THE EFFECTS OF INTERLOCUTORS ON STUDENT PERFORMANCE ON CONSTRUCTED DIALOGUE TASKS ASSESSING PRIMARY PHRASE STRESS PRODUCTION

BY

RYAN D BOYD

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Arts in Teaching of English as a Second Language
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Advisers:

Professor Fred Davidson
Professor Wayne Dickerson

Abstract

This study attempted to identify the effects that performance variation of interlocutors has on examinees in constructed dialogue tasks assessing primary phrase stress and intonation usage in American English by working with the University of Illinois at Urbana-Champaign's (UIUC) oral section of its English placement test (EPT). This study developed a comparable version of a specific section of oral section of the EPT's specification which was administered to 47 incoming UIUC students. The participants were split into two groups and tested using two different types of interlocutor performance—one which adapts features identified in studies on spontaneous into the interlocutor performance and one which requires the interlocutor to read aloud. The two groups were further homogenized through separating the participants who were required to participate in phase II of UIUC's oral section of the EPT from the participants who were not required to participate in phase II. By calculating mean difficulty from the number of correctly produced targets out of the total number of targets for each target, linguistic focus, and task, this study found that the specific group of participants required to take phase II of the oral section of the EPT receiving the adapted performance had a lower mean difficulty overall indicating that the test was easier for them while the groups who were not required to participate in phase II performed similarly in terms of mean difficulty. This study recommends future research for the EPT to further understand these findings.

Acknowledgements

This study would not have been possible without the help and support of my adviser, Fred Davidson. Feedback on multiple versions from both Wayne Dickerson and Fred Davidson proved invaluable in developing this thesis. I owe them much for their time and effort. I am also sincerely grateful for their confidence in allowing me to have the unique experience to serve as a rater for the EPT during my time at the University of Illinois. Without that opportunity, this work would not exist.

Table of Contents

Chapter 1: Introduction 1

Chapter 2: Literature Review 4

Chapter 3: Methodology 19

Chapter 4: Results 31

Chapter 5: Discussion 51

References 67

Appendix A: IRB Submission 70

Appendix B: Test Specification 80

Appendix C: Experimental Test 88

Appendix D: Linguistic Foci in Context 90

Appendix E: Figures 91

Chapter 1: Introduction

The University of Illinois at Urbana-Champaign has seen an increase in its international student population in recent years. In fact, currently, international students comprise 14% of the undergraduate population, 43% of the graduate population, and contribute to 20.3% of the entire student population at the university (International and Student Scholar Services, 2012). The Fall semester of 2012 the university saw 2,678 international students admitted which is significantly higher than the 1,534 five years ago in the 2007 (International Student and Scholar Services, 2012). For many of these international students, English is not their native language. In these cases a campus-based English placement test (EPT) is used to place students into service courses to acclimate them to their new academic environment. This increasing number of students taking the university's placement test draws much attention to the exam, which continues to serve as a successful tool to place students into English as a second language (ESL) courses suitable for their needs.

As a rater for both the written and oral parts of the EPT, I have been working with this test for nearly three years. It was my experience as a rater which motivated me to pursue this study. The oral section of the EPT typically begins at 1:00 PM and proceeds until all the students scheduled for the test day have completed the exam. Some test days require raters to work more hours than others. During one long day of rating for the oral section of the EPT, I became fatigued. While I did not sense that this affected my ability to rate, I did, however, speak to the students and read text within the test with less enthusiasm. One student noticed this, and during a section of the test in which he and I read scripted lines in constructed dialogues, he performed exceptionally and spoke his lines as if he were an actor on stage. I found that his

performance compelled me to perform with just as much vigor. When the test finished, I asked him if he had studied theater in the past. He told me that he had not and asked why I asked such a strange question. I responded by praising his performance on the scripted dialogue section of the test and explaining to him that it was the reason I had asked the question.

After the testing for the day had finished, I reflected on that experience and thought that certainly if he could influence me through his performance, I could influence the performance of others as well. In later ratings I tried different performances and found that in some cases students seemed to resonate with the performance and in other cases it seemed there was no effect. At that time I had only my informal observations and no sturdy support for whether the dramatized performance enhanced student performance or whether it had indeed had no effect. This curiosity is what prompted me to study this situation more formally.

This research is focused on the university's oral component of the EPT. The oral part of the EPT is composed of two phases, phase I and phase II. Phase II contains several sections, one of which contains a constructed dialogue reading task in which both the candidate and the rater read scripted lines on the test. Both the rater's lines and the student's lines are used to make a dialogue. The student's lines contain specific primary phrase stress and intonation pattern targets in English. Only the rater's dialogue sheet indicates the targets. The student does not know what the targets are nor where they are. In phase II of the oral part of the EPT, the rater becomes an interlocutor for the purpose of conducting the oral component of the EPT. Thus, this research explores the effects of the role of the interlocutor on student performance. Because the current oral part of the EPT test specification does not provide information regarding the manner in which an interlocutor should perform in the constructed dialogue task, it is necessary to determine what effects, if any, variable interlocutor performance has on student response in order

to explore the validity of such tasks. This research is exploratory in nature and seeks to understand whether a phenomenon exists or not and how, if at all, the findings of this study affect the oral part of the EPT. To accomplish this, two research questions are posed which this study attempts to answer.

1. What effects, if any, does an interlocutor's performance have on a test taker's response in constructed dialogue tasks?
2. How, if at all, do the findings for the first research question affect the oral section of the EPT?

Chapter 2: Literature Review

As this study focuses on a niche of testing pronunciation, there is little related literature. However, this literature review will review criticisms of interlocutors in oral proficiency interviews to establish a basis for identifying interlocutors as potential variables in testing language. It will then consider the interlocutor as a test facet which could be controlled and evaluate the consequences of controlling such a facet to provide a background for discussing the effects the effects this study may have on the oral part of the EPT's test specification. Dynamic assessment's interaction between examiner and examinee is reviewed to establish possible areas of future research on examinee performance on scripted dialogue pronunciation tests resulting from this study. Lastly, this literature review contains a description of the oral section of the EPT to develop a greater level of transparency for this study and to identify why this type of pronunciation test has been both criticized and praised.

Oral Proficiency Interview

There has been much discussion surrounding the validity of oral proficiency interviews (OPI) and often times the discussion is juxtaposed with an evaluation of simulated oral proficiency interviews (SOPI) (Stansfield and Kenyon, 1992a; Stansfield and Kenyon, 1992b). The focal point of the discourse as it relates to this research becomes the direct nature of the OPI. By definition, the OPI requires an interlocutor rater to directly interact with an examinee (Chalhoub-Deville and Fulcher, 2003; Fulcher, 2003). However, a movement for independence from interlocutor raters arose in part from a low supply of OPI raters trained to administer

interviews for less commonly taught languages (Stansfield and Kenyon, 1992b).[1] With independence from interlocutor raters oral proficiency testing of less commonly taught languages would be more accessible to examinees. The independence came in the form of the SOPI, but with the development of the SOPI, came a need to evaluate its ability to measure an examinee's speaking proficiency as a "surrogate" of the OPI (Stansfield and Kenyon, 1992b).

The OPI and the SOPI differ in that the SOPI does not include an interlocutor as a means to elicit spoken language from an examinee. Instead, tape recordings along with visuals are used to elicit responses from examinees (Stansfield and Kenyon, 1992b). Therefore, the interlocutor assumes a questionable role in the reliability of the OPI in relationship to the SOPI which lacks an interlocutor, and as a result, the contribution of this role to the validity of a speaking assessment comes under scrutiny.

Fulcher (2003) indicated that the interlocutor may actually function as a confounding variable in the OPI, which suggests that the inclusion of an interlocutor into a speaking assessment could affect its reliability. This results from the unique experience examinees have when interacting with different interlocutors. Ideally, each examinee's OPI would result in comparable discourse contributed by the interlocutor, however, this is not always the case (Chalhoub-Deville and Fulcher, 2003; Stansfield and Kenyon, 1992b), which surfaces in Stansfield and Kenyon's (1992a) analysis of the comparability of the OPI and the SOPI's reliability. They reveal the possibility that examinees can produce divergent OPI performances as a result of the variation in questions supplied by the interlocutor for the examinee (Stansfield and Kenyon, 1992a). Examinee OPI performances could then vary when rated multiple times by the same interlocutor, when rated by two different interlocutors, and when examinees at

[1] *The OPI* is often used to refer to the adaptive one-on-one testing developed by the United States government. However, *OPIs* is used generically to refer to any face-to-face oral test.

comparable language proficiency are rated differently by the same interlocutor (Stansfield and Kenyon, 1992a).

The SOPI may mitigate this effect to some extent by eliminating the variation in discourse by providing each examinee with comparably the same testing experience that is less assured by the OPI. The SOPI does this by using alternative methods of eliciting examinee response to interview questions. Stansfield and Kenyon tested this by comparing the inter-rater reliability of SOPIs of six less commonly taught languages (and the test-retest reliability by using different raters and different SOPIs for each language) to the test-retest reliability among three government agencies in the United States that use the OPI for French and German (1992a). They found that the six less commonly taught language SOPIs exhibited greater test-retest reliability coefficients overall than that of the French and German OPIs used in a comparable study (Stansfield and Kenyon, 1992a). The SOPI coefficients ranged from .84 to .98 with an average of .92 whereas the OPI coefficients ranged from .84 to .92. with an average of .88 (Stansfield and Kenyon, 1992a).

These results serve as an indication that the SOPI is capable of producing greater reliability results when compared to the OPI as a result of the absence of an interlocutor as a rater (Stansfield and Kenyon, 1992a). However, the increased reliability does not come without a cost. By standardizing many facets in an SOPI, it allows a greater amount of what Bachman calls "systematic errors" to affect validity for all test takers (Bachman, 1988, p. 153).

A second study by Stansfield and Kenyon (1992b) on the validation of the SOPI presented findings that did not align with the results presented in the study on the comparability of the OPI and the SOPI. The study focused on the Indonesian Speaking Test (IST), a semi-direct SOPI, developed by the Center for Applied Linguistics (CAL). The IST consists of five

sections. In the first part the examinee listens to short recorded questions in the target language about things related to his life. The next part asks the examinee to give directions from one location to another as marked on a map. Part three requires the examinee to provide a narrative explanation of a series of three pictures representing past, present, and future. In part four, the examinee is instructed to talk to six different Indonesians. The examinee is given a short prompt for each Indonesian explaining the topic of that talk. In the final task, the examinee is given an audience and a speaking task. The examinee must then complete the task.

This test was used in a validation study by CAL. The study used two comparable forms of the IST and one OPI. The tests were administered to 16 learners of Indonesian with varying exposure to the language and each test was scored by two raters, so each learner had four SOPI scores (two scores for each test form) and two scores for each OPI. Using generalizability theory framework, this study found that there was a strong indication that the SOPIs "can be rated reliably and consistently" (Stansfield and Kenyon, 1992b). However, the results also indicated that examinees' scores varied across different versions of the IST while the variation between the IST and the OPI was minimal. These two factors suggest that the much of the difference in scores came as a result of examinees performing differently on all three tests and not from the difference in testing method between the OPI and SOPI.

The results showed that score variation was most likely a product of the different versions of the test and not necessarily a result of the difference between the OPI and the SOPI.

Based on the findings from research conducted by Stansfield and Kenyon, it is not clear to what extent the interlocutor affects examinee scores. Their study on the comparison of reliability between the OPI and the SOPI yielded results indicating that the interlocutor acts as a confounding variable in the reliability scores of the OPI (Stansfield and Kenyon, 1992a).

However, a study assessing the comparability of the OPI with an SOPI, the IST, by Stansfield and Kenyon suggests that the interlocutor is not the source of variance in reliability scores. In fact, Stansfield and Kenyon claim that the variance found in their study "is not due to inconsistencies between the raters; there were no real consistent differences in raters or in raters' interaction with the three tests. Raters consistently applied the same standards across tests" (1992b, p. 140-141). Instead, the multiple versions of the IST and OPI are thought to act as the confounding variable in this study (Stansfield and Kenyon, 1992b). They posit that this is expected in such a study given that three long tests were administered to participants. Because of this, the many factors may affect the results such as test taker motivation to perform well on the test and test taker fatigue (Stansfield and Kenyon, 1992b). The findings of these two studies seem to conflict somewhat, thereby highlighting the need for further study into the effects of interlocutor behavior on examinee performance in oral assessment.

Test Facets

Aside from comparison with the SOPI, the OPI is still criticized for reasons similar to those that arose out of the comparison of the two tests (Bachman, 1988; Chalhoub-Deville and Fulcher, 2003). Bachman (1988) identified gaps in the validation of OPI testing practices which could affect measured abilities and provided a framework for which these gaps could be addressed. He claimed that because of these gaps, the validity of the OPI could not be assessed. Bachman supported this by arguing that in order to evaluate content validity, the conditions, or facets, including methods and practices which are imposed upon the examinee and administered to elicit responses must be made transparent in order separate content from test design (1988). This distinction is necessary because test design facets should not be measured and examinee

performance on content should be measured (Bachman, 1988). This transparency is necessary in evaluating tests because even slight differences in elicitation methods or interaction between rater and examinee have potential to affect individual examinee performances (Bachman, 1988). Thus, when "method facets…vary from interview to interview in an uncontrolled manner, they are sources of random measurement error" (Bachman, 1988, p. 153). He suggests training raters on specific elicitation practices for oral interviews is one method to reduce such measurement errors and make the test more similar across examinees, but warns that by standardizing such procedures, "systematic errors" become engrained in the test (Bachman, 1988, p. 153). Therefore, a clear understanding of the types of error stemming from testing procedure and the extent to which they may affect measurement is necessary to determine what to control.

Bachman's suggested framework for developing an OPI that addresses validity concerns related to test design focuses on five distinct categories, "(a) the testing environment, (b) the nature of the test instructions, (c) the nature of the input the test taker receives, (d) the nature of the response to that input, and (e) the interaction between input and response" (1988, p. 157). By developing an OPI through the detailing of the categories presented in this framework, Bachman believes such tests can take strides towards accountability for their validation.

