CONFIRMING THE IMPACT OF PERFORMANCE TASKS ON LATENT CLASS MEMBERSHIP AND PLACEMENT DECISIONS

A Dissertation in
Applied Linguistics
by
Jeremy R. Gevara

© 2016 Jeremy R. Gevara

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2016
The dissertation of Jeremy R. Gevara was reviewed and approved* by the following:

Robert W. Schrauf  
Professor of Applied Linguistics  
Dissertation Advisor

Hoi K. Suen  
Distinguished Professor of Education (Educational Psychology)

Xiaofei Lu  
Associate Professor of Applied Linguistics and Asian Studies  
Gil Watz Early Career Professor in Language and Linguistics  
Department of Applied Linguistics Chair

Pui-Wa Lei  
Associate Professor of Education (Educational Psychology)

*Signatures are on file in the Graduate School.
Abstract

Pennsylvania law requires that all international students on graduate teaching assistantships (ITAs) pass an in-house test of English proficiency in addition to a standardized test. To this end, Penn State developed and administers the American English Oral Communication Proficiency Test (AEOCPT), and the general purpose of this dissertation is to validate the AEOCPT. I focus on three components of the AEOCPT. The first component (Scoring Rubric) consists of eight indicators that operationalize a four factor framework of language knowledge. Although this language knowledge framework is an accepted framework for ITA tests in general, this dissertation is the first attempt to empirically confirm it (RQ1). The second component (Placement decisions) is currently made by adding up the scores of the eight indicators across the four performance tasks. This summative score communicates placement into one of four classes. However, this score cannot be used to empirically confirm that students are best placed into four classes or identify the strengths and weaknesses that group them together (RQ2). The final component (Performance Tasks) is assumed by ITA administrators to elicit similar rubric scores. However, Instructors believe that the four performance tasks are not equally salient to informing English ability in the American classroom (RQ3). The dataset that I used to answer the three research questions consists of the results from 498 ITA candidates that took the test between the 2012 Summer and 2014 Fall semesters. The results from the Confirmatory Factor Analysis show that the four factor language knowledge model (RQ1) has acceptable fit to the dataset (Using Global and Componential Fit statistics). In addition, the four factor model better fits the dataset than alternative models of language knowledge. The results from Latent Class Analysis show that a three class model for placement decisions better fits the dataset than a four class model (RQ2). The results also show that the four tasks contribute unique information to understanding candidates’ most probable placements (RQ3). These results have concrete implications for the local interpretation of AEOCPT results by the department, and they contribute to larger questions about interaction-based theories in language testing research.
# TABLE OF CONTENTS

List of Figures .................................................................................................................... vi
List of Tables ....................................................................................................................... vii
List of Abbreviations ......................................................................................................... ix
Acknowledgments ............................................................................................................... xi

Chapter 1. INTRODUCTION ........................................................................................ 1
  A brief history of ITA programs in the United States ...................................................... 1
  Adhering to ITA testing laws ......................................................................................... 5
  Penn State’s development of the AEOCPT .................................................................... 9
  Motivation for this dissertation ..................................................................................... 13
  Summary ......................................................................................................................... 18

Chapter 2. LITERATURE REVIEW ............................................................................. 20
  The Scoring Rubric: From a construct to observable language use .............................. 20
  Score: Establishing its relationship to intended decisions ............................................ 31
  Task: Striking a balance between Cognitive and Contextual loads ............................ 37
  Summary and research questions ................................................................................. 43

Chapter 3. METHODS .................................................................................................. 45
  Participants: The AEOCPT Dataset .............................................................................. 45
  Procedures: AEOCPT Administration ......................................................................... 47
  Materials: Scoring the four AEOCPT tasks ................................................................. 51
  Data Analysis: Analyzing the AEOCPT dataset ......................................................... 53
  Notes on Validity: Keeping the study framed within the AEOCPT’s stated purposes .... 66
  Summary ......................................................................................................................... 68

Chapter 4. CFA RESULTS .......................................................................................... 70
  Data Screening Results ................................................................................................. 70
  Configural Invariance: Identifying an initial model ...................................................... 72
  Metric Invariance: Collecting results for answering RQ3 ........................................... 78
LIST OF FIGURES

Figure 1-1. Foundational tests of the AEOCPT………………………………………………………6
Figure 2-1. Communicative language ability framework (Bachman, 1990)…………………………26
Figure 2-2. Information Processing Model (Martinez, 2010)…………………………………………27
Figure 2-3. Language Knowledge framework (Bachman & Palmer, 1996, 2010)…………………29
Figure 2-4. Framework of Communicative Activities (Cummins, 1983)……………………………40
Figure 3-1. Four factor language knowledge CFA model…………………………………………54
Figure 3-2. Two factor language knowledge CFA model…………………………………………56
Figure 3-3. One factor language knowledge CFA model…………………………………………57
Figure 3-4. Conceptual LCA model with eight indicators…………………………………………58
Figure 4-1. Example frequency charts tested for an underlying distribution……………………72
Figure 4-2. Standardized estimates of the language knowledge model for the Mini-lecture task…………………………………………………………………………………………………..79
Figure 4-3. Standardized estimates of the language knowledge model for the Role-play task……………………………………………………………………………………………………………79
Figure 4-4. Standardized estimates of the language knowledge model for the Opinion task………………………………………………………………………………………………………………80
Figure 4-5. Standardized estimates of the language knowledge model for the Announcement task………………………………………………………………………………………………………………80
Figure 6-1. Four factor language knowledge CFA model submitted for analysis………………..111
LIST OF TABLES

Table 1-1. Summary of AEOCPT development history.........................................................13
Table 1-2. Summary of results from contrasting groups standard setting approach..............16
Table 1-3. Updated AEOCPT development history.................................................................18
Table 2-1. Bachman’s (2007) categories of language testing theories.................................20
Table 2-2. Bachman’s (2007) three categories of interactionalist theories.............................24
Table 3-1. Demographic information of AEOCPT test takers..............................................46
Table 3-2. Revised AEOPCT scoring rubric...........................................................................50
Table 3-3. Current ITA placement decisions with brief course descriptions.....................62
Table 4-1. Descriptive statistics, Mean (SD) for AEOCPT data screening..........................71
Table 4-2. Global fit statistics for four-factor model across four tasks...............................74
Table 4-3. $\chi^2$ difference test for nested alternative models............................................75
Table 4-4. Component fit statistics for the four factor hierarchical model.........................77
Table 4-5. Global fit and component statistics for configural and metric invariance of the four AEOCPT tasks.................................................................78
Table 4-6. Global fit and component fit statistics for the Configural Invariance and CU models........................................................................................................79
Table 5-1. Observed and expected response patterns for the eight AEOCPT rubric items......87
Table 5-2. Frequency of ITA candidates receiving values across all four tasks.....................88
Table 5-3. Information from PROC LCA used for model identification.............................90
Table 5-4. Comparative fit statistics for 1st half of the dataset............................................92
Table 5-5. Comparative fit statistics for 2nd half of the dataset...........................................92
Table 5-6. Indicator response patterns for each class on the Mini-lecture task.....................95
Table 5-7. Indicator response patterns for each class on the Role-play task........................96
Table 5-8. Indicator response patterns for each class on the Opinion task..........................96
Table 5-9. Indicator response patterns for each class on the Announcement task..................97
Table 5-10. Comparison between summative and LCA informed placements....................98
Table 5-11. Example posterior probability results...............................................................99
Table 5-12. Example feedback form for a candidate placed in Class 2 (Advanced)..............100
Table 5-13. Example feedback form for a candidate placed in Class 1 (High-Intermediate)….100
Table 6-1. Current ITA placement decision with brief course descriptions..........................111
Table 6-2. Indicator response patterns for Class 1 on each task........................................114
Table 6-3. Indicator response patterns for Class 2 on each task........................................116
Table 6-4. Indicator response patterns for Class 3 on each task........................................117
Table 7-1. Foundational tests of the AEOCPT.................................................................122
Table 7-2. Observed and expected response patterns for the eight AEOCPT rubric items......126
LIST OF ABBREVIATIONS

American English Oral Communicative Placement Test……………………………………AEOCPT
International Teaching Assistant……………………………………………………………ITA
Educational Testing Service………………………………………………………………………ETS
Test of English as a Foreign Language Paper-Based Test……………………………TOEFL PBT
Test of Spoken English……………………………………………………………………….TSE
Test of Written English……………………………………………………………………….TWE
Spoken Proficiency English Assessment Kit…………………………………………………SPEAK
Oral Proficiency Interview……………………………………………………………………OPI
Communicative Language Ability…………………………………………………………….CLA
Norn-Referenced Test…………………………………………………………………………NRT
Cognitive Diagnostic Assessment…………………………………………………………….CDA
Test of Oral Proficiency……………………………………………………………………….TOP
Latent Class Analysis…………………………………………………………………………LCA
Latent Profile Analysis………………………………………………………………………..LPA
Latent Transition Analysis……………………………………………………………………LTA
Basic Interpersonal Communicative Skills…………………………………………………BICS
Cognitive/Academic Language Proficiency………………………………………………CALP
Confirmatory Factor Analysis………………………………………………………………CFA
Item Response Theory………………………………………………………………………..IRT
Principal Component Analysis…………………………………………………………….PCA
Akaike Information Criterion………………………………………………………………AIC
Bayesian Information Criterion……………………………………………………………..BIC
Corrected Akaike Information Criterion…………………………………………………CAIC
Evidence Centered Design………………………………………………………………….ECD
Standardized Root Mean Square Residual………………………………………………SRMR
Root Mean Square Error of Approximation………………………………………………RMSEA
Comparative Fit Index………………………………………………………………………..CFI
Non-normed Fit Index………………………………………………………………………………..NNFI
Multitrait-Multimethod…………………………………………………………………………..MTMM
Correlated Trait-Correlated Method…………………………………………………………..CTCM
Correlated Uniqueness…………………………………………………………………………..CU
Weighted Root Mean Residual……………………………………………………………………..WRMR
Weighted Least Squares Means and Variances………………………………………………..WLSMV
Acknowledgments

I would like to use this part of the dissertation to thank people in my professional and personal life. I would like to thank the professors that have guided me through my time as a graduate student at both Texas Tech University and Pennsylvania State University. I am sure the experience of guiding a first generation college student through graduate school has been a challenge. I would also like to thank my wife and family for whom without I would surely not be at this university or stage of my life.
Chapter 1 Introduction

For testing speaking ability in a foreign or second language, performance tests have become an increasingly popular option. Brown and Abeywickrama (2010) define performance tests as measures of “doing” language (p. 351). “Doing” language means that test takers provide extended responses to the test questions (tasks) that are replications of a target context (Robinson, 1996). McNamara (1996) categorizes the target context as one of two influences (Contextual Influence and Theoretical influence) for developing performance tests. Contextual influences are variables within an environment that support the need for a performance test. Theoretical influences are variables within Applied Linguistics that allow test administrators to best measure language.

For this dissertation, I am using a test given at Pennsylvania State University as a case study of performance tests. The performance test at Pennsylvania State University is used to certify or place International Teaching Assistants (ITAs) into one of three courses. I chose this test as the focus of this dissertation because the contextual influences are well documented at this university as well as across the United States. In addition, the theoretical influences behind the test have changed in accordance with developments in research about language knowledge frameworks. In this chapter, I will document the contextual influences that led to the need for an ITA performance test at Pennsylvania State University.

1.1 A brief history of ITA programs in the United States

At Pennsylvania State University, every graduate international student eligible for an assistantship must take an oral performance test of English proficiency. This test is given in addition to a standardized English assessment, e.g., TOEFL or IELTS, that is provided by every
student prior to being accepted into a graduate program. Penn State’s current in-house English proficiency test is called the American English Oral Communicative Proficiency Test (AEOCPT). The AEOCPT’s development history is unique but draws on materials that are used by several testing and training programs of International Teaching Assistants across the United States. As with the common materials used to create the AEOCPT, the ITA program at Penn State shares some history with other programs from around the country.

Wilkening (1991) notes that the national development of ITA programs and tests started in the 1980s. An increase in international graduate student enrollment actually began in the 1970s, but the 1980s is when state governments began to respond. Each state had a different impetus for pushing legislation to regulate the awarding of assistantships to international students, but Kaplan (1989) summarizes these cases as addressing the “ITA problem.” The problem derived from anecdotal evidence that undergraduates were not able to pass a course largely due to inefficiencies of their ITA instructors. Thomas and Monoson (1993) added empirical evidence identifying these inefficiencies as language related. To date, 20 states have passed laws requiring the development of ITA training programs and the administration of an additional test of English proficiency. The fact that states create their own laws on education (Rather than laws passed at a federal level) presents a unique opportunity for comparing strategies to addressing ITA training and assessment.

Based on my personal experiences with ITA programs in Texas and Pennsylvania, I will discuss these two states’ laws and how public universities within them responded to these laws. Both Texas and Pennsylvania are among the 20 states cited by Thomas and Monoson as passing laws requiring the development of ITA training programs and assessment procedures. Texas’s law was passed during the 71st legislative meeting, 1989, as HB 638 (Texas, 1989), and
Pennsylvania’s law was passed during the 1990 session as SB 539 (Pennsylvania, 1990). HB 638 states that every public university in Texas must establish both short courses and assessment procedures for evaluating ITAs. Short courses and assessment procedures must be focused on English proficiency. Texas Tech University and the University of Texas at Austin are two examples of universities that have working ITA programs with in-house tests. Both universities offer a single semester course for students that do not pass their respective assessments.

Pennsylvania’s SB 539 contains similar wording to Texas’s HB 638, but differences exist that likely contribute to the development of different ITA programs. One key difference between the two laws is that Pennsylvania’s SB 539 requires the assessment of English proficiency as used in a classroom. The addition of context in Pennsylvania’s SB 539 would appear to be a minor addition to the common voter, but it has a more profound impact in Applied Linguistics research. Does the context where a language is used have an impact on perceived proficiency? In addition, Pennsylvania’s SB 539 does not contain any wording on how ITA candidates who do not pass the proficiency test should be trained. This freedom for interpretation differs slightly from Texas’s HB 638, which requires a short course that I have shown to be interpreted by two Texas public universities as one semester long.

Administrators at Pennsylvania State University have met the certification section of SB 539 by designing three courses. These courses are linked to each other through varying degrees of language and American classroom instruction. The lowest level is called ESL 115G, American Oral English for ITAs I, and focuses exclusively on English knowledge development. The next level, ESL 117G or American Oral English for ITAs II, continues the development of language knowledge but also introduces some aspects of teaching in an American classroom (e.g., Introducing a topic, Explaining a visual, and Handling questions). One example of an
American classroom lesson is instructing students on how to introduce a syllabus on the first day of class. The last course is ESL 118G, American Oral English for ITAs III. This course focuses mainly on developing additional skills within the American university classroom context.

Readers of this dissertation will notice that the two states of Texas and Pennsylvania have two different effects in testing the English proficiencies of ITA candidates. Although the goal of both states’ laws is to certify candidates with a high level of proficiency, the effects of not being certified have different consequences. In Texas, students are funneled back into a single course. Whereas in Pennsylvania, students are divided into courses according to their ranging scores. In the following, I use Brown’s (2005) categorization of language testing purposes to contrast the ITA program at Texas Tech University and at Pennsylvania State University.

Because the ITA program at Texas Tech University offers one course for students who do not pass its test, this test is used for achievement purposes. Brown describes tests for achievement purposes as closely related to program objectives (p. 7). A single cut-score is used in these kinds of tests to divide students into those who passed and those who didn’t. Although Brown describes achievement tests as being commonly administered at the end of the course, the ITA test at Texas Tech University is used before and after classroom instruction.

The purpose of the ITA test at Pennsylvania State University, on the other hand, is to divide students into more than two possible categories. The highest category certifies students to teach at the university without any additional instruction. The lowest category, however, requires students to pass ESL 115G, 117G, and 118G, before they are certified to teach. Although the two tests at Texas Tech and Penn State are used for different decisions, they share a common developmental history.
1.2 Adhering to ITA testing laws

After several states passed laws requiring the additional English proficiency testing of ITA candidates (Thomas & Monoson, 1993), a problem arose at many universities of identifying options for adhering to these new laws. Spolsky (1995) describes the challenge of this time period and Educational Testing Service’s (ETS) attempt to provide a solution. During the late 1980s, ETS offered the Test of English as a Foreign Language Paper-Based Test (TOEFL PBT) to measure English proficiency for university admissions. This test consists of three sections, listening comprehension, structure and written expression, and reading comprehension. One of the weaknesses of this test is that it only measures receptive English skills.

One reason for ETS’s decision to measure English proficiency through receptive skills is that the entire test can be designed using selected response items such as multiple choice, matching, and true/false items. This choice of items allows for adherence to a single dimension of English knowledge and easier measurement of reliability. Spolsky referred to these kinds of tests as following a psychometric-structuralist trend in language testing. However, it is difficult to design selected response items to assess productive language skills (Speaking and Writing) following the psychometric-structuralist trend because of the complex nature of language knowledge. Constructed response items, by contrast, are commonly used to elicit an extended response that contains enough evidence of language knowledge to make a decision (Luoma, 2004; Weigle, 2002). Because selected and constructed response items are challenging to connect following the psychometric-structuralist trend, ETS decided to create optional tests of productive skills.
Along with several state laws requiring oral proficiency test of ITAs, universities themselves challenged ETS to design a test of written English proficiency. The company’s response to these two challenges was the Test of Spoken English (TSE) and the Test of Written English (TWE). These two tests were designed as optional additions to the TOEFL PBT for any universities that needed measurements of productive English skills. The TSE is the first of several ITA tests (Shown in Figure 1-1) that Pennsylvania State University ITA administrators used to create the AEOCPT. In addition, researchers cite the TSE as source material for the other three tests included in Figure 1-1 (Smith, Meyers, & Burkhalter, 1992; Powers, Schedl, Leung, & Butler, 1999)

Figure 1-1. Foundational tests of the AEOCPT
The TSE consists of 15 tasks that test takers answer through extended spoken responses. Tasks are contrasted from items in that more context is needed in the former to construct an extended response (McNamara, 1996). Test takers’ responses to all 15 tasks are recorded and sent to raters
to be scored on a holistic scale. The holistic rubrics starts at a minimum score of 20 and increases in increments of five points to the maximum of 60.

The TSE became widely used in a variety of institutions, e.g., government, business, education, etc., but this test did not meet all of the state laws passed for ITA testing. For instance, Pennsylvania’s SB 539 (1990) law states that ITA candidates’ English should be test for use in the classroom (p.2). Arguing that the TSE meets this test of English for this specific purpose is difficult to do since a variety of institutions use the scores for a variety of hiring practices.

Lazaraton and Wagner’s (1996) study of the TSE provides some brief descriptions of the 15 tasks used in the test. Two of these tasks required test takers suggest a gift and recommend a place to visit with reasons to support these two decisions. Other tasks are more closely related to a classroom context, such as defining a technical term, but readers can see the weak argument made that this test can be used for certifying English proficiency in a classroom context.

One challenge for ETS in creating a test measuring ITA English proficiency was that no two states laws were exactly alike. Some states explicitly stated that the test be a measure of English knowledge in the classroom while others focused on general proficiency. In addition, the fact that public universities in Texas could write interpretations of the law into their operating policies required further nuances for creating an ITA test. ETS’s (1982) proposed solution for this variety was to offer a test that served as a framework for universities to create their own tests. This solution is called the Spoken Proficiency English Assessment Kit (SPEAK) test. This test is the subject of several studies with a focus on ITA candidates (Davies, Tyler, & Koran, 1989; Dick & Robinson, 1994; Gorsuch, 2006; Smith et al., 1992). The SPEAK test is also still used in ITA programs such as Texas Tech University, the University of Kansas, and the University of Florida.
Although the SPEAK test is a collection of retired TSE forms (Kaplan, 1989), the advantage for universities in adopting this test is that they are given all the materials for adapting and administering the test in-house. This condition allowed many universities to adhere to state laws requiring an in-house test to be administered without having to spend the resources to develop it from scratch. Two versions of the SPEAK test are available, following a revision to the TSE.

The first version of the SPEAK consists of six tasks: Reading aloud, Sentence completion, Storytelling, Single picture description, Opinion questions, and Describing a schedule (Kim, 2001). Responses to these tasks are recorded and scored by raters using a holistic rubric. This rubric consists of four points, a range of 0 to 3, and includes the descriptors of grammar, pronunciation, and fluency.

The second version of the SPEAK test (ETS, 1996), contains additions to the tasks given and an expanded scoring rubric. The tasks were doubled to a total of 12, but no new tasks were introduced. An increase in the same types of tasks from the previous version would be evidence to support the content validity of the test. Administrators have more information to support that students were tested in a variety of simulated contexts. For the scoring rubric, a holistic scoring method was retained for this newer version. Raters assign a single score to each of the 12 tasks, and these scores are then averaged to make a decision. Measuring language knowledge through responding to several tasks is known as construct-centered performance tests by Messick (1994) and mean-focused by task-based language testing researchers (Long & Crookes, 1992; Robinson, 1996).
1.3 Penn State’s development of the AEOCPT

In order to meet the requirements of Pennsylvania’s SB 539, faculty at Pennsylvania State University purchased the 1982 version of the SPEAK test to use for certifying ITA candidates. First administered in 1990, the faculty were quick to begin revising the test to better match the wording of SB 539. As discussed earlier, the state legislature passed SB 539 so that ITA tests must be of English proficiency in the classroom. Halleck and Moder (1995) point out one weakness of the SPEAK test is that the items to not replicate situations in a university classroom. Kaplan (1989) offers a similar critique of the test but states his argument that results from test are ambiguous and difficult to interpret for certifying ITAs.

An example that illustrates these criticisms comes from one question on the test that was memorable to me as a rater. Test takers are prompted to answer a question about whether they think animals should be kept in a zoo. From my personal experiences in rating this item, the responses given to this task varied widely, ensuring that the answers were not just routine phrases. One issue with this item (Following Halleck and Moder’s (1995) concern is that it has little to do with the context of a university classroom. If a student does not answer this kind of question appropriately, what does a lower score on item inform us about their ability to teach in a classroom? An argument could be made that this question is more a test of online cognition for which test takers are not likely to prepare an answer before the test.

*From SPEAK to AEOCPT.* Because Pennsylvania State University ITA administrators started their testing with the SPEAK, I start the AEOCPT development timeline from this stage. One of the first revisions that occurred was reducing the amount of tasks from six to four. The four tasks that remained on the test were a single picture description, picture story, an opinion question, and reading a schedule. Whereas up until this point students’ responses were recorded
and graded later, administrators revised this method to having raters administer and score the test simultaneously. This change in testing procedures followed the methods of The ITA test (Smith, Meyers, & Burkhalter, 1992), which is part of the textbook used by the Pennsylvania State University ITA program.

By changing the administration of the test from recorded performances to ones that are scored live by raters, the AEOCPT administrators effectively created the AEOCPT as an assessment instrument independent of (but based on) the SPEAK test. The AEOCPT not only requires raters to score candidates in real time but also to interact with them. This interaction transforms the test from static to dynamic.

Sternberg and Grigorenko (2002) describe two possible types of dynamic tests. One is interventionist, called the “sandwich model” by Sternberg and Grigorenko. The interventionist approach is similar to the experimental research design in social sciences. A pretest and posttest are given in between an intervention, typically classroom instruction. The two tests are intended to inform teachers on the effectiveness of the instruction and whether enough students have mastered the classroom material. One disadvantage of interventionist dynamic testing is that it requires twice the time to design and administer the two tests. In addition, teachers are only able to know whether their students need additional instruction after the intervention has ended.

Sternberg and Grigorenko introduce an alternative dynamic testing approach that addresses the two previously stated concerns. The interactionist approach, called the “cake model” by Sternberg and Grigorenko, requires the design of only one test. The difference between this approach and interventionist dynamic testing is that students receive instruction immediately after giving an incorrect response. The advantage to using this approach is that teachers are able to know immediately where students struggle and what kind of instruction can help them to learn.
Administrators transformed the AEOCPT to dynamic interactionist by changing the raters’ roles from retrospective observers to active participants. This transformation was achieved by training raters to ask follow-up questions to task responses. These follow-up questions were intended to provide test takers an opportunity to correct any mistakes noticed by raters. If a candidate were able to successfully repair an error, they were seen as having more advanced language knowledge than a test taker who could not correct a mistake. An interactionist AEOCPT follows a theory of language teaching and acquisition researched by faculty at Pennsylvania State University.

**AEOCPT Task Revisions.** After revising the AEOCPT to become a dynamic interactionist test, the next stage of revisions occurred when faculty further revised the tasks themselves. The number of tasks was not changed, but the prompt given to test takers was better aligned with an academic context. The *opinion* questions were changed to elicit information about ITA candidates’ reasons for choosing Penn State or what their career goals were after completing their graduate degree. The task that requires *reading from a schedule* was also changed to have information more closely aligning with the classroom environment. Test takers were now asked to read announcements such as test review sessions or job search seminars. Finally, the *single picture task* was replaced with a task that required candidates to give a *mini-lecture* on a term from their field of study. All of these tasks became dynamic because raters were required to ask follow up questions. (As a minor note, one variable of the scoring rubric was given more weight during this round of revisions, *pronunciation*. I note this revision as minor because this weight was not retained through subsequent revisions but changes to the task were kept.)

**AEOCPT Scoring Rubric Revisions.** So far, I have discussed revisions to the AEOCPT that have been primarily focused on the tasks of the test. Revisions to the scoring rubric did occur
after changing the tasks to better fit the classroom context. The scoring rubric from the SPEAK test is holistic with a score ranging from 0 to 3 and awarded by raters based on descriptions of grammar, pronunciation, and fluency for each level (Kim, 2001). Stated previously, pronunciation was momentarily given more weight, but this stage of revisions transformed the rubric into an analytic one. Raters were now instructed to give a score, still ranging from 0 to 3, for each descriptor. Researchers in language testing have provided evidence supporting the use of analytic scoring rubrics over holistic ones (Hamp-Lyons, 1991; Weigle, 2002).

One advantage to using an analytic rubric is that more information can be gained on raters’ perceptions of test takers. Although this additional information could introduce more error variance that contributes to lower reliability, these scores could be used by administrators to understand where differences are occurring. Scores from a holistic rubric, by contrast, can only be investigated for low reliability by interviewing raters outside of the testing moment. Faculty used the additional information from the analytic AEOCPT rubric to revise the test’s cut-score. Four cut-scores are used for making placement decisions with the test, but this revision focused on changing the cut-score that exempts students. Out of a total score of 300, the previous cut-score for being exempt from coursework was 230. This standard was raised by 20 points to 250 after investigating the influence of the analytic rubric.

