?**

Very little effort 1 2 3 4 5 6 A lot of effort

CO2. *How much did the TA's oral English language abilities interfere with your understanding?**

Did not interfere at all 1 2 3 4 5 6 Interfered completely

CO3. *How certain/confident are you that you understood the TA?*

Extremely uncertain 1 2 3 4 5 6 Extremely certain/confident

CO4. *How easy or difficult was it for you to understand the TA?*

Very difficult 1 2 3 4 5 6 Very easy

CO5. *How comprehensible was the TA's speech?**

Highly comprehensible 1 2 3 4 5 6 Incomprehensible

* Reverse-keyed items

Appendix E – Oral skills to TA rating scales

Adapted from Clark and Swinton (1980).

Please answer the following questions about the TA's oral English proficiency.

TA1. *When the TA was lecturing to the class, his or her oral English-language ability interfered with my understanding of what was being said.**

Rarely or never 1 2 3 4 5 6 Always or almost always

TA2. *The TA appeared to easily understand questions asked by students.*

Rarely or never 1 2 3 4 5 6 Always or almost always

TA3. *When the TA responded to student questions, his or her oral English-language ability made the answers unclear or difficult to understand.**

Rarely or never 1 2 3 4 5 6 Always or almost always

TA4. *Do you think the TA has adequate oral English language skills to effectively TA a typical undergraduate discussion or lab section?*

Strong No 1 2 3 4 5 6 Strong Yes

* Reverse-keyed items

Appendix F – Familiarity with speaker's accent

Please indicate how familiar you are with accented English from speakers of the following languages. Please think about your interaction with people across a variety of contexts, including school, work, social events, etc. Please use the following scale:

- **Not at all familiar**
- **A little familiar**
- **Somewhat familiar**
- **Familiar**
- **Very familiar**

	Not at all familiar	A little familiar	Somewhat familiar	Familiar	Very familiar
British-accented	1	2	3	4	5
Chinese-accented	1	2	3	4	5
French-accented	1	2	3	4	5
Spanish-accented	1	2	3	4	5
Korean-accented	1	2	3	4	5
Indian-accented	1	2	3	4	5

(etc.)

Appendix G – Familiarity with speaker's native language (L1)

Please indicate how familiar you are with each of the following languages. Please use the following scale:

- **Not at all familiar** (no knowledge of vocabulary, grammar, etc.)
- **A little familiar** (a little knowledge of vocabulary and/or grammar; might be able to listen, read, speak, or write to a very limited extent)
- **Somewhat familiar** (some knowledge of vocabulary and/or grammar, through study or other exposure; might be able to listen, read, speak, or write but with limited proficiency)
- **Familiar** (knowledge of vocabulary and/or grammar, through study or other exposure; able to listen, read, speak, or write for communication in some contexts)
- **Very familiar** (wider knowledge of vocabulary and grammar, through study or other exposure; able to listen, read, speak, or write for communication in various contexts)

	Not at all familiar	A little familiar	Somewhat familiar	Familiar	Very familiar
English	1	2	3	4	5
Chinese (Mandarin)	1	2	3	4	5
French	1	2	3	4	5
Spanish	1	2	3	4	5
Korean	1	2	3	4	5
Hindi	1	2	3	4	5

(etc.)

Appendix H – Attitude homophily scale

Adapted from McCroskey, Richmond, & Daly (1975)

Please indicate your general perception of how similar or different the TA is from you in the following ways. Please make an impressionistic judgment. It could be based on the TA's personality, appearance, discussion style, etc.

AH1.	Doesn't think like me	1	2	3	4	5	6	Thinks like me
AH2.	Similar to me	1	2	3	4	5	6	Different from me*
AH3.	Behaves like me	1	2	3	4	5	6	Doesn't behave like me*
AH4.	Is unlike me	1	2	3	4	5	6	Is like me

* Reverse-keyed items

Appendix I – Teacher Personality scale

Adapted from Coetzee-Van Rooy (2009).

Please indicate your general perception of the TA with respect to each of the following attributes.

P1.	Friendly	1	2	3	4	5	6	Unfriendly*
P2.	Uninformed	1	2	3	4	5	6	Knowledgeable
P3.	Unhelpful	1	2	3	4	5	6	Helpful
P4.	Active	1	2	3	4	5	6	Passive*
P5.	Experienced	1	2	3	4	5	6	Inexperienced*

* Reverse-keyed items

Appendix J – Overview and instructions for participants in the main study

OVERVIEW

In this survey, you will watch two video clips of a teaching assistant (TA) presenting a short lesson to a small group of undergraduates at UCLA.

In each video clip, the TA will present a short lecture and answer any questions. The topic of the lecture is chosen by the TA, but should be at the level of an introductory undergraduate course in the TA's department. Thus, the topic should be accessible to UCLA undergraduates.