Similar concerns are expressed by Chalhoub-Deville and Fulcher (2003). They emphasize the demand for OPI scores and argue that ACTFL has a responsibility to stakeholders to "provide high-quality information and…to provide evidence that validates the intended interpretations and uses of these ratings" (Chalhoub-Deville and Fulcher, 2003, p. 501).

They note that reliability is an important element in furthering research on ACTFL's OPI and that by creating a research plan to improve its test's reliability, measurement errors resulting

from inconsistencies in test facets can be minimized (Bachman, 1988; Chalhoub-Deville and Fulcher, 2003). Thus, a test can be closer to capturing an examinee's true score (Chalhoub-Deville and Fulcher, 2003).

In order to control facets such as interlocutors and OPI scoring system, Chalhoub-Deville and Fulcher (2003) suggest that generalizability theory framework should be applied to such research. This framework will allow for a simultaneous comparison of variations in test facets as seen in Stansfield and Kenyon's (1992b) research on the validation of the SOPI. Error sources can then be represented side-by-side so that their effect on measurement can be weighted and assessed more appropriately which allows test developers to discern which facets to control to balance systematic measurement errors and random measurement errors (Bachman, 1988). Generalizability theory will allow ACTFL's research to develop its OPI further instead of simply scratching the surface with inter-rater reliability research. By utilizing such an analysis on research, a deeper understanding of the effects and interaction of test facets in the OPI can be understood (Chalhoub-Deville and Fulcher, 2003).

Bachman (1988) and Chalhoub-Deville and Fulcher (2003) conclude that ACTFL needs to devote more resources to developing substantial research to support the interpretations of its OPI and improve its validity.

Dynamic Assessment

Transparency of test practices, as urged by Bachman (1988) and Chalhoub-Deville and Fulcher (2003), can be intrinsically beneficial for a test. Additionally, by making testing tasks transparent to stakeholders such as examinees, a richer view of examinee ability is obtainable. This can be accomplished through interaction between an examiner and an examinee as

displayed in Antón's work on dynamic assessment in an undergraduate Spanish foreign language program in the United States (Antón, 2009). A broader look at test facets affecting student performance reveal that dynamic assessment methods lend themselves to fostering interaction between examiner and examinee and yields benefits for both parties (Antón, 2009).

Antón looked at the implementation of dynamic assessment in the context of a United States university's diagnostic exam for Spanish major students. The exam is administered to students as they enter into their third year of the program and can be compared with program entrance exam data to assess student learning. The third-year diagnostic exam consists of five sections focused on the following areas of Spanish ability: grammar and vocabulary, listening comprehension, reading comprehension, writing, and speaking. The writing and speaking sections are adapted to dynamic assessment practices.

The writing section gives students 20 minutes to write an essay about their experiences with the Spanish language. After 20 minutes, they are told to make revisions and are given a Spanish language dictionary, a Spanish language reference grammar, and the opportunity to ask the test examiner questions about their composition.

The speaking section consists of four parts. The first section functions like an interview. The examiner asks the student questions and the student responds. The second section requires the student to iterate a story in the past based on a series of pictures. This section is broken down into three phases dependant on student performance. Phase one is where the student first details the story. If the student does not meet examinee expectations, phase two will follow. In phase two, the student is able to attempt the task a second time. However, before beginning, the student receives suggestions for improvement from the examiner. If the student is still not performing well enough, phase three begins. In phase three, the examiner narrates the picture

story as a model and then offers one last opportunity for the student to retry the task. In section three, the student must assume the role of a character from the previous section and say something that would be appropriate for the chosen character. The final section asks students to select one of the provided topics and speak for three minutes about the topic. After the student's turn, the examiner asks questions related to the content of the student's speech.

Five students participated in the study resulting in detailed qualitative scoring narratives. Each of the students participating in the study took section two of the speaking section of the Spanish test. The oral interview was videotaped and later scored.

The student responses to section two of the speaking section provide interesting results related to dynamic assessment. Analysis of these responses illustrates a level of transparency about the test task achieved between the examiner and the examinee as a result of the dynamic assessment. In this case, the task required students to narrate a story based on a series of pictures. This task was designed to elicit use of the past tense in Spanish by requiring examinee's to begin their story with the word *yesterday*. One student began section two by speaking in past tense and then switched to present tense in the middle of the narration of the pictures. Another student began in present tense and continued using present tense throughout the picture narration task. Antón stated that on the surface, it may seem as if both students lack control over their usage of past tense in Spanish, but because of the interactive nature of dynamic assessment, a deeper understanding of their performance using past tense was obtained (Antón, 2009). In both cases the examiner identified the problem in the performance of the student after they had completed phase one of speaking section two. The students then moved on to phase two in which they attempted the task a second time. The student who switched to present tense halfway through the task in phase one was able to sustain a description of a series of events in

past tense in phase two. The student who narrated the picture sequence in present tense in phase one did change narration to past tense in phase two although the student's overall ability to control verb tense usage was lower than that of the other student.

Interaction occupies a major role in dynamic assessment (Antón, 2009). As seen through a comparison of student performance of phase two of speaking section two with phase one of speaking section two, an increase in ability seems to occur after interacting with the examiner. Thus, it may be concluded that the initial assessment did not provide an accurate representation of the students' language abilities, and because of the interaction in phase two of speaking section two, the test task became more transparent to the students, which allowed a more accurate representation of the their language ability to be obtained (Antón, 2009). This is particularly prominent in the case of the student who did not use past tense in phase one but did use past tense in phase two after the examiner indicated that the student did not meet the requirements of the task because past tense was not used. By providing examinees with the opportunity to improve performance when mistakes occur, a richer display of language ability emerges, which acts as data that an advisor would then be able to review in order to provide more suitable, individualized methods with which the student can improve language ability (Antón, 2009; 2012).

However, as previously indicated, oral interviews with such an individualized test experience as a result of interaction between examiner and examinee would face issues with reliability (Bachman, 1988; Chalhoub-Deville and Fulcher, 2003; Fulcher, 2003; Stansfield and Kenyon, 1992a). This produces an apparent tradeoff: test developers can encourage transparency among the examiner, the examinee, and the test through individualized interaction at the expense that such interaction would act as a confounding variable in reliability analyses, or test

developers can restrict individualized interaction between examiner and student in fear that such interaction would then become a confounding variable in reliability testing at the expense of reducing transparency between the examiner, the examinee, and the test (Antón, 2012).

This tradeoff between reliability and richness of performance is well-illustrated even within the methodology of dynamic assessment. There are two major types of dynamic assessment—interactionist and interventionist. In the interactionist approach to dynamic assessment the examiner (mediator) must make judgments regarding the extent of assistance that is believed to be necessary for the examinee. The interventionist approach to dynamic assessment is the type of dynamic assessment used in the assessment of advanced Spanish learners. Although there is interaction between mediator and examinee, the intervention (phase two) is given a strict formula and does not vary in its effects on student performance to the extent that the interaction method does. It is clear that the tradeoff between reliability and richness of performance exists between dynamic assessment and non dynamic assessment. However, given that the same issue also emerges across different types of dynamic assessment, it may not be unreasonable to attempt to incorporate finer shades of when balancing test characteristics.

This discussion reveals one clear and present issue to the surface which is that when human raters are involved in rating and facilitating an oral interview, there is room for performance variation. It can be seen in Fulcher's criticism of the OPI (2003). This theme also appears in Stansfield and Kenyon's reliability tests of the SOPI and the OPI in which the SOPI produced overall greater reliability coefficients (1992a). Rater performance variation was also targeted by Bachman (1988) in which he warned that method facets, if left unaccounted for, can negatively affect the reliability of assessments. ACTFL faced similar criticisms in from

Chalhoub-Deville and Fulcher (2003) who targeted ACTFL's OPI reliability on the basis of its test facets.

Whether a high large scale test is in question or a university placement test, it is an organization's responsibility to respond to such criticisms and questions with evidence from a research-based test development agenda. Without such a response, tests are left vulnerable to criticism from not only testers and researchers but also malcontent test takers.

The EPT

This research targets one of the currently uncontrolled facets in the oral part of the EPT to determine whether the rater's performance while reading scripted lines can affect a student's performance while reading scripted lines which are written as responses to what the rater as an interlocutor is required to read.

While much literature on OPIs focus on what raters say in less controlled interviews, there is little which focuses specifically on how a rater's performance can affect the candidate's responses in more controlled interviews where raters and candidates both follow a script. Reading aloud scripted pronunciation tests is problematic in that speaking ability and reading ability are distinct from each other, which would mean that taker's ability to read aloud would affect that test taker's performance (Lado, 1964). Similarly, reading aloud and speaking naturally require different skills (Madsen, 1983). Additionally, pronunciation by reading aloud is less likely to yield the same reductions (Lado, 1964) and other natural speech phenomena such as linking (Madsen, 1983) that are present in speaking. This means that tasks involving reading of pronunciation targets could provide a source of construct underrepresentation in pronunciation tests for communicative purposes. Nevertheless, these issues cannot totally discount scripted

pronunciation tests. Scripted pronunciation tests are highly efficient at testing specific pronunciation targets (Lado, 1964; Madsen, 1983) making them an extremely useful tool in large-scale testing of pronunciation. They may be particularly useful as part of the oral EPT used in this study as it is used to test many students in a relatively short amount of time.

The current oral part of the EPT interview used by the University of Illinois is used to identify international students whose pronunciation of English impedes their ability to successfully communicate in English in an academic setting. Their performance on the oral part of the EPT determines whether they are required or recommended to take a pronunciation course or whether they are exempt from taking a pronunciation course.

The oral part of the EPT is composed of two phases. In phase I, the examinee responds to the rater's general discussion question for three to five minutes. During this time the rater must listen to the examinee's speech and avoid participating which would transform the interview into a conversation. If the rater is unable to determine any part of the examinee's speech, the test proceeds onto phase II. If the rater could easily understand everything the examinee said, the examinee is exempted from the pronunciation course and not required to take phase II of the oral part of the EPT.

Phase II of the oral part of the EPT tests specific pronunciation targets which are taught in the pronunciation service courses. Thus, the content of the pronunciation course determines the content of the oral part of the EPT. It can be said that the oral part of the EPT is validated against the pronunciation service course, so examinees who score low on the oral part of the EPT would be taught pronunciation targets on which they would likely have difficulty in the pronunciation service course and examinees who score high on the oral part of the EPT would be taught pronunciation targets over which they likely have a high degree of proficiency. Phase II

of the oral part of the EPT is anchored in a well-established tradition. It follows a testing technique termed *reading* presented by Lado (1964) and later called *reading aloud* by Madsen (1983). In each section of the test, examinees are instructed to read aloud sentences which contain the pronunciation targets. The examinee is not informed of which elements of pronunciation are being tested and is only instructed to read each sentence twice as smoothly and as naturally as possible. Only the second reading is rated. Although this method is not flawless as previously discussed, Lado describes it as "The most uniform, precise, and simple method for testing production of the sound segments of a language…" (1964, p. 83). During the test, the examiner simply makes a mark on the scoring sheet through a hole above the respective target on an overlaid rater's exam sheet for each missed target. The marks are counted and the examinee is placed one of three ways, exempt from a pronunciation class, recommended to take a pronunciation class, or required to take a pronunciation class, according to the number of marks on the scoring sheet.

Raters for the oral part of the EPT are chosen based on several factors. A rater must have successfully completed the course English Phonology and Morphology for Language Teachers (EIL 488). Raters must also receive the approval of the instructor who leads the rater training as well as teaches the EIL 488 course on English phonology. Lastly, raters must annually participate in an oral part of the EPT recalibration session and complete the recalibration session prior to rating the oral part of the EPT.

This study focuses on the niche of how an interlocutor's performance can affect an examinee's performance in a section of phase II of the oral part of the EPT in which scripted dialogues are read aloud by the rater and the examinee to assess primary phrase stress and intonation usage. This section of the test was specifically selected because it is the only section

in which the rater takes on the role of an interlocutor and engages the examinee in a scripted conversation constructed to test primary phrase stress and intonation usage, which contributes to their score, so if the participants do not recognize that meaning is being created through the interaction in the discourse and read the script as if it were isolated lines of text, the performance of the examinees on these tasks may not represent their true abilities. Ayers describes this phenomenon by saying that "Even if the read speech is in the form of a multispeaker dialogue based on spontaneous conversation, it develops more as a series of monologues instead of as a true dialogue like the original" (1994, p. 3). The ability to recreate the feeling of a spontaneous conversation comes from the performance of the readers (Ayers, 1994). Certainly the instructions given are another important variable in this section of the test. However, due to the fact that this study was conducted by a single researcher, only the interlocutor performance variable is explored in order to control the scope of the study. Phase I of the oral section of the EPT was not selected for study because the raters should not engage the examinee in conversation during this part. Instead they are required to simply pose a question to the examinee and listen to his or her response for three to five minutes. Therefore, it is the aim of this research to identify whether variations in rater's performance in scripted dialogues like those used in the operational oral component of the EPT could result in examinee performance variation.

Chapter 3: Methodology

Participant Recruitment

The EPT research assistant notified international students required to take the university's EPT of an opportunity to participate in this research study through email. The email advertised participation in this research study as practice for the university's oral component of the EPT in that the tasks in this research study are reverse engineered (Fulcher & Davidson, 2007) from tasks that appear in the university's oral component of the EPT. The recruitment email is included in Appendix A. This helped to acquire participants and helped to more closely tether the results of the study to student performance on the University's EPT.

In the email students were instructed to contact this researcher who was conducting the study in order to schedule an interview. After establishing contact, students received a consent form to review and an interview availability spreadsheet. If students wished to participate after reviewing the consent form, they were instructed respond to the email and provide as much availability for open time slots as possible. This researcher then manipulated to schedule to maximize the number of available participants and emailed individual participants to instruct them where and when the research study would take place.