After the AEOCPT rubric was changed to analytic, a revision in 2005 was made to the storytelling task. Administrators replaced this task with one that required candidates to act out an office hour scenario with one rater. This role play task was developed from classroom materials that will be covered in a later chapter. Chiang’s (2009, 2011) research also offers support for an office hour role play task. Chiang found that office hour contexts are encountered by all ITAs but can present challenges to language performance. The challenges can be overcome, however,
through negotiation with an interlocutor. Because Pennsylvania State University requires office hours to be held by all instructors teaching classes, this role play task is important for understanding candidates’ abilities to communicate and resolve a problem with the rater.

1.4 Motivation for this dissertation

In Table 1-1 below, I briefly summarize the history of the AEOCPT from its original SPEAK version to its addition of the office hour task.

<table>
<thead>
<tr>
<th>Revision</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEAK test license acquired</td>
<td>1990</td>
</tr>
<tr>
<td>Face to face interaction and rating replaces recorded rating</td>
<td>1991</td>
</tr>
<tr>
<td>Tasks revised to better match academic context and pronunciation is weighted</td>
<td>1995</td>
</tr>
<tr>
<td>Scoring rubric is revised to analytic, cut-scores are adjusted</td>
<td>2000</td>
</tr>
<tr>
<td>Role play task replaces picture story task</td>
<td>2005</td>
</tr>
</tbody>
</table>

Readers will notice from the information above that the contextual influences impact multiple facet of the AEOCPT. Although test designers are able to integrate contextual influences into test administration, researchers argue that these influences are not accounted for when scoring the test (Bachman, 2002, 2007; Chapelle, 1998; Leung & Lewkowicz, 2006). In other words, test takers may be asked to act out multiple scenarios of a target context, but the test scores do not reflect the different performances. McNamara (1997, 2001) identifies test design, raters, and test partner (e.g. paired testing) as some ways to operationalize contextual influences in language tests. So far in language testing, rater variation has been the most researched of the three (Carey, Mannell, & Dunn, 2010; Isaacs & Thomson, 2013; Johnson & Lim, 2009; Kim, 2009; Xi & Mollaun, 2011).
The information in the table above, however, shows that contextual influences impact the AEOPCT’s tasks the most. In order to address this gap in research, I propose using a language testing theory that integrates contextual and theoretical influences. I took the first step to addressing this gap by identifying the three major components of the AEOCPT: Tasks, Rubric, and Score. According to McNamara’s (1996) description of performance tests, tasks would represent the contextual influences, the rubric represents the theoretical influences, and I propose score as a third influence (Testing influence). Brown (2005) describes four possible uses for language tests: 1) Proficiency, 2) Placement, 3) Diagnostic, or 4) Achievement. Because the AEOCPT is a placement test, I argue that it is important to consider how the rubric scores match the intended purpose of the test.

**Tasks.** As is evident in Table 1-1, the tasks have gone through the greatest revision in the test’s development history. The four tasks of the AEOCPT take around 30 minutes for test takers to complete, and this is comparable to commercial tests. TOEFL iBT’s speaking section has six tasks, and IELTS speaking section has three tasks. What has not been made clear is how the four tasks relate to each other.

Luoma (2004) notes that different tasks can have different levels of difficulty and the level differences can negatively impact scores if not taken into account by test developers. Previous AEOCPT raters support this statement by stating that they believed that not all four tasks were equally valuable for informing placement decisions. From a survey I administered to raters, I was able to identify that the fourth task, Reading an announcement was especially problematic. The concerns that raters had with this task was that it was too routine. In other words, raters did not observe enough variance between test takers to distinguish them. I attempted to resolve this issue by revising the task to Making changes to an announcement. This
change copies the revision ETS made to the exact same task in the 1996 version of the SPEAK test. I will discuss the development of the four tasks in the next chapter and discuss why testing their contribution to placement decisions is important.

**Scoring Rubric.** The second component of the AEOCPT that I focus on is the *scoring rubric*, and this received increased attention toward the end of the development history. The purpose of the scoring rubric is to focus raters on variables that can be observed across the four tasks. Because scores from the AEOCPT are used by administrators to place ITA candidates into a class for English instruction, the scoring rubric contains descriptions of language knowledge. These descriptions are intended to help raters distinguish between the different placement points. I have provided the scoring rubric that was used before my revisions in Appendix A. One issue that raters expressed with the rubric in Appendix A is that criteria and descriptions were difficult to understand. My interpretation of this feedback is that the rubric contains double-barreled descriptions (Olson, 2008).

Although the term double-barreled is more commonly used to describe an issue in questionnaire design, this concept can be used to describe the AEOCPT’s rubric descriptions as well. One reason for this is that the test is an instrument that surveys raters about the performance they observe from test takers, a survey research design. The definition of a double-barreled description is wording that raters perceive as two independent ideas. Thus, making one decision about two independent ideas would be an extraneous error variable. In the description of Fluency from the rubric in Appendix A, raters give one score for thought groups, routine lexical phrases, and their ability to negotiate with the test taker. Rater disagreement on this item would be difficult to resolve without dividing the item into three separate ones. The revisions that I
implemented for the AEOCPT scoring rubric were expanding the criteria and reducing the descriptions for each criterion.

**Placement Score.** Finally, the last component I focused on within the AEOCPT is the score used for making placement decisions. From Table 1-1, readers will notice that only one revision was made in relation to the score, moving the cut score from 230 to 250. Although this component of the AEOCPT received the least attention, it carries the most power. The score is the only piece of information given to other departments for understanding candidates’ placements. The decision to move the cut-score from 230 to 250 was likely made by administrators, but there is no paperwork documenting the process that went into the change. Because I was making changes to the other two components of the AEOCPT, I followed recommendations from Cizek (2012) to check the stability of all four cut-scores.

The revision to the cut-score in the AEOCPT only addressed the decision of whether a candidate should be placed in a course or immediately certified to teach. Two other decisions, however, are made from scores that are lower than 250. I conducted a study with raters to determine if the same cut-scores were maintained when the new analytic scoring rubric was used. Using a contrasting-groups standard setting approach (Brown & Hudson, 2002), identified by Livingston and Zieky (1982) as student centered, I instructed raters to identify where they believe every candidate should be placed during a fall administration of the AEOCPT. The reason this approach could work without raters being biased by their knowledge of the cut-scores was that raters began using the new scoring rubric during this time. Because the new rubric contained more criteria to score, raters did not know the relationship between the old cut-scores and the rubric.
I was able to use the results from the contrasting-groups method to confirm the score for deciding between candidates who are certified to teach and those that need one semester of coursework, 118G. More meaningful contributions I was able to make to AEOCPT-based decisions were changing the cut-scores for the other two decisions.

Table 1-2. Summary of results from contrasting-groups standard setting approach

<table>
<thead>
<tr>
<th>Old Score Band</th>
<th>New Score Band</th>
<th>Decision</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>250-300</td>
<td>250-300</td>
<td>Exempt</td>
<td>Certified to Teach</td>
</tr>
<tr>
<td>230-249</td>
<td>200-249</td>
<td>118G</td>
<td>One semester of coursework</td>
</tr>
<tr>
<td>200-229</td>
<td>150-199</td>
<td>117G</td>
<td>Two semesters of coursework</td>
</tr>
<tr>
<td>&lt;200</td>
<td>&lt;150</td>
<td>115G</td>
<td>Three semesters of coursework</td>
</tr>
</tbody>
</table>

For example, the cut-score for ITA candidates requiring two semesters of coursework, 117G, was 200. Raters’ judgments, however, supported that this standard was too high. My proposal was to change the 117G cut-scores to 150. This change in cut-score then meant that the cut-score for the lowest level, requiring three semester of coursework, was changed to a score below 150.

Table 1-2 above shows a summary of the contrasting-groups approach with information about the previous cut-scores and what they were changed to now.

Administrators of the ITA program noted that there was an adjustment period for client departments to understand the reasons for seeing overall lower scores. Many department representatives did eventually note, however, that they were able to use the new scores to better understand the relationship between their candidates. This increase in transparency is good, but the granularity of information given by the test is still too coarse. Granularity is commonly associated with Cognitive Diagnostic Assessments (Leighton & Gierl, 2007), but Kunnan and Jang (2009) have connected the term with communicating information for placement tests.
Granularity refers to the amount of information that can be gained about a test taker from the scores of the test.

Kunnan and Jang provide feedback from the TOEFL as statements that are provided when test takers receive their scores. These statements are broad characteristics of test takers’ strengths and weaknesses who score within the same band. Currently, ITA candidates at Pennsylvania State University are given their score on the AEOCPT along with their required placement within the program. Despite the fact that the test consists of eight rubric items scores by raters across four tasks, a total of 32 data points, no other information is given to help stakeholders understand the placement decisions. Providing more information to stakeholders would require a way to make placement decisions beyond a summative score. This challenge within the scoring component of the AEOCPT is one I will address in this dissertation.

1.5 Summary

I used this introductory chapter to document the history of states requiring an additional test of English proficiency for any international student expecting to receive a teaching assistantship. More specifically, I focused on two states’ laws, Pennsylvania and Texas, due to my personal connection with creating tests to meet these laws requirements. I then discussed ETS’s attempt to create the TSE and SPEAK tests for meeting the various state requirements for certifying ITAs. My discussion of SB 539 and HB 638 show the challenge for both testing companies and universities to create tests that comply with every detail within one state’s law. After setting up the challenge of ITA testing across the United States, I shifted the focus to the ITA test given at Pennsylvania State University.
The AEOCPT started off as an early version of the SPEAK test, but administrators soon began revising the test to better align with their theories of language teaching. Table 1-2 above shows a brief timeline of the AEOCPT’s development history. Table 1-3 is an update to the timeline that includes my revisions.

**Table 1-3. Updated AEOCPT development history**

<table>
<thead>
<tr>
<th>Year</th>
<th>Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>SPEAK test license acquired</td>
</tr>
<tr>
<td>1991</td>
<td>Face to face interaction and rating replaces recorded rating</td>
</tr>
<tr>
<td>1995</td>
<td>Task revised to better match academic context and pronunciation is weighted</td>
</tr>
<tr>
<td>2000</td>
<td>Scoring rubric is revised to analytic, cut-scores are adjusted</td>
</tr>
<tr>
<td>2005</td>
<td>Role play task replaces picture story task</td>
</tr>
<tr>
<td>2012</td>
<td>Rubric criteria expanded and descriptions reduced</td>
</tr>
<tr>
<td>2013</td>
<td>Task 4 modified and placement cut-scores adjusted</td>
</tr>
</tbody>
</table>

In order to answer a gap in research that context is insufficiently taken into account in language tests, I identified three major components (Tasks, Rubric, and Score) with the intent of connecting with a language testing theory. I gave some details about these components and the amount of work that went into revising them. My intentions in this dissertation are to: 1) Confirm a theory of language ability raters can use, 2) Confirm the AEOCPT is making sound placement decisions into graduated categories of distinct language abilities, and 3) Confirm the integration of task, rubric, and score as reflecting the complexity of testing language in particular. In the next chapter, I will review previous research that went into developing my research questions and proposed revisions with the three components.
Chapter 2 Literature Review

In this chapter, I will review previous research used to revise the three components of the AEOCPT identified in Chapter 1. Research supporting revisions to the rubric, score, and tasks will be discussed in this order. I will discuss previous research on each component and identify gaps in each that I propose to answer with my three research questions. I will review frameworks and coordinating key concepts from selected theories to develop what I am going to call a moderate interactionalist theory of language testing.

2.1 The Scoring Rubric: From a construct to observable language use

The “P” in AEOCPT refers to the fact that this is a performance test; defined by McNamara (1996) as consisting of one or more tasks that elicit an extended performance from test takers. These performances are then scored by raters either live or through the use of an audio or video recording device. McNamara gives two advantages for test developers to use performance tests. One advantage is that the tasks can be representations of real world situations that test takers will encounter. For the AEOCPT, these situations are either in an American university classroom or office. The second advantage given by McNamara is that a performance test allows for the measurement of complex variables, such as language knowledge.

Describing the complex nature of language knowledge is a challenge that has been a prime focus of language testing research. In terms of performance test development, Brown and Abeywickrama (2010) give three language knowledge frameworks that are commonly used to develop performance test scoring rubrics. These are discrete-point, integrative, and communicative language testing. Hoekje and Williams (1992) provide language knowledge frameworks within communicative language testing that were used in ITA performance testing. Since the writing of this article, researchers have proposed revisions and alternatives that I will
discuss in this chapter. In addition, I will use a different categorization from the three previously stated as associated with language testing. One limitation with the discrete-point, integrative, and communicative language testing labels is that they describe the kinds of items associated with each test rather than how developers operationalize language knowledge.

**Table 2-1. Bachman’s (2007) categories of language testing theories**

<table>
<thead>
<tr>
<th>Focus</th>
<th>Definition</th>
<th>Example Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ability</td>
<td>Tests that isolates measurement to knowledge within test takers</td>
<td>TOEFL (Farnsworth, 2013)</td>
</tr>
<tr>
<td>Task</td>
<td>Tests that measure the accomplishment of a single task</td>
<td>TEACH (Abraham &amp; Plakans, 1988)</td>
</tr>
<tr>
<td>Interaction</td>
<td>Tests that measure test takers’ abilities to use knowledge appropriate for a target context</td>
<td>AEOCPT</td>
</tr>
</tbody>
</table>

Bachman’s (2007) chapter organizes and discusses language knowledge frameworks used in testing. His chapter begins by labelling the three categories as, ability-focused, task-focused, and interaction-focused. **Table 2-1** above shows a definition of the three categories as well as ITA tests that were developed from each of the theories. ETS and UCLES use an ability-focused language framework proposed by Lado (1964). The language knowledge framework used here at Pennsylvania State University was proposed earlier than 1964, but it has only been recently applied to testing. Using Bachman’s categorization, I will organize previous research on language frameworks and connect them to the AEOCPT components.

I argue that the AEOCPT reflects an interactionist language framework. Bachman (2007) traces his category of interaction-focused language frameworks to Kramsch’s (1986) interactional competence. Kramsch’s discussion of language proficiency focused on the ACTFL
Oral Proficiency Interview (OPI), an ability-focused test. She defined interaction as a temporary space where individuals’ abilities and contextual resources come together. Bachman further divides the category into minimalist, moderate, and strong interactionalist. I will briefly describe these three interaction-focus categories and support the one that I propose the AEOCPT scoring rubric is measuring.

2.1.1 Proposing a moderate interactionalist language ability framework

The first type of interactionalist language construct is **minimalist**, explained by Bachman (2007) as including Kramsch’s (1986) interactional competence theory. A feature of Kramsch’s theory that places it into the category of **minimalist interactionalist** is the separation of ability and context as independent. When an individual (e.g., a teacher) is placed in a situation, he/she strategically selects dimensions of their cognitive ability (e.g., lexical memory) in conjunction with contextual resources (e.g., info on a PowerPoint) in order to accomplish their communicative goal (e.g., teaching). Kramsch argues that these selected abilities and contextual resources should be the focus of testing decisions. Bachman differentiates **minimalist interactionalist** from tests that use non-interactionalist language constructs by emphasizing that environment is an important dimension to consider rather than an extraneous variable that must be reduced. Bachman’s category of **minimalist interactionalist** suggests that the focus of the construct is still on an individual’s cognitive ability.

On the opposite end of the interactionalist spectrum, Young and He (1998) propose a theory of language that Bachman identifies as **strong interactionalist**. Young and He argue that language knowledge can only be constructed from the interaction of individuals within a target context. In other words, every situation produces different dimensions of language knowledge
that must be identified in order to successfully test. Young and He’s theory was developed from the results of research in Conversation Analysis and Discourse Analysis. Young and He propose testing interactional competence through the use of actual discursive practices rather than constructed tasks or items. A test developed from a strong interactionalist theory has the potential to provide a wealth of information for making decisions, but there are many challenges. Two of these challenges are creating discursive practices within a testing situation and connecting these practices with scores to infer individual ability for making decisions. Young and He acknowledge at the beginning of their book the challenge of proposing a language knowledge framework that is the same across discursive practices.

The last type of the interactionalist category is described by Bachman as being in between the minimalist and strong types. The theory Bachman proposes as being moderate interactionalist is Chalhoub-Deville’s (2003) ability-in language user-in context. Chalhoub-Deville describes this theory as a person’s language knowledge only being understood through a set of tasks developed from one context. This theory parallels the minimalist theory in that the same language knowledge framework is measured for accomplishing a communicative goal. Where they differ is that Chalhoub-Deville proposes that the test’s tasks must come from the same context so that the measurement of language knowledge can be best understood. A difference between the moderate and strong theories would also be related to the identification of a common set of language knowledge. The moderate interactionalist theory has a common language knowledge framework that is measured in every task, but each task in the strong interactionalist theory has a unique framework that cannot generalize. Table 2-2 below contains summaries of each theory along with visualizations of the relations between tasks and language knowledge.
<table>
<thead>
<tr>
<th>Type of Interactionalist Theory</th>
<th>Description</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimalist (Kramsch, 1986)</td>
<td>Task(s)-in-Context is important to consider when measuring language knowledge, tasks are not designed to be interrelated</td>
<td><img src="image1" alt="Minimalist Model" /></td>
</tr>
<tr>
<td>Moderate (Chalhoub-Deville, 2003)</td>
<td>Task(s)-in-Context is used to better infer language knowledge, task share some aspect of common context</td>
<td><img src="image2" alt="Moderate Model" /></td>
</tr>
<tr>
<td>Strong (Young &amp; He, 1998)</td>
<td>Task-in-Context is inseparable from language knowledge, task and language knowledge are completely interdependent</td>
<td><img src="image3" alt="Strong Model" /></td>
</tr>
</tbody>
</table>

The claims associated with the moderate interactionalist theory would make interpreting test results more meaningful for their intended purposes.

### 2.1.1.1 Application to the AEOCPT

Because I identified three components of the AEOCPT, Chalhoub-Deville’s (2003) moderate interactionalist theory is the only that has similar components. In addition, the purpose
of the test is to measure the extent that candidates can use English as instructors in an American college classroom. This purpose aligns with Chalhoub-Deville’s argument that language knowledge is best understood when a test measures what individuals can do in a target context. The model in Table 2-2 explicitly shows only two of three components, but score is also present in the model. The overlap between each task and the language knowledge factor represents a portion of the factor that raters can measure with the AEOCPT. In addition, the overlap between tasks represents that they are from the same target context. Results for answering RQ1 will also provide information on the extent that the four tasks are unique from each other but capable of measuring the same language knowledge framework.

2.1.2 Operationalizing language user for the scoring rubric

One challenge of the moderate interactionalist theory is that Chalhoub-Deville does not provide any tests that could serve as operational examples. One of my goals in adapting this theory to the AEOCPT is to operationalize the theory in actual testing practice. Chalhoub-Deville’s ability-in language user-in context theory matches with the AEOCPT’s score, rubric, and tasks. In addition to there being no operational examples of the theory, there is no information about descriptions that could fit into the theory’s three components. Therefore, I will need to use additional research to construct operational definitions of the three components.

Chalhoub-Deville and Deville (2006) propose their moderate interactionalist theory as an advanced interactionist language ability framework to the Communicative Language Ability (CLA) framework proposed by Bachman (1990). His interactionist language ability framework has been critiqued by several researchers (Bachman & Palmer, 1996; Chapelle, 1998; McNamara, 1996; Oller, 1983). One challenge is that Bachman distinguishes two frameworks that are commonly conflated. Given that prominence of Bachman’s CLA theory in the literature, I will
take a moment to summarize it. I refer readers to Figure 2-1 below for an accompanying visual of my description. The CLA theory shows language production from an individual as the combination of two knowledge systems. Information from the two systems are brought together into the strategic competence system and physically produced into the environment. The strategic competence variable is most critiqued by researchers.

![Figure 2-1. Communicative language ability framework (Bachman, 1990)](image)

One reason that strategic competence has been discussed by notable scholars like McNamara (1996) and Chapelle (1998) is that term appeared in another theory. Strategic competence first appears in Canale and Swain’s (1980) seminal work on communicative competence. Their framework of language knowledge consisted of grammatical competence, sociolinguistic competence, and strategic competence. Here strategic competence is defined as an individual selecting only linguistic resources to best function in context. Bachman (1990)
modifies the definition of strategic competence to be the ability of an individual to strategically select world knowledge and language knowledge to accomplish a goal within the target context.

My interpretation for this move of strategic competence is similar to that of McNamara’s (1996) explanation. Bachman argues that strategic competence involves more cognitive processes than those assigned only to language knowledge. This revised definition of strategic competence more closely matches Cognitive Psychology’s definition of working memory.

![Diagram: Information Processing Model (Martinez, 2010)](image)

**Figure 2-2. Information Processing Model (Martinez, 2010)**

Martinez (2010), an Educational Psychologist, defines working memory as a structure within human cognition that carries out work on information from an individual’s external environment and long-term memory. An illustration of this definition is given above in Figure 2-2 for readers to compare with the CLA framework. Because long term memory would contain language knowledge and world knowledge, dividing it into two variables shows the near identical structure of Figures 2-1 and 2-2. The work done with working memory is diverse, e.g.,
understanding an environment, changing an environment from information in long-term memory, modifying long-term memory based on information within the environment, etc. A study done by Phakiti (2008) confirms that strategic competence is a complex structure that can be measured with the same instruments used for working memory, e.g., a memory capacity test.

This additional information about working memory, or strategic competence, closely matches the description I previously gave about interactional competence (Kramsch, 1983). Individuals are using resources within their environment along with their own knowledge to accomplish desired goals. In fact, Bachman (1990) does refer to Kramsch’s article when introducing strategic competence.

Because Bachman’s research (Bachman, 1990; Bachman & Palmer, 1996, 2010) has focused on the cognitive elements of CLA, I agree with Chapelle’s (1998) classification of the CLA framework as minimalist interactionalist. Figure 2-1 shows that context is important to consider in the model but the majority of the variables are within an individual’s cognition. In a testing situation, the previous sentence means context is important to consider when making a task but the relation between tasks is not a major concern. After establishing Bachman’s CLA framework is similar to Kramsch’s (1983) minimalist interactional competence, I am now able to connect the language knowledge framework to the moderate interactionalist framework I used for the AEOCPT scoring rubric.

One limitation of Chalhoub-Deville’s (2003) moderate interactionalist framework is that the three elements are not described in great detail. Bachman’s (1990) language knowledge framework is a component of his minimalist interactionalist framework and can fit into Chalchoub-Deville’s (2003) language-user component. The language knowledge framework comes from previous research on communicative competence (Canale, 1983a, 1983b; Canale &
Swain, 1980; Hymes, 1972). Figure 2-3 below is a visualization of the framework from its most recent revisions (Bachman & Palmer, 1996).

![Language knowledge framework](image)

**Figure 2-3. Language knowledge framework (Bachman & Palmer, 1996, 2010)**

Although not a major revision, Bachman and Palmer (2010) revised the variables’ names from competence to knowledge. My description of the language knowledge model in Figure 2-3 will start from the language knowledge variable also in Figure 2-1. Bachman first divides language knowledge into organizational knowledge and pragmatic knowledge.

*Organizational knowledge* is defined as the abilities that are expressed in the formation of language at local and global points. *Grammatical knowledge* represents the local level and *textual knowledge* represents the global level. *Grammatical knowledge* was retained from previous theories of communicative competence (Canale, 1983a, 1983b; Canale & Swain, 1980) and is defined as having lexical, morphological, syntactic, sentence-level semantic, and phonological abilities. *Textual knowledge* is defined as the connection of two or more utterances or sentences.
The second variable that Bachman distinguishes within language knowledge is *pragmatic knowledge*, defined as an individual’s selection of appropriate language to use in context. Bachman’s division of pragmatic knowledge into sociolinguistic knowledge and functional knowledge comes from pragmatic research done by van Dijk (1977). Sociolinguistic knowledge is defined as the appropriate utterance for a context of interest. The definition of this variable is language used to alter the information within a context in order to achieve a desired goal. Bachman and Palmer (1996) retained the definition of this variable but changed the word to functional knowledge.

### 2.1.2.1 Application to the AEOCPT

Bachman and Palmer’s language knowledge framework is the most recent and is part of the CLA framework that I identified as minimalist interactionalist. Using the language knowledge framework fills in the gap with the moderate interactionalist theory having no information about what comprises the language user. In addition, using the language knowledge gets the AEOCPT rubric closer to other ITA tests. The previous rubric in Appendix A contains descriptors that come from the SPEAK and ITA test. This information operationalizes both organizational knowledge and pragmatic knowledge but into the same categories.

Because the textbooks used in ITA courses have lessons that divide language knowledge into Bachman and Palmer’s (1996) four categories, the rubric should reflect this as well. The revised rubric, Appendix B, contains information that better divides the four first-level variables. I use first-level variables here to propose that Figure 2-3 serves as the structural portion of a Confirmatory Factor Analysis model (Kline, 2011). The scoring rubric would then be the measurement portion, allowing the model’s fit to be tested with AEOCPT response data.
2.2 Score: Establishing its relationship to intended decisions

The second major component of the AEOCPT I wish to discuss is score. Once raters have completed the scoring rubrics for all tasks in the AEOCPT, the selected Likert points must be interpreted as a score for making a placement decision. This decision is a function of the number of points within the rubric and how different categories of performance are calculated. For the AEOCPT rubric, the eight items are currently weighted evenly with four possible points. The possible total of 32 points is converted to a 300 point scale that is reported to university stakeholders. The only revision made to score during the AEOCPT’s development history was changing the cut-score for one decision from 230 to 250. This cut-score, however, is only one of four decisions, 115G, 117G, 118G, or certified, made by administrators using the AEOCPT score. I will organize this section by discussing previous work done with placement tests, points, and standards as they relate to test scores.

Brown (2005) identifies placement tests as norm-referenced in which students’ abilities are compared to each other or a norming sample. This means that placement decisions are made by grouping students by similar abilities. His discussion of the term within the same chapter, however, states that these tests group students together by similar strengths and weaknesses. The identification of strengths and weaknesses is done by comparing students to performance descriptions within a program of interest. Green (2012) notes that individualized student information has the additional benefit of informing teachers about their incoming class.