Video clips vary in length from 5-10 minutes, but are generally closer to 10 minutes long. Please listen to the video carefully as if you were a student in the class. Many of the video clips are edited, and may appear to end suddenly or be cut off without reaching a conclusion. Do not rewind the video to try to re-listen to something you might have had trouble understanding.

Before you begin, please listen to each of the following samples to become familiar with the format of the TA lectures, and the range of oral proficiency of TAs you might encounter.

- CONTINUE -

Please play Sample Video #1, below.

Sample video #1, embedded

In general, undergraduates reported that this TA's oral English was very easy to understand. His oral proficiency in English did not interfere with his duties as a TA at all.

- CONTINUE -

Next, please play Sample Video #2, below.

Sample video #2, embedded

In general, undergraduates reported that this TA's oral English was difficult to understand, and required a lot of effort to understand. His oral proficiency in English frequently interfered with his duties as a TA.

-CONTINUE-

Next, please listen carefully to the first video clip.

After the video clip finishes, please immediately move on to the survey questions. The survey will include five groups of questions:

- The topic and main points of the lecture
- Overall, how comprehensible the TA's oral English was
- The extent to which the TA's oral proficiency in English helped or hindered classroom tasks
- Your familiarity with the topic of the lecture
- General perceptions of the TA

Please provide your careful, honest opinion!

Please be careful: read each survey question carefully, as similar questions may be phrased differently and require you to use the rating scale differently.

Please be honest: your responses will be anonymous and will not in any way impact the TAs in the video clips you view. The questions are largely impressionistic, so after you've read the question carefully please provide your instinctual reply – don't overthink it!

Your careful and honest response is a critical component in helping us ensure that the Test of Oral Proficiency helps promote quality undergraduate education at UCLA.

- CONTINUE -

References

Abrami, P. C. , & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of N=1 research: Comment on Marsh (1991). *Journal of Educational Psychology, 30*, 221-227.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*, 561-593.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*, 529-555.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bailey, K. M. (1983). *Teaching in a second language: The communicative competence of non-native speaking teaching assistants* (Unpublished doctoral dissertation). University of California, Los Angeles.

Bangbose, A. (1998). Torn between the norms: Innovations in world Englishes. *World Englishes, 17*, 1-14.

Basow, S. A. (1990). Effects of teacher expressiveness: Mediated by teacher sex-typing? *Journal of Educational Psychology, 82*(3), 599-602.

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). TOEFL 2000 Listening Framework: A Working Paper. *TOEFL Monograph Series No. 19*. Princeton, New Jersey: Educational Testing Service.

Bentler, P. M. (2006). *EQS 6 structural equation program manual*. Encino, CA: Multivariate Software, Inc.

Berns, M. (2008). World Englishes, English as a lingua franca, and intelligibility. *World Englishes, 27*(3/4), 327-334.

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag: New York.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29*(1), 91-108.

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning* (22), 203-220.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying Generalizability theory using Edu-G*. Routledge: New York.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Catano, V. M., & Harvey, S. (2011). Student perception of teaching effectiveness: Development and validation of the Evaluation of Teaching Competencies Scale (ETCS). *Assessment and Evaluation in Higher Education, 36*(6), 701-717.

Chesebro, J. L. (2003). Effects of teacher clarity and nonverbal immediacy on student learning, receiver apprehension, and affect. *Communication Education, 52*, 135-147.

Choi, I. K., & Schmidgall, J. E. (2011, June). *A survey of methodological approaches employed to validate language assessments*. Paper presented at the 33rd annual conference of the Language Testing Research Colloquium, Ann Arbor, MI.

Coetzee-Van Rooy, S. (2009). Intelligibility and perceptions of English proficiency. *World Englishes, 28*(1), 15-34.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity. *Review of Educational Research, 51*, 281-309.

Deardorff, D. K. (2008). Intercultural competence: A consensus definition, model, and implications for assessment. The University of Arizona Center for Educational Resources in Culture, Language, and Literacy. PowerPoint. http://cercll.arizona.edu/icc_materials/Deardorff_presentation.ppt#1.

Derwing, Munro, & Thomson (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359-380.

Derwing, T. M., & Munro, M. J. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1-16.

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. *The Canadian Modern Language Review, 66(*2), 181-202.

Douglas, D. (2000). *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Myers, C. (1989). TAs on TV: Demonstrating communication strategies for international teaching assistants. *English for Specific Purposes, 8*, 169-179.

Entwistle, N. J., & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education, 19*, 169-194.

Erdle, S., Murray, H. G., Rushton, J. P. (1985). Personality, classroom behavior, and student ratings of college teaching effectiveness: A path analysis. *Journal of Educational Psychology, 77*(4), 394-407.

Farnsworth, T. (2004). *The effect of teaching skills on holistic ratings of language ability in performance tests for international teaching assistant selection* (Unpublished masters thesis). University of California, Los Angeles.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.