Test Development

The test used in this study was generated from reverse engineering (Fulcher & Davidson, 2007) tasks from the constructed dialogue section of the oral part of the EPT. A fine-grained test specification (Davidson, 2012) was developed for the tasks in the oral part of the EPT to identify the phonological rule being tested in each target (see Appendix B). This was accomplished

through developing each task to closely mimic a task from the university's oral component of the EPT in linguistic structure and pronunciation targets. Both the text that the interlocutor reads and the text that the participants read are similar in structure. These were developed through communication with a professor who teaches pronunciation service courses, trains raters for the oral section of the EPT, and serves as an instructor for a phonology course. This method was used to ensure tasks used in the research study are comparable to the original tasks. Targets which had acceptable variation in production were not marked for scoring in the study. Targets developed for this study which did not test the same phonological rule as the corresponding targets after which they were developed were revised based on feedback from the expert.

The communication consisted of email exchanges between this researcher and the expert. In the emails this researcher asked questions about what types the specific linguistic foci were being tested in each task of the original items in the constructed dialogue section of the EPT and requested feedback to determine whether the tasks developed for this study elicited the same targets in the text to be read by participants. Only the thematic content was changed significantly in order to maintain test security and to ensure that students who participate in the research study would not have a significant advantage over students who chose not to participate. The resulting changes based on the feedback are identified as multiple versions of the task in Appendix B.

The test specification itself contains several parts. A general description is provided to detail the mechanics of how the rater and examinee exchanges proceed. This section is followed by a listing of prompt attributes. This lists all of the primary phrase stress and intonation phenomena tested in the constructed dialogue section of the oral part of the EPT. However, this portion has been redacted from this study to maintain test security. A description of two types of

performances is included, which is unique to the test specification in this study. This identifies general characteristics of the performance based on spontaneous speech and the performance based on read speech. Each test task is developed through three parts: guiding language and revisions, operational test task, and experimental test task and revisions. There are four tasks used in this study and each contains guiding language which details which linguistic focus that each target in the operational test task and experimental test task assess. It also specifically identifies how the performance should be modified to fit the experimental conditions of this study. The development can be tracked across multiple versions of guiding language presented in the test specification. The operational test tasks were removed from this version to maintain test security. The experimental test tasks represent the tasks used in this study. Their development can be seen through changes across multiple versions.

Test specifications tend to contain sensitive information about the tests for which they were developed. Many are never released in a public domains. However, in order to achieve a level of transparency about the content used in this study and its development, the test specification is provide in Appendix B. By providing a test specification stakeholders can have a greater understanding the test and its validity. However, there is a trade off with test security. Finocchiaro and Sako clearly identify the importance of test security by stating, "Tests must be protected at all times from review by possible future examinees. This protection is required to provide an equal and fair chance of success on each test" (1983, p. 256). This statement assigns two roles to test administrators. First, they must protect tests from being reviewed by those who may take the test in the future. Secondly, test administrators must ensure that each examinee has an equal opportunity to perform well. For the purposes of this discussion, the latter responsibility will be limited to the extent of ensuring a test's security is not compromised.

It is important to note that the operational EPT is a secure test. Therefore, this study does not intend to violate the security of the oral section of the EPT. As a result, much information is redacted from this test specification.

Specifically, prompt attributes were removed from the test specification. This section was removed in order to prevent current students, who might have access to this document, from transmitting information about all of the pronunciation patterns tested in the constructed dialogue section of the oral part of the EPT to future students. This would be especially harmful to the security of the test as it identifies specific patterns of primary phrase stress tested in this section. If it were acquired by future students, it may allow them to prepare for this section of the test, which would give them an advantage over other students.

Most importantly, the operational test tasks were removed from this test specification. Releasing the operational test tasks publically represents one of the most significant threats to the security of this test. Students who may have access to this test specification would be able to be to rehearse specific parts of the script or listen to a native speaker's rendition of the script before the interview to become proficient at producing those parts native-like. Additionally, students would then be able to identify which specific targets are being tested in each dialogue and focus on producing those targets accurately thus reducing the ability of the oral part of the EPT to place students into pronunciation service courses accurately and providing an unfair advantage to some students.

To develop the test to be as similar to the oral part of the EPT as possible, the testing materials consisted of two versions of the test like the operational oral part of the EPT—a rater version and a participant version. The student version contained the same instructions at the top, same script line indicators, and same text organization and formatting as the original. There was,

however, a difference in the physical properties of the test. The student version of the test was printed on plain white paper instead of sturdy blue paper. The rater's version of the test contained the same text organization and formatting as well as the same target indicators and intonation target symbols. There are two differences to be noted. The first is that each word containing a target is preceded by a number in order to facilitate target identification. The second difference is that the rater version was printed on plain white paper instead of laminated orange paper. Models of rater and student versions of the test are included in Appendix C.

The variable in this study was the way in which the interlocutor's lines were performed. This resulted in two types of conditions, control conditions and experimental conditions. Features such as "pitch range, intonational contour, declination patterns, utterance duration, preboundary lengthening phenomena, pause patterns, speaking rate, and energy patterns" have been subjects of study of in discourse description fields and also in utterance detection in computational areas of research (Shriberg et al., 1998, p. 445). Blaauw (1994), however, notes that there is still some uncertainty as to which acoustic features allow for the distinction between spontaneous and read speech, but surmises, similarly to what Shriberg et al. identified, that prosodic information such as speed, intonation, and intensity are needed to distinguish between spontaneous and read speech. The conditions in the experimental group adapted some of these features studied in discourse to the script to be read by a rater. More specifically, pitch, lengthening, pauses, speech rate, and energy patterns were manipulated and added to the experimental performance to make it distinct from the control performance. The lengthening used in the experimental conditions does not refer to the lengthening which arises out of natural American English speech but rather a dramatized lengthening not tied to the same strict phonological environment. The performance used under the control conditions lacked these

features insofar as the examiner scripts were read aloud with American English Midwestern pronunciation. Pauses were only included at sentence boundaries.

Each of the four tasks contained at least one of the discourse features to distinguish it from the control performance. The differences are specifically identified here. In the first task the examiner script was spoken at a high pitch from the beginning and transitioned to a low pitch by the end of the script under experimental conditions. In the second task the examiner script was spoken with vowel lengthening on *why* under experimental conditions. In the third task the examiner script was spoken *three* and *cheeseburgers* loud and slow with a pause between the two words in both of parts of the examiner's script. In the fourth task *let's watch a movie tonight* was spoken quickly and included vowel lengthening on the last syllable, *night,* and the examiner's second scripted line should be spoken with a lengthening on the first *I* under the experimental conditions.

This researcher served as the rater and interlocutor. As a senior rater, I have rated the rated the oral section of the EPT for nearly three years. This means that I have participated in multiple recalibration sessions and even assisted in leading a recalibration session. In addition to this, I have conducted phase one of the oral section of the EPT online through voice chat. Multiple raters were not used because the operational oral part of the EPT is rated by a single rater/interlocutor. This research is conducted under the same operational model that the oral part of the EPT uses. It is important to remember that this research is exploratory and subject to future research with multiple raters as the timeline of this study did not allow for multiple ratings.

Participant Description

This research study was open to incoming University of Illinois at Urbana-Champaign students who were required to take the oral part of the EPT and were at least 18 years of age at the time they participated in the study. The participants must not have previously taken the oral component of the EPT. This is essentially the population that participates in the university's English placement tests each fall semester with the exception that EPT can test students who are under the age of 18 and admitted to the university whereas this research could not.

The participant sample for this study consisted of 47 University of Illinois at Urbana-Champaign students who self selected as participants in this study. The sample consisted of 32 males and 15 females where 43 were graduate students and four were undergraduate students.

The 47 students participated in either a control group or an experimental group. These participants were systematically placed into either the control group or experimental group depending on their scheduled interview time. The first participant experienced the experimental conditions. The second participant experienced the control conditions. The pattern continued in this manner. This method helped to balance the number of participants in each group as it was not possible to foresee the exact number of participants.

In the control group 16 spoke Chinese as their native language, two were native speakers of Spanish and there was one native speaker each for Thai, Vietnamese, Korean, Russian, and Hindi. In the experimental group there were 17 native Chinese speakers, two natives speakers of Portuguese, and there was one native speaker each for Spanish, Korean, Russian, Serbian, and Kannada.

In the control group 20 participants had been in the United States for one month or less. Two participants were in the country for two to three months and only one participant was in the

country for over one year. In the experimental group 16 participants were in the United States for one month or less. Two were in the country for two to three months. Three participants were in the country for one year, and three were in the country for over one year. Most students participating in the study had been in the United States for one month or less. The most significant difference between the two groups in this aspect is that five more participants had been in the experimental group had been in the United States for one year or more than in the control group.

The English learning background was similar between the two groups. Most participants had studied English for 6 to 10 years before participating in the study. In the control group, four participants had studied English for 0-5 years, 11 had studied for 6-10 years, five had studied for 11-15 years, two had studied for 16-20 years, and one had studied between 21 and 25 years. In the experimental group six participants had studied English for 0-5 years, 10 had studied for 6-10 years, five had studied for 11-15 years, two had studied for 16-20 years, and one did not respond.

The two groups were asked to rate their English speaking ability as either beginner, intermediate, advanced, or near-native. In the control group four rated their English speaking ability as advanced, 18 rated their English speaking ability as intermediate, and one rated it as beginner. In the experimental group five rated their English speaking ability as advanced, 13 rated their English speaking ability as intermediate, five rated their English speaking ability as beginner, and one did not respond.

The TOEFL speaking scores of the participants ranged from 17 to 23 for 36 of the participants. Seven participants reported their TOEFL scores, which ranged from 89 to 98. One student reported an IELTS speaking score of 6, and one student did not provide any large scale English test score.

This study also looked at the participants' performance on the university's phase I of the oral part of the EPT. After participating in this study, the students took the university's oral part of the EPT. Trained raters other than this researcher rated students who participated in this research study. From their ratings, the participants in this study can be broken down into four groups: participants not required to take phase II of the oral section of the EPT who experienced control conditions in the study, participants not required to take phase II of the oral section of the EPT who experienced experimental conditions in the study, participants required to take phase II of the oral section of the EPT who experienced control conditions in the study, and participants required to take phase II of the oral section of the EPT who experienced experimental conditions in the study.

The group which experienced control conditions and was required to take phase II of the oral section of the EPT was composed of seven Chinese speakers, one Hindi speaker, one Korean speaker, and one Thai speaker. One had studied English from 0-5 years, three from 6-10 years, three from 11-15 years, two from 16-20 years, and one from 21-25 years. Two of the participants rated their English speaking ability as advanced, and eight rated their English speaking ability as intermediate.

The group which experienced experimental conditions and was required to take phase II of the oral section of the EPT was composed of nine Chinese speakers, one Kannada speaker, one Portuguese speaker, and one Serbian speaker. Two had studied English from 0-5 years, six from 6-10 years, three from 11-15 years, and one from 16-20 years. Three of the participants rated their English speaking ability as advanced, five rated their English speaking ability as intermediate, three rated their English speaking ability as beginner, and one did not respond.

The group which experienced control conditions and was not required to take phase II of the oral section of the EPT was composed of nine Chinese speakers, two Spanish speakers, one Russian speaker, and one Vietnamese speaker. Three had studied English from 0-5 years, eight from 6-10 years, and two from 11-15 years. Two of the participants rated their English speaking ability as advanced, 10 rated their English speaking ability as intermediate, and one rated it as beginner.

The group which experienced experimental conditions and was not required to take phase II of the oral section of the EPT was composed of eight Chinese speakers, one Portuguese speaker, one Spanish speaker, one Korean speaker, and one Russian speaker. Four had studied English from 0-5 years, five from 6-10 years, one from 11-15 years, one from 16-20 years, and one did not respond. Two of the participants rated their English speaking ability as advanced, eight rated their English speaking ability as intermediate, and two rated their English speaking ability as beginner.

Administering the Test

Participants in the study met the researcher in a collaborative study room in a university building. Each participant first completed a language background questionnaire (see Appendix A). Afterward, they were briefed on the format of the oral part of the EPT before beginning the test. The participant was then given their test. The test was one piece of paper with the instructions printed at the top and four tasks consisting of 18 targets. The rater and student versions of the test can be found in Appendix C.

This researcher reviewed the instructions with each participant before beginning. During this process, this researcher read the instructions aloud and then pointed to background

information of each task and the line that says *I say*. These parts were read by this researcher during the tests. This researcher also indicated to each participant which lines he should read by pointing to the lines that contain the words *You say*. Each participant was given the opportunity to ask questions before beginning the test. Just before beginning the test, the researcher initialized audio recording software through a laptop which was placed near the researcher and the participant.

During the process of interviewing participants, the researcher and the participant sat across from each other at a table. For each task, this researcher first read the background information, then read the rater/interlocutor dialogue script. Participants would then read their portion of the dialogue script which responded to the examiner dialogue script. Longer dialogues contained multiple alternations between examiner script and examinee script.

The examiner script was read once following either the control or experimental conditions. The participant responded by reading the examinee script. The examiner script was read again under the same conditions, and the participant responded by reading the examinee script again. Only the final reading was scored. This means that even if the examinee produced the correct target on the first attempt but did not produce the correct target on the second attempt, the target was marked as incorrect. Each turn in the examinee script contained at least one target. Each target assessed whether the participant could produce a correct primary phrase stress or intonation change. Missed targets were noted by checking a small circle above the target. If the target was produced correctly on the second attempt, the circle was left empty above the target.

An initial rating took place during the interview. However, audio recordings were made and reviewed by this researcher to maintain consistency in ratings and performance as the annual

recalibration session occurred during the middle of the data collection process. The recordings were reviewed two times to check the rater performance and the participant response and a third time to justify targets marked as incorrect. Since the control conditions and experimental conditions were administered systematically by alternating between participants, it was clear which performance should be administered in each case. Additionally, the presence (or absence) of control conditions in audio recordings made it possible for this researcher to know under which conditions the participants were being rated. Performances which did not follow the systematically determined conditions were not included in this study. This was an issue in only one instance. This researcher administered the experimental conditions to a participant which should have received the control conditions. However, in this case, the experimental conditions were implemented on the first part of the examiner script. Having realized this, this researcher continued with the experimental conditions for the duration of the test. Due to the fact that the conditions did not change in during the test, the data from this interview was ultimately included in this study. The systematic alternating pattern of administering experimental and control conditions was reversed as a result of this.