The use of norm-referenced tests (NRTs) in language programs for placement purposes other than initial class organization extends its description beyond traditional norm-referenced. True diagnostic tests would be those that measure the current knowledge state of test takers and how much is left to learn for achieving a desired goal or standard. NRT placement tests, on the
other hand, would appear to possess qualities given to diagnostic tests, but Brown (2005) requires that diagnostic tests be criterion-referenced. Kunnan and Jang (2009) refute this by arguing that diagnosis is not a type a test but instead a potential feature. They describe score reports from norm-referenced tests such as TOEFL and IELTS to show that diagnostic feedback is currently being given with proficiency tests. One difference between norm-referenced and criterion-referenced diagnostic feedback is that the latter is more closely connected to classroom instruction that can more directly address weaknesses.

Scores from the AEOCPT are currently being used by administrators to inform placement decisions, but candidates’ strengths and weaknesses are not communicated to teachers. This lack of communication is not a limitation of the test but rather an underuse of the rubric. Like many other performance tests in language testing (Brown, 2012; Luoma, 2004; McNamara, 1996), the AEOCPT has Likert scales (Likert, 1932) for raters to score candidate performance. These scales should provide more information about strengths and weaknesses than dichotomous items traditionally used in diagnostic testing. The current problem is that the Likert scores are summed to make placement decisions, reducing the 32 possible scores to one. In order to achieve the goal of adding diagnostic information to the AEOCPT, one step is to look at previous recommendation on optimal creation and use of Likert scales.

Likert scales provide raters or respondents an alternative scoring method for giving information, gradations rather than dichotomies. Winke (2014) provides a general overview of Likert scales and how they are used in latent trait research. In her discussion of Likert scales, Winke gives a recommendation of five to seven points based on previous discussions by researchers. Brown (2012) confirms the same recommendation of using a Likert scale that has five to seven points. What neither Winke nor Brown provide in their discussion is arguments for
test developers to choose one set of points over another. Choosing either five or seven points is immaterial from a psychometric perspective. Miller (1956) shows that both options produce similar reliabilities.

One reason for there being little information about selecting the optimal Likert scale is previous research has focused on other qualities. Betts and Hartley (2012) conducted a study that looked at how the positioning of points affects reported scores. In other words, does having the higher points on the left or right side of a paper affect participants’ selections? Cook, Heath, Thompson, and Thompson (2001) looked at the impact of alternative Likert scales (e.g., sliders) administered on computers. An additional reason that little research has been done on identifying an optimal amount of points is that the summed scores are interpreted equivalently. Scores from Likert scales are commonly summed together and used for interpretation or further analysis (Albaum, 1997).

However, interpreting a summative score is problematic for placement tests because test takers could be provided only general information about what is in the rubric. In order to operationalize a moderate interactionalist framework within the AEOCPT, a scoring model needs to be used that takes into account the impact of every descriptor in the rubric. Researchers of Cognitive Diagnostic Assessments (CDA) have proposed several methods for getting information from a variety of tests purposes. Leighton and Gierl (2007) define CDA, alternatively known as Diagnostic Classification Models (Rupp, Templin, & Henson, 2010), as instruments that provide information about knowledge and processes within an individual. Rupp and Mislevy (2007) discuss one type of model that I will use in this dissertation, mixture models.

Mixture models are probabilistic models that analyze instrument response patterns for subsets within a population of interest. In other words, mixture model analysis could be used
with AEOCPT data to identify candidate subgroups that make up placement categories. The type of mixture model that I will use in this study is finite mixture models, also known as Latent Class Analysis (Goodman, 1974a, 1974b; Lazarsfeld & Henry, 1968). One benefit of using this model for informing AEOCPT placement decisions is that scores are candidates’ posterior probabilities of class placement. After determining the highest probability of class placement, responses can be viewed to identify strengths, items that contributed to current class placement or higher, and weaknesses, items that contributed to lower class placement. The methods for identifying and measuring the fit of Latent Class models will be discussed more in the next chapter.

2.2.1 A Cognitive Diagnostic Model used on an ITA performance test

The type of mixture modeling I intend to use in this dissertation is new to language testing, but a variation of it was done by Choi (2013, 2015). In addition to analyzing the data using mixture modeling, Choi’s dataset was also an ITA performance test (Test of Oral Proficiency). Three key differences exist between Choi’s study and the one I am proposing in this dissertation. The first difference between Choi’s study and this dissertation is that developers of the Test of Oral Proficiency (TOP) conceptualize it as an ability-focused language test. In other words, raters are instructed to focus on English ability (e.g., pronunciation, vocabulary, grammar, rhetorical organization, and question handling) and not on how test takers complete the two tasks of syllabus presentation and presentation (University of California Los Angeles, 2013). Although the AEOCPT also has question handling as a rubric item, the descriptor in TOP focuses only on understanding questions while the descriptor in the AEOCPT focuses on how questions are answered.

The second difference between the two studies is that the TOP serves as an achievement test. Although test takers are divided into three categories (Pass, Provisional Pass, and Non Pass),
those labeled as Provisional Pass and Non Pass can choose from the same courses. Decisions using the AEOCPT, on the other hand, place students into one of four categories. Three categories are connected with specific courses that must be completed to pass (See Chapter 1 for more information). Because of this direct connection between placement decisions and coursework, LCA has practical implications for course development in the ITA program. The results from Choi’s study provide information on why test takers could have passed or failed the TOP.

The third difference between Choi’s study and this dissertation involves how information about test takers’ strengths and weaknesses is gained from the two tests. Choi identifies the data analysis method as finite mixture modeling. Collins and Lanza (2010) further divide finite mixture modeling into three analyses: Latent Class Analysis (LCA), Latent Profile Analysis (LPA), and Latent Transition Analysis (LTA). The formulas for carrying out the analyses are similar with the key difference being assumptions about the indicators (rubric items). Because LCA is used in this dissertation and LPA is used in Choi’s study, I will focus here on describing the differences between these two analyses.

Because LCA is a cross-sectional data analysis method, it has the advantage not requiring a large number of items (Less than 15). This number prevents many large scale assessments from using this analysis but is ideal for performance tests. Performance tests elicit extended performances from test takers but are limited by what raters can grade in real time (Luoma, 2004). LPA is a variation of LCA that assumes indicators are continuous. Tests are analyzed using LPA by adding up number of questions correct in each subtest. Using TOEFL as an example, LPA could be used to analyze the mixture of scores for the Speaking, Listening, Reading, and Writing subtests.
Choi’s (2013, 2015) dataset was originally categorical (Four rubric items across two tasks), but the data was converged to an IRT model before being submitted to LPA. Scoring test takers using an IRT model produces an ability score (Theta) for each test taker. The number of indicators is still four, but ability scores are on a continuous scale instead of categorical. Choi’s stated advantage for converting the data to continuous ability scores is that rater variation could be accounted for in the analysis. In other words, the program running the IRT model could adjust a test taker’s score for strict raters (Move ability up) or lenient raters (Move ability down). The results from LPA would then be test taker’s strength and weaknesses in language ability with rater variation controlled.

One disadvantage to using LPA is that selecting the best fitting class model (model selection) is not as clear as the process in LCA. Bauer and Curran (2003) found that class overextraction can occur if normality is not confirmed within indicators. Overextraction could be observed in model selection when seeing comparative fit statistics that are unable to identify the best fitting class. This happens in LCA and LPA when the comparative fit statistics continually decrease (Indicating better fit) past the number of classes that are theoretically supported. After a researcher selects a latent profile model, another limitation with LPA is interpreting the classes.

After identifying the best fitting class model, the next step is to interpreting the classes based on the response to each indicator. In this dissertation, interpretation is labelling a class based on the scores for each indicator. LPA, on the other hand, uses indicators that are a composite of items. Interpreting LPA classes is done through the means and variances the results each class would likely receive for each subtest. The assumption of LCA and LPA that indicators scores show high means to one class (homogeneity) would be easy to support. The assumption that the same indicator scores show different means between classes, however, is difficult to
support. In other words, is there a meaningful difference in classes when one subtest divides them by less than an item? In addition to the difficulty with meeting assumptions, the interpretation of classes can be difficult when average items correct are reported. What is the meaning of students placed in a class because they got 8.3 items correct?

Choi (2013, 2015) does not have the challenge of interpreting classes by the average amount of item correct. Instead, the interpretation of class is based on the average ability values for the four TOP indicators. Ability values in IRT can be interpreted similarly to items correct, a higher value means higher language ability and vice versa. The challenge that emerges from using IRT is connecting the ability values back to observable performance. The typical range for IRT theta values is from -3.00 to +3.00 (Hambilton & Swaminathan, 1984). An ability value of 0 represents the average performance of the dataset, but this value does not represent one value on the TOP’s four-point Likert scale. Choi interpreted the classes by creating summary performances for each class that he confirmed by checking a corpus. The corpus, however, consists of classroom lectures, not video performances of Choi’s dataset.

The LCA results from this study can be positioned against Choi’s to create a body of research on mixture modeling in language testing. LCA and LPA are able to identify mixtures of students within the test scores of small and large item tests. Using LCA, I can communicate class placement along with keeping the feedback tied directly to the AEOCPT.

2.3 Task: Striking a balance between Cognitive and Contextual loads

The final component of the three I identified from the AEOCPT is Task, which received the most revisions during the test’s developmental history. The tasks from the original SPEAK test were reduced from six to four. Administrators further revised the tasks to better replicate an
American university classroom context and enable the test to be administered as an interactionist dynamic test. I will review previous research on tasks and connect it to this dissertation’s testing of the moderate interactionalist framework.

While interaction-focused language tests attempt to assign a score based on an interaction applied knowledge state is in a target context, scores from task-based tests focus only on the success or failure of the communication (whether linguistic or not). In other words, was a test taker able to accomplish the given task regardless of the method chosen? Task-focused tests are possible to score because the variability comes from how an individual completes the task. Skehan (1996, 1998) argues that task-focused tests measure the entire cognitive construct of an individual, not just the language portion. This belief from test developers using this language knowledge framework led to the amount of detail within performance test task descriptions.

### 2.3.1 Defining performance test tasks

Performance tests are defined as tests that measure individuals through simulations of real world situations (AERA, APA, & NCME, 1999). Messick (1994) divides performance assessments into construct-centered or task-centered. In language testing, the terms means-focused and ends-focused have also been used (Long & Crookes, 1992; Robinson, 1996). Construct-centered, means-focused, performance tests measure the language construct through multiple tasks. An example of a construct-centered performance test given by Lane and Stone (2008) is an extended writing assignment. The assignment is divided into three tasks, the initial draft, a revised draft from peer reviews, and proofreading the final draft. All three drafts address the same topic and feedback is used for increased language efficiency within the one assignment.
Task-centered or ends-focused performance assessments measure whether an individual can accomplish a single goal. The variability for these kinds of tests comes from the task being accomplishable through a series of prescribed steps. An individual who masters all of the steps will have the highest chance of success while the probability decreases with every step still left to learn. The feedback that is given from task-centered tests is both catered to the individual’s performance and framed only within the task given. An example of this using Lane and Stone’s assignment is one writing assignment where students are graded on the extent that they write and revise the essay in one administration.

One limitation to using multiple tasks in a performance test, however, is the assumption that they all have equal weights to measuring the same construct (Messick, 1994; Norris, Brown, Hudson, & Yoshioka, 1998; Robinson, 1996). The problem is that more familiar tasks will have a higher chance for success and compensate for lower language ability (Skehan, 1996, 1998).

A situation with fewer resources in the environment to process could also impact the linguistic measurement of an individual. Cummins’s (1983) framework of Communicative Activities shows a possible way to strike a balance between the amount of resources available and the amount of time given to an individual for processing.
Figure 2-4 above shows that Cummins divides Communicative Activities into one of four possible categories. Each category is described by two properties, whether the activity is contextually complicated and whether it is cognitively demanding (p. 121). This framework was developed from Cummins’s (1980) previous work on Basic Interpersonal Communicative Skills (BICS) and Cognitive/Academic Language Proficiency (CALP). Context in this figure means the amount of common knowledge shared by all parties in the conversation or writing. Cognition refers to the amount of information processing required.

2.3.2 Using the Communicative Activities Framework to support the AEOCPT tasks

As I was working on the AEOCPT, one of the most pressing questions I came up with was why did test designers choose four tasks? Although there is no documentation to explain the tasks’ first revisions, evidence is still available now that could have been used at that time. Two textbooks still used by the ITA program have chapters that instruct students on how to best answer the AEOCPT’s tasks (Meyers & Holt, 2002; Smith, Meyers, & Burkhalter, 1992).
Cummins’s Communicative Activities Framework has also been around since the first revision, providing further evidence to support the number of tasks chosen by administrators.

Because Cummins discussed his Communicative Activities Framework by the four quadrants shown in Figure 2-4, I will start my description of the tasks at quadrant A. Test takers will successfully respond to a task from this quadrant by focusing more on the contextual resources presented to them. In the AEOCPT, the first task, mini-lecture, best fits in quadrant A because candidates are instructed to prepare the lecture before the administration date. ITA program administrators expect candidates to collaborate with members of their respective departments to present a lesson that best replicates their future teaching assignments. The cognitive demands for this task are remembering the key terms of the lecture and answering follow-up questions that stay within the topic of the lesson.

Moving on to quadrant B, test takers are able to successfully answer a task from this quadrant by being able to process contextual information in a relatively shorter amount of time than is available for the other tasks. The conditions of a task from quadrant B would appear to be the most difficult for test takers but most closely match the moderate interactionalistic language framework. The AEOCPT task that most closely fits the description of quadrant B is Task 2, office hour role-play. Chiang’s (2009, 2011) research supports that this context presents unique challenges to ITAs. In the AEOCPT, test takers are presented with the situation they will work with raters to resolve. Test takers are given 30 seconds to think of possible answers they can give to resolve the situation, but they must also attend to answers given to them by a rater who is serving as the student.

A task that fits in quadrant C requires routine cognitive processing by test takers with little contextual information available. The AEOCPT task that best fits in this quadrant is task 3,
giving your opinion. During this task, raters ask test takers a question related to teaching, their major, or their future career aspirations. I describe the response to this task as routine because candidates provide memories that raters cannot check for correctness. This task would at first appear to be only an extended response question, but it is made into interactionist through raters’ follow-up questions. Task 3 is given to simulate a classroom situation where students ask a question ITA candidates could not anticipate. Responding to the question relies solely on test takers’ memories but still must be clearly communicated to an audience.

The final quadrant of Cummins’s Communicative Activities framework, quadrant D, will require test takers to use few available contextual resources to construct a routine answer. The task that best fits this description from the AEOCPT is the final one, announcing changes to an event. This final task of the AEOCPT requires test takers to make a class announcement about changes that were fictitiously made in a previous class meeting. Test takers are given only necessary information to talk about on a card. They must use the information to create the announcement that highlights the old information that was changed to new information. I revised this last task to discuss changes to announcement in response to raters recommending that it be removed from the test. This recommendation conflicted with views from administrative stakeholders that all four tasks were important to accurate placement decisions.

Because each task of the AEOCPT corresponds to a quadrant of Cummins’s Communicative Activities Framework, unique information can be gained about ITA candidates’ language knowledge. In terms of the moderate interactional framework, each task provides information about how test takers use contextual information and with their own language knowledge to communicate with interlocutors. What I was not able to answer in this literature review is finding empirical evidence to support that tasks are able to contribute unique
information. Although faculty at Pennsylvania State University support the informative potential of the four tasks, ITA administrators sum the rubric score to inform placement decisions. A method has not yet been proposed to show the impact of tasks on rubric scores.

2.4 Summary and research questions

In this literature review chapter, I discussed previous research that supported AEOCPT revisions and guided the next stage of development. I labelled the major sections of the chapter to follow the three components of the test I identified in the first chapter. For the scoring rubric, I discussed communicative language frameworks that have been previously used in ITA performance testing (Hoekje & Williams, 1992). I then introduced Chalhoub-Deville’s score-in-language user-context framework (2003; Chalhoub-Deville & Deville, 2006) and how this could mediate faculty’s theoretical beliefs with the information needed for making placement decisions. This framework, however, does not contain information on the three components, but they are similar to the components I identified for the AEOCPT.

Following the history of communicative competence, I identified Bachman’s (1990; Bachman & Palmer, 1996) language knowledge framework as usable for constructing the scoring rubric, or language user. This framework is still being used in ITA test development (Gorsuch et al, 2013) but has been extensively discussed by researchers (Chalhoub-Deville, 1997, 2003; Chalhoub-Deville & Deville, 2006; Chapelle, 1998; McNamara, 1996). Harding (2014) summarizes the research done on the language knowledge framework and highlights a major weakness. Harding argues that both the CLA and language knowledge frameworks are too complex to operationalize. Phakiti (2008) is the only research I could find that attempted to test either one of the frameworks.
For the scoring component, I discussed that a model would need to be proposed that allows for placement decisions to be made and provide diagnostic feedback to stakeholders. This feedback would help teachers understand the strengths and weaknesses of students placed in their class. In addition, ITA candidates will be able to know what language knowledge variables they need to work on with every task. Finally, the opportunity for task analysis in this dissertation is determining if the tasks elicit unique candidate performances that cause raters to score the rubrics differently. Answering this research question would validate the four tasks and answer a limitation to all interactionist theories. Bachman (2007) and Chapelle (1998) discuss that no research has yet tested whether there is a significant interaction between tasks and individual knowledge.

In this dissertation, I will answer the following three research questions

RQ1: Does the AEOCPT data support the language knowledge framework being operationalized in the scoring rubric?

RQ2: Does the AEOCPT data support a four class latent model for making placement decisions?

RQ3: How do the four AEOCPT tasks contribute information to understanding placement, language knowledge strengths, and language knowledge weaknesses?

In the next chapter, I will explain the data analysis methods I used to answer the three research questions. I will also identify a validity model that I will use to explain the results and keep the discussion framed around the AEOCPT and its intended purpose of informing placement decisions.
Chapter 3 Methods

In this chapter, I will describe the methods I used to answer the three research questions in this dissertation. The first part of this chapter includes a description of participants who took the AEOCPT and whose response data were analyzed. In the second part, I will describe the AEOCPT’s administration procedures. I described the content within the AEOCPT in the previous chapter, but this description will have information on how the test is administered. Finally, I will describe the analysis procedures for answering the three research questions and propose a validity model for discussing the results as they relate to the AEOCPT’s placement decisions.

3.1 Participants: The AEOCPT Dataset

The dataset that I collected for answering this dissertation comprises AEOCPT results from the 2013 Summer semester to the 2014 Fall semester. I collected a total of 598 responses from five semesters of administrations at Pennsylvania State University. Demographic information for the 598 ITA candidates that took the AEOCPT are shown below in Table 3-1.
Table 3-1. Demographic information of AEOCPT test takers

<table>
<thead>
<tr>
<th>Language</th>
<th>Language Percentage</th>
<th>Language Number</th>
<th>Majors</th>
<th>Major Percentage</th>
<th>Major Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>49.0% (293)</td>
<td></td>
<td>Engineering</td>
<td>40.5% (242)</td>
<td></td>
</tr>
<tr>
<td>Korean</td>
<td>10.1% (60)</td>
<td></td>
<td>IST</td>
<td>7.8% (47)</td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>5.2% (31)</td>
<td></td>
<td>Economics</td>
<td>7.8% (47)</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>5.2% (31)</td>
<td></td>
<td>Life Sciences</td>
<td>5.2% (31)</td>
<td></td>
</tr>
<tr>
<td>Persian</td>
<td>4.6% (27)</td>
<td></td>
<td>Business</td>
<td>4.6% (27)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>25.9% (156)</td>
<td></td>
<td>Biology</td>
<td>4.6% (27)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Architecture</td>
<td>4.6% (27)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Other</td>
<td>24.9% (150)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total</td>
<td>100% (598)</td>
<td></td>
</tr>
</tbody>
</table>

I divided this demographic information into two categories: a) candidates reported first language and b) majors. For categories with fewer than 10 candidates, I combined these as “Other.” A total of 31 languages and 32 majors were identified by ITA candidates before I condensed them to the above labels.

No candidates were excluded from this dissertation’s dataset and analysis. One reason for this is that Pennsylvania State University’s guidelines require that every department in the university registers their respective candidates for the test. The only criterion for exclusion singles out candidates that come from an inner circle English speaking country (McKay, 1992). These countries are Australia, Canada, New Zealand, the United Kingdom, and the United States.
3.2 Procedures: AEOCPT Administration

After ITA candidates’ verify their identity with an administrator associated with the program, they wait in a room until two raters are available to administer the AEOCPT. Raters are randomly paired and assigned to administration rooms. One challenge with this procedure, however, is that selecting two raters from a larger group will affect reliability calculations. I will discuss how I addressed this reliability concern in the analysis section of this chapter.

Once the pairing of ITA candidate and raters are made by the scheduling administrator, raters lead the candidate to an available classroom. The test begins with a simple warm-up where raters ask questions about what the candidate is studying, how long they have been in town, or what their plans are for after they complete their program of study. The warm-up phase concludes with raters reading a script that informs candidates that the test will be recorded for a possible third rater to score. After the warm-up phase comes Task 1, Giving a Mini-lecture. All four tasks of the AEOCPT are given in the same order since the latest revisions and were developed from ITA course materials (Gorsuch, Meyers, Pickering, & Griffie, 2013; Meyers & Holt, 2002; Smith, Meyers, & Burhalter, 1992).

For Task 1, Giving a Mini-lecture, raters instruct ITA candidates to lecture on a term or process in their field of study. This term or process should be drawn from an undergraduate level course, expectedly first year courses. This requirement ensures that ITA candidates are delivering a lecture that closely matches their future teaching assignment. This requirement also ensures that raters are able to understand the lecture well enough to ask questions. Raters are allowed to ask questions any time during the lecture or after it is finished. ITA program administrators recommend that candidates should take between 7 to 10 minutes to complete this task.
During Task 2, the Office Hour Role Play, candidates are instructed that they will act out a situation with one of the raters. The situation is one from a list of five possible scenarios that simulate common issues presented by students to teaching assistants. An example of one situation is a student asking for an extension on an assignment because he or she is still getting used to college life. ITA candidates are given the scenario and 30 seconds to think of possible answers to give the rater. For the majority of situations, the rater begins the role-play by presenting the situation request and rationale for it. This role play presents a challenge to candidates that is unique because raters present a reason for the situation that is not scripted, and candidates must be able to quickly process the new information and incorporate it into their responses. ITA administrators recommend that this task last 3 to 4 minutes, but it ends when the rater and candidate have resolved the problem.

In Task 3, Stating an Opinion, raters select one question from a list of seven possible questions requiring that the candidate explain an opinion. Candidates are given 30 seconds to think of their responses and are instructed to deliver an answer about two minutes long. The questions within this task elicit information about candidates’ future goals, an inspirational teacher in their life, challenges they will face in the next few months, etc. Although it might seem that these questions do not closely match likely questions from the classroom, the purpose is to measure candidates’ abilities to construct a response with little time to prepare. Another way to think of this task is that it is more online while Task 1 would be described as more offline. Follow up questions are asked by raters to clarify any miscommunications during the candidate’s response.

Announcing Changes to an Event is the final task. The directions for this task are that students look at a card that contains information for making one of four possible classroom
announcements. Raters inform ITA candidates that an event was announced in a previous class, but that details from this event have been revised, and candidates must make a new announcement to the class. The new announcement must contain the previous information that was not revised, the previous information that was revised, and the new information that replaced the old. Candidates are given one minute to review the card and organize their announcement into a 1 to 2 minute response. The purpose of this task is to measure a candidate’s ability to recall a common genre, making an announcement, and apply new information that modifies it. Follow-up questions from raters clarify any information that was missed or not clearly communicated.
Table 3-2. Revised AEOCPT scoring rubric

<table>
<thead>
<tr>
<th></th>
<th>1 (115G)</th>
<th>2 (117G)</th>
<th>3 (118G)</th>
<th>4 (Pass)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammatical</strong></td>
<td>Generally incorrect and labored structures</td>
<td>Consistently incorrect constructions; some self-correction</td>
<td>Noticeable incorrect structures; successful self-correction if necessary</td>
<td>Infrequent incorrect structures; successful self-correction if needed</td>
</tr>
<tr>
<td><strong>Structures</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Thought</strong></td>
<td>Unintelligible thought groups and/or volume</td>
<td>Limited effectiveness of thought groups and/or inappropriate volume</td>
<td>Occasional difficulties with thought groups, or appropriate volume</td>
<td>Uses thought groups adequately with appropriate volume</td>
</tr>
<tr>
<td><strong>Groups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tone Choice</strong></td>
<td>Monotone</td>
<td>Mostly falling, rising, or level tones</td>
<td>Uses rising and falling tones but occasionally misleading</td>
<td>Uses rising, falling, and level tones in appropriate manner</td>
</tr>
<tr>
<td><strong>Multimodality</strong></td>
<td>Minimal eye contact, consistent body orientation away from audience; poor use of nonverbal resources</td>
<td>Frequent retreat to body position away from audience and insufficient use of nonverbal resources</td>
<td>Infrequent periods of orientation away from audience supplemented by effective use of nonverbal resources</td>
<td>Negligible orientation away from audience and effective use of nonverbal resources</td>
</tr>
<tr>
<td><strong>Transition</strong></td>
<td>Minimal transitions used</td>
<td>Task answer organized with restricted use of transition</td>
<td>Task answer organized but varied transition are lacking</td>
<td>Task answer organized with appropriate and varied transition</td>
</tr>
<tr>
<td><strong>Prominence</strong></td>
<td>No key words identified through stress or non-verbal resources</td>
<td>Minimally effective identification of key words through stress and/or non-verbal resources</td>
<td>Consistent identification of key words through stress and/or non-verbal resources</td>
<td>Effective identification of key words via stress throughout task</td>
</tr>
<tr>
<td><strong>Task Response</strong></td>
<td>Initial answer does not address the task</td>
<td>Task response pragmatically inappropriate and/or several gaps of information</td>
<td>Completes task with pragmatic appropriateness; any gap of information requires minimal repair</td>
<td>Completes task in pragmatically appropriate manner with no repair needed</td>
</tr>
<tr>
<td><strong>Question Response</strong></td>
<td>Answer does not address raters' concerns; other-initiated repair unsuccessful</td>
<td>Partial answer even after two turns of negotiation</td>
<td>Answer negotiation across two or more repair sequences; successful completion</td>
<td>Resolved with minimal repair</td>
</tr>
</tbody>
</table>
3.3 Materials: Scoring the four AEOCPT tasks

While candidates give their responses to each task and answer follow up questions, raters score each task using the AEOCPT scoring rubric. The rubric is made up of eight items, each containing four points of performance ratings and descriptions. More detailed information about the descriptions in the AEOCPT rubric can be found in Table 3-2. Items 1 and 2 operationalize Grammatical Knowledge. *Grammatical structure* is not limited to syntax but includes all knowledge associated with creating comprehensible utterance. The focus for raters with this item is on pronunciation, but it is also used for syntax and vocabulary. Item 2, *Thought Groups*, is not explicitly described in Bachman and Palmer (2010) but follows a line of research that associates grammatical knowledge and fluency (Gorsuch, Meyers, Pickering, & Griffee, 2013; Lin & Francis, 2014). A noticeable and unnatural pause between utterances can be indicative of limitations in Grammatical Knowledge.