Field, J. (2003). The fuzzy notion of 'intelligibility': A headache for pronunciation teachers and oral testers. *IATEFL Special Interest Groups Newsletter*, 34-38.

Field, J. (2011). Into the mind of the academic listener. *Journal of English for Academic Purposes, 10*, 102-112.

Flege, J., & Fletcher, K. (1992). Talker and listener effects on the degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370-389.

Flege, J., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97*, 3125-3134.

Fulton, W. (1996). How can we use course evaluation to improve teaching and the curriculum. http://www.oeghd.or.at/zeitschrift/1996h1-2/07g_art.html (accessed January 20, 2002).

Gao, S., Mokhtarian, P. L., & Johnston, R. A. (2008). Nonnormality of data in structural equation models. *Transportation Research Record: Jounral of the Transportation Research Board, 2082*, 116-124.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*(1), 65-89.

Gibbs, G., & Coffey, M. (2004). The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active Learning in Higher Education, 5*(1), 87-100.

Guyton, E., & Farokhi, E. (1987). Relationships among academic performance, basic skills, subject matter knowledge, and teaching skills of teacher education graduates. *Journal of Teacher Education, 38*, 37-42.

Harding, L. (2008). Accent and academic listening assessment: A study of test-taker perceptions. *Melbourne Papers in Language Testing, 13*(1), 1-33.

Hardman, J. B. (2010). *The intelligibility of Chinese-accented English to international and American students at a U.S. university* (Unpublished doctoral dissertation). Ohio State University, Columbus.

Harlow, L. L. (1985). *Behavior of some elliptical theory estimators with non-normality data in a covariance structures framework: A Monte Carlo study* (Unpublished doctoral dissertation). University of California, Los Angeles.

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings of teaching effectiveness. *College Student Journal, 40*(1), 195-203.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistants. *TESOL Quarterly, 26*(2), 243-269.

Hosek, A. (2011). *Extending intergroup theorizing to the instructional context: Testing a model of teacher communication behaviors, credibility, group-based categorization, and*

instructional outcomes (Unpublished doctoral dissertation). University of Nebraska, Lincoln.

Houser, M. L., & Frymier, A. B. (2009). The role of student characteristics and teacher behaviors in students' learner empowerment. *Communication Education, 58*(1), 35-53.

Hsieh, C-N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47-74.

Hsieh, C.-N. (2011b). *Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review, 64*(4), 555-580.

Isaacs, T. (2010). *Issues and arguments in the measurement of second language pronunciation* (Unpublished doctoral dissertation). McGill University, Montreal.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*, 83-103.

Jordan, M. (2010, April 30). Arizona grades teachers on fluency: State pushes school district to reassign instructors with heavy accents or other shortcomings in their English. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424052748703572504575213883276427528.html

Junker, B., Weisberg, Y., Matsumura, L. C., Crosson, A., Kim Wolfe, M., Levison, A., & Resnick, A. (2006). Overview of the Instructional Quality Assessment. *CSE Technical Report No. 671*, 1-80.

Kachru, Y. (2008). Cultures, contexts, and interpretability. *World Englishes, 27*(3/4), 309-318.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Measures of Effective Teaching Project Research Paper*.

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 181-205.

Kang, Z. (2008). *Advanced models for assessing perceived instructional quality of university faculty* (Unpublished doctoral dissertation). University of Toledo.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *The Canadian Modern Language Review, 64*(3), 459-489.

Kavas, A., & Kavas, A. (2008). An exploratory study of undergraduate college students' perceptions and attitudes toward foreign accented faculty. *College Student Journal, 42*(3), 879-890.

Hanna, P. L., & Allen, A. (2012). Educator assessment: Accent as a measure of fluency in Arizona. *Educational Policy,* 1-28. doi:10.1177/0895904811429293

LaRocco, M. J. F. (2011). *International teaching assistants and the essence of the development of intellectual competence* (Doctoral dissertation). Retrieved from http://digitalcommons.ric.edu/etd/40

Lev-Ari, S. (2010). *Variability in language processing: Processing non-native speech* (Unpublished doctoral dissertation). University of Chicago.

Li, L., Mazer, J. P., & Ju, R. (2011). Resolving international teaching assistant language inadequacy through dialogue: Challenges and opportunities for clarity and credibility. *Communication Education, 60*(4), 461-478.

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics, 7*(3), 348-364.

Lindemann, S. (2011). Who's "unintelligible"? The perceiver's role. *Issues in Applied Linguistics, 18*(2), 223-232.

Londe, Z. C. (2008). *Working memory and English as a second language listening comprehension tests: A latent variable approach* (Unpublished doctoral dissertation). University of California, Los Angeles.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187-1197.