Chapter 4: Results

The results of this study will be viewed from several perspectives. First, the results for each target will be discussed followed by the results of linguistic focus, task, and a descriptive analysis of the trends that emerged after closely reviewing the recorded data. The results seek to identify differences between the experimental and control group. In order to highlight the differences in performance, mean difficulty derived from correct participant responses to total participant responses is used. These scores are calculated for the target, linguistic focus, and task results analysis on a scale from zero to one where one represents 100% of participants responding correctly identifying the target as easier, and zero represents 0% of participants responding correctly identifying the target as more difficult. When a group is referred to as having "a lower mean difficulty," it indicates that more participants in the group produced the target correctly. Thus, the number used to represent mean difficulty may be higher (closer to one), but this indicates that the difficulty was actually lower (easier).

Target analysis views each target on the test as separate from other targets. Thus, each target is recorded individually resulting in a total of 18 targets. Each one of the 18 targets is numbered in Appendix D.

Figure 1 shows the target difficulty across all 18 targets with the control group and the experimental group separated. Both the control group and the experimental group performed nearly the same on the first six targets and targets 9 and 15 with less than .4 difference in mean difficulty. Targets 7, 10, 11, 13, 14, and 17 displayed the largest difference in performance. These targets produced mean difficulty differences ranging from .19 to .28. Targets 13 and 14

yielded the greatest difference in mean difficulty which was .28 easier for the experimental group.

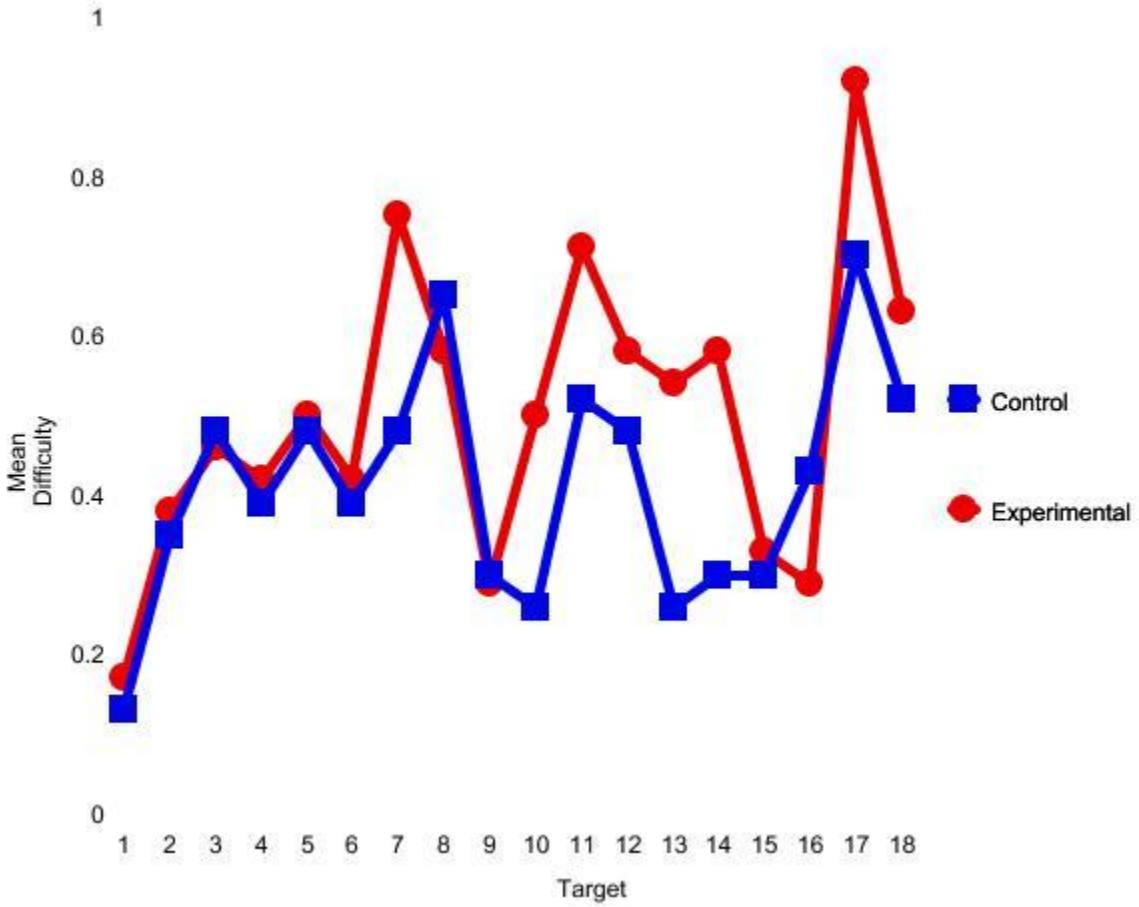


Figure 1. Mean Target difficulty. This figure shows the mean difficulty for each target for both the control group and the experimental group.

In the case of target seven, 11 participants from the control group missed the target as a result of maintaining equal stress throughout the phrase in which there is a contrast among parallel phrases. Target 10 was easier for the experimental group because 14 participants in the control group did not change intonation on a rising intonation target. Target 11 was easier for the experimental group because six participants in the control group used equal stress on the two words that compose the narrowing question. Target 13 was easier for the experimental group

because 15 participants from the control group used equal stress on the first part of a choice question. Target 14 was easier for the experimental group because 15 participants from the control group did not change their intonation for the first choice in a choice question. There was no clear trend in erroneous production of target 17 for the control group. The control group performed better than the experimental group on target 8 by .07 in which seven participants in the experimental group stressed both words in a two word tag question equally and target 16 by .14 in which 14 participants in the experimental did not change intonation on a target which requires falling intonation.

Linguistic focus organizes multiple targets into groups based on their linguistic similarity. Target suprasegmental features which work together to create a meaningful linguistic effect are grouped together. This means that single or multiple targets may comprise one of the eight linguistic foci. Linguistic focus is measured in through two methods, a fine-grained atomic level of measurement and a molecular level of measurement, presented by Davidson (1996). The practice of recording data at the level of an atom helps to better understand the changes in the molecules. Therefore, without recording atomic level data along with molecular level data, it is impossible to measure small variations within the molecules because the molecules are composed of the atoms. In this case, linguistic focus mean difficulty is calculated through two methods: one which accounts for correct production of parts of a linguistic focus (atoms) and one which requires that a participant correctly produce all parts of a of a linguistic focus in order for that participant's performance on the linguistic focus to be correct (molecule). Both methods of scoring will be explored here. Each of the eight linguistic foci are identified in Appendix D.

Figure 2 shows that overall the experimental group produced each linguistic focus more correctly than the control group. This means that they performed better on the targets which

make up each linguistic focus with the exception of true tag questions (linguistic focus 3). True tag questions were .07 easier for the control group. The greatest difference in mean difficulty arose from answers to a choice questions (linguistic focus 7) in which the experimental group scored .22 higher than the control group.

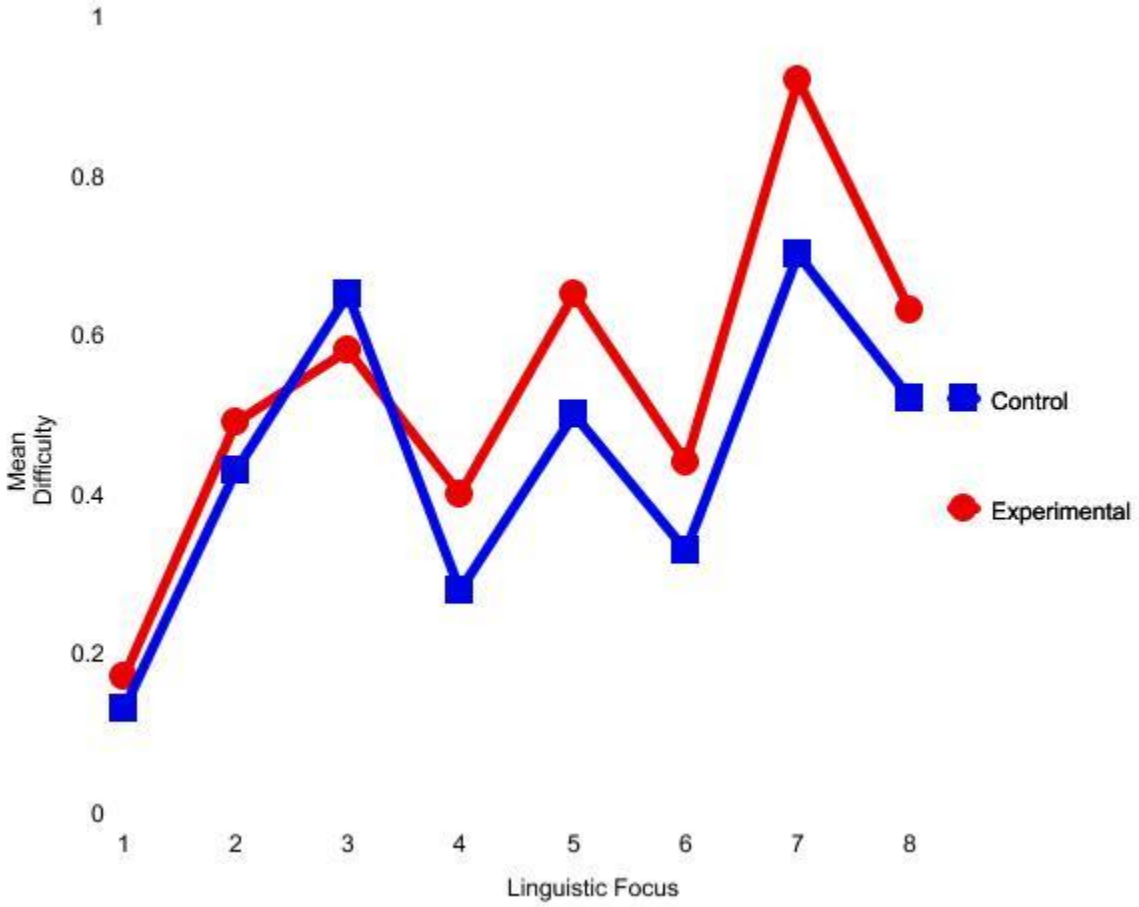


Figure 2. Atomic level mean linguistic focus difficulty. This figure shows the mean difficulty for each linguistic focus at the atomic level for both the control group and the experimental group.

The molecular view of linguistic focus mean difficulty, as seen in figure 3, is similar to the atomic view. The most noticeable differences is the increase in mean difficulty for each linguistic focus which is composed of more than one target which may be partially a result from the requirement that the participants must correctly produce all targets within a linguistic focus in

order to for the linguistic focus to be assessed as correct and that some linguistic foci have more targets than others which provides more opportunities for participants to miss a target. In the previous linguistic focus analysis, scores could be partially correct if not all of the targets were correct. This affects contrasts among parallel phrases (linguistic focus 2), repetition questions (linguistic focus 4), narrowing questions (linguistic focus 5), and choice questions (linguistic focus 6). From this perspective contrasts among parallel phrases and true tag questions were easier for the control group than the experimental group indicating that more participants were able to correctly produce all target suprasegmental features in a list of parallel phrases and to correctly produce true tag questions if they were in the control group. Repetition questions, narrowing questions, and choice questions were especially difficult for the control group to perform correctly at the molecular level which can be seen by a .20 increase in mean difficulty for repetition questions and a .24 increase in mean difficulty for narrowing questions and choice questions. These same linguistic foci were not as difficult for the experimental group to produce entirely correct based on a .11, .04, and .08 increase in mean difficulty for repetition questions, narrowing questions, and choice questions respectively. Instead, contrasts among parallel phrases was the most difficult for the experimental group to produce entirely correct based on a .29 increase in mean difficulty.

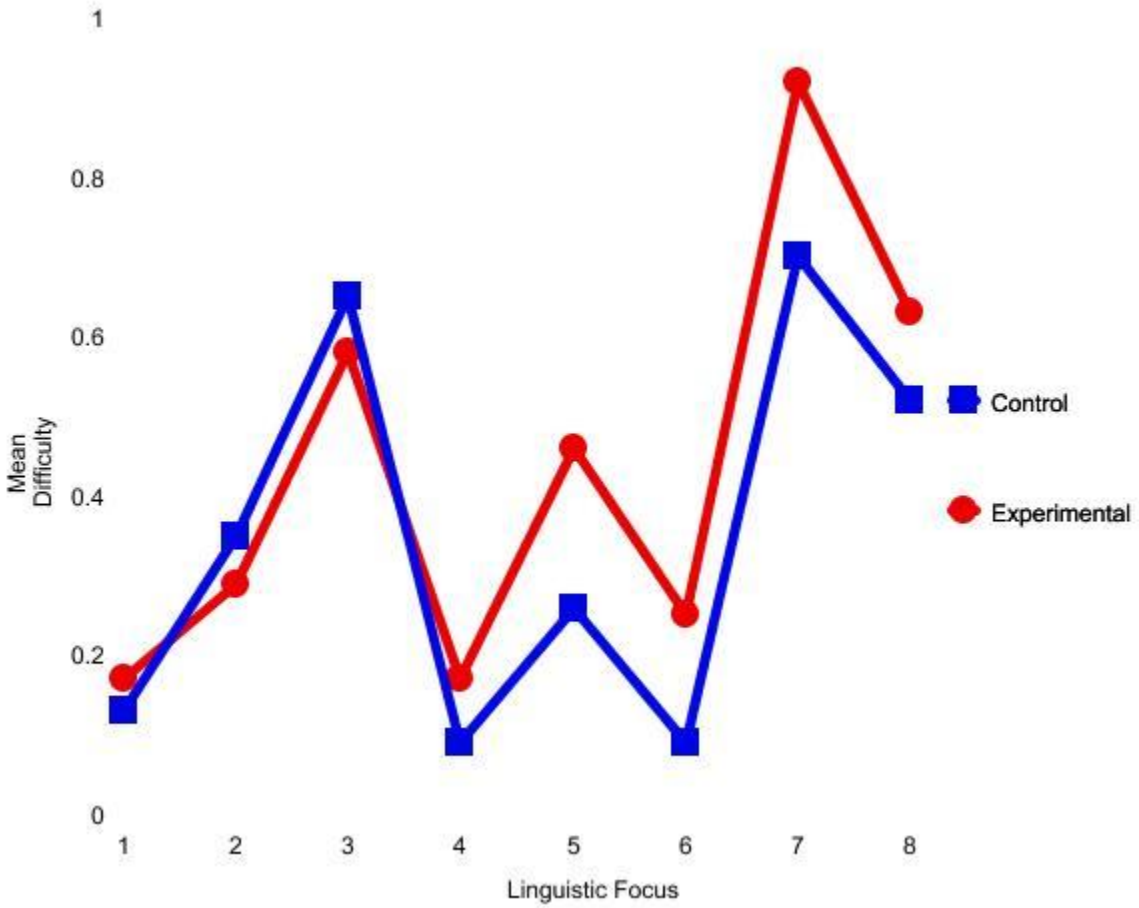


Figure 3. Molecular-level mean linguistic focus difficulty. This figure shows the mean difficulty for each linguistic focus at the molecular level for both the control group and the experimental

Viewing the results from the perspective of a task also organizes multiple targets into groups. There are four discreet situations each with its own context provided in the test which create four tasks. Each task is composed of a single target or multiple targets. Individual task focus scores are calculated from the mean score of the targets of which it is composed. Appendix D identifies the targets which comprise each of the four tasks.