*Item 3, Tone Choice*, and *Item 4, Multimodality*, are operationalize Sociolinguistic Knowledge. Bachman and Palmer (1996) provide several descriptions within this knowledge: genres, dialects/varieties, register, natural/idiomatic expressions, and cultural references/figures of speech. *Tone Choice* is not mentioned in Bachman and Palmer’s language knowledge framework, but it is given as a teaching point in ITA instructional materials (Gorsuch et al., 2013; Meyers & Holt, 2002; Smith, Meyers, & Burkhalter, 1992). Tone choice relates to Sociolinguistic Knowledge because in the classroom environment, lecturing requires a variety of tones in order to communicate information correctly. Also, *Multimodality* is not mentioned in Bachman and Palmer, but researchers view it as important forms of expressions and cultural references for highlighting information. The rubric focuses on eye contact as the most important
form of multimodality, aligning with ITA classroom material (Gorsuch et al., 2013; Meyers & Holt, 2002; Smith et al., 1992).

*Item 5, Transitions,* is part of Textual Knowledge and is a paraphrased operationalization of Bachman and Palmer’s description of Cohesion (p.45). Raters are instructed to look for key words that link details together in their response. *Prominence, Item 6,* is considered one of the most important skills in the classroom environment but is hard to develop outside this setting (Gorsuch et al., 2013). Prominence is not given in Bachman and Palmer’s description but referenced in their brief description of conversational organization (Brown, Anderson, Shillcock, & Yule, 1984; Luoma, 2004). Item 6 measures the extent to which candidates are able to highlight information they perceive as important in their response. Raters are trained to give full credit for this option if candidates are able to make words salient through voice or gestures combine with visual cues, e.g. pointing to a key word.

Lastly, Items 7 and 8 are the operationalizations of Functional Knowledge, one of the hardest to define because it is the interpretation of relationships between utterances and understanding the intent of language users (Bachman, 1990; Bachman & Palmer, 2010) Because the AEOCPT is a construct-centered performance test, Item 7 elicits the extent to which ITA candidates are able to utilize their language knowledge across multiple situations encountered in a classroom context. Using Bachman and Palmer’s descriptions of Functional Knowledge, *Task Response* requires imaginative functions, performances based on their prior observations and experiences, and heuristic functions, using information provided by the test along with their knowledge to construct an answer. Item 8 is the only one given by Gorsuch et al. (2013) for Functional Knowledge, and it is taught by all instructional material previously referenced. *Question Response* requires ITA candidates to use ideational functions to provide an answer that
correctly conveys their intended meaning and manipulative functions to better understand the question asked or clarify the response given.

In sum, the AEOCPT scoring rubric operationalizes two items for each factor of the four factor language knowledge model. *Grammatical Structures* (Item 1) and *Thought Groups* (Item 2) are the items that measure *Grammatical Knowledge*. *Items 3 (Tone Choice) and 4 (Multimodality)* measure *Sociolinguistic Knowledge*. *Textual Knowledge* is measured by the next two items in the rubric, *Transition* (Item 5) and *Prominence* (Item 6). Finally, the last two items are *Task Response* and *Question Response*, and they measure *Functional Knowledge*.

3.4 Data Analysis: Analyzing the AEOCPT dataset

In order to answer the first two research questions of this dissertation, I propose two latent trait statistical models. For determining whether the language knowledge framework is operationalized in the scoring rubric, I test the fit between the AEOCPT dataset and a Confirmatory Factor Analysis (CFA) model. The dataset to be analyzed for answering this question is a covariance matrix calculated from observed dataset. For answering whether a four class latent class is present in the AEOCPT results, I identify the best fitting Latent Class Analysis (LCA) model. This identification is done by comparing multiple LCA models that are measured for fit using response patterns in the AEOCPT. I discuss how I chose the models for both questions in the following sections in addition to describing how answering the third research question is a variation of the two latent trait models.
3.4.1 Identifying and fitting language knowledge CFA models

The first step to answering RQ1 is screening the data and determining whether it meets the multivariate normality assumption of CFA. Although language knowledge is considered continuous, the AEOCPT scoring rubric in Table 3-2 is categorical. Joreskog and Sorbom (2013) define a categorical variable for their CFA program, Lisrel 9.10, as a variable having seven or fewer points, the AEOCPT scoring rubric has four. Screening data that uses categorical variables requires that an underlying normal distribution must be tested. This is done by requesting a Polychoric correlation matrix and looking at chi-square tests for all possible item pairings.

Figure 3-1. Four factor language knowledge CFA model

Readers will notice that the CFA model I am proposing above differs from the language knowledge framework shown in Figure 2-3. The reason for this difference is so the model passes
the second step conducting CFA, identification. Identification is the determination of whether a statistical program is able to analyze a proposed CFA model (Kline, 2011). For the language knowledge framework in Figure 2-3 (a hierarchical model), Kline states that at least two lower order latent variables must load onto the higher order variable (p.249). Readers can see from Figure 2-3 that only one second order general language latent variable can be identified from the four first order variables. This omission of Organizational and Pragmatic knowledge does not suggest that they do not exist. Rather, the model is not theoretically likely to converge without revisions.

In addition to the hierarchical rule, the language knowledge model must meet an additional rule for recursive CFA models. The identification rules that will be used for the model above are the minimum degrees of freedom and the two-indicator rule (Kline, 2011). The minimum degrees of freedom is met when the following equation is true:

\[ t \leq q(q + 1)/2 \]

In the equation above, \( t \) represent the number of free parameters, and \( q \) represents the number of indicators (observable variables). The two-indicator rule applies if the CFA model being identified has more than two factors and at least two indicators load onto each factor. The language knowledge model in Figure 3-1 above meets the minimum degrees of freedom rules and has four factors that each have two indicators loading onto them. Thus, this CFA model is identified and can produce interpretable results when submitted for analysis.

Studies that use CFA, however, do not commonly stop at identifying and measuring the fit of one model. For the language knowledge framework, there are two other models that can be
proposed for analysis and comparison to the one in Figure 3-1. One of the additional models would operationalize older language knowledge research.

**Figure 3-2. Two factor language knowledge CFA model**

By submitting the above model to analysis, I am proposing that AEOCPT rubric items have a relationship that is different from the four factor model. **Figure 3-2** suggests that the descriptions given by Bachman and Palmer (1996, 2010) better fit Organizational and Pragmatic knowledge with AEOCPT data. This proposed model would fit the communicative competence framework researched by Hymes (1972) and Campbell and Wales (1970). Because the same dataset will be used to analyze this model, data screening does not have to be done again. Identification, however, does need to be re-assessed since the amount of factors have changed. This two factor model is also recursive but is able to be identified by both the two and three indicator rules. Kline (2011) applies the three indicator rule if a model contains at least three indicators. A single
factor model is able to meet the three indicator rule although the model above contains two factors.

![Diagram of One factor language knowledge CFA model]

**Figure 3-3. One factor language knowledge CFA model**

The second model that I will propose for comparing fit comes from discussions about whether language knowledge is a unidimensional or multidimensional variable. Most of the research that contributed to these discussions occurred during the late 1980s and early 1990s. An edited book by Oller (1983) is a seminal piece of work that reflects the contributions of several authors. Bachman and Palmer’s (1983) contribution to this book is the source of the language knowledge model used in this dissertation. Although the conclusion of Oller’s book supported that language is multidimensional, the discussion came up again when researchers began using Item Response Theory to score test takers (McNamara & Knoch, 2012). Item Response Theory
(IRT) currently has a strong unidimensional assumption for analyzing test results (Hambleton & Swaminathan, 1984).

In addition, the data analysis method used by researchers in Oller’s (1983) book was Principal Component Analysis (PCA). Although researchers believed that PCA could be used for Factor Analysis (Exploratory or Confirmatory), current research identifies it as a data reduction method (Kline, 2011). In order to submit the model in Figure 3-3 to CFA, it must be identified by similar rules to the previous two since it is also recursive. The one factor model meets the three indicator rule described earlier for identifying the two factor model. After identifying all three models, they were submitted to Lisrel 9.1 for estimation.

*Estimation* is the algorithm applied to dataset with the goal of reproducing an observed covariance or correlation matrix. Model goodness-of-fit is measured by the extent to which the observed and expected matrices differ. The most common estimator for CFA is Maximum Likelihood Estimation (Kline, 2011). This estimation, however, works best with continuous indicators. Because the AEOCPT dataset has categorical indicators, an alternative estimation procedure must be used. Diagonally Weighted Least Squares is an estimation method recommended by several researchers for models with categorical indicators (Du Toit & Du Toit, 2001; Kline, 2011). Although I am using an estimator that is an alternative to Maximum Likelihood Estimation, the results for the models will have the same fit statistics as those analyzed by Maximum Likelihood Estimation.

For measuring fit, I will look at both global and component fit statistics provided by Lisrel for all estimators. (I will give more details about these statistics along with their standards in the next chapter). In order to determine which of the three CFA models best fits the data, comparative fit statistics will be interpreted. Because the latent variables are changed instead of
the factor loadings, the three models are nested. The fit statistics that I will use to compare the three models are AIC and BIC, discussed more in the next section.

### 3.4.2 Exploring and confirming ITA program LCA models

When using test results to make placement decisions, administrative stakeholders attempt to organize students based on the strengths and weaknesses that are measured in the placement test. Placement categories have been determined by ITA programs through standard setting procedures. Cizek’s (2012) provides several methods for calculating standards using both Classical Test Theory and IRT. One challenge with these methods, however, is that they require individuals (e.g., raters, instructors, administrators, etc.) to serve as Subject Matter Experts. These experts organize individuals based on their test scores or latent ability into the placement classes. Because the ITA program at Pennsylvania State University trains ESL instructors as part of the Applied Linguistics program, possible Subject Matter Experts are constantly changing after every semester. My attempt to resolve this problem with standard setting is to use a data driven approach, a latent trait statistic that identifies latent classes.

Latent Class Analysis (LCA) is the method I will use to answer RQ2 and is defined by Collins and Lanza (2010) as a person-oriented approach to modeling a latent variable and its indicators. LCA is also known as finite mixture modeling and has other variations such a Latent Profile Analysis and Latent Transition Analysis (Muthen, 2001). One feature of LCA that distinguishes it from the other analyses is that the indicators are categorical and the latent variable is assumed to be categorical as well. At first, this assumption would appear to conflict with the CFA in RQ1. Collins and Lanza resolve this conflict by distinguishing variable-oriented approaches from person-oriented.
The purpose of conducting CFA in RQ1 is to measure the relationships between rubric items and to test whether they operationalize the language knowledge latent model. LCA, on the other hand, is a person-oriented approach, focusing the analysis on the response patterns within a population of interest.

**Figure 3-4. Conceptual LCA model with eight indicators**

Figure 3-4 above conceptualizes the LCA model as indicators (the rubric items) that load onto the same latent variable. The latent variable is shown as singular, but LCA tests whether the variable is divisible into multiple classes. The formula for calculating a Latent Class Analysis is the following (Collins & Lanza, 2010):

$$P(Y_i = y_i | X_i = x_i) = \sum_{c=1}^{C} Y_c(x_i) \prod_{l=1}^{J} \prod_{r_1=1}^{R_j} \rho_{j,r_j}^{I(y_j=r_j)}$$
In this formula $\gamma_c$ represents the probability of belonging to each latent classe proposed in the model, $J$ represents the indicators, $R$ represents the response options, and $\rho_{j,r_j|c}$ represents the probability of getting response $r_j$ for indicator $j$ conditional on latent class $c$. One interpretation of this formula is that the probability of latent class membership is the product of response probabilities to every item conditional on the proposed latent class membership. In other words, identification of a class depends on how participants were scored by raters on every item in the test. One advantage to using LCA is that posterior probabilities can be produced for every test taker. These probabilities would allow ITA administrators to make placement decisions and create feedback forms for teachers and students based on item level performance. Before this operation can be carried out, several steps must be completed to discover which LCA model best fits the AEOCPT dataset.

The first step in conducting a Latent Class Analysis is identifying an initial model using an estimation method, similar to those in CFA. The program that I use to conduct the Latent Class Analysis is PROC LCA (Lanza, Dziak, Huang, Wagner, & Collins, 2014), which is run through SAS. This program uses Maximum Likelihood Estimation as the estimator for measuring model fit. This estimation is appropriate for LCA with the same AEOCPT dataset because PROC LCA is estimating the fit of response probabilities (normally distributed) to latent class prevalence (Collins & Lanza, 2010). The results produced by PROC LCA include statistics for measuring identification and statistics for measuring fit. Most of these fit statistics are comparative, but a likelihood-ratio statistic ($G^2$) is calculated by PROC LCA. This statistic is similar to $X^2$ but measures the exact fit of contingency table methods (Dayton, 1998). Collins and Lanza (2010) note for $G^2$ statistics that sparseness becomes an issue for complex LCA
models. This means that a test of absolute fit can be difficult to interpret when the sample size is not large enough to endorse every option on all indicators.

Lanza et al. (2014) propose an alternative for testing LCA identification by running multiple Maximum Likelihood Estimation calculations from different starting points in the log-likelihood distribution. This test is done through PROC LCA by writing the NSTARTS statement in the program’s syntax. PROC LCA requires a SEED value to start the MLE process and produce results. If the optimal SEED value is not known (i.e., the model has not previously been identified), NSTARTS produces random seeds that each run MLE. The results from PROC LCA that include the NSTARTS statement include a line that gives the percentage of random seeds that successfully converged to the same maximum likelihood. For the question of model identification, more than 50% of seeds finding the same maximum likelihood support the model as identified. Because none of the models in this dissertation have been tested, I requested 100 random seeds to be created from the NSTARTS statement.

After I have identified a model using PROC LCA, the next step is to compare the identified model with alternative ones. This step is similar to the one taken for answering RQ1 using CFA. For this dissertation, the initial LCA model I tested for identification is a four class model. This model replicates the current placement decisions that are in the ITA program. These placement decisions are given in Table 5 including the cut-scores that are currently applied to making each decision.

Table 3-3. Current ITA placement decisions with brief course descriptions

<table>
<thead>
<tr>
<th>Score Band</th>
<th>Decision</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>250-300</td>
<td>Exempt</td>
<td>Certified to Teach</td>
</tr>
<tr>
<td>200-249</td>
<td>118G</td>
<td>Advanced English ability, Focus on classroom communication</td>
</tr>
<tr>
<td>150-199</td>
<td>117G</td>
<td>Intermediate English ability, Focus</td>
</tr>
</tbody>
</table>
Included in the placement decisions in **Table 3-3** is a general description of what kind of candidates are placed in each course and what classroom instruction focuses on for English development in an American university classroom. I will compare a four class LCA model against three and five class models. Because these models are non-nested, I will use the Akaike Information Criterion (AIC) and the Schwarz Bayesian Information Criterion (BIC) for fit comparison. I will look at both of these fit statistics in order to account for any variations that may exist between their calculations (Dayton, 1998). I will also use the Corrected AIC (CAIC) and adjusted BIC, which make changes based on a finite sample size. For these statistics, I will interpret a lower value as evidence for a model that better fits the AEOCPT dataset.

Finally, I will use an Entropy statistic as the final piece of evidence for comparing the LCA models. The Entropy statistic has been typically used with cluster analysis to measure the extent that the clusters are well separated. The Entropy statistic’s range is from 0 to 1.00 with a higher number interpreted as a model that has better cluster separation. Celeux and Soromeaho’s (1996) study found that the Entropy statistic is able to identify the correct mixture model as do the other relative fit statistics I discussed earlier. After I select a LCA model as the best fitting, there is one more step in the process.

The final step in LCA is to interpret the classes that are in the chosen model, which is done by using evidence from rubric item descriptions and the points selected by raters. In current practice, placement decisions are made on a cumulative score that can be met through multiple
ways. In LCA, however, each class is made up of a single response pattern that informs administrators about the kinds of students that are in each class. Because it is possible for a candidate to have a response pattern that does not exactly match a class’s response pattern, the posterior probability results give the possible membership probabilities for all classes of each candidate. Interpreting the rubric item descriptions in order to understand each class requires that the items have been developed from a well operationalized theory. This requirement provides an additional need for answering RQ1. Lanza, Savage, and Birch (2010) used factor analysis before LCA as well in order to better interpret their identified classes.

3.4.3 Testing the impact of task on CFA and LCA models

To this point in my discussion of procedures for answering the first two research questions, I have focused on the rubric items and placement decisions. The last component of the three I mentioned in earlier chapters is tasks. My reason for not addressing this yet is that testing the impact of tasks does not require the addition of another latent trait model. Instead, a statement has to be added to the syntax of both CFA and LCA models to include the four tasks. This statement is called a grouping variable and was originally intended to test whether distinct groups of people responded to items differently. For the AEOCPT, the hypothesis that I am testing is whether raters select different categories for the same candidate on different tasks.

For Confirmatory Factor Analysis, I answer the impact of tasks by testing the measurement invariance of the accepted model (Du Toit & Du Toit, 2001). The first step in testing measurement invariance is determining whether there is acceptable fit for each task. This test is done using the procedures described above for RQ1 and interpreting the global and component fit statistics. Once the model fit for each task is verified as acceptable, the next step is
to test configural invariance. Configural invariance tests whether the language knowledge model has acceptable fit when all four tasks are analyzed at once. Global and component fit statistics are used here as well to determine fit.

If configural invariance is supported through acceptable model fit, the next step to test weak invariance, fixing factor loadings as the same across groups. Because the weak invariant and configural invariant models are nested, I will use a chi-square difference test to identify the better fitting model. If the weak invariant model is the better fitting one, the final test of measurement invariance is strong invariance. I will test this invariance by fixing factor loadings and means across tasks. I would also use a chi-square difference test here to determine which nested model is a better fit to the data. In order to support that the AEOCPT tasks are not contributing unique information to placement decisions, either weak or strong measurement invariance would need to be found.

In order to test the impact of tasks on the final LCA model, I will include a GROUPS statement in the LCA PROC syntax, a simpler procedure than for CFA. The first step to adding a grouping variable is identifying the best fitting model without the GROUPS statement added. This means going through the steps earlier that compare the relative fit of numerous class models. After I identify the best fitting LCA model, I conduct additional analyses with the GROUPS statement added. Because the difference between the two models is freeing some parameters, the models are considered nested. A chi-square difference test can be conducted in this case to determine whether identifying tasks contributes to a better fitting LCA model.

By adding a grouping variable to the analysis, the model becomes more complex and more likely to be unidentifiable. Hence, Lanza et al. (2014) recommend the inclusion of a RHO
PRIOR statement. Including this statement will add several pseudo-cases to the dataset, biasing solutions away from all 0s or 1s (p.20). In other words, the model could fail to identify due to several candidates receiving all 4s or all 1s on a task that is now separated from the rest. In addition to the model being more likely to be identified, including a RHO PRIOR statement produces standard errors that can be used to construct confidence intervals (Dayton, 1998).

3.5 Notes on Validity: Keeping this study framed within the AEOCPT’s stated purposes

While results from this dissertation contribute to language testing research that seeks to better operationalize language knowledge, my primary motivation is about the AEOCPT. First and foremost, the results from this dissertation are supposed to determine if the AEOCPT adheres to administrator’s claims. In other words, the primary aim of this dissertation is to validate the AEOCPT.

Validity is defined as a body of evidence that supports inferences or claims that are made about the test. For the AEOCPT, administrators’ claims are that the test measures English knowledge in the context of an American university classroom. The scores from this test are intended for making placement decisions. An additional claim that ITA administrators would like to make about the AEOCPT is that it provides feedback information to all levels of stakeholders. Determining the best way to collect and discuss validity evidence is an ongoing discussion in language testing and educational psychology research (Chapelle, Enright, & Jamieson, 2008; Cronbach & Meehl, 1955; Kane, 2006; Moss, Girard, & Haniford, 2006). Weir (2005) proposes one validity model in language testing that he calls the Socio-Cognitive framework. This model consists of three variables, context, cognition, and scoring, which are similar in description to the three components I identified in the AEOCPT. One
difference between Weir’s Socio-Cognitive model and the validity model I am looking to use is that Weir does not use latent trait statistics. Two publications that use his model produce either qualitative arguments or correlations to validate a large commercial test published by the University of Cambridge (Khalifa & Weir, 2009; Taylor, 2011). One problem with this validity evidence for my dissertation is that the three components are difficult to connect. The evidence produced from the Socio-Cognitive model also supports a different relationship between the latent and observed variables.

Because I am answering all three of my research questions by using latent trait statistics, my discussion of test validity will focus on the interpretation of the CFA and LCA models. Borsboom (2006) and Borsboom, Mellenbergh and van Heerden (2003, 2004) support the use of latent trait statistics in validity studies. They argue that the results from these studies are a better representation of the relationship between tests and test takers. Chapelle, Enright, and Jamieson (2008), on the other hand, use correlations as evidence to construct their validity arguments. Borsboom et al. (2003, 2004) argue that the use of correlations is a constructivist perspective, suggesting that language knowledge is constructed from the items given. This perspective would make comparative validity studies difficult to carry out.

In addition to the use of latent trait statistics to support a realist view of validity, Mislevy, Sternberg, and Almond (2002) support the use of these statistics in language testing validation. Their model, called Evidence Centered Design (ECD), was initially proposed to support task-focused language knowledge. Mislevy et al.’s argument is that students’ abilities cannot be measured without having them complete tasks that are close replications of the target context. Their way of combing student ability and task is through evidence models, closely matching factor analysis models but identified by Miselvy et al. as CTT, factor analysis, IRT, or LCA. If
either student ability or tasks is not well established, the evidence model will not converge and must be evaluated.

Based on the descriptions provided by Mislevy et al., I am using ECD to produce evidence that validates the AEOCPT. The complexity that I am adding to this model is that I am using a performance test that consists of four evidence models. The first is an individual student knowledge that is combined with rubric items. The second evidence model combines ITA placement categories and the rubric items. The third and fourth evidence models combine the same student knowledge and ITA categories as student models but add the complexity of four AEOCPT tasks. This evidence would support whether I have successfully captured candidates’ language knowledge strengths and weaknesses within the context of an American university classroom. This context exists outside of the AEOCPT, and this test acts as an attempted replication, adopting a realist perspective.

3.6 Summary

In this chapter, I provide demographic information on the ITA candidates that make up the dataset used in my analysis. I then described the administrative procedures for giving the AEOCPT, and I described the process for how I answered this dissertation’s three research questions. For RQ1 (Confirmatory Factor Analysis), I stated that the dataset was screened for meeting an underlying normality assumption. I then cited two and three indicator rules for recursive CFA models as support that the language knowledge model is identified. Finally, I identified an estimation procedure for extracting global and component fit statistics when using categorical indicators.
For RQ2 (Placement Categories), I am using LCA, and I described the first analysis step as identifying an initial model (Four class model) using the NSTARTS statement with 100 iterations. After identifying an initial model, the next step is to compare the initial model with alternative class models. I will then compare relative fit statistics to decide on the best fitting model. For answering RQ3, I described two sets of procedures for use with CFA and LCA models. I will conduct a measurement invariance test on the CFA model with tasks as the grouping variable. For the LCA model, I will label the four tasks as groups using the GROUP statement and produce fit statistics I can compare to the previously chosen model. In the next chapter, I will provide my results organized by research questions.
Chapter 4 CFA Results

For the results section of this dissertation, I divide the information gained from CFA and LCA into their own chapters. The content of this chapter will primarily be information collected for answering RQ1 (Does the AEOCPT data support the language knowledge framework being operationalized in the scoring rubric)? In addition, I partially answer RQ3 (How do the four AEOCPT tasks contribute information to understanding placement)? The organization of this chapter is the results from procedures I outlined in the previous chapters and discussed by Kline (2011). I show the results from data screening, overall model fit, component model fit, and comparative model fit. I discuss the interpretations of these statistics as they relate to language testing research in a future chapter, but I provide some information on why the statistics were included. Finally, I answer part of RQ3 by going through DuToit and Dutoit’s (2001) procedures for testing the impact of tasks (measurement invariance).

4.1 Data Screening Results

Before analyzing the AEOCPT data using CFA, I screened the data to make sure it met the normality assumption (Kline, 2011). Because the observed variables within the AEOCPT dataset are ordinal, I screened the data by requesting a Polychoric matrix from Lisrel 9.1. Table 4-1 below shows the means and standard deviations for all eight rubric items across the four tasks.
Table 4-1. Descriptive statistics, Mean (SD), for AEOCPT data screening.