Matsuura, H., Chiba, R., & Fujieda, M. (1999). Intelligibility and comprehensibility of American and Irish Englishes in Japan. *World Englishes, 18*, 49-62.

Mazer, J. P. (2012). Development and validation of the student interest and engagement scales. *Communication Methods and Measures, 6*, 99-125.

McCroskey, J. C. (1994). Assessment of affect toward communication and affect toward instruction in communication. In S. Morreale & M. Brooks (Eds.), *1994 SCA summer conference proceedings and prepared remarks: Assessing college students' competency in speech communication*. Annandale, VA: Speech Communication Association.

McCroskey, J. C., Richmond, V. P., & Daly, J. A. (1975). The development of a measure of perceived homophily in interpersonal communication. *Human Communication Research, 1*, 321-332.

Meierkord, C. (2004). Syntactic variation in interactions across international Englishes. *English World-Wide, 25*, 109-132.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 192-218). Amsterdam: John Benjamins.

Munro, M. J. (2013). Intelligibility. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*(2), 159-182.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*, 451-468.

Munro, M.J., Derwing, T.M., & Morton, S.L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111–131.

Muthén, B., & Kaplan, D. (1985). A comparison of methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38(1)*, 171–189.

Nelson, C. L. (2008). Intelligibility since 1969. *World Englishes, 27*(3/4), 297-308.

Patrick, J., & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: The emergence of three critical factors. *Assessment and Evaluation in Higher Education, 23*(2), 165-178.

Pickering, L. (2006). Current research on intelligibility in English as a Lingua Franca. *Annual Review of Applied Linguistics, 26*, 219-233.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly, 31*(1), 95-119.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Polk, J. A. (2006). Traits of effective teachers. *Arts Education Policy Review, 107*(4), 23-29.

Powers, D. E., Schedl, M. A., Wilson-Leung, S., & Butler, F. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*, 399-425.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes, 26*(1), 87-98.

Riney, T., & Flege, J. E. (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition, 20*, 213-243.

Riney, T., Takada, M., & Ota, M. (2000). Segmentals and foreign language accent: The Japanese flap in EFL. *TESOL Quarterly, 34*, 711-737.

Roelofs, E., & Sanders, P. (2007). Towards a framework for assessing teacher competence. *European Journal of Vocational Training, 40*, 123-139.

Rost, M. (2005). L2 Listening. In E. Hinkel (ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: LEA.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET project*. National Education Policy Center: Boulder, CO.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Ryan, J. M., & Harrison, P. D. (1995). The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education, 36*(5), 577-594.

Schmidgall, J.E. (2011, March). *Confidence in the cut score: Reliability and conditional standard errors for a test of oral English*. Paper presented at the annual conference of the American Association of Applied Linguistics, Chicago, IL.

Schmidgall, J. E. (2012, April). *The relationships between speaker proficiency variables and contextualized listener perceptions of oral language use*. Paper presented at the 15th annual conference of the Southern California Association of Language Testing Researchers Conference, Los Angeles, CA.

Shaw, D., Young, S., Shaffer, J., & Mundfrom, D. (2003). A strategy for addressing the validity of teacher effectiveness instrument. *Multiple Linear Regression Viewpoints, 29*(1), 44-48.

Smith, J., Meyers, C. M., & Burkhalter, A. J. (1992). *Communicate: Strategies for international teaching assistants*. Prentice Hall.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. Kachru (ed.), *The Other Tongue: Englishes across Cultures* (pp. 75-90). Urbana: University of Illinois Press.

Smith, L. E., & Christopher, E. (2001). "Why can't they understand me when I speak English so clearly?" In E. Thumboo (ed.), *The Three Circles of English: Language Specialists Talk about the English Language* (pp. 91-100). Singapore: UniPress.

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: Directions and resources. *World Englishes, 4*, 333-342.

Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly, 13*(3), 371-380.

Swartz, C. W., White, K. P., Stuck, G. B., & Patterson, T. (1990). The factorial structure of the North Carolina teaching performance appraisal instrument. *Educational and psychological measurement*, *50*(1), 175-182.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics, 25*, 1-24.

Ting, K. (2000a). Cross-level effects of class characteristics on students' perceptions of teaching quality. *Journal of Educational Psychology, 92*(4), 818-825.

Ting, K.-F. (2000b). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*(5), 637-661.

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 4*, 114-136.

Varonis, E., & Gass, S. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition, 4*, 114-136.

Venkatagiri, H. S., & Levis, J. M. (2007). Phonological awareness and speech comprehensibility: An exploratory study. *Language Awareness, 16*(4), 263-277.

Wells, R. (2007). International faculty in U.S. community colleges. *New Directions for Community Colleges, 138*, 77-82.

Yule, G., & Hoffman, P. (1993). Enlisting the help of U.S. undergraduates in evaluating International Teaching Assistants. *TESOL Quarterly, 27*(2), 323-327.