As shown in figure 4, the job search task (task 1) and preparing dinner task (task 2) were of similar mean difficulty for both the experimental and control group with only a .04 difference for each. The test appears to become easier for the experimental group with an overall increase in performance throughout the test. However, the test became more difficult after the preparing

dinner task in which the mean difficulty increased from .46 to .39 in the office workers chatting task (task 3) for the control group, but then decreased slightly to .42 in the planning a get-together task (task four).

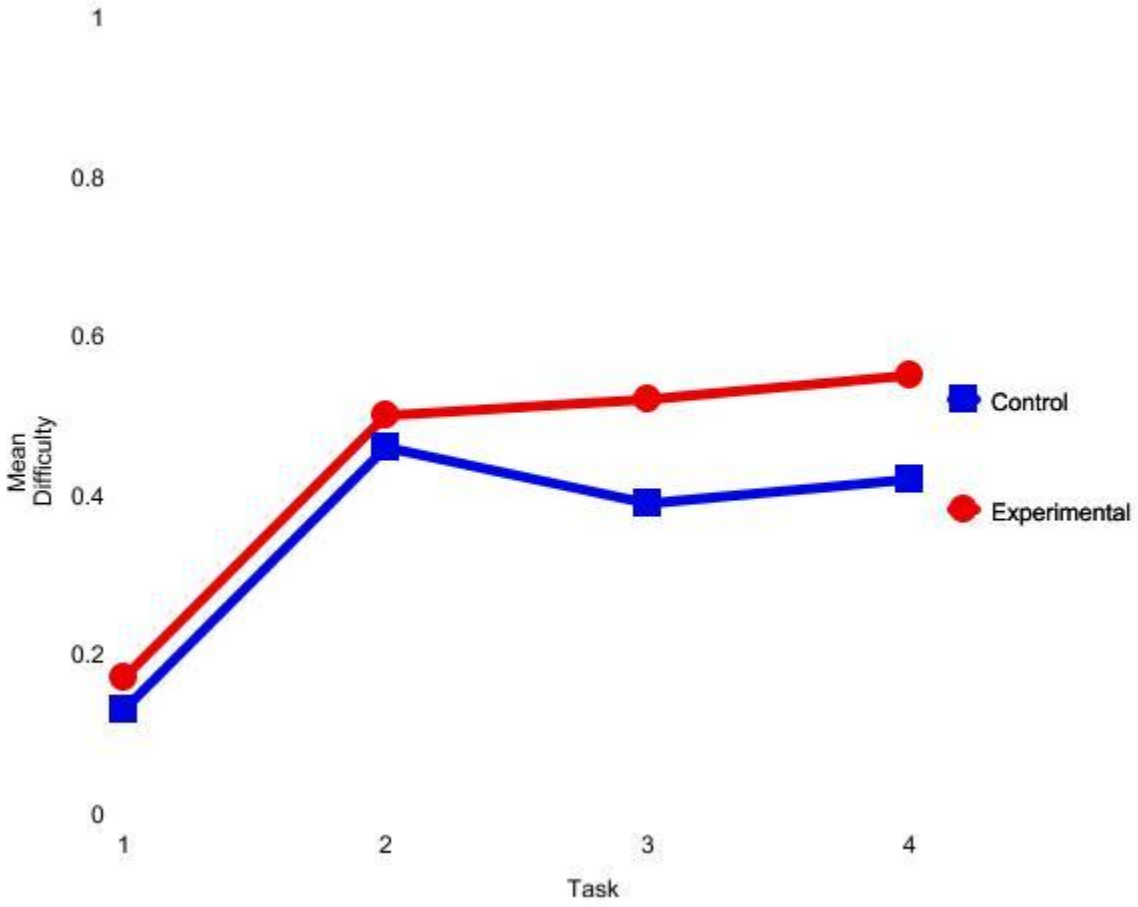


Figure 4. Atomic level mean task difficulty. This figure shows the mean difficulty for each task at the atomic level for both the control group and the experimental group.

From the molecular view of mean task difficulty in figure 5, the test appears very difficult. The easiest task for the control group was the preparing dinner task in which mean difficulty was .35. This is .14 easier than the same task for the experimental group. The planning a get-together task was the easiest for the experimental group with a mean difficulty of .21, which is .17 easier than the same task for the control group. The increase in mean difficulty

from the atomic level view of the task to the molecular level view of the task is most clearly seen in the experimental group whose mean difficulty increased by .29 in the preparing dinner task, .39 in the office workers chatting task and .34 in the planning a get-together task. The control group also displayed an increase in mean difficulty by .16 in the preparing dinner task, .35 in the office workers chatting task, and .38 in the planning a get-together task.

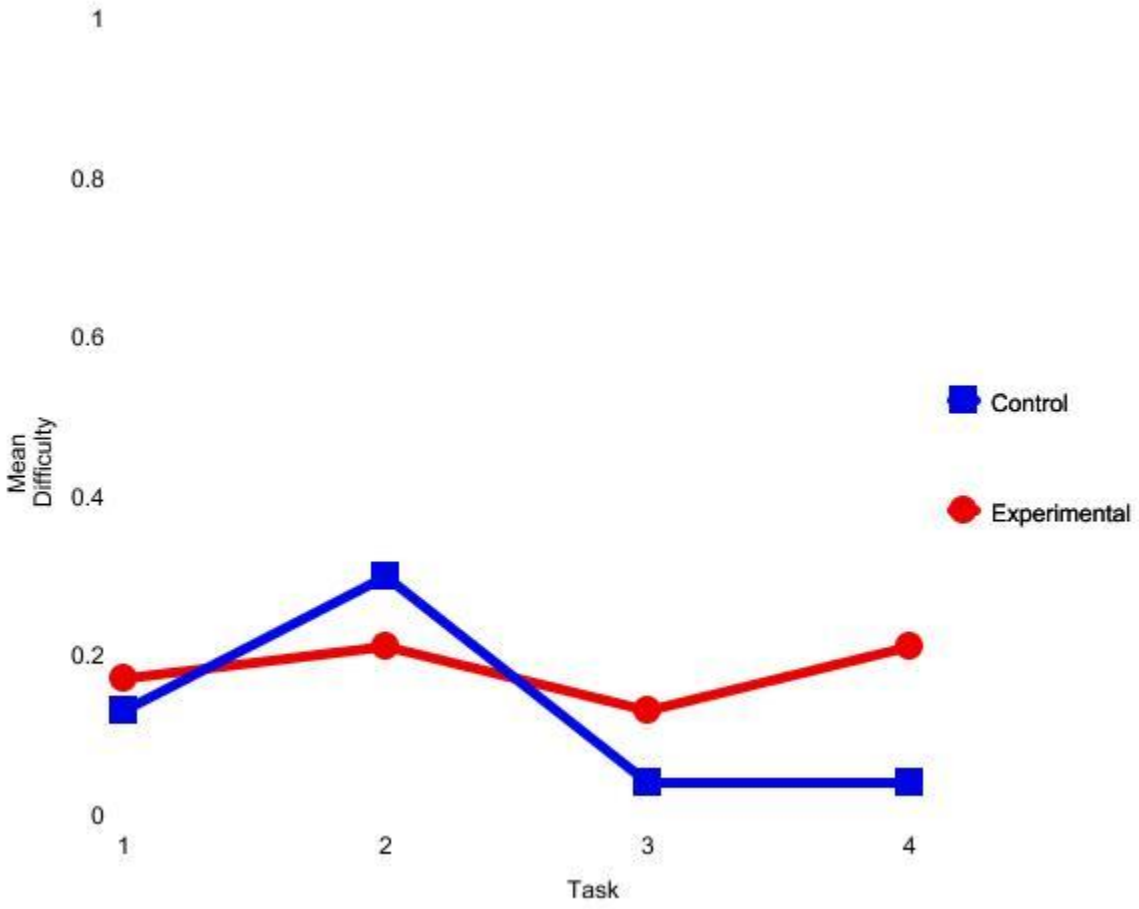


Figure 5. Molecular level mean task difficulty. This figure shows the mean difficulty for each task at the molecular level for both the control group and the experimental group.

In summary, the previous descriptive analysis of the recordings of candidates identified several diverging trends in the performance from each of two groups. Target one was close in

mean difficulty for both groups with a mean difficulty of .17 for the experimental group and .13 for the control group. However, the frequency of errors produced by the two groups is not as similar. Participants in the control group used equal stress throughout the contradicting phrase 11 times and stressed an incorrect word nine times. Participants in the experimental group used equal stress throughout the contradicting phrase six times and stressed an incorrect word 14 times. There was also a difference in intonation usage. There are four scored intonation targets: repetition question, narrowing question, and two choices in a choice question. There is also a fifth intonation target which appears on a tag question. However, this tag question can be interpreted as a known answer tag question or as a seeking agreement tag question. Therefore, rising and falling intonation would both be acceptable so this intonation target was removed from the test specification used in this study—this target does not contribute to a participants score. Based on these five intonation targets, the experimental group produced both correct and incorrect changes in intonation on 55% of the targets whereas the control group produces an intonation change in 43% of the targets.

The score frequencies for the control group are concentrated mostly at the low numbers which range from two to four producing a positive skew with a slight spike in frequency at the score of 12. The score frequencies for the experimental group are concentrated in the range of six to nine, slightly below the midpoint, with a spike in frequency at the score of four. This indicates that there is a greater amount of low-performing participants in the control group and a greater amount of mid-performing participants in the experimental group. The experimental group had slightly higher concentration of high scores when compared with the control group but were very close in frequency in the 15-17 score range. The spread of the scores can be seen in figure 6.

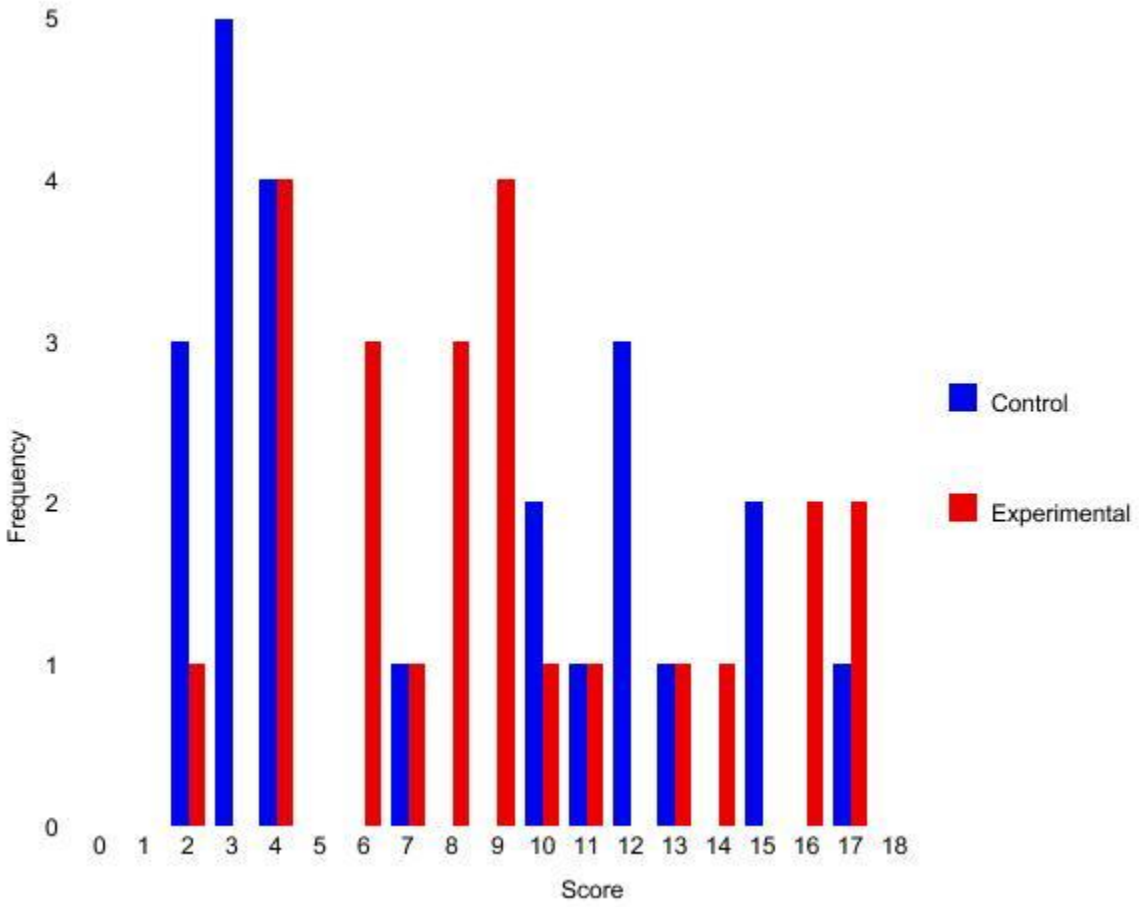


Figure 6. Participant score frequency This figure shows the frequency of scores for both the control group and the experimental group. The maximum possible score is 18.