<table>
<thead>
<tr>
<th></th>
<th>Mini-Lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>2.12 (0.65)</td>
<td>2.19 (0.64)</td>
<td>2.15 (0.67)</td>
<td>2.17 (0.68)</td>
</tr>
<tr>
<td>Thought</td>
<td>2.06 (0.66)</td>
<td>2.21 (0.65)</td>
<td>2.18 (0.67)</td>
<td>2.14 (0.67)</td>
</tr>
<tr>
<td>Tone</td>
<td>2.08 (0.64)</td>
<td>2.19 (0.63)</td>
<td>2.19 (0.63)</td>
<td>2.18 (0.66)</td>
</tr>
<tr>
<td>Modality</td>
<td>2.38 (0.66)</td>
<td>2.57 (0.58)</td>
<td>2.50 (0.60)</td>
<td>2.26 (0.70)</td>
</tr>
<tr>
<td>Transition</td>
<td>2.23 (0.69)</td>
<td>2.35 (0.66)</td>
<td>2.33 (0.66)</td>
<td>2.36 (0.66)</td>
</tr>
<tr>
<td>Prominence</td>
<td>2.16 (0.70)</td>
<td>2.21 (0.70)</td>
<td>2.29 (0.66)</td>
<td>2.30 (0.70)</td>
</tr>
<tr>
<td>Response</td>
<td>2.41 (0.63)</td>
<td>2.35 (0.67)</td>
<td>2.53 (0.62)</td>
<td>2.45 (0.64)</td>
</tr>
<tr>
<td>Question</td>
<td>2.46 (0.65)</td>
<td>2.53 (0.62)</td>
<td>2.61 (0.59)</td>
<td>2.54 (0.62)</td>
</tr>
</tbody>
</table>

The descriptive statistics above show that raters gave most candidates between two and three points on the four point Likert scale. The data would appear to be normally distributed, but Lisrel informs users that more than 15 points is needed to classify data as continuous (Joreskog & Sorbom, 2013). Thus the AEOCPT data is categorical, but I assume the language knowledge being measured has an underlying continuous latent variable. For ordinal variables with an underlying normal distribution (language knowledge in this case) a polychoric matrix is the most appropriate (Appendix C). Using a Pearson correlation, Lisrel produces statistics to test multivariate and univariate normality. For the polychoric matrix, Lisrel tests the normality for the underlying response variables by comparing every pair of items. A significant chi-square between items means that they are not likely normally distributed. Figure 4-1 below is an example of two items that meet the underlying normality assumption. The frequency distributions for the four Likert points, the boxes, are about the same across the two items. These frequency distributions are not normally distributed but a nonsignificant chi-square test suggests that the assumption of underlying normal distribution is not significantly violated.
Figure 4-1. Example frequency charts tested for an underlying distribution

The chi-square test results show that all the rubric items across tasks support an underlying normal distribution. With this assumption met, the entire dataset can be tested for model fit with the three models proposed in the previous chapter. The polychoric matrix given in Appendix C is extensive but can be used by readers of this dissertation to replicate the following results. The next step to answering RQ1 is determining which of the three proposed models best fits the dataset.

4.2 Configural Invariance: Identifying an initial model

DuToit and DuToit (2001) state the first step to determining measurement invariance is confirming that each of the four tasks individually fit the desired CFA model. One reason for doing this is that other variables could compensate for the poor fit of one grouping variable. Because I have two layers of complexity in this dissertation, comparing three language knowledge frameworks and the four AEOCPT tasks, I will describe how I determine configural invariance. First, I provide the fit statistics for the four factor language knowledge model (Bachman & Palmer, 1996, 2010) analyzed for each task individually. I have chosen this model
as my starting point because this is the framework that I used for making the most recent revisions to the AEOCPT scoring rubric. If the four factor model shows acceptable fit for both global and component fit statistics, I will then compare the fit with the two alternative models described earlier.

4.2.1 Global Fit Indices

Global fit indices, or Approximate Fit Indexes (Kline, 2011), are measurements of how well the observed covariance matrix matches the one expected from the specified CFA model. Brown (2006) describes three categories of global fit indices that I report in this dissertation, absolute fit, parsimony correction, and comparative fit. *Absolute fit* is defined by Brown as a strict measurement of the difference between observed and expected covariances (p.82). The statistics that I will use to measure absolute fit is the $X^2$ test and the Standardized Root Mean Square Residual (SRMR). $X^2$ is commonly reported with a p-value that I will interpret against an alpha level of .05. In addition to this statistic, Kline (2011) recommends reporting the Satorra-Bentler Scaled $X^2$ when using ordinal observed variables. For interpreting SRMR, I will use Hu and Bentler’s (1999) recommendation of a value less than .08.

The next category of global fit indices, *parsimony correction*, is defined by Brown (2006) as statistics that include a penalty for the amount of parameters freely estimated. The statistic that I report for this index is the Root Mean Error of Approximation (RMSEA). Because a 90% confidence interval with this statistic is included in the Lisrel output, I include it in my results as well. The interpretation of this statistic derives from a three category system created from the results of two studies. Hu and Bentler (1999) recommend a RMSEA value below .06 for good model fit, and Browne and Cudek (1993) recommend a value below .10 for acceptable model fit.
The last category of global fit indices are those of *comparative fit*, evaluating the fit of a model compared to a more restricted, or baseline model (Brown, 2006). The statistics that I will use for measuring this global fit index are the Comparative Fit Index (CFI) and the Non-normed Fit Index (NNFI). The NNFI statistic is used for the same reason as the Satorra-Bentler Scaled $X^2$ statistic. For the comparative fit statistics, a higher value is indicative of a better fitting model. This is because the statistic shows how much the target model improves fit to the dataset compared to the baseline model. Hu and Bentler (1999) recommend CFI and NNFI values above .95 as supporting good model fit.

**Table 4-2. Global Fit statistics for four-factor hierarchical model across four tasks**

<table>
<thead>
<tr>
<th></th>
<th>Mini-Lecture (df = 16)</th>
<th>Role-Play (df = 16)</th>
<th>Opinion (df = 16)</th>
<th>Announcement (df = 16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S\cdot B$ $X^2$ (p)</td>
<td>53.07 (p&lt;.01)</td>
<td>43.38 (p&lt;.01)</td>
<td>61.08 (p&lt;.01)</td>
<td>39.42 (p&lt;.01)</td>
</tr>
<tr>
<td>SRMR</td>
<td>.023</td>
<td>.020</td>
<td>.026</td>
<td>.015</td>
</tr>
<tr>
<td>RMSEA (90% CI)</td>
<td>.108 (.091 ; .126)</td>
<td>.070 (.052 ; .089)</td>
<td>.079 (.062 ; .098)</td>
<td>.049 (.030 ; .069)</td>
</tr>
<tr>
<td>CFI</td>
<td>.97</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>NNFI</td>
<td>.95</td>
<td>.99</td>
<td>.99</td>
<td>.99</td>
</tr>
</tbody>
</table>

**Table 4-2** above shows that the four-factor model has an acceptable fit to the AEOCPT dataset for all four tasks. The Satorra-Bentler $X^2$ values across the four tasks are significant, suggesting poor fit between the observed and expected covariance matrices. The statistics for parsimony correction and comparative fit, however, are within the acceptable limits that I described earlier. In fact, the 90% RMSEA confidence interval supports acceptable fit across the four tasks. Although these results support using the four-factor model to determine measurement invariance next, I still must show that this is the best solution compared to alternative models.
4.2.2 Comparing nested alternative models

When comparing alternative models, there are two categories of statistics that can be used depending on whether the models are nested or non-nested. Kline (2011) defines nested models as one being a subset of the other (e.g., fixing a parameter that was freed in another model). Although the alternative language knowledge models appear to have different numbers of latent variables, these can be created through fixing or freeing covariances between them. This fact makes the models nested. The statistic that I report is the Satorra-Bentler $\chi^2$ difference test (Satorra & Bentler, 2001) calculated using a program created by Crawford and Henry (2003).

Table 4-3. $\chi^2$ difference tests for nested alternative models

<table>
<thead>
<tr>
<th>Task</th>
<th>One-Four Comparison (df = 6)</th>
<th>Two-Four Comparison (df = 5)</th>
<th>Four-Four Hierarchical (df= 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini-lecture</td>
<td>63.988 (p&lt;.001)</td>
<td>64.751 (p&lt;.001)</td>
<td>6.406 (p=.041)</td>
</tr>
<tr>
<td>Role-Play</td>
<td>80.10 (p&lt;.001)</td>
<td>71.910 (p&lt;.001)</td>
<td>11.764 (p=.003)</td>
</tr>
<tr>
<td>Opinion</td>
<td>103.42 (p&lt;.001)</td>
<td>92.492 (p&lt;.001)</td>
<td>12.738 (p=.002)</td>
</tr>
<tr>
<td>Announcement</td>
<td>102.14 (p&lt;.001)</td>
<td>99.441 (p&lt;.001)</td>
<td>10.293 (p=.006)</td>
</tr>
</tbody>
</table>

Table 4-3 above shows the chi-squared scale difference results for the three language knowledge models across the four AEOCPT tasks. Because the one factor model is a standard CFA model, I made the two and four factor models standard as well. In other words, I did not include a second-order structural portion for the two and four factor models. Across all four tasks, the four factor model is the better fitting one. In addition, the hierarchical four factor model, the one with a second-order structural portion, better fits the dataset than the standard four factor CFA model. These results mean that the global fit statistics do not need to be recalculated for one of the alternative models. I cannot accept the hierarchical four-factor model, however, with just
global fit indices and alternative models. Global fit indices and comparative fit tests do not provide any information on whether every loading is significant. The last step to confirming this model is to look at component fit indices.

4.2.3 Component fit statistics

Although not called as such by Kline (2011), he does reference component fit statistics to support model fit. The benefit of viewing these statistics is that problems can be traced to a single item and resolved, improving the fit of the model to the dataset. The statistics that I report for component fit are the standardized residuals, parameter estimates, and R² values. For standardized residuals, any parameter with a value greater than an absolute value of 1.96 will be noted. For parameter estimates, Lisrel provides z-values with significance tests for all observed variables loading onto latent ones. Any parameter that does not have a significant z-value, also larger than 1.96, will be noted for possible revision. Finally, the R² values are similar to those given when a researcher runs a correlation or regression analysis. Byrne (1998) recommends a minimum R² value of .50 to support component fit for CFA models.

Table 4-4. Component fit statistics for the four factor hierarchical model

<table>
<thead>
<tr>
<th></th>
<th>Mini-Lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized Residuals above</td>
<td>3.57%</td>
<td>10.71%</td>
<td>17.86%</td>
<td>3.57%</td>
</tr>
<tr>
<td>Z-values below 2</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Average R² (Range)</td>
<td>.80 (.48 – 1.03)</td>
<td>.82 (.61-. 98)</td>
<td>.82 (.57-.95)</td>
<td>.83 (.66-.95)</td>
</tr>
</tbody>
</table>

Because the component fit statistics are given for either every observable variable or in a matrix form for every relationship between observable variables, Table 4-4 above shows
condensed information. I provided the percent of standardized residuals above an absolute value of two and the percentage of z-values below two for every task. Large percentages of these values indicate problems with the proposed relations among variables. Reducing the large standardized residuals across the four tasks would require freeing a parameter between the two observable variables that produced the residual. For example, the largest standardized residual in Task1 is between rubric items five and seven. Freeing the error covariance between these items, however, is not supported from these component fit statistics alone. Freeing the error covariance should be supported from the language knowledge framework. Correcting a Z-value below 2 or a low $R^2$ value would require fixing a parameter that is freed.

After identifying parameters that could either be freed or fixed, model respecification (Kline, 2011), modification indices are run through Lisrel to estimate model improvement. From the results of the modification indices, none of the standardized residuals identified in Table 9 would significantly improve model fit. In addition to these results, Kline also notes that model respecification should only be done if it fits the theory supporting the model. There are no discussions in Bachman and Palmer (1996, 2010) supporting a relationship between any of the observed variables noted from the component fit statistics. The fact that modification indices show freeing error covariances as improving model fit suggest that the relationship is outside the language knowledge framework.

4.3 Metric invariance: Collecting results for answering RQ3

After confirming configural invariance, the same CFA model is shown to have good fit across the four tasks, and DuToit and DuToit (2001) mandate metric invariance as the next step. Metric invariance is tested in Lisrel by fixing the factor loadings to be the same. Because the difference between these two models is freeing some parameters, I identify the better fitting
model using the Satorra-Bentler chi-square difference test (Satorra & Bentler, 2001). A significant difference will mean that the less constrained model, configural invariance, better fits the dataset.

**Table 4-5. Global fit and component statistics for configural and metric invariance of the four AEOCPT tasks**

<table>
<thead>
<tr>
<th></th>
<th>Configural Invariance (df = 64)</th>
<th>Metric Invariance (df = 88)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-B $X^2$ (p)</td>
<td>250.37 (p&lt;.01)</td>
<td>590.14 (p&lt;.01)</td>
</tr>
<tr>
<td>$X^2_{diff}$ (24)</td>
<td></td>
<td>73.69 (p&lt;.001)</td>
</tr>
<tr>
<td>SRMR</td>
<td>.023</td>
<td>.042</td>
</tr>
<tr>
<td>RMSEA (90% CI)</td>
<td>.070 (.061 ; .079)</td>
<td>.069 (.061 ; .077)</td>
</tr>
<tr>
<td>CFI</td>
<td>.99</td>
<td>.99</td>
</tr>
<tr>
<td>Standardized Residuals above $</td>
<td>2</td>
<td>$</td>
</tr>
<tr>
<td>Z-values below 2</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Average $R^2$ (Range)</td>
<td>.82 (.48-.1.03)</td>
<td>.82 (.50-1.09)</td>
</tr>
</tbody>
</table>

Global fit statistics for both measurement invariance models, shown above in **Table 4-5**, support good fit. Although the chi-squared tests are significant, it is still possible to run the Satorra-Bentler chi-squared difference test. The component fit statistics are interpreted the same as when I ran the four tasks independently. The chi-squared difference test result supports that the less constrained model, configural invariance, better fits the AEOCPT dataset.
Figure 4-2. Standardized estimates of the language knowledge model for the Mini-lecture task.

Figure 4-3. Standardized estimates of the language knowledge model for the Role-play task.
Figures 4-2 through 4-5 above show the models that I have accepted as supporting configural invariance across the four AEOCPT tasks. The numbers included in each figure are the standardized estimates of the relationship between latent and observed variables. These values are interpreted as changes in standard deviation. A one standard deviation change to grammar knowledge, for example, will cause a .85 standard deviation change in item 1 and a .94 standard deviation change in item 2. (Readers will notice a parameter estimate in Figure 4-1 that is slightly out of bounds but worth investigating through an additional model).
4.4 Confirming a Multitrait-Multimethod CFA model

In Figure 4-2, a parameter estimate between sociolinguistic knowledge that loads onto the second order language knowledge is larger than one, whereas all parameter estimates from a CFA model should be between 0 and 1. Kline (2011) explains that the illogical parameter estimate is called a Heywood case (Heywood, 1931), and gives six possible reasons for such an occurrence (p. 158). These possible reasons are:

1) the relation between the factor structure and observable variables was incorrectly specified
2) the model was not identified prior to analysis
3) outliers in the dataset
4) The sample size was less than 200 or there were only two observable variables loading onto each factor
5) bad starting values
6) correlations that are either extremely high or low

In this dissertation, an illogical parameter estimated was produced from a low correlation between the two observable variables that loaded onto sociolinguistic knowledge. This correlation between the two variables is .58, below Byrne’s (1998) recommendation of at least .60.

This is not the first time that I have observed a Heywood case from Task 1’s CFA model, but this case is less serious. A previous analysis that I did with a smaller dataset was intended to determine which items. However, the AOECPT scoring rubric should be removed. I observed a Heywood case from Task 1 as well but for a different reason. Two items that loaded onto
functional knowledge were too highly correlated, resulting in a Heywood case which Lisrel could not interpret. The solution for that case was to eliminate the problematic items. However, a possible solution for this dissertation is to test the fit of all 32 items loading onto the language knowledge model.

The four AECOPT tasks loading onto the same structural model could be referred to as a single Multitrait-Multimethod (MTMM) model (Kline, 2011). Bachman and Palmer (1983) used this kind of model to confirm their first version of the language knowledge framework. However, they did not analyze their revised frameworks using CFA, an opportunity that I can address in this dissertation. The MTMM model of Bachman and Palmer is Correlated trait-correlated method (CTCM). This model contains observable variables that load onto two types of latent variables, one set that represents the mental traits of interest and the other representing methods.

While attempting to analyze the AEOCPT data using a Correlated trait-Correlated Method model, however, I found that the model failed to converge, probably due to the fact that most recent versions of the language knowledge are too complex. The final accepted model given by Bachman and Palmer and the example model given by Kline (2011) are standard CFA models. Given this fatal error, it would not be possible to test the new language knowledge using a CTCM model, but Kline cites an additional MTMM model. Marsh and Grayson (1995) propose an alternative MTMM model they call correlated uniqueness.

The difference between the correlated uniqueness (CU) and CTCM models is that the former measures methods through freeing error covariances whereas the latter measures methods through latent variables.
Table 4-6. Global fit and component statistics for the Configural Invariance and CU models

<table>
<thead>
<tr>
<th></th>
<th>Configural Invariance (df = 64)</th>
<th>Correlated Uniqueness (df = 348)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-B X² (p)</td>
<td>250.37 (p&lt;.01)</td>
<td>798.78 (p&lt;.01)</td>
</tr>
<tr>
<td>SRMR</td>
<td>.023</td>
<td>WRMR</td>
</tr>
<tr>
<td>RMSEA (90% CI)</td>
<td>.070 (.061 ; .079)</td>
<td>RMSEA (90% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.047 (.042 ; .051)</td>
</tr>
<tr>
<td>CFI</td>
<td>.99</td>
<td>CFI</td>
</tr>
<tr>
<td>Standardized</td>
<td>14.29%</td>
<td>Residuals above</td>
</tr>
<tr>
<td>Residuals above</td>
<td>2]</td>
<td></td>
</tr>
<tr>
<td>Z-values below 2</td>
<td>0.00%</td>
<td>Z-values below 2</td>
</tr>
<tr>
<td>Average R² (Range)</td>
<td>.82 (.48-.1.03)</td>
<td>Average R² (Range)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.77 (.48-.96)</td>
</tr>
</tbody>
</table>

Table 4-6 above shows the global and component fit statistics for the CU model and the previously given Configural Invariance model. Because comparative fit statistics cannot be used to compare the two models, the information in the table above will be used to answer whether the CU model also has acceptable fit to the dataset. I should note here that I analyzed the CU model using Mplus (Muthen & Muthen, 2015). The estimator I used in Mplus was WLSMV, which Muthen and Muthen cite as also being acceptable for categorical variables. In addition some fit statistics are different in the table. For global fit, Mplus provides a Weighted Root Mean Residual (WRMR) when using the Weighted Least Squares Means and Variances (WLSMV) estimator. Muthen (2004) states a good fitting model as having a WRMR value less than .90 and acceptable at less than 1.00 (Muthen, 2010).

In terms of component fit indices, Mplus does not produce standardized residuals for the WLSMV estimator. Instead, I will use Kline’s (2011) standard of residuals being above an absolute value of .10. According to Table 4-6, the majority of variance are below Kline’s standard and the z-values for the parameters between latent and observable variables are all above 2.00. The error covariances have their own parameter estimates as well and some of these
estimates were not significant. I do not consider this to be an issue, however, since the resolution to this would be fixing the parameters to 0. I previously showed that a solution with fixed error covariances produces an acceptable solution, but I do not have support from the language knowledge framework to fix individual error covariances. Finally, the average $R^2$ value from the CU model is lower than the configural invariance model, but the out of bounds estimates are not in the CU model.

The global and component fit statistics for the CU model support that it has acceptable fit to the AEOCPT dataset. Modification indices do not support freeing any parameters that I could support from previous research. The CU model resolves the issues that were present in the configural invariance model. Previously stated, a minor concern with the configural invariance model is that one of the loadings in Task 1 was identified as a Heywood case. The results from the CU model, however, show that this Heywood case has been resolved. The reason for this is that more observable variables are loading onto sociolinguistic knowledge. These variables provide more information about the relations between the rubric items loading onto the latent variable as they vary across the four tasks. In addition, the CU model better replicates administrators’ intent to use the AEOCPT scores as feedback separately and placement together. The Configural Invariance model shows that the language knowledge model acceptably fits the four tasks. The CU model, however, shows that the now 32 items (Eight items across four tasks) acceptably fit the language knowledge model.

4.5 Summary

In this chapter, I showed the results I collected for answering RQ 1 (Does the AEOCPT data support the language knowledge framework being operationalized in the scoring rubric?)
and RQ3 (How do the four AEOCPT tasks contribute information to understanding placement, language knowledge strengths, and language knowledge weaknesses?) in this dissertation. The results from this chapter come from the five CFA models I analyzed using Lisrel 9.2. For producing an answer that I will provide more information on in the discussion chapter, I provided global and component fit statistics. After determining whether the models are an acceptable fit to the AEOCPT dataset, I used comparative fit statistics to answer which model fits the best. A general conclusion from these results is that the four factor language knowledge model best fits the data. In addition, the four tasks produce rubric item factor loadings that are unique from each other. This can best be expressed through a CU model, a type of MTMM model, which shows the relation of items within a task through error covariances.

In the next chapter, I will answer RQ2 and continue to answer RQ3 through the results produced by a different latent trait statistic, Latent Class Analysis (LCA). I will first discuss the information within the syntax that I submitted to PROC LCA. I will then discuss the results for answering whether the initial LCA model is identified. Next, I will show the comparative statistics I need for finding the best fitting class model. Finally, I will test whether the AEOCPT tasks produce statistically different class probabilities.
Chapter 5 LCA Results

In this chapter, I will report the results for answering RQ2 and the second part of RQ3. Most of the results in this chapter come from reports produced by PROC LCA, but I also used Lisrel for the first step of answering RQ2. In order, I present first the data screening results that I did before submitting the data to PROC LCA. I then provide the results for the model identification phase followed by model selection. Finally, I interpret the classes from the chosen LCA model and provide some information on the model’s impact to AEOCPT placement.

5.1 Data screening results and data modification

Because LCA is a cross-sectional data analysis method (Muthen, 2001), Collins and Lanza (2010) advise screening the data via the analysis of frequency tables. The purpose of these tables is to help determine whether all points of the AEOCPT rubric have been equally chosen by raters. The current number of points within the rubric is four, one for each placement decision made from the test scores. This number is below Winke’s (2014) recommendation for Likert scales used in language testing (five to seven points), but raters and administrative stakeholders have suggested that even fewer points could be appropriate. One way to determine the optimal amount of points needed within a rubric is by comparing the observed total response patterns to the maximum possible.
Table 5-1. Observed and expected response patterns for the eight AEOCPT rubric items

<table>
<thead>
<tr>
<th>Possible Likert scales within a rubric</th>
<th>Maximum possible response patterns</th>
<th>Tasks</th>
<th>Observed (Rated) response patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two</td>
<td>256</td>
<td>Mini-lecture</td>
<td>286</td>
</tr>
<tr>
<td>Three</td>
<td>6,561</td>
<td>Role-Play</td>
<td>236</td>
</tr>
<tr>
<td>Four</td>
<td>65,536</td>
<td>Opinion</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Making an Announcement</td>
<td>273</td>
</tr>
</tbody>
</table>

Table 5-1 shows the results from the Lisrel output that I used to answer RQ1 in the previous chapter. The results for the third column in the table come from a simple calculation of the desired rubric level raised to the power of eight, the number of items in the rubric. Using the information in Table 5-1, I can support stakeholders’ previous suspicions that not all four points of the AEOCPT scoring rubric are equally endorsed. The number of points supported from the information above is between two and three. I decided to reduce the number of points within the rubric to three in order to not limit the number of choices made by raters for two of the tasks. Although stakeholders did communicate which level they believe has been endorsed the least, the table above does not contain the necessary information to justify a revision.

Determining which level was not endorsed enough by raters can be investigated by requesting PROC LCA to produce frequency tables.
Table 5-2. Frequency of ITA candidates receiving values across all four tasks

<table>
<thead>
<tr>
<th>Rubric Items</th>
<th>Level 1 (Intermediate)</th>
<th>Level 2 (High Intermediate)</th>
<th>Level 3 (Advanced)</th>
<th>Level 4 (Certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>17</td>
<td>344</td>
<td>1291</td>
<td>740</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>23</td>
<td>357</td>
<td>1278</td>
<td>734</td>
</tr>
<tr>
<td>Transition</td>
<td>15</td>
<td>263</td>
<td>1075</td>
<td>1039</td>
</tr>
<tr>
<td>Prominence</td>
<td>23</td>
<td>329</td>
<td>1116</td>
<td>924</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>34</td>
<td>298</td>
<td>1346</td>
<td>714</td>
</tr>
<tr>
<td>Multimodality</td>
<td>24</td>
<td>185</td>
<td>952</td>
<td>1231</td>
</tr>
<tr>
<td>Task Response</td>
<td>9</td>
<td>194</td>
<td>945</td>
<td>1244</td>
</tr>
<tr>
<td>Question Response</td>
<td>11</td>
<td>153</td>
<td>792</td>
<td>1436</td>
</tr>
</tbody>
</table>

Table 5-2 above shows how raters scored candidates using the four points for all eight rubric items. Readers should note that the values within the table above are larger than the stated sample because they are summed across the four AEOCPT tasks. As a result, the frequency totals for each row are 2,392. It is from this information in the table that I can confirm the suggestions made by stakeholders. Descriptions for an intermediate level of performance are not endorsed enough by raters to warrant keeping them in the rubric.

Based on the information I gained from data screening, I revised the dataset by dropping the lowest level in the AEOCPT rubric. The benefit of revising the rubric and dataset at this stage of analysis is that it better improves the chances of successful model identification. Without the revisions, the dataset could be too sparse for PROC LCA to find well identified classes. This revision does not mean, however, that the four placement decisions currently made within the ITA program are no longer supported. Answering RQs 2 and 3 will be done through the next steps in LCA, model identification and selection.
5.2 Finding a starting point through model identification

In the previous chapter, I conducted model identification in CFA by adhering to a set of rules created by previous research (Berry, 1984; Bollen, 1986; Kenny, 1979; Kenny, Kashy, & Bolger, 1998; Kline, 2011; O’Brien, 1994; Wagner, Torgeson, & Rashotte, 1994). A standard set of rules has not been developed for LCA yet, but Lanza et al. (2014) developed a process for identifying an initial LCA model in PROC LCA. Using seed selection and the NSTARTS syntax that I discussed in the previous chapter, model selection is driven by the number of seeds that identified the same maximum likelihood value. Counting the seeds that find the same maximum likelihood value ensures the likelihood distribution is unimodal. Because LCA has never been run on the AEOCPT dataset before, I set the value of NSTARTS to 100, shown in Appendix D. This value is more computationally demanding for the computer running PROC LCA, but it provides more information for model identification.