The results of this study were further analyzed based on whether or not examinees were required to take phase II of the university's oral section of the EPT.

The mean difficulty for each target for the groups which were not required to take phase II of the university's oral section of the EPT showed that the mean difficulty was lower for the control group than the experimental group on nine targets while the difficulty was lower for the experimental group than the control group on the other nine targets. This can be seen in figure 7. Targets three and eight were .19 and .35 easier for the control group respectively. Target 10 was .21 easier for the experimental group and targets 13, 14, and18 were all .28 easier for the experimental group.

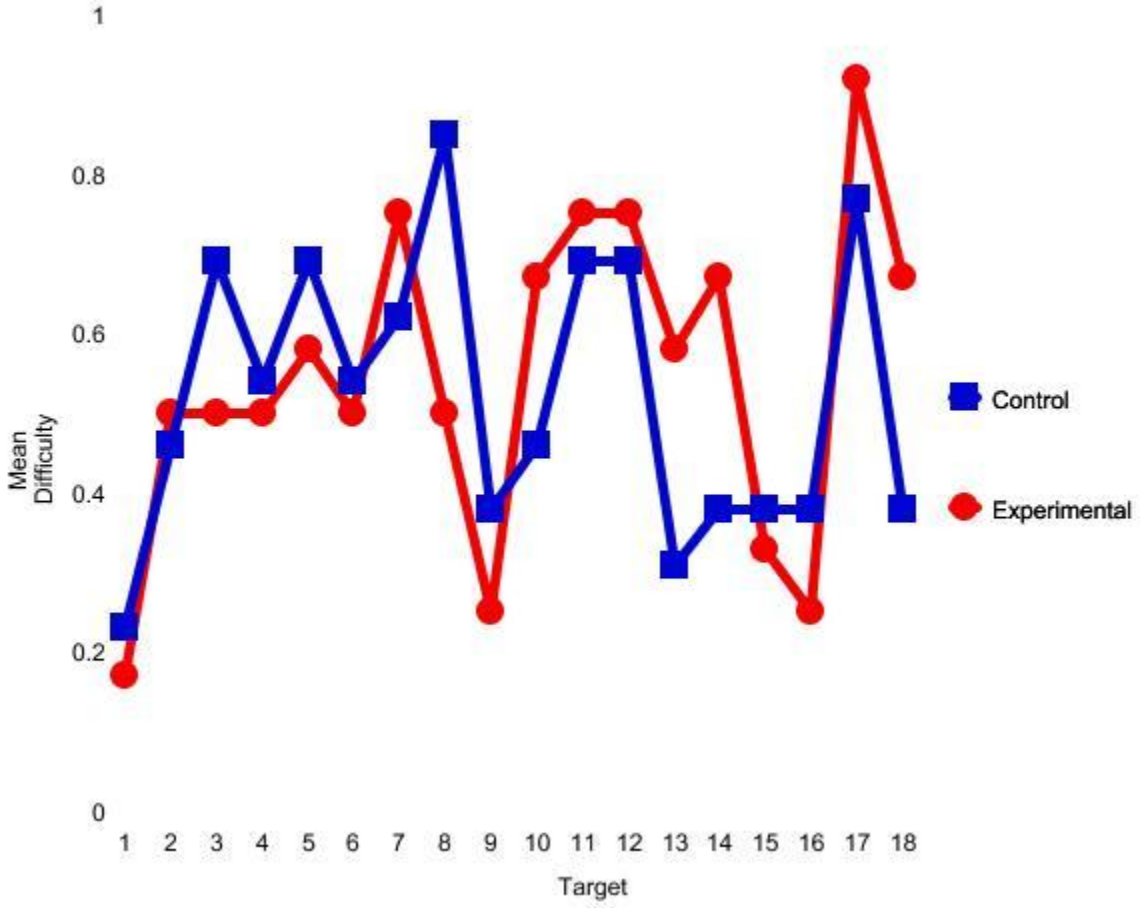


Figure 7. Target mean difficulty for groups not required to take phase II. This figure shows the mean difficulty for each target for both the control group and the experimental group that were not required to take phase II.

The atomic view of linguistic focus mean difficulty for groups which were not required to take phase II shows the experimental group had a lower mean difficulty on all linguistic foci except one, two, and three as seen in figure 8. In these cases the control group had a lower mean difficulty. The mean difficulty for linguistic focus three was .35 easier for the control group, and the mean difficulty for linguistic focus eight was .29 easier for the experimental conditions. Other linguistic foci did not exhibit much difference in performance.

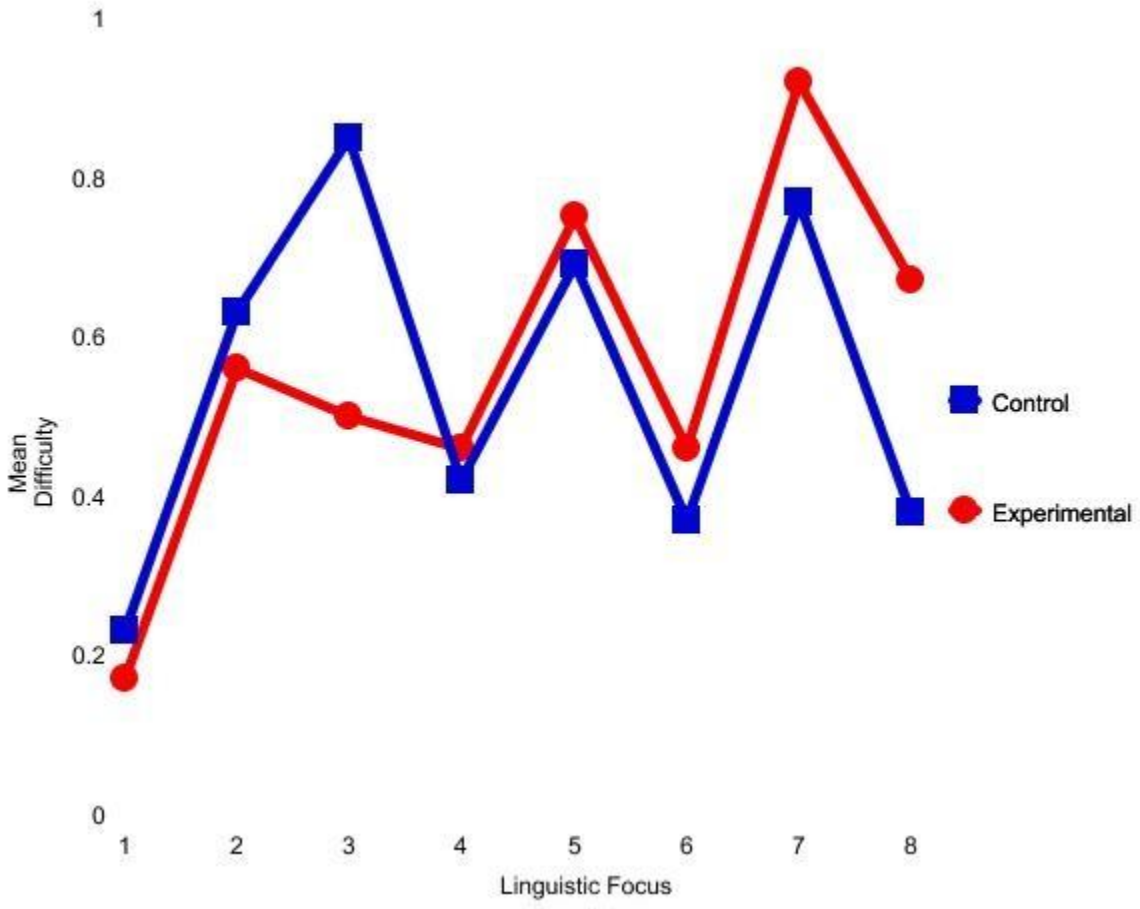


Figure 8. Atomic level linguistic focus mean difficulty for groups not required to take phase II. This figure shows the atomic level linguistic focus mean difficulty for both the control group and the experimental group that were not required to take phase II.

The molecular view of linguistic focus mean difficulty for groups which were not required to take phase II exhibits a similar pattern as the atomic view does. Figure 9 shows that the first control group has a lower mean difficulty on linguistic foci one, two, and three while experimental group has a lower mean difficulty on all other targets. Linguistic focus two had a lower mean difficulty for the control group by .13 and linguistic foci seven and eight had a lower mean difficulty for the experimental group by .15 and .29 respectively.

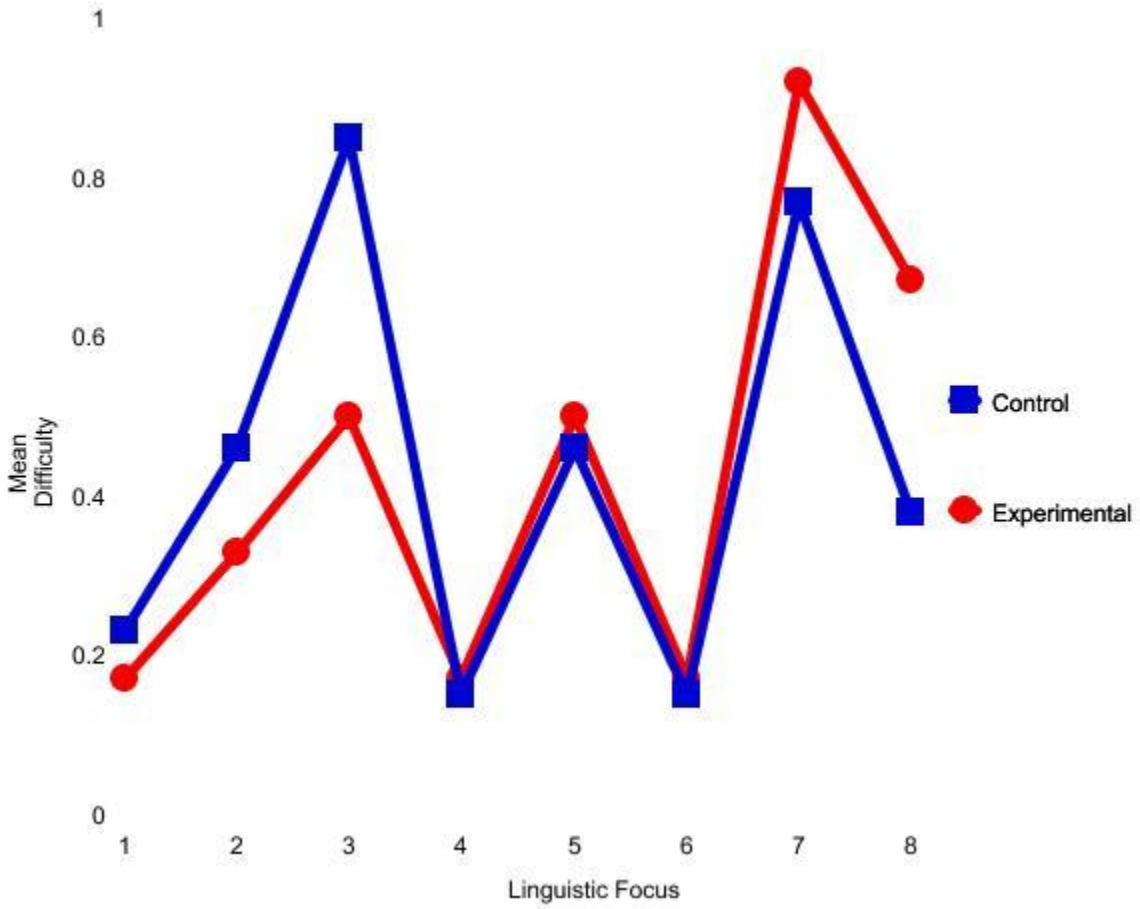


Figure 9. Molecular level linguistic focus difficulty for groups not required to take phase II. This figure shows the mean molecular level linguistic focus difficulty for both the control group and the experimental group that were not required to take phase II.

The atomic view of task difficulty for groups which were not required to take phase II shows that the control group has a lower mean difficulty on the first two tasks while the experimental group has a lower mean difficulty on the last two tasks. This can be seen in figure 10. In all four tasks, little difference in mean difficulty between the experimental group and the control group is seen.

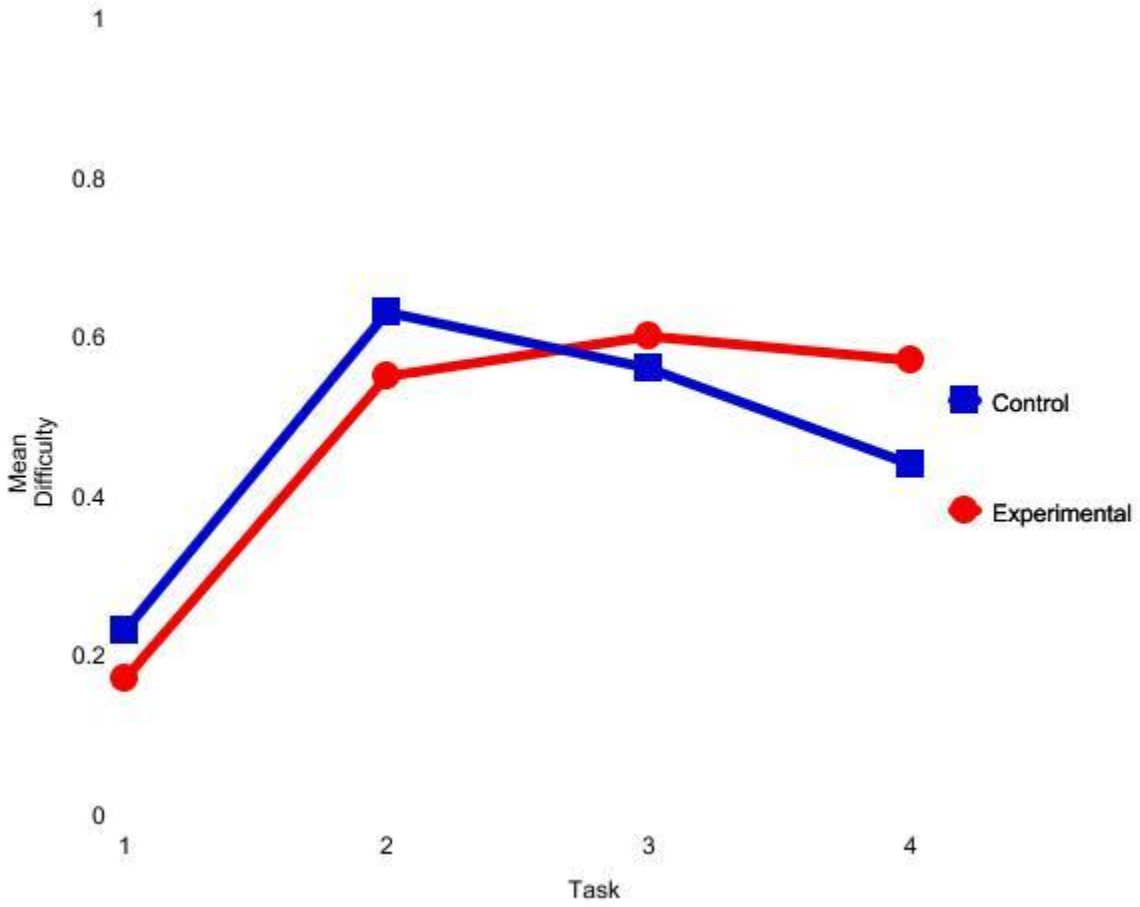


Figure 10. Atomic level task mean difficulty for groups not required to take phase II. This figure shows the atomic level task mean difficulty for both the control group and the experimental group that were not required to take phase II.

The molecular view of task difficulty for groups which were not required to take phase II shows a similar pattern as the atomic view. The mean difficulty for the control group was lower in tasks one and two and the mean difficulty for the experimental group was lower in tasks three and four. The experimental groups mean difficulty for each task remained consistent throughout the four tasks but the control group created a spike in task two as its mean difficulty was .21 lower than experimental group. This can be seen in figure 11.

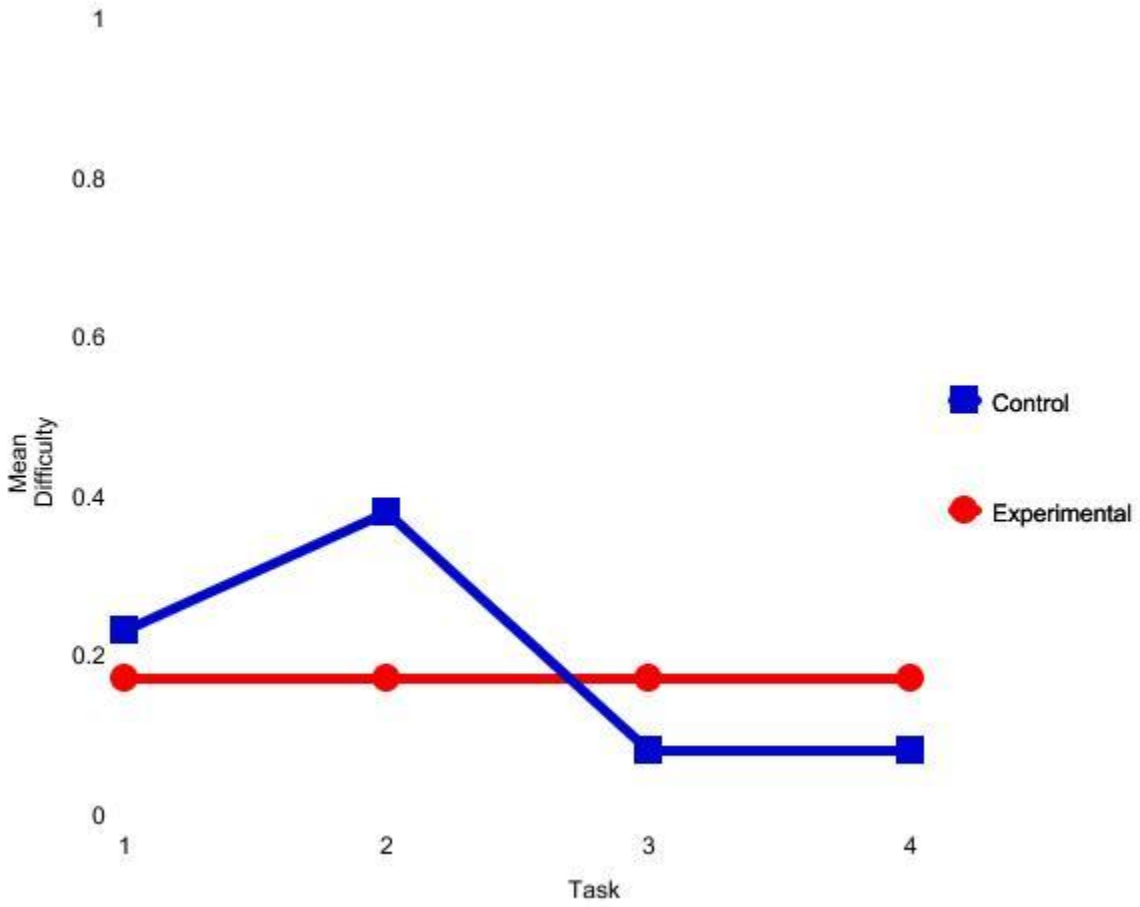


Figure 11. Molecular level task mean difficulty for groups not required to take phase II. This figure shows the molecular level task mean difficulty for both the control group and the experimental group that were not required to take phase II.

The mean difficulty of each target for the groups which were required to take phase II show an overall tendency for the experimental group to have a lower mean difficulty than the control group as seen in figure 12. The experimental group had a lower mean difficulty on targets 3 by .22, 5 by .22, 7 by .45, 8 by .27, 10 by .33, 11 by .37, 12 by .22, 13 by .30, 14 by .30, and 17 by .32. The control group had a lower mean difficult on target 16 by .17 and target 18 by .12.

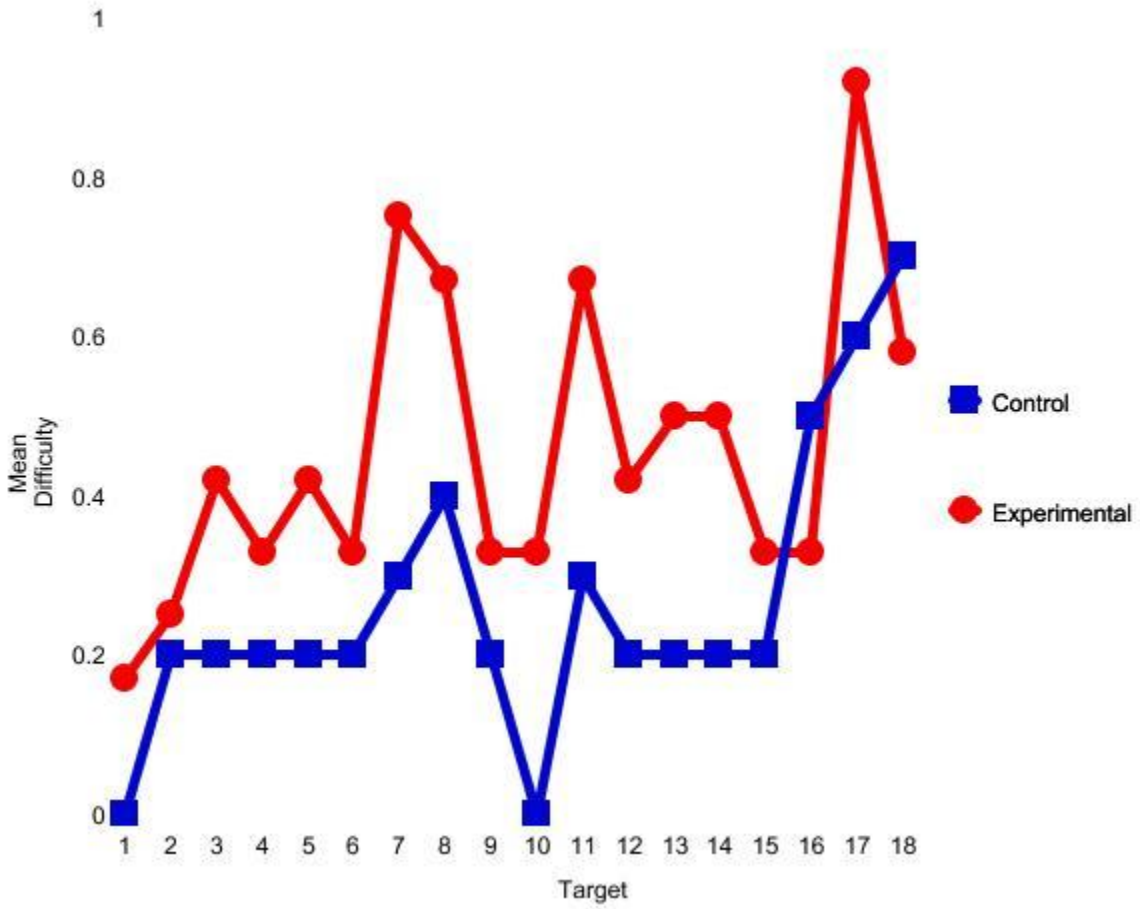


Figure 12. Target mean difficulty for groups required to take phase II. This figure shows target mean difficulty for both the control group and the experimental group that were required to take phase II.

The atomic view of mean difficulty for each linguistic focus for the groups which were required to take phase II of the university's oral section of the EPT shows that the Experimental group had a lower mean difficulty for each linguistic foci except for linguistic focus eight. In this case the control group had a slightly lower mean difficulty than the experimental group. Linguistic foci two, three, four, five, and seven all proved to have a lower mean difficulty for the experimental group than the control group by at least.20. This can be seen in figure 13.

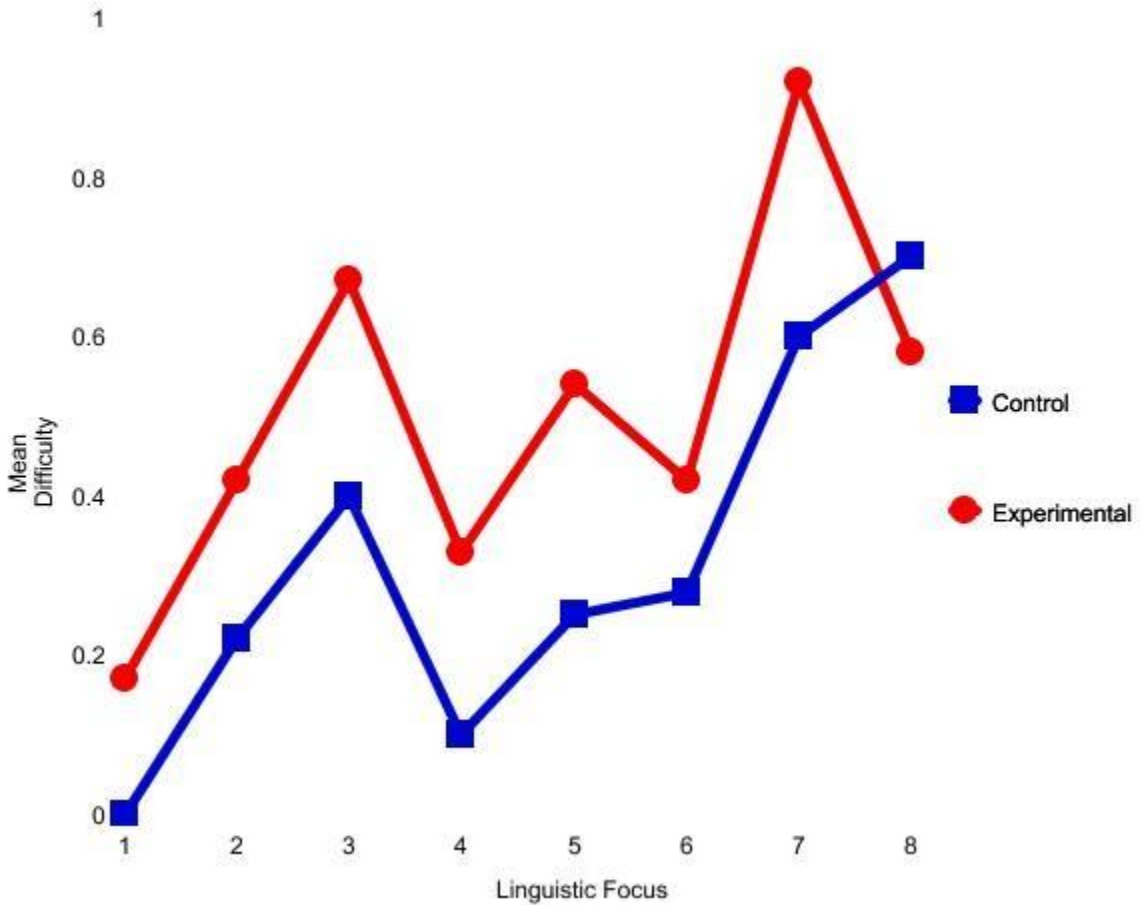


Figure 13. Atomic level linguistic focus mean difficulty for groups required to take phase II. This figure shows the atomic level linguistic focus mean difficulty for both the control group and the experimental group that were required to take phase II.