An important consideration when running PROC LCA for model identification is selecting the initial model. Because LCA assumes that true class membership is unknown (Collins & Lanza, 2010), an initial model can be difficult to determine. Researchers can choose any starting point, but the disadvantage to this is that model selection could become more difficult to resolve. Because the AEOCPT began as a variable-oriented approach (Bergman & Magnusson, 1997; Bergman, Magnusson, & EL-Khour, 2003; Collins & Lanza, 2010) to placement, a four class model is a practical starting point. An additional reason to start with this model is that I can better justify an alternative model to administrators by showing a comparison to the current four class one.
Table 5-3. Information from PROC LCA used for model identification

<table>
<thead>
<tr>
<th>Task</th>
<th>Percentage of seeds identifying the same maximum likelihood</th>
<th>Seed for best fitted model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mini-Lecture</td>
<td>69.00%</td>
<td>269289921</td>
</tr>
<tr>
<td>Role Play</td>
<td>83.00%</td>
<td>2038873619</td>
</tr>
<tr>
<td>Opinion</td>
<td>66.00%</td>
<td>838136479</td>
</tr>
<tr>
<td>Making an announcement</td>
<td>96.50%</td>
<td>1635894722</td>
</tr>
</tbody>
</table>

Table 5-3 above contains information I used from the results output in PROC LCA to determine model identification. I included a third column containing information about the seed associated with the best fitted model to show an important feature when using PROC LCA. In addition to using seeds for model identification, Lanza et al. (2014) also instruct users to note the seed for the best fitted model in order to reproduce results. This value has no substantive interpretation other than starting the analysis, but providing the seed produces classes with the same label. Readers can think of the seed as a key for decoding an encrypted message correctly.

Similar to the measurement invariance process in the previous chapter, I am using the information in Table 5-3 to determine whether the model is identified in each task separately. Because I am making a dichotomous decision, I will use a percentage of seeds greater than 50% as the standard for identification. Using this standard, a four class model is identified for all four tasks and can be used as a starting point for model selection. I will now add complexity to model selection by including a grouping variable, seen as GROUPS in the LCA syntax in Appendix D. Because LCA is an exploratory procedure (Collins & Lanza, 2010), I also randomly divided the dataset into two halves in order to confirm a final model during model selection.
5.3 Selecting the best fitting LCA model

After identifying an initial LCA model, the next step in answering RQ3 and RQ4 is determining which latent class model fits best. I will answer RQ3 by adding a GROUPS statement to PROC LCA and run two, three, four, and five class models. I chose a three class model since it is one class fewer than the current placement categories. I chose a two class model because the total response patterns for two tasks supported using a dichotomous rubric. Finally, I chose a five class model because several administrative stakeholders suspected that another class existed within the ITA population. They believe that this new class is between the 118G and certified decisions, a group of students that have weaknesses warranting some intervention that is not a semester long course. Because LCA is a person-oriented approach (Bergman & Magnusson, 1997; Bergman, Magnusson, & EL-Khoury, 2003; Collins & Lanza, 2010), the results from this study are well suited for addressing stakeholders’ suspicions with empirical evidence.

Lanza et al. (2014) provide several measures for researchers to use when deciding on the best fitting LCA model. Two of these are the Akaike Information Criterion (AIC; Akaike, 1987; Lin & Dayton, 1997) and the Bayesian Information Criterion (BIC; Schwarz, 1978; Lin & Dayton, 1997). The AIC and BIC values are used when comparing non-nested models (models that have theoretically different latent variables). In addition to these, Lanza et al. also provide the consistent AIC (Bozdogan, 1987; Lin & Dayton, 1997) and adjusted BIC (Sclove, 1987). The last measure for model comparison given by Lanza et al. is one unique to cross-sectional analysis techniques. Collins and Lanza (2010) use a measure of entropy designed by Ramaswamy, DeSarbo, Reibstein, and Robinson (1993). Although entropy is defined as the disorder within a system, Collins and Lanza define the entropy statistic as a measure of latent class separation.
Using Ramaswamy et al’s statistic, a value between 0 and 1 is given. I will interpret a higher value as a LCA model that gives clearer separation between classes.

**Table 5-4. Comparative fit statistics for 1st half of the dataset**

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Two Class</th>
<th>Three Class</th>
<th>Four Class</th>
<th>Five Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>3825.21</td>
<td>2928.50</td>
<td>2720.59</td>
<td><strong>2674.68</strong></td>
</tr>
<tr>
<td>BIC</td>
<td>4496.66</td>
<td><strong>3945.85</strong></td>
<td>4083.84</td>
<td>4383.83</td>
</tr>
<tr>
<td>CAIC</td>
<td>4628.66</td>
<td><strong>4145.85</strong></td>
<td>4351.84</td>
<td>4719.83</td>
</tr>
<tr>
<td>Adjusted BIC</td>
<td>4077.38</td>
<td>3310.57</td>
<td><strong>3232.57</strong></td>
<td>3316.56</td>
</tr>
<tr>
<td>Entropy</td>
<td>.92</td>
<td>.89</td>
<td>.86</td>
<td>.85</td>
</tr>
</tbody>
</table>

**Table 5-5. Comparative fit statistics for 2nd half of the dataset**

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>Two Class</th>
<th>Three Class</th>
<th>Four Class</th>
<th>Five Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>3791.36</td>
<td>2945.12</td>
<td>2746.71</td>
<td><strong>2694.79</strong></td>
</tr>
<tr>
<td>BIC</td>
<td>4462.81</td>
<td><strong>3962.47</strong></td>
<td>4109.95</td>
<td>4403.93</td>
</tr>
<tr>
<td>CAIC</td>
<td>4594.81</td>
<td><strong>4162.47</strong></td>
<td>4377.95</td>
<td>4739.93</td>
</tr>
<tr>
<td>Adjusted BIC</td>
<td>4043.53</td>
<td>3327.19</td>
<td><strong>3258.68</strong></td>
<td>3336.67</td>
</tr>
<tr>
<td>Entropy</td>
<td>.90</td>
<td>.89</td>
<td>.86</td>
<td>.86</td>
</tr>
</tbody>
</table>

Tables 5-4 and 5-5 show the results for the comparative fit statistics I described earlier across all four latent class models. The mixed interpretations from the five statistics in the tables highlights the importance of using all the information given from a statistics program. To help readers quickly interpret the information in Tables 5-4 and 5-5, bolded values denote the best fitting models for each statistic. For the first half of the dataset, the results show the three class latent model as the best fitting. This result is replicated nearly the same with the second half of the dataset, supporting the choice of a three class model.

Although the majority of the comparative fit statistics do not support my choice of a three class model, this is not yet a cause for concern. Similar to the model selection process in factor analysis, the step following the assessment of global fit is determining component fit.

Component fit is assessed in LCA by checking two characteristics, homogeneity and separation.
(Collins & Lanza, 2010). In the next section of this chapter, I will define these two terms and provide evidence that shows the extent that these two conditions are met. After selecting the best fitting model, I will recombine the dataset to in order to best measure component fit.

5.4 Remaining questions about the chosen LCA model

Recombining the two halves of the dataset is necessary after selecting the best fitting model because the next steps of LCA involve reporting component fit statistics and interpreting the three classes. Because the selected model is different from the one I identified, it is prudent that I go through all the same checks for this new model. The percentage of seeds finding the same maximum likelihood in the three class model is 48.70%. This is below the standard I stated previously, but Lanza et al. (2014) provide a solution in PROC LCA, shown in Appendix D. By adding a Rho prior as an additional line of syntax, the percentage of seeds finding the same maximum likelihood increases to 61.70%, making this model identified.

Lanza et al. (2014) state that adding a rho prior is appropriate when a researcher believes that several individuals are receiving marginal scores (i.e., receiving the maximum or minimum scores possible). Table 5-2 shows that the highest level of the AEOCPT rubric was selected by raters the 2nd most amount of times. By adding pseudo-cases to the dataset, candidates receiving the maximum score by raters are drawn away from the maximum (Lanza et al., 2014). An additional advantage to adding this line of syntax to PROC LCA is an increased posterior probability of class membership. A similar procedure is done in IRT to bias candidates away from the extreme ends of the distribution so that ability estimation can be calculated (Hambleton & Swaminathan, 1985). Standard errors are also produced by PROC LCA when a Rho Prior is given, which can be used to construct confidence intervals.
5.4.1 Testing measurement invariance for RQ3

To this point, I have reported results from PROC LCA that have freed the grouping variable. In order to answer RQ3, I will report the result that compares the freed group model to the fixed group model. Fixing the grouping variable is done in PROC LCA by including the MEASUREMENT line with the name of the grouping variable, which is Tasks in this dissertation. Because the two models are nested, comparing the two cannot be done using the comparative fit statistics that I used in the previous section. As with testing measurement invariance in the factor analysis chapter, I need to conduct a difference test.

The statistic that I use to conduct the difference test is the likelihood-ratio statistic, \( G^2 \) (Agresti, 1990, 2007). Collins and Lanza (2010) explain that this statistic has a similar distribution to \( X^2 \) when sparseness is checked. (The difference between the two is that \( G^2 \) is used for cross-sectional analysis methods). Sparseness is calculated by dividing a study’s sample size by the size of the contingency table. Previous research shows that a sparseness value less than 5 will result in a \( G^2 \) distribution no longer matching that of \( X^2 \) (Koehler, 1986; Koehler & Larntz, 1980; Larntz, 1978). If the grouping variable is freed, this would produce a contingency table of 96 cells. The scores from 598 candidates divided by the contingency table value results in a sparseness value of 6.23. This value is above the standard of 5.00, supporting that a \( G^2 \) difference test can be used to test measurement invariance. The results from the \( G^2 \) difference test are \( G^2 (144) = 331.16, p<.001 \), which supports that the freed three class model better fits the dataset.

5.4.2 Interpreting the three classes and testing assumptions

After confirming that the LCA model with a grouping variable provides the best fit to the dataset, the next step is to interpret the classes. Collins and Lanza (2010) explain that this is done
by looking at the points within each class that received the highest probability. Researchers must then infer what each level means in accordance with the theory that was used to develop the instrument. This step is what makes the factor analysis portion of the dissertation so important and needing to be answered first. The ability to interpret classes is dependent on the information in every indicator, the score given for the class, and how the indicators relate to each other.

Using the AEOCPT scoring rubric and tasks, I label the three classes by the scores that are the most highly associated with them. **Tables 5-6 to 5-9** contain these two pieces of information. Under the three class labels are the rubric scores most highly associated with each class and the exact probability of those associations.

**Table 5-6. Indicator response patterns for each class on the Mini-Lecture task**

<table>
<thead>
<tr>
<th>Mini-Lecture</th>
<th>Class 1 (High Intermediate)</th>
<th>Class 2 (Advanced)</th>
<th>Class 3 (Certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item Score</td>
<td>Probability</td>
<td>Item Score</td>
</tr>
<tr>
<td>Grammar</td>
<td>1</td>
<td>0.59</td>
<td>2</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>1</td>
<td>0.69</td>
<td>2</td>
</tr>
<tr>
<td>Transition</td>
<td>1</td>
<td>0.65</td>
<td>2</td>
</tr>
<tr>
<td>Prominence</td>
<td>1</td>
<td>0.65</td>
<td>2</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>1</td>
<td>0.64</td>
<td>2</td>
</tr>
<tr>
<td>Multimodality</td>
<td>2</td>
<td>0.60</td>
<td>2</td>
</tr>
<tr>
<td>Task Response</td>
<td>2</td>
<td>0.66</td>
<td>2</td>
</tr>
<tr>
<td>Question Response</td>
<td>2</td>
<td>0.63</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 5-7. Indicator response patterns for each class on the Role-Play task

<table>
<thead>
<tr>
<th>Office Hour</th>
<th>Class 1 (High Intermediate)</th>
<th>Class 2 (Advanced)</th>
<th>Class 3 (Certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item Score</td>
<td>Probability</td>
<td>Item Score</td>
</tr>
<tr>
<td>Grammar</td>
<td>1</td>
<td>0.60</td>
<td>2</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>1</td>
<td>0.70</td>
<td>2</td>
</tr>
<tr>
<td>Transition</td>
<td>1</td>
<td>0.66</td>
<td>2</td>
</tr>
<tr>
<td>Prominence</td>
<td>1</td>
<td>0.90</td>
<td>2</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>1</td>
<td>0.61</td>
<td>2</td>
</tr>
<tr>
<td>Multimodality</td>
<td>2</td>
<td>0.67</td>
<td>3</td>
</tr>
<tr>
<td>Task Response</td>
<td>1</td>
<td>0.53</td>
<td>2</td>
</tr>
<tr>
<td>Question Response</td>
<td>2</td>
<td>0.56</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 5-8. Indicator response patterns for each class on the Opinion task

<table>
<thead>
<tr>
<th>Opinion</th>
<th>Class 1 (High Intermediate)</th>
<th>Class 2 (Advanced)</th>
<th>Class 3 (Certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item Score</td>
<td>Probability</td>
<td>Item Score</td>
</tr>
<tr>
<td>Grammar</td>
<td>1</td>
<td>0.73</td>
<td>2</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>1</td>
<td>0.75</td>
<td>2</td>
</tr>
<tr>
<td>Transition</td>
<td>1</td>
<td>0.77</td>
<td>2</td>
</tr>
<tr>
<td>Prominence</td>
<td>1</td>
<td>0.59</td>
<td>2</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>1</td>
<td>0.70</td>
<td>2</td>
</tr>
<tr>
<td>Multimodality</td>
<td>2</td>
<td>0.62</td>
<td>2</td>
</tr>
<tr>
<td>Task Response</td>
<td>1</td>
<td>0.52</td>
<td>2</td>
</tr>
<tr>
<td>Question Response</td>
<td>2</td>
<td>0.57</td>
<td>3</td>
</tr>
</tbody>
</table>
Tables 5-6 through 5-9 show the three latent classes with the points that have the highest probability of being associated with each one. Because the AEOCPT scoring rubric contains detailed descriptions associated with the courses ITA students take, I have included labels with each class that I will justify in the next chapter. These labels, however, are similar to the ones I described in the first chapter.

Collins and Lanza (2010) recommend that researchers check two assumptions from their selected class model: homogeneity and separation. Homogeneity is defined as the extent to which the indicators within each classes’ response patterns have high probabilities. For an instrument with dichotomous items, acceptable homogeneity are probabilities above .50. Because I reduced the points within the AEOCPT scoring rubric to three, probabilities for each variable should be higher than .33. Tables 5-6 to 5-9 shows that all the probabilities are acceptable for supporting homogeneity.

The next assumption Collins and Lanza describe is separation, defined as the item response probabilities showing clear differences between classes. In other words, the three points
of the rubric should be most indicative of only one class for each indicator across the four tasks. This assumption is acceptably met for the majority of indicators across the four tasks, but three items are problematic. The task response indicator in Table 5-9, for example, shows an almost equal probability of two classes receiving a score of two. In order to determine whether these indicators with low separation are problematic, posterior probabilities need to be analyzed.

5.4.3 Analyzing posterior class probabilities

Similar to IRT, one strength of LCA is that posterior probabilities can be calculated to score test takers. In the case of LCA, scoring means determining each candidates’ highest class membership probability. For the model in Appendix D, every candidate receives four class memberships for every task. The mean of all membership probabilities for all four tasks is 0.94, reflecting that the indicators with poor separation are not significantly impacting overall membership. Because the results from this study are comparable to placement decisions that already have been made, I can compare the results from the summative and LCA informed placement methods.

Table 5-10. Comparison between summative and LCA informed placements

<table>
<thead>
<tr>
<th>Prior Class Placement</th>
<th>Class 1 (High Intermediate)</th>
<th>Class 2 (Advanced)</th>
<th>Class 3 (Certified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>128</td>
<td>200</td>
<td>270</td>
<td></td>
</tr>
<tr>
<td>Mean Posterior Probability</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Posterior Class Placement</td>
<td>100</td>
<td>233</td>
<td>265</td>
</tr>
</tbody>
</table>

Overall Posterior Probability Mean = 0.94

The information in Table 5-10 above shows the frequency numbers for the summative and LCA inspired placements as well as the mean posterior probabilities for each class. Readers will notice that all of the frequency numbers have changed. This change likely represents
differences between the two placement methods for borderline students. In addition to overall placement decisions, posterior probabilities produced by PROC LCA also have membership probabilities for every test taker. Table 5-11 below is a sample of PROC LCA’s posterior probability output for two ITA candidates on the same task.

Table 5-11. Example posterior probability results

<table>
<thead>
<tr>
<th></th>
<th>ITA Candidate #1</th>
<th>ITA Candidate #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Multimodality</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Transition</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Prominence</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Task Response</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Question Response</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>High Intermediate (117G)</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Advanced (118G)</td>
<td>0.000</td>
<td>0.910</td>
</tr>
<tr>
<td>Exempt</td>
<td>0.999</td>
<td>0.089</td>
</tr>
<tr>
<td>BEST</td>
<td>1 (0.999)</td>
<td>3 (0.910)</td>
</tr>
</tbody>
</table>

The first eight rows of the table above are the average scores given by the two raters for both ITA candidates’ performances on one task. Lanza et al.’s (2014) program inserted the last four rows, which represent the membership probabilities each candidate has to the selected three latent classes. They also included the last row as a quick summary of which class the test taker has the highest probability of membership. This information helps me to answer RQ2, but it is not ready to be implemented within the ITA program. Because every ITA class has a maximum of 15 students, a large amount of information (a total of 60 tables like Table 5-11) needs to be processed by instructors in a short amount of time. In order to ensure that the AEOCPT are useful and practical (Bachman & Palmer, 1996), I propose a way of reporting placement results other than numerically.
Mertens (2008) points out that evaluation of a program or test does not only depend on stakeholders that are directly involved either the program or test. Although I can interpret the posterior probabilities from the PROC LCA output, this information must accommodate ITA candidates, their respective program administrators, and ITA instructors. All of these stakeholders have a wide range of statistical knowledge, requiring me to greatly simplify the LCA informed placements. I transformed the information from tables like Table 5-11 into the symbols shown in Tables 5-12 and 5-13.

Table 5-12. Example feedback form for a candidate placed in Class 2 (Advanced)

<table>
<thead>
<tr>
<th>Example #1</th>
<th>Class 2 = 0.98</th>
<th>Mini-lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Changes to an Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thought Groups</td>
<td>↑</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prominence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tone Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multimodality</td>
<td>↑</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Response</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question Response</td>
<td>X</td>
<td></td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
</tbody>
</table>

Table 5-13. Example feedback form for a candidate placed in Class 1 (High-Intermediate)

<table>
<thead>
<tr>
<th>Example #2</th>
<th>Class 1 = 0.89</th>
<th>Mini-lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Changes to an Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thought Groups</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Transition</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Prominence</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Tone Choice</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Multimodality</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Response</td>
<td>↑</td>
<td>↑</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question Response</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The two tables above are feedback forms I propose to use for reporting placement results for all points of stakeholders. Researchers of diagnostic testing have proposed feedback forms similar to the ones I provide here (Kunnan & Jang, 2011; Li, 2011; Rupp, Templin, & Henson, 2010).

My reason for making this decision is that the majority of stakeholders that will use this form will only look at it for several seconds. Candidates will likely use the form to understand what they need to work on for certification. Teachers, on the other hand, will likely look over several feedback forms to better understand their class composition for lesson planning purposes. By choosing to use symbols instead of an 8 x 4 grid of numbers, I am ensuring that both teachers and students are only given essential information.

I determine which symbol to use by comparing a candidate’s response pattern with the ones associated with each class. Although there appears to be only two symbols, I utilize a blank square to be informative as a third one. A blank symbol represents a candidate’s performance that is at the level fitting their class membership. In other words, the score given by raters matches the response pattern associated with the placed class. A “X” symbol means that a candidate’s performance is below expectations for their recommended class. For teachers, a “X” symbol is an indication that additional instruction would be needed to get that candidate to the next level.

The last symbol I chose for the feedback form is an upward arrow, ↑, representing a candidate’s performance that is above the recommended class. This would not indicate that the candidate is misplaced. Instead, this symbol should communicate to both teachers and ITA candidates that performance associated with that indicator is not a major concern for promotion or certification. Along with the symbols associated with the comparison of candidates’ response patterns and the class associated ones, I also included the posterior probability of class
membership. I propose that this value be given in the top left box along with the candidate’s identifying information. By including the posterior class probability in the feedback form, stakeholders can see be assured placement was based on the highest probability.

5.5 Summary

In this chapter, I reported the results from PROC LCA using a series of steps prescribed by Collins and Lanza (2010). After splitting the dataset into two halves in order to confirm the final model, I analyzed both halves using a four class model. The number of seeds that identified the same maximum likelihood was acceptable for all four tasks. After identifying an initial model, the next step was to compare the four class model against the two, three, and five class ones. Using non-nested comparative fit statistics, the results supported that the three class model is best fitting for both data halves.

After selecting the three class model as the best fitting, I answered RQ3 by fixing the grouping variable, tasks, and comparing it to the freed three class model. Because these two models are nested, the chi-squared difference test showed the freed grouping variable model as better fitting. A basic interpretation of the three classes in the selected model are high-intermediate language knowledge, advanced language knowledge, and certified to teach. These labels are similar to the ones I described in Chapter 1. In terms of the two assumptions Collins and Lanza suggest checking both homogeneity and separation. The results showed that both assumptions are met overall, but some indicators do not have good separation on every task.

In order to check whether the poor separation in some indicators affects placement decisions, I compared LCA informed placements from posterior probabilities against the summative placements. I showed that the differences in placement are not greatly different, but
the average posterior probability of class membership is 94 percent. This kind of information cannot be gained from the summative placement. In addition, I proposed a feedback form that shows the strengths and weaknesses of each candidate along with their class placement. I converted the diagnostic feedback into symbols so that the information could be more easily processed by ITA candidates and instructors.

In the next chapter, I will describe the results as they relate to the four evidence models (Mislevy, Steinberg, & Almond, 2002) I described earlier. I will go into more detail about the connection of the results to the three components of the AEOCPT. In addition, I will narrate how the results impact the ITA program. I will then describe any limitations I noticed from conducting this dissertation. Finally, I will propose future studies that I or colleagues could conduct to continue collecting validity evidence for the AEOCPT.
Chapter 6 Discussion

In this chapter, I discuss the results from CFA and LCA as they impact the ITA program that administers the AEOCPT. The validation model that I chose for the AEOCPT is Mislevy, Sternberg, and Almond’s (2002) Evidence-Centered Design. From this model, I identified three components of the AEOCPT to validate. I created four evidence models from the test’s rubric, tasks, and placement decisions. These four models are: 1) Rubric to language knowledge framework, 2) Tasks to language knowledge framework, 3) Rubric to placement decisions, and 4) Tasks to placement decisions.

6.1 Evidence model 1: Fit of the rubric to a language knowledge framework

The first evidence model that I discuss is one that operationalizes a language knowledge framework through the AEOCPT scoring rubric. The results from this model answered RQ1 (Four Factor Language Model). I used Bachman and Palmer’s (1996, 2010) language knowledge framework after modifying it so that it could be identified as a structural model. Figure 6-1 below is a reminder of the model that I submitted for analysis and provided in a previous chapter.
I fitted this model to the four tasks individually first before fitting them to a single model with tasks as a grouping variable. Aside from following DuToit and DuToit’s (2001) procedures for testing measurement invariance, an additional advantage to fitting the model to each task is that it makes it able to interpret differences in fit.

Comparative fit statistics confirm that the four factor language knowledge model shows the best fitting compared to one and two factor models. The one factor model represents the unitary competence hypothesis (Irvine, Atai, & Oller, 1974; Oller, 1979; Oller & Hinofotis, 1980). The two factor model, on the other hand, represents Organizational and Pragmatic knowledge, proposed by previous researchers of communicative competence frameworks (Bialystok & Smith, 1985; Canale, 1983a, 1983b Chomsky, 1980; Davies, 1989; Hymes, 1972). What this result means for the AEOCPT is that raters are capable of identifying candidates’
performances using the rubric items as I intended with the revisions. If raters had problems using the rubric items, this could be identified using the component fit statistics and correlation matrix.

Although the four factor model has acceptable fit to the AEOCPT dataset, the differences in fit among the four tasks are important for providing meaningful information for the ITA program. Before I started this dissertation, stakeholders informed me of their opinions on which tasks they perceived as useful and which they did not. The most frequent opinion was that the *Mini-lecture* task was the best and *Making Changes to an Announcement* was the worst. Stakeholders intended “best” and “worst” to mean how likely candidates would encounter the situations in the classroom. I can use the differences in fit as evidence to confirm or contest their perceptions.

As I noted previously in the CFA results chapter, the *Mini-lecture* task had the worst global fit statistics among the four tasks. Although acceptable, the reason for the poorer fit is because of a Heywood case. As is evident in the correlation matrix that I provided in Appendix C, the likely reason for the Heywood case (Heywood, 1931) is the correlation between the two observable variables loading onto sociolinguistic knowledge. The correlation value between Tone Choice and Multimodality is the lowest among all four pairs at 0.584. Byrne (1998) notes that the correlation between variables loading onto the same latent variable should be at least 0.60.

The rubric descriptions of multimodality focus on how eye contact is used to maintain connection and awareness with the audience. Candidates also use gestures to focus the audience on visuals that correspond with their speaking (Gorsuch et al., 2013; Meyers & Holt, 2002; Smith et al., 1992). This gesturing is part of Multimodality’s description is what I believe is
I believe raters’ confusion between multimodality and prominence for the Mini-lecture task result from rater training, because trainers have told raters that gestures can count as prominence or multimodality. Information that should be given along with the previous instruction is that raters should discern why candidates use each gesture. A gesture to begin using visuals is multimodality while one isolating a piece of information is prominence. To resolve this issue, changes could either be made with the rater training procedures or within the rubric descriptors to explicitly state the differences.

Also, although stakeholders believe the announcement task has less relation to actual classroom performance, the task show good fit to the language knowledge model. There are two main reasons for this: revisions to the task and administration order. This task is a slight revision of its previous version from the SPEAK test (ETS, 1982, 1986). The original item instructs test takers to make an announcement to a fictional audience. Stakeholders expressed that they thought the task was too straightforward and did not contribute any additional information to placement. By revising the task to make changes to an announcement, I sought to push candidates to highlight the information that was changed.

Secondly, I assume that the tasks are independent of each other, and that there is no information candidates can transfer from one task to the next. What I have yet to answer, however, is the extent to which raters transfer their judgments from one task to another. It is not surprising that raters use the multiple tasks to confirm their placement. Bachman (2002)
discusses several limitations of task-based language tests, but the order of tasks is not mentioned. A future study might test whether differences in model fit are attributable to either task or administration order. This is not a topic I thought of in this dissertation because I had not yet confirmed that a difference between tasks existed.

6.2 Evidence model 2: Testing the contribution of the four tasks to measuring language knowledge

Having confirmed in the previous section that the four factor language knowledge is the best fitting, the next evidence model I proposed for validating the AEOCPT involved including a grouping variable in the CFA model. Results from this evidence help to answer part of RQ3 (Assessing the Role of the Four Tasks). A group analysis is important for testing whether the tasks give unique factor loadings that still have acceptable fit to the language knowledge model. I tested two models for answering RQ 3, a measurement invariance model and a Multitrait-Multimethod model. Results supporting good model fit for either model will have different interpretations that impact scoring and placement decisions.

As I discussed in Chapter 2, ITA administrators designed the AEOCPT to be a construct-centered performance test. Raters measure candidates on the extent to which they show acceptable language knowledge on each of the four tasks (Long & Crookes, 1992; Messick, 1994; Robinson, 1996). This type of performance test differs from ends-focused tests, which measures test takers’ abilities to accomplish a single goal. The results from the measurement invariance confirm that the performance test is construct-centered. One interpretation of the results is that raters’ scoring of a candidate change for each task.

That scores vary by task contradicts the summative method of placement decisions because the summative model assumes all rubric items are weighted equally across tasks.
Because the results did not support measurement invariance, however, every task contributes unique information about candidates’ language knowledge. This variation needs to be taken into account for both the ITA program and ITA instructors.

For the program, the additional information about language knowledge will help resolve borderline placement decisions. Because the AEOCPT is a high-stakes placement test, the decision about borderline candidates will have financial implications for both candidates and their respective departments. Another reason for needing to account for task variation is for instructors’ class preparations. Currently, instructors are given candidates’ summative scores. These scores, however, can be the same for several candidates but summed differently in the rubric. Giving instructors more fine grained information about candidates’ placement will help them to prepare for the class. In addition, the additional information can be used to better understand misplaced students.

Results supporting measurement invariance using CFA also offer some insight into the LCA results. Accepting the configural invariance model as final, however, still leaves a challenge for making placement decisions. Having the ability to provide information about language performance for each task is beneficial but a single placement decision still needs to be made. In other words, a candidate cannot attend one class for some tasks and a more advanced class for the other. Given that the same eight rubric items are applied across all four tasks, the question we need to ask is how does the complete test fit to the language knowledge model?

The MTMM model tests whether the loadings across tasks each contribute unique information to model fit. In other words, model fit is measured using the same structural model but increasing the indicators from 8 to 32. Readers should note that the MTMM I used
(Correlated Uniqueness) measures the four tasks as freed error covariances between the 32 rubric items. This MTMM model is different from Bachman and Palmer’s (1982) Multitrait-Multimethod, which conceptualizes method as additional latent variables. The fact that this model is acceptable supports that the items within tasks are weighted differently, but administrators are able to use the information for making a single placement decision.

6.3 Evidence model 3: Testing and confirming the number of placement classes

After confirming the operationalization of a language knowledge framework within the AEOCPT, the next step is to switch focus to another latent variable. The ITA population at Pennsylvania State University is considered a latent variable in the model because true class membership is not known. In other words, the AEOCPT is supposed provide information about where candidates are most likely organized, but there is no way to determine whether the placement is correct or not. ITA program administrators decided that four classes are present within the candidate population. Collins and Lanza (2010) would label this discovery of class membership as variable-oriented (Expert Judgments). LCA, on the other hand, is person-oriented (Empirical Logic), identifying classes based on the number of unique response patterns given to candidates. I briefly noted in the previous chapter a difference between the LCA results and current placement categories.

One significant finding from the LCA results of RQ2 (Testing the Four Class Placement Model) is that a three class model fits the AEOCPT dataset better than the four class model. According to the response patterns associated with the three confirmed classes, the lowest class was not supported within the dataset.
Table 6-1. Current ITA placement decisions with brief course descriptions

<table>
<thead>
<tr>
<th>Score Band</th>
<th>Decision</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>250-300</td>
<td>Exempt</td>
<td>Certified to Teach</td>
</tr>
<tr>
<td>200-249</td>
<td>118G</td>
<td>Advanced English ability&lt;br&gt;Focus on classroom communication</td>
</tr>
<tr>
<td>150-199</td>
<td>117G</td>
<td>Intermediate English ability&lt;br&gt;Focus split between language knowledge and classroom communication</td>
</tr>
<tr>
<td>&lt;150</td>
<td>115G</td>
<td>Low-Intermediate English ability&lt;br&gt;Focus on developing language knowledge with emphasis placed on pronunciation</td>
</tr>
</tbody>
</table>

Table 6-1 appears here to remind readers of the placement categories in effect prior to this dissertation. One reason that the 115G was not supported in the LCA results is that the course’s focus is on one aspect of language knowledge (i.e., pronunciation). If PROC LCA identified the 115G class, it is likely that the Level 1 rubric would have been retained to show the significant weakness in one part of language knowledge.

The interpretation of the three confirmed classes are the same as the descriptions in Table 6-1. The highest class is exempt from any additional coursework and candidates are immediately Certified to Teach at the university. Thus, candidates in this class show consistently advance English proficiency in the classroom setting. Individuals within this class do not perform error free, but the errors produced are minimal or corrected immediately. ITA administrators can still give candidates information about their performance to ensure teacher development continues outside the program.

Candidates in the Advanced English ability class are placed in the ESL 118G course, and they show advanced English knowledge through the majority of the AEOCPT. Candidates
produce errors, however, that are noticeable to raters. These errors are either not corrected by candidates or produced incorrectly even after correction. Instructors teaching in this course provide students several opportunities to address weaknesses by practicing their future roles in the classroom.

The lowest class from the LCA model is **High Intermediate**, and candidates who show a balance of positive and negative performances on the AEOCPT. Candidates within the high intermediate class would be assigned to take ESL 117G followed with 118G the next semester. A description of students enrolled in 117G is that they struggle with accomplishing tasks and communicating with raters through most of the test. Instructors teaching this course provide more detailed feedback on grammatical and textual knowledge. In addition, the tasks from the AEOCPT are introduced in more detail so that students better understand how to communicate in these contexts for the next class and future teaching assignments.

I will discuss the change in item response probabilities across for the four tasks for the three classes in Evidence Model 4. Meanwhile, one might ask what happened to the fourth class. Noted earlier, the LCA results show that a fourth response pattern does not appear within the AEOCPT dataset with enough likelihood to support the model as better fitting than the three class one. This means that the **Low Intermediate** class (the 115G course) does not need to be offered at this time.

ITA stakeholders confirm this conclusion through statements that the 115G course is not well defined as to what students should achieve while enrolled. I can also use these results to address another comment from stakeholders. That comment was the existence of an additional class that is not described among the four categories. Readers will remember that I ran a four
class model when conducting the model selection procedure of LCA. Although the four class model did not fit the best, I can look at its results to propose where a fourth class could emerge from in the response patterns. If test developers confirm a four class appearing in future analyses of AEOCPT data, it is likely that this class will be between certified and high intermediate. In other words, candidates with the highest probability of placement in a fourth class will have a response pattern that is a mixture of 3s and 2s. ITA stakeholders have reported observing candidates that fit into this class. Stakeholders are currently working on a workshop proposal that would address the needs of these candidates. Because 118G is a one semester course, administrators foresee a class in-between 118G and certified as a workshop that take less than a semester to complete.

6.4 Evidence model 4: Testing the impact of tasks on placement decisions

The last evidence model I will talk about is similar to evidence model 2 with the difference that I am analyzing variation in response probability (the likelihood of class membership dependent on the indicator score) rather than factor loadings. Because the freed grouping variable model fits the dataset better than the fixed one, interpreting the three classes requires a discussion of how item response responsibilities change across tasks. In order for readers to more easily follow the interpretation of classes across the four tasks, I will reorganize the tables from my presentation of LCA results in Chapter 5. I will give one table for each class along with commentary on what the responses mean for language performance. Table 6-2 below contains the class response patterns for Class 1.
Table 6-2. Indicator response patterns for Class 1 on each task

<table>
<thead>
<tr>
<th>Class 1: High-Intermediate</th>
<th>Mini-Lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1 (0.59)</td>
<td>1 (0.60)</td>
<td>1 (0.73)</td>
<td>1 (0.66)</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>1 (0.69)</td>
<td>1 (0.70)</td>
<td>1 (0.75)</td>
<td>1 (0.70)</td>
</tr>
<tr>
<td>Transition</td>
<td>1 (0.65)</td>
<td>1 (0.66)</td>
<td>1 (0.77)</td>
<td>1 (0.54)</td>
</tr>
<tr>
<td>Prominence</td>
<td>1 (0.65)</td>
<td>1 (0.90)</td>
<td>1 (0.59)</td>
<td>1 (0.59)</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>1 (0.64)</td>
<td>1 (0.61)</td>
<td>1 (0.70)</td>
<td>1 (0.64)</td>
</tr>
<tr>
<td>Multimodality</td>
<td>2 (0.60)</td>
<td>2 (0.67)</td>
<td>2 (0.62)</td>
<td>1 (0.58)</td>
</tr>
<tr>
<td>Task Response</td>
<td>2 (0.66)</td>
<td>1 (0.53)</td>
<td>1 (0.52)</td>
<td>2 (0.57)</td>
</tr>
<tr>
<td>Question Response</td>
<td>2 (0.63)</td>
<td>2 (0.56)</td>
<td>2 (0.57)</td>
<td>2 (0.62)</td>
</tr>
</tbody>
</table>

The information in the table above shows that the level 1 descriptors work fairly well for describing Class 1’s performance. In terms of Grammatical and Textual knowledge, candidates in this class produce errors through a large portion of their responses. This behavior is consistent through the four tasks. What readers should note, however, is that some responses from the rubric indicators are more indicative of class membership. Candidates receiving a score of 1 for Prominence on the Role-Play task are 90% likely to be in Class 1. One possible reason for this is that candidates need to highlight key information in order to resolve the problem presented by raters. Candidates can use the blackboard to highlight key information, but most will use their voices in order to better simulate an office setting. Candidates in Class 1 will struggle with highlighting key information during the fast-paced communication of the task.

Some areas of the rubric to continue developing in future revisions are the descriptions for Sociolinguistic and Functional knowledge. Readers can see that for the items associated with Sociolinguistic and Functional knowledge, candidates in class 1 receive scores of 2. This is not normally a problem in LCA, but I noted in the LCA results chapter that the response probability is nearly the same for classes 1 and 2. Although this overlap is not a problem when looking at posterior probabilities, I would suggest solutions for improving the AEOCPT’s
validity. I previously noted the problem involving Multimodality for the Mini-lecture task, but the information in Table 6-1 shows an additional issue with its descriptor.

For Multimodality, I recommend that the descriptor for Level 1 be revised to better operationalize class 1’s performance. The Changes to an Announcement task is the one task in which candidates are likely to receive a score of 1. One possible reason is that time is a factor to include in the descriptors. From my personal experiences rating the AEOCPT, many candidates will break eye contact to either write on the blackboard or read the information on the card. The revision to the descriptor might include a description of breaking eye contact from raters.

Class 1 not receiving scores of 1 for Task Response is not as clear as the one for Multimodality, but there is a pattern that administrators can address. My interpretation of the anomalous result for Task Response is that class 1 candidates are performing better on rehearsed tasks. The Mini-lecture and Changes to an Announcement task require candidates to prepare more extended speeches than the other two tasks. Conklin and Schmitt (2008) show that sequences of memorized words are processed faster than those that are not memorized. I recommend that the description within this item be revised to make raters aware of rehearsed speech.

Finally, Question Response appears to have the same issue as Multimodality, raters are not seeing candidate behavior that matches the descriptor for a score of one. My proposed solution is to break the question answering process into four parts. Researchers working with ITAs have proposed a four step process to answering questions (Gorsuch et al., 2013; Griffee, 2011; Smith, Meyers, & Burkhalter, 1992). The four steps to answering a question are: 1) acknowledging the person, 2) rephrasing the question, 3) answering the question, and 4)
checking for understanding. By revising the descriptor to focus on completing the four steps, the three classes will likely get better separation.

### Table 6-3. Indicator response patterns for Class 2 on each task

<table>
<thead>
<tr>
<th>Class 2: Advanced</th>
<th>Mini-Lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>2 (0.78)</td>
<td>2 (0.79)</td>
<td>2 (0.75)</td>
<td>2 (0.74)</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>2 (0.76)</td>
<td>2 (0.83)</td>
<td>2 (0.80)</td>
<td>2 (0.83)</td>
</tr>
<tr>
<td>Transition</td>
<td>2 (0.78)</td>
<td>2 (0.83)</td>
<td>2 (0.84)</td>
<td>2 (0.76)</td>
</tr>
<tr>
<td>Prominence</td>
<td>2 (0.84)</td>
<td>2 (0.86)</td>
<td>2 (0.81)</td>
<td>2 (0.74)</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>2 (0.86)</td>
<td>2 (0.84)</td>
<td>2 (0.83)</td>
<td>2 (0.85)</td>
</tr>
<tr>
<td>Multimodality</td>
<td>2 (0.53)</td>
<td>3 (0.49)</td>
<td>2 (0.55)</td>
<td>2 (0.65)</td>
</tr>
<tr>
<td>Task Response</td>
<td>2 (0.59)</td>
<td>2 (0.68)</td>
<td>2 (0.60)</td>
<td>2 (0.59)</td>
</tr>
<tr>
<td>Question Response</td>
<td>2 (0.49)</td>
<td>2 (0.55)</td>
<td>3 (0.50)</td>
<td>3 (0.52)</td>
</tr>
</tbody>
</table>

Table 6-3 shows the response patterns associated with Class 2, candidates who show advance language knowledge but produce some noticeable errors. Because a three class model is the best fitting to the AEOCPT dataset, I expect scores that match their respective classes. Class 2 meets this expectation because these points are well matched to candidates’ performances.

One issue within class 2 is that candidates are most likely to receive a score of 3 for Multimodality on the Role-Play task. A likely reason for this is that candidates are oriented to raters through most of their responses since the interaction is mostly resolved verbally. I propose resolving this issue by interviewing raters and instructors for Multimodality behavior that separates class 2 from 3. Similar to my suggestions for class 1, including descriptors of eye contact could be key to separating the two classes. The only other issue to resolve in future revisions to the AEOCPT is with Question Response.

The Question Response item shows a pattern in class 2 that divides candidate performance into at level (Score of 2) for the first two tasks and above level (Score of 3) for the
last two tasks. My interpretation of this is that raters are more able to ask questions for Mini-lecture and Role-play but are reluctant to pose questions with the Opinion and Making Changes to an Announcement tasks. Candidates normally provide enough information for raters to ask a variety of questions. For the opinion task, the questions are personal enough that raters may not want to challenge the answers. The likely reason for the high score on the announcement task, on the other hand, is that candidates are given all the information on a card. Raters may find it difficult to ask questions if everything said by candidates is correct and from the card. My recommendation for this item is not to revise the descriptor but address the issue in rater training.

**Table 6-4. Indicator response patterns for Class 3 on each task**

<table>
<thead>
<tr>
<th>Class 3: Exempt</th>
<th>Mini-Lecture</th>
<th>Role-Play</th>
<th>Opinion</th>
<th>Announcement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>3 (.57)</td>
<td>3 (.60)</td>
<td>3 (.57)</td>
<td>3 (.63)</td>
</tr>
<tr>
<td>Thought Groups</td>
<td>3 (.52)</td>
<td>3 (.65)</td>
<td>3 (.64)</td>
<td>3 (.65)</td>
</tr>
<tr>
<td>Transition</td>
<td>3 (.77)</td>
<td>3 (.89)</td>
<td>3 (.78)</td>
<td>3 (.90)</td>
</tr>
<tr>
<td>Prominence</td>
<td>3 (.78)</td>
<td>3 (.76)</td>
<td>3 (.74)</td>
<td>3 (.88)</td>
</tr>
<tr>
<td>Tone Choice</td>
<td>3 (.56)</td>
<td>3 (.62)</td>
<td>3 (.56)</td>
<td>3 (.70)</td>
</tr>
<tr>
<td>Multimodality</td>
<td>3 (.77)</td>
<td>3 (.88)</td>
<td>3 (.80)</td>
<td>3 (.70)</td>
</tr>
<tr>
<td>Task Response</td>
<td>3 (.85)</td>
<td>3 (.81)</td>
<td>3 (.91)</td>
<td>3 (.90)</td>
</tr>
<tr>
<td>Question Response</td>
<td>3 (.88)</td>
<td>3 (.92)</td>
<td>3 (.94)</td>
<td>3 (.95)</td>
</tr>
</tbody>
</table>

Candidates having the highest probability of belonging to Class 3 are certified to teach immediately and are exempt from any additional ITA coursework. These candidates show language ability that is largely appropriate for the American classroom. The information in **Table 6-4** shows that Class 3 has good homogeneity and separation. In addition, the expected response patterns for all four tasks are all 3s. One possible explanation for these patterns is that raters agreed with the descriptions of exempt students. The only recommendation I have for class 3 is maintaining homogeneity and separation during future administrations.
The new standards that I am proposing for placement decisions appear to have nicely reduced the AEOCPT dataset into three response categories, but readers should keep in mind LCA’s class membership assumption. This assumption is that researchers do not know ITA candidates’ true class membership. What this means from the LCA results is that raters will not score every candidate according to the classes’ expected response patterns. Raters observed a range of 236 response patterns at the low end and 286 response patterns at the high end. From these response patterns, PROC LCA identified three classes that occurred with the highest likelihood. From a total of 598 candidates in the dataset, raters gave 94 candidates the expected response patterns for one of the three classes. The remaining 504 candidates still need to be placed, however, and PROC LCA’s posterior probabilities is important to identifying the most likely class membership.

6.5 Summary

In this chapter, I discussed the results from the previous two chapters as they relate to the AEOCPT. Because this dissertation is also a validity study for the performance test, it is important that I connect the results to the AEOCPT’s administration and score use. Using an Evidenced-Centered Design, I organized this chapter into four evidence models. Two models address the fit of the AEOCPT dataset to the language knowledge and ITA frameworks. The other two models answer RQ3, whether the four tasks produce significant variation with the scoring rubric.

The first evidence model confirmed that the scoring rubric is a successful operational version of Bachman and Palmer’s (1996, 2010) language knowledge framework. I noted one issue for administrators to monitor from the Mini-lecture task. I also pointed out the unexpected result that the Changes to an Announcement task fits the best. Analyzing the four tasks loading
onto the same factor structure produces results that acceptably fit the AEOCPT dataset. In addition, the Heywood case with the *Mini-lecture* task is resolved when more observable variables load onto the same latent variable. What this means for the test is that administrators are better able to understand candidates’ sociolinguistic knowledge when raters score four tasks instead of one.

For answering the first half of RQ3 (Assessing the Role of the Four Tasks), I ran a measurement invariant model and two MTMM models. The results from the analyses show that the configural invariance and Correlated Uniqueness models acceptably fit the dataset. The information that the Correlated Uniqueness model provides for the AEOCPT is that rubric items for the four tasks can be used to make a single placement decision. In other words, the 32 items (Eight rubric items across the four tasks) contribute unique information to placing ITA candidates based on language knowledge.

I proposed the next two evidence models to discuss which classes were discovered and confirmed from the AEOCPT dataset. The general LCA results, the third evidence model and results for RQ2, show that a three class model fits best. I gave an interpretation of each class along with an analysis of their respective response probabilities. Results from testing LCA measurement invariance, RQ3, also confirm that the eight rubric items nested in four tasks contribute unique information. In this case, the 32 items contribute unique probabilities for informing class membership. This amount of information is more than the previous summative placement score and can be organized to better inform stakeholders about candidates’ performances.

I concluded each section of this chapter by suggesting modifications and future research questions from possible limitations that I discovered while conducting this study. From the CFA
results, I recommended revisions to rater training procedures. I also suggested that the order of the tasks be investigated to confirm variation can be associated with tasks. From the LCA results, I recommended revisions for some of the rubric descriptors. I also identified some rubric items that should be watched for increased class separation when more data is added.

In the next chapter, I will broaden the scope of this study by discussing the implications of the results for language testing and Applied Linguistics. I will propose contributions I can make from the results of this study for understanding the language knowledge framework. In addition, I will articulate contributions that the LCA results make to scoring placement tests. I will compare and contrast my results with previous research to better define opportunities where contributions could be made. Along with providing validation evidence to the AEOS, this dissertation can be a contribution to theory development.
Chapter 7 Implications and Contributions to Language Testing Literature

In the previous chapter, I interpreted the results as they relate to the use of AEOCPT scores by the ITA program. In this final chapter, I propose contributions that I can make from the results of this dissertation to the field of Applied Linguistics and language testing research. The organization of this chapter is based on four contributions I propose making to language testing literature. These four contributions are: 1) Empirically confirming Bachman and Palmer’s (1996) language knowledge framework, 2) Providing fine grained information for placement and pedagogical purposes, 3) Confirming the impact of performance tasks, and 4) Empirically identifying the optimal Likert scale. Finally, I include a section about limitations I found while conducting the data analysis and future studies that could address these limitations.

7.1 Empirically confirming Bachman and Palmer’s language knowledge framework

In RQ1, I confirm that a four factor language knowledge framework can be operationalized in a language assessment. Developers of other ITA performance tests (See Table 7-1) also noted that their tests operationalized theories of communicative competence. In this section, I contrast the implications of the CFA results for AEOCPT against the performance tests that inspired its development. In addition, I discuss how the AEOCPT relates to ITA tests developed from communicative competence theories.

The information in Table 7-1 shows the tests that inspired Pennsylvania State University’s ITA administrators to create its first version.
Because all three tests in the table above are still being administered at ITA programs across the United States, it is important that I justify developing the AEOCPT from the four factor language knowledge framework. McNamara’s (1996) development history of communicative competence ends with the conclusion that communicative competence is a multi-dimensional construct. If a multi-dimensional construct is operationalized, several scores should be produced about each dimension of the construct. The SPEAK, TEACH, and ITA Test, however, calculate scores using a holistic rubric. This single score can only provide general (coarse grained) information about all factors being measured. In addition, the information from a holistic rubric suggests that the multiple factors change in unison, identical to a unidimensional construct.

Smith, Meyers, and Burkhalter (1992) state that their ITA Test was inspired from the SPEAK and TEACH tests. The ITA Test score sheet contains three categories, Presentation Language Skills, Teaching Skills, and Interactive Language Skills. Each of these skills contain descriptors that raters score on a seven point Likert scale. The performance descriptors, however, are not given for every point in the rubric. Descriptors are only given for the whole point values.
and not the half-point values that are part of the seven point scale. In addition, raters provide a score for overall impression of the performance that is based on a 15-point Likert scale. This Likert scale, however, does not have any descriptors to guide raters’ choice. The lack of descriptors and violation of independence for all items means the ITA Test rubric serves best as holistic.

The ITA Performance Test 9.0 (Gorsuch et al., 2013) is not one of the foundational pieces that Pennsylvania State University ITA administrators used to create the first AEOCPT. It is foundational, however, for the most recent revisions. Because Gorsuch et al., developed their performance test using Bachman and Palmer’s (1996) language knowledge framework, this gave precedence for me to do the same to the AEOCPT scoring rubric. The difference between the AEOCPT and ITA Performance Test 9.0 is that the latter has only one task. In addition, Gorsuch et al. have not shown that the Bachman and Palmer language knowledge framework was successfully operationalized. The results from this dissertation show such operationalization is possible and can be done with a MTMM model.

Confirming that the four factor language knowledge model better fits the AEOCPT dataset than the one factor model has implications for applying the parsimony principle to test development. In CFA literature, the parsimony principle is selecting the simpler model of two given they have similar fit statistics (Kline, 2011). In terms of test development, this principle means developers should design a test that provides the most information with the fewest construct variables possible. Abraham and Plakans (1988) developed the TEACH test from a two factor language knowledge framework, but only one score is given to administrators as information. Smith, Meyers, & Burkhalter’s (1992) ITA Test has the potential to report scores
according to their two factor language knowledge framework, but they decided to transform the score into a holistic one.

One possible reason for developers of the TEACH and ITA tests to report a single score is the belief that one score is easier for stakeholders to understand. Green and Weir (2004) note a problem with this belief by showing that a placement test with a holistic grammar score accounted for only 25% of the sample’s variance. The authors concluded that there was not enough information for the test to be useful for instructional decisions. Kokhan (2013) confirms the limitation of a single holistic score by showing that test scores from large-scale standardized tests are difficult to use for ESL class placement.

7.2 Providing fine grained information for placement and pedagogical purposes

The CFA results show that the AEOCPT’s four tasks affect the eight rubric items differently, but factor loadings do not directly connect with placement decisions. What I am answering with the results of RQ2 (Testing the Four Class Placement Model) is whether the variation among the four language knowledge factors can better inform placement decisions. Researchers working with placement tests have shown that stakeholders would like to use these tests for purposes that are different from previous tradition (Green & Weir, 2004; Kokhan, 2012, 2013). The stakeholders in both of these studies as well as this dissertation reported that they would like more information to better understand incoming students’ strengths and weaknesses. The traditional purpose of placement tests is to group students into broad categories of similar ability concerning the target knowledge (Brown & Abeywickrama, 2010; Brown, 2005). Brown (2005) notes that placement decisions are made by either comparing test takers’ scores to each other or to a larger sample previously measured. Placing students according to group performance is a norm-referenced test. He distinguishes norm-referenced tests from those made from program standards (criterion-referenced).
Brown suggests that constructing a test that measures classroom objectives allows administrators to provide feedback. He labels these tests as diagnostic and categorizes them as criterion-referenced. I discussed in the Section 7.2 that researchers found it difficult to get diagnostic information from placement tests. Using Brown’s definitions of the two test types, the likely reason for this difficulty is that placement test items are created to separate students instead of operationalizing classroom objectives. Brown acknowledges this later in his chapter by stating that placement tests can be created to provide information about classroom objectives. Researchers have even shown that diagnostic information can be gained from norm-referenced tests (Buck, Tatsuoka, & Kostin, 1997; Kunnan & Jang, 2009; Li, 2011).

The results from RQ2 show that LCA models can be fit to the AEOCPT, which adds a cognitive diagnostic feature to the placement test. In addition, Latent Class Analysis is able to analyze partial credit scoring for mixtures of ability. The ability to account for partial credit scoring make LCA appropriate for performance tests and different from other Cognitive Diagnostic Models (Chapter 2) which only account for dichotomous scoring (Buck, Tatsuoka, & Kostin, 1997; Kunnan & Jang, 2009; Li, 2011). Fine grained (Item level) feedback information from latent trait statistics is difficult to get from performance tests because of the number of items required for model convergence. With the results from this analysis, I was able to create a feedback form similar to ones created by researchers using dichotomous CDA models (Kunnan & Jang, 2009; Li, 2011). Stakeholders (e.g., ITA candidates, instructors, department advisors, etc.) are now able to quickly scan the feedback form for information on candidates’ strengths and weaknesses.
7.3 Confirming the impact of performance tasks

So far in this chapter, I have argued that the results from RQ1 and RQ2 support the operationalization of Ability (Score) and Language User (Language Knowledge) in the moderate interactional framework (Bachman, 2007; Chalhoub-Deville, 2003; Chalhoub-Deville & Deville, 2006). The final piece of the framework, Context (Task), is operationalized in the AEOCPT by RQ3’s (Assessing the Role of the Four Tasks) results. The importance of operationalizing Task is noted by previous research. Chapelle (1998) argues that ability cannot be context independent in the interaction-focused language framework. Bachman (2007) contends, however, that researchers of the framework have not conceptualized context. The LCA results that I used to answer RQ3 show that the four tasks are important to understanding placement. These results begin to conceptualize context but still show the tasks as separate. The addition of the MTMM results further frames the tasks as separate but capable of being brought together to measure the same framework of language knowledge.

I would argue that the RQ3 results support an operational interaction-focused language construct but not a complete measurement of interaction. McNamara (1997, 2001) points out several variables (e.g., test partner, raters, or test designer) that he argues are part of a social dimension that is not well captured in language testing. Several researchers (Choi, 2013, 2015; Wang, 2010; Weigle, 1994; Xi & Molluan, 2011) have measured one variable (rater differences), and I capture another in my LCA model (task variability). Readers can see that the two variables are key components of the ability- in language user- in context framework. Since I did not account for the leniency or strictness of raters, I cannot offer a complete picture of context from the results of this study. I do argue, however, that this study offers a meaningful piece to compliment previous research.
7.4 Identifying the optimal Likert scale for an instrument

Before submitting the AEOCPT dataset to Confirmatory Factor Analysis and Latent Class Analysis, I collapsed the scoring rubric’s four point scale into a three point scale. The reason that I investigated this issue in the first place resulted from ITA stakeholder feedback on the test. Stakeholders reported that they did not recall giving very many candidates a score of one (Empirically confirmed in Chapter 5). Having a rubric point that raters rarely selected is not a problem for LCA (Lanza and Collins, 2010) but reducing Likert scale points would help raters’ fatigue during grading. Reducing the rubric’s Likert scale from four points to three eliminates eight descriptors that raters would have to keep in mind while grading.

Using the information organized into Table 7-1 below, I show that the optimal rubric is a three level Likert scale. In the first two columns, the maximum possible response patterns are shown for each Likert scale. The last two columns show the number of response patterns observed for the four tasks. The optimal Likert scale is the one that has a maximum possible response pattern that is greater than what was observed. The information in Table 7-1 show different optimal Likert scales for each task, but the same scale needs to be chosen to make a placement decision. Thus, a three point Likert scale allows raters to vary but not provide more information than they need.

Table 7-2. Observed and expected response patterns for the eight AEOCPT rubric items

<table>
<thead>
<tr>
<th>Possible Likert scales within a rubric</th>
<th>Maximum possible response patterns</th>
<th>Tasks</th>
<th>Observed (Rated) response patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two</td>
<td>256</td>
<td>Mini-lecture</td>
<td>286</td>
</tr>
<tr>
<td>Three</td>
<td>6,561</td>
<td>Role-Play</td>
<td>236</td>
</tr>
<tr>
<td>Four</td>
<td>65,536</td>
<td>Opinion</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Making an Announcement</td>
<td>273</td>
</tr>
</tbody>
</table>
Although I found evidence to support revising the AEOCPT scoring rubric to three points, this decision contradicts other researchers’ recommendations. For example, Winke (2014) and North (2000) recommend that Likert scales should have five to seven points for research and language testing purposes. These recommendations, however, come from previous research that neither Winke nor North empirically checked. The origin of the Likert scale recommendation is from Miller’s (1956) Cognitive Psychology experiment that found people, on average, can process seven plus or minus two units of information.

One challenge with Miller’s rule in relation to developing Likert scales is that the results are meant to explain cognitive load. Cognitive load is an issue in making ratings because human raters are holding several pieces of information in mind while making decisions. However, what this rule does not account for is that raters typically have physical written rubrics for reference. In addition, the rule concerns the maximum amount of information individuals can hold in their working memory. Using a Likert scale with the maximum amount of points possible to process would likely fatigue raters more quickly than a Likert scale with fewer points. If this argument is true, test developers would favor Likert scales with fewer than five points.

Research conducted around the same time as Miller’s study show that Likert scales between 3 and 11 points produce similar reliability values (Bendig, 1953, 1954; Lissitz & Green, 1975). Because these experiments involved raters making judgments, they are more applicable to the findings of this dissertation. The AEOCPT’s three point Likert scale is within the range of previous research and produces good reliability (Above .80). By comparing the observed response patterns to maximum response patterns for several Likert scales, I propose an empirical method for selecting the optimal Likert scale. I argue that this method addresses a concern for providing just enough information to raters. Response pattern comparison is not as precise as
Wakita, Ueshima, and Noguchi’s (2012) method using IRT, but the method I am proposing is a small sample option for test developers.

7.5 Limitations and future studies

In the process of conducting this study, I identified some limitations that do not diminish the contributions but rather present opportunities for future studies. One limitation is that I am not accounting for variation among raters in LCA. Accounting for rater strictness and leniency is a research question answered in Choi’s (2013, 2015) study using Latent Profile Analysis. In addition, explaining and controlling for rater variation is a popular topic in language testing research (Carey, Mannell, & Dunn, 2010; Isaacs & Thomson, 2013; Johnson & Lim, 2009; Kim, 2009; Xi & Mollaun, 2011). In fact, Rasch modeling is preferred over other forms of IRT because of a computer program that accounts for rater differences (McNamara & Knoch, 2012).

I addressed rater differences in this study by collapsing the two scores for each candidate into one. Lanza et al. (2014) suggest that in addition to using indicators to inform class membership, researchers can also use covariates to predict latent class membership. The analysis for including covariates is logistic regression with covariates serving as predictor variables and latent class membership serving as the outcome. One possible way to control for rater variation as a covariate is with the Credibility Index (H. K. Suen, personal communication, March 14, 2013), which is a measure calculated from rater accuracy, consistency, and transferability. The methods for measuring these three variables is similar to what administrators do during rater training. Raters watch several videos of previous performances (Consistency) that they score and compare to accepted scores (Accuracy). Because the AEOCPT consists of four tasks, administrators expect standardization to increase with every task (Transferability). Other
Another limitation that I did not address in this dissertation concerns assumptions about how ITA candidates move across classes after instruction (117G to 118 G to Certification). The Pennsylvania State University ITA program does not use the AEOCPT to promote students. According to Brown (2005), achievement tests should be used for promotional purposes. The question left unanswered from the AEOCPT, however, is whether students move classes after subsequent instruction qualifies them for promotion. Messick (1980) labels this question as one of predictive validity, which he defines as judging the relevance of the test to a job-related domain. The job-related domain for the AEOCPT is the predictable progression of candidates through the ITA courses toward teaching certification.

Lee and Greene (2007) tested predictive validity by correlating the test of interest with a more recognized measure. Specifically, they correlated the ESL placement test with students’ GPAs and faculty evaluations. Borsboom, Mellenbergh, and van Heerden (2003, 2004) argue that researchers using correlations as validity evidence are adopting a constructivist view of psychological testing. The constructivist view is that latent variables are manmade and created through the test items. The limitation with this line of thinking can be seen from the Lee and Greene’s (2007) results. What does a correlation of .05 between the ESL test and GPA infer is being constructed between the two?

Following Borsboom, Mellenbergh, and van Heerden’s argument (2003, 2004), I propose adopting a realist view of psychological testing. My suggestion for testing the predictive validity of the AEOCPT is through a third variation of LCA, Latent Transition Analysis. Collins and Lanza (2010) describe Latent Transition Analysis (LTA) as a special version of LCA that
analyzes longitudinal data. The results from LTA are used by researchers to measure the extent to which members of one latent class move to the next in time. LTA fits a realist view because the analysis tests whether candidates move through classes in the way I am proposing (117G to 118G to Certification).

An additional benefit of conducting a future LTA study on the AEOCPT is that the results could be used to evaluate the ITA courses. Because two placement decisions require candidates to enroll in either ESL 117G or 118G, many candidates will receive instruction after placement. The AEOCPT could be administered after candidates have received instruction to determine if the test captures class movement. If the LTA results show no class movement, an additional benefit of the analysis is item level information. Researchers can view where class movement did or did not occur for each indicator.

7.6 Summary

In this chapter, I suggested possible contributions of my results to language testing literature. From RQ1’s (Four Factor Language Model) results, I confirmed a language knowledge framework that has been used in another ITA test but not assessed for fit. I also argued that reported test scores should closely match the structural model that best fits the data. From answering RQ2 (Testing the Four Class Placement Model), I showed that LCA is capable of adding diagnostic information to performance test. The results from RQ3 (Assessing the Role of the Four Tasks) validate the operationalization of an interaction-focused language framework. Along with operationalizing the framework, I also showed that context can be operationalized through performance tasks. This result contributes an additional piece to measuring the social dimension through language testing (Chapelle, 1998; McNamara, 1997, 2001).
After discussing the implications of three research questions, I proposed a method for determining the optimal amount of Likert scale points. The results from this method helped refine the data so that PROC LCA could better identify the best fitting class model. I concluded this chapter by noting some limitations in this dissertation that are opportunities for future studies. One limitation was that I did not account for rater variability like other researchers of performance tests. I proposed, however, that this could be addressed in LCA by the addition of covariates. Another limitation I did not address in this dissertation is the assumption of how students move across classes over time. The ITA program and I assume that students will progress from 117G to 118G and 118G to certification, but this could be empirically tested through Latent Transition Analysis.
Appendix A: Old AEOCPT Rubric

Oral Proficiency Grading Rubric

Pronunciation

0 Speaker is often not easily comprehensible due to one or a combination of the following: Consistent phonemic differences; Consistent inappropriate pitch falls or rises, or a lack of pitch differentiation in thought groups; Consistent inappropriate pattern of stress and unstressed syllables at the word and/or sentence level. Little to no evidence of monitoring. Requires significant and consistent negotiation by listener.

1 Speaker is occasionally not easily comprehensible due to one or a combination of the following: Consistent phonemic differences, especially for key words; Consistent inappropriate pitch falls or rises, or a lack of pitch differentiation in thought groups; Consistent inappropriate pattern of stress and unstressed syllables at the word and/or sentence level, especially in key phrases. Little to no evidence of monitoring. Requires consistent negotiation by listener, especially if key words are not comprehensible.

2 Speaker is occasionally not comprehensible due to one or a combination of the following: Some consistent phonemic differences; Some consistent inappropriate pitch falls or rises, or a lack of pitch differentiation in thought groups; Some inappropriate pattern of stress and unstressed syllables at the word and/or sentence level. Little to no evidence of monitoring. Requires some negotiation by listener, especially if key words are not comprehensible.

3 The speaker is comprehensible even though there may be some consistent phonemic differences, and/or some consistent inappropriate pitch falls or rises, or lack of pitch differentiation in thought groups, and/or some inappropriate pattern of stresses and unstressed syllables at the word and/or sentence level. Negotiation required by listener is minimal.

Fluency

0 Speaker is often not easily comprehensible due to one or a combination of the following: numerous pauses and/or inappropriate pauses within thought groups; overuse of fillers; slow and strained speaking. Requires significant and consistent negotiation by listener.

1 Speaker is occasionally not easily comprehensible due to one or a combination of the following: numerous pauses and/or inappropriate pauses within thought groups; overuse of fillers; slow and hesitant speaking but for routine lexical phrases. Requires consistent negotiation by listener.
2 Speaker is **occasionally not comprehensible** due to one or a combination of the following: some pauses and/or inappropriate pauses within thought groups; some distracting use of fillers; choppy speaking. Requires **some** negotiation by listener.

3 Speaker is **comprehensible** even though there may be some pauses and fillers; Generally smooth. Negotiation required by listener is **minimal**.

**Comprehensibility**

0 Speaker is **often not easily comprehensible** due to one or a combination of the following: limited grasp of vocabulary. Speaker uses inappropriate vocabulary to accomplish task, simple vocabulary or repetition or circumlocution frequently; lack of grammatical control or use of very simple and/or repetitive grammatical structures. Frequent global and/or local errors. Requires **significant and consistent** negotiation by listener.

1 Speaker is **occasionally not easily comprehensible** due to one or a combination of the following: inappropriate vocabulary to accomplish task, simple vocabulary, or repetition or circumlocution; lack of grammatical control; complex grammatical structures are not used or are misused. Requires **consistent** negotiation by listener, especially at key words.

2 Speaker is **occasionally not comprehensible** due to one or a combination of the following: inappropriate vocabulary to accomplish task, circumlocution, or redundancy; occasional errors in choice of vocabulary items. May evidence some ability to paraphrase. Requires **some** negotiation on part of listener, and listener may **impose** meaning at times. May show evidence of monitoring.

3 Speaker is **comprehensible** even though there may be some local errors in grammar. Speaker uses generally appropriate and extensive vocabulary to accomplish task successfully. Uses varies and complex grammatical structures. May show evidence of monitoring. May evidence ability to paraphrase. Requires **minimal** negotiation by listener.
<table>
<thead>
<tr>
<th></th>
<th>1 (115G)</th>
<th>2 (117G)</th>
<th>3 (118G)</th>
<th>4 (Pass)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammatical Structures</strong></td>
<td>Generally incorrect and labored structures</td>
<td>Consistently incorrect constructions; some self-correction</td>
<td>Noticeable incorrect structures; successful self-correction if necessary</td>
<td>Infrequent incorrect structures; successful self-correction if needed</td>
</tr>
<tr>
<td><strong>Thought Groups</strong></td>
<td>Unintelligible thought groups and/or volume</td>
<td>Limited effectiveness of thought groups and/or inappropriate volume</td>
<td>Occasional difficulties with thought groups, or appropriate volume</td>
<td>Uses thought groups adequately with appropriate volume</td>
</tr>
<tr>
<td><strong>Tone Choice</strong></td>
<td>Monotone</td>
<td>Mostly falling, rising, or level tones</td>
<td>Uses rising and falling tones but occasionally misleading</td>
<td>Uses rising, falling, and level tones in appropriate manner</td>
</tr>
<tr>
<td><strong>Multimodality</strong></td>
<td>Minimal eye contact, consistent body orientation away from audience; poor use of nonverbal resources</td>
<td>Frequent retreat to body position away from audience and insufficient use of nonverbal resources</td>
<td>Infrequent periods of orientation away from audience and supplemented by effective use of nonverbal resources</td>
<td>Negligible orientation away from audience and effective use of nonverbal resources</td>
</tr>
<tr>
<td><strong>Transition</strong></td>
<td>Minimal transitions used</td>
<td>Task answer organized with restricted use of transition</td>
<td>Task answer organized but varied transition are lacking</td>
<td>Task answer organized with appropriate and varied transition</td>
</tr>
<tr>
<td><strong>Prominence</strong></td>
<td>No key words identified through stress or non-verbal resources</td>
<td>Minimally effective identification of key words through stress and/or non-verbal resources</td>
<td>Consistent identification of key words through stress and/or non-verbal resources</td>
<td>Effective identification of key words via stress throughout task</td>
</tr>
<tr>
<td><strong>Task Response</strong></td>
<td>Initial answer does not address the task</td>
<td>Task response pragmatically inappropriate and/or several gaps of information</td>
<td>Completes task with pragmatic appropriateness; any gap of information requires minimal repair</td>
<td>Completes task in pragmatically appropriate manner with no repair needed</td>
</tr>
<tr>
<td><strong>Question Response</strong></td>
<td>Answer does not address raters' concerns; other-initiated repair unsuccessful</td>
<td>Partial answer even after two turns of negotiation</td>
<td>Answer negotiation across two or more repair sequences; successful completion</td>
<td>Resolved with minimal repair</td>
</tr>
</tbody>
</table>
### Appendix C: Polychoric Correlation Matrix

Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Item1</th>
<th>Item2</th>
<th>Item5</th>
<th>Item6</th>
<th>Item3</th>
<th>Item4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item2</td>
<td>0.71</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item5</td>
<td>0.64</td>
<td>0.63</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item6</td>
<td>0.66</td>
<td>0.68</td>
<td>0.77</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item3</td>
<td>0.64</td>
<td>0.73</td>
<td>0.70</td>
<td>0.77</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Item4</td>
<td>0.48</td>
<td>0.56</td>
<td>0.59</td>
<td>0.60</td>
<td>0.58</td>
<td>1.00</td>
</tr>
<tr>
<td>Item7</td>
<td>0.63</td>
<td>0.67</td>
<td>0.78</td>
<td>0.72</td>
<td>0.68</td>
<td>0.63</td>
</tr>
<tr>
<td>Item8</td>
<td>0.68</td>
<td>0.67</td>
<td>0.69</td>
<td>0.71</td>
<td>0.68</td>
<td>0.58</td>
</tr>
<tr>
<td>Item9</td>
<td>0.82</td>
<td>0.68</td>
<td>0.59</td>
<td>0.60</td>
<td>0.62</td>
<td>0.44</td>
</tr>
<tr>
<td>Item10</td>
<td>0.70</td>
<td>0.79</td>
<td>0.65</td>
<td>0.66</td>
<td>0.72</td>
<td>0.49</td>
</tr>
<tr>
<td>Item13</td>
<td>0.70</td>
<td>0.68</td>
<td>0.72</td>
<td>0.70</td>
<td>0.72</td>
<td>0.54</td>
</tr>
<tr>
<td>Item14</td>
<td>0.65</td>
<td>0.70</td>
<td>0.70</td>
<td>0.78</td>
<td>0.77</td>
<td>0.55</td>
</tr>
<tr>
<td>Item11</td>
<td>0.61</td>
<td>0.63</td>
<td>0.68</td>
<td>0.72</td>
<td>0.82</td>
<td>0.51</td>
</tr>
<tr>
<td>Item12</td>
<td>0.53</td>
<td>0.55</td>
<td>0.56</td>
<td>0.60</td>
<td>0.62</td>
<td>0.67</td>
</tr>
<tr>
<td>Item15</td>
<td>0.58</td>
<td>0.58</td>
<td>0.62</td>
<td>0.59</td>
<td>0.58</td>
<td>0.46</td>
</tr>
<tr>
<td>Item16</td>
<td>0.63</td>
<td>0.67</td>
<td>0.59</td>
<td>0.57</td>
<td>0.61</td>
<td>0.51</td>
</tr>
<tr>
<td>Item17</td>
<td>0.83</td>
<td>0.64</td>
<td>0.60</td>
<td>0.59</td>
<td>0.64</td>
<td>0.43</td>
</tr>
<tr>
<td>Item18</td>
<td>0.70</td>
<td>0.80</td>
<td>0.66</td>
<td>0.66</td>
<td>0.73</td>
<td>0.48</td>
</tr>
<tr>
<td>Item21</td>
<td>0.67</td>
<td>0.61</td>
<td>0.75</td>
<td>0.67</td>
<td>0.70</td>
<td>0.54</td>
</tr>
<tr>
<td>Item22</td>
<td>0.66</td>
<td>0.69</td>
<td>0.70</td>
<td>0.79</td>
<td>0.76</td>
<td>0.56</td>
</tr>
<tr>
<td>Item19</td>
<td>0.63</td>
<td>0.65</td>
<td>0.68</td>
<td>0.74</td>
<td>0.83</td>
<td>0.52</td>
</tr>
<tr>
<td>Item20</td>
<td>0.51</td>
<td>0.55</td>
<td>0.52</td>
<td>0.58</td>
<td>0.61</td>
<td>0.69</td>
</tr>
<tr>
<td>Item23</td>
<td>0.64</td>
<td>0.62</td>
<td>0.63</td>
<td>0.63</td>
<td>0.62</td>
<td>0.52</td>
</tr>
<tr>
<td>Item24</td>
<td>0.69</td>
<td>0.68</td>
<td>0.61</td>
<td>0.62</td>
<td>0.64</td>
<td>0.55</td>
</tr>
<tr>
<td>Item</td>
<td>Item7</td>
<td>Item8</td>
<td>Item9</td>
<td>Item10</td>
<td>Item13</td>
<td>Item14</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Item7</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item8</td>
<td>0.81</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item9</td>
<td>0.59</td>
<td>0.63</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item10</td>
<td>0.65</td>
<td>0.64</td>
<td>0.74</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item13</td>
<td>0.71</td>
<td>0.69</td>
<td>0.74</td>
<td>0.76</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Item14</td>
<td>0.69</td>
<td>0.64</td>
<td>0.68</td>
<td>0.77</td>
<td>0.87</td>
<td>1.00</td>
</tr>
<tr>
<td>Item11</td>
<td>0.62</td>
<td>0.57</td>
<td>0.62</td>
<td>0.74</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>Item12</td>
<td>0.55</td>
<td>0.58</td>
<td>0.57</td>
<td>0.64</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td>Item15</td>
<td>0.61</td>
<td>0.59</td>
<td>0.67</td>
<td>0.64</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>Item16</td>
<td>0.63</td>
<td>0.67</td>
<td>0.70</td>
<td>0.69</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>Item17</td>
<td>0.60</td>
<td>0.66</td>
<td>0.84</td>
<td>0.72</td>
<td>0.70</td>
<td>0.65</td>
</tr>
<tr>
<td>Item18</td>
<td>0.65</td>
<td>0.67</td>
<td>0.71</td>
<td>0.84</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Item19</td>
<td>0.64</td>
<td>0.64</td>
<td>0.63</td>
<td>0.71</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>Item20</td>
<td>0.54</td>
<td>0.59</td>
<td>0.51</td>
<td>0.58</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Item21</td>
<td>0.66</td>
<td>0.67</td>
<td>0.67</td>
<td>0.70</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td>Item22</td>
<td>0.65</td>
<td>0.65</td>
<td>0.66</td>
<td>0.76</td>
<td>0.76</td>
<td>0.86</td>
</tr>
<tr>
<td>Item23</td>
<td>0.64</td>
<td>0.64</td>
<td>0.63</td>
<td>0.71</td>
<td>0.71</td>
<td>0.76</td>
</tr>
<tr>
<td>Item24</td>
<td>0.54</td>
<td>0.59</td>
<td>0.51</td>
<td>0.58</td>
<td>0.65</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Correlation Matrix
<table>
<thead>
<tr>
<th>Item</th>
<th>0.70</th>
<th>0.69</th>
<th>0.68</th>
<th>0.68</th>
<th>0.75</th>
<th>0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item23</td>
<td>0.70</td>
<td>0.72</td>
<td>0.70</td>
<td>0.70</td>
<td>0.75</td>
<td>0.68</td>
</tr>
<tr>
<td>Item24</td>
<td>0.56</td>
<td>0.59</td>
<td>0.82</td>
<td>0.68</td>
<td>0.67</td>
<td>0.62</td>
</tr>
<tr>
<td>Item25</td>
<td>0.61</td>
<td>0.61</td>
<td>0.75</td>
<td>0.80</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Item26</td>
<td>0.68</td>
<td>0.67</td>
<td>0.66</td>
<td>0.71</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>Item27</td>
<td>0.67</td>
<td>0.65</td>
<td>0.70</td>
<td>0.74</td>
<td>0.77</td>
<td>0.85</td>
</tr>
<tr>
<td>Item28</td>
<td>0.68</td>
<td>0.67</td>
<td>0.68</td>
<td>0.75</td>
<td>0.75</td>
<td>0.80</td>
</tr>
<tr>
<td>Item29</td>
<td>0.59</td>
<td>0.59</td>
<td>0.55</td>
<td>0.63</td>
<td>0.63</td>
<td>0.66</td>
</tr>
<tr>
<td>Item30</td>
<td>0.64</td>
<td>0.61</td>
<td>0.66</td>
<td>0.71</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>Item31</td>
<td>0.61</td>
<td>0.65</td>
<td>0.69</td>
<td>0.71</td>
<td>0.71</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Appendix D: LCA Syntax

PROC LCA DATA=WORK.DISSertation;
NCLASS 3;
ITEMS Item_1 Item_2 Item_3 Item_4 Item_5 Item_6 Item_7 Item_8;
CATEGORIES 3 3 3 3 3 3 3 3;
SEED 4893;
NSTARTS 100;
GROUPS Tasks;
RHO PRIOR = 1;
RUN;
References


Cook, C., Heath, F., Thompson, R. L., & Thompson, B. (2001). Score reliability in web-


Jeremy Gevara  
Curriculum Vitae

Education

2012-Aug. 2016 Pennsylvania State University, State College PA  
PhD. in Applied Linguistics Minor in Educational Psychology

2010-2011 Texas Tech University, Lubbock TX  
M.A. in Applied Linguistics

2003-2008 Texas Tech University, Lubbock TX  
B.A. in Psychology, Minor: Chemistry, Magna Cum Laude

Teaching Experience

Aug. 2012-Present Pennsylvania State University Graduate Assistant  
Courses Taught: IECP Level 4 OC, IECP Level 4 Reading, IECP Levels 3/4 Humanities, IECP Level 3 Reading, APLNG 583 Methods of Language Assessment, APLNG 593 Experimental Research on Language, ESL 118G American Oral English for ITAs III

Jul. 2010-Aug. 2012 Texas Tech University Teaching Assistant  
Courses Taught: IEP 100 Listening/Speaking, IEP 100 Reading/Writing, IEP 200 Listening/Speaking, IEP 300 Listening/Speaking, IEP 400 Listening/Speaking, IEP 500 Listening/Speaking, IEP TOEFL/IETS Prep, ITA Summer Workshop, ESL 5310 Spoken English Fluency

Select Refereed Publications