The molecular view of linguistic focus mean difficulty for groups which were required to take phase II of the university's oral section of the EPT can be seen in figure 14. This pattern is similar to the atomic view except that no students in the control group were able to correctly produce all the targets in linguistic foci four, five, or six.

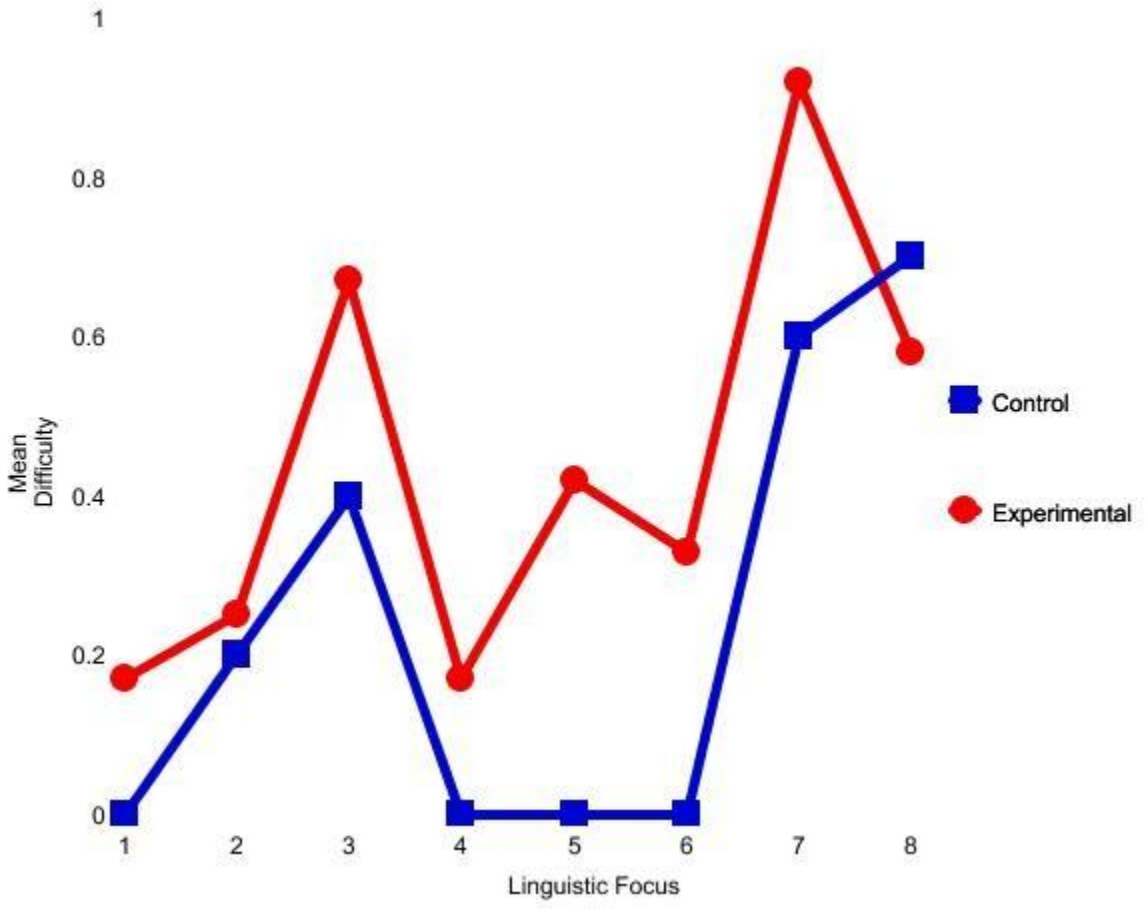


Figure 14. Molecular level linguistic focus mean difficulty for groups required to take phase II. This figure shows the molecular level linguistic focus mean difficulty for both the control group and the experimental group that were required to take phase II.

The atomic view of mean difficulty across tasks for groups which were required to take phase II can be seen in figure 15. This shows that in every task, the experimental group had a lower mean difficulty. The experimental group performed particularly better than the control group on tasks two and three in which the experimental group had a lower mean difficulty than the control group by .21 and .26 respectively.

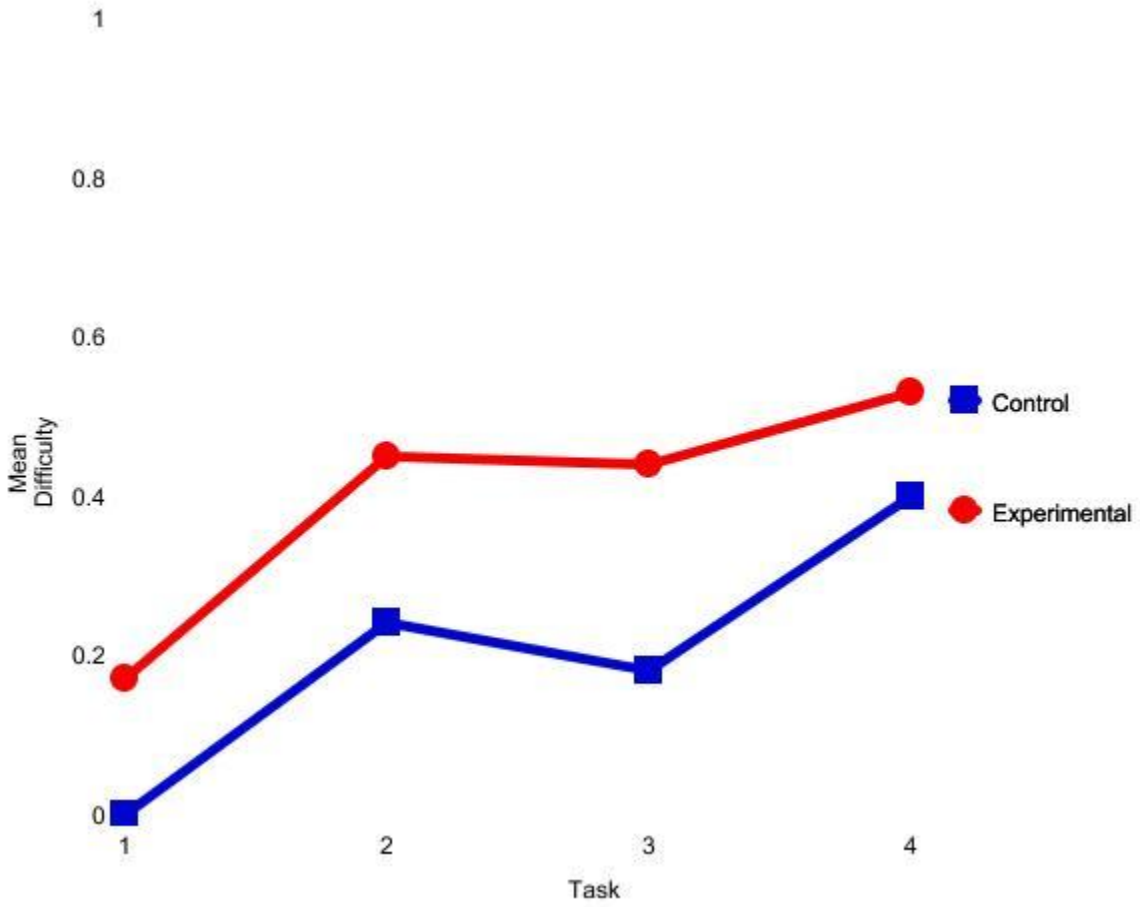


Figure 15. Atomic level task mean difficulty for groups required to take phase II. This figure shows the atomic level task mean difficulty for both the control group and the experimental group that were required to take phase II.

The molecular view of mean difficulty across tasks for groups which were required to take phase II can be seen in figure 16. Again, the experimental group has a lower mean difficulty than the control group across all tasks. However, the molecular view shows that experimental group and the control group had similar mean difficulties for the tasks two and three, and that the biggest difference in mean difficulty came from task four in which the experimental group had a lower mean difficulty by .25.

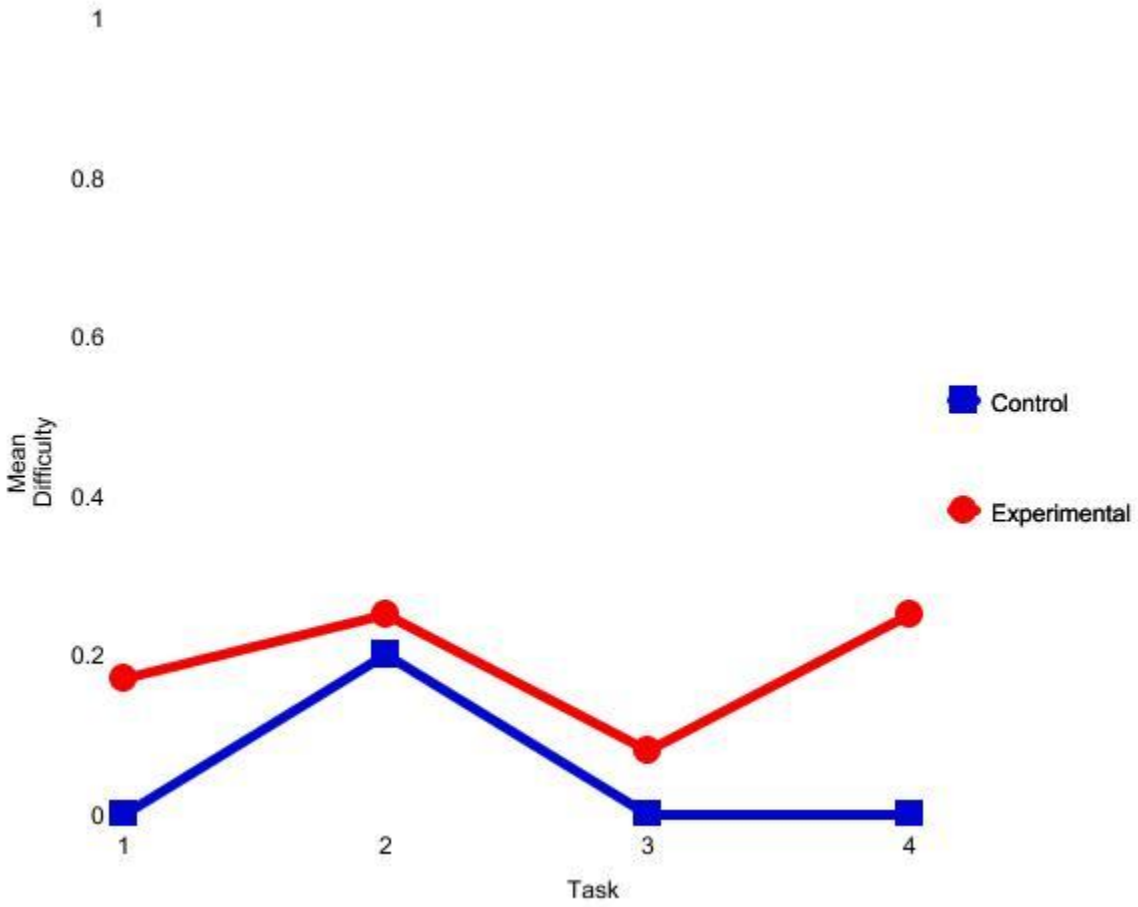


Figure 16. Molecular level task mean difficulty for groups required to take phase II. This figure shows the molecular level task mean difficulty for both the control group and the experimental group that were required to take phase II.

Chapter 5: Discussion

Limitations

It is important to take into account that this research is exploratory and subject to limitations. In interpreting the findings, these limitations should be considered.

One of the limitations this study faces concerns the sample size. The control group contained 23 participants. The experimental group contained 24 participants. The control group required to take phase II contained 10 participants. The experimental group required to take phase II contained 12 participants. The control group not required to take phase II contained 13 participants. The experimental group not required to take phase II contained 12 participants. The numbers of participants phase II required and the phase II which can lead the results to be more easily swayed by measurement errors. To minimize this effect on results as rater recalibration occurred during the middle of the data collection timeframe, audio recordings were reviewed as a check after this rater had participated in rater recalibration. However, this check cannot account for construct irrelevant variables such as participant fatigue as many of them had arrived in the country fewer than three days prior to participating in the study.

This study did not attempt to measure the anxiety of participants to see if this potential variable may have affected the participants performance on the test. Additionally, this study did not analyze individual performances of outliers. Anxiety and outlier performances are areas which may produce a deeper understanding of the phenomenon in this study and are later discussed in more detail as recommended future research directions.

A limitation of this study is that in the initial ratings which occurred during live interviews, required this researcher to focus on consistent alternation of performance and as well

as rating multiple targets whereas there is no performance requirement or alternation of performance requirement. Therefore, this researcher listened to audio recordings to achieve rating consistency. An average of two changes were made for each participant between the first and the final rating. These changes are likely a result of an attention burden affecting scoring accuracy. This researcher focused on maintaining consistent alternating performances, while scoring multiple pronunciation targets. However, in order to maintain accuracy in scoring reading aloud pronunciation tasks, a rater's attention should be focused on one target (Lado, 1964; Madsen 1983). This might explain an average of two adjustments were made after listening to the audio recordings.

Listening to the audio recordings introduces the problem of the researcher as rater. This researcher fulfilled the role of rater and interlocutor. The operational model of the oral part of the EPT only requires one trained rater. From this perspective, this rater is senior. There was little corroboration with other raters. Corroboration form other raters came from comparison of mean difficulty with the phase I/phase II decisions made by other raters. The mean difficulty of the control and experimental groups containing participants who were not required to take phase II was lower than the mean difficulty of the control and experimental groups containing participants required to take phase II. This suggests that this researcher's ratings were aligned with other rater's perceptions of the performance of participants in this study.

One method to avoid the researcher as rater limitation would require editing and coding of the sound files. In each sound file a code was spoken at the beginning so that the sound file could be connected to the participant information. This along with the performance of the rater would need to be edited out of each sound file so that they contained only the speech of the participants. The sound files would need to be coded and information about whether each

participant was in the experimental or control group and was required to take phase II or not required to take phase II would need to be stored in the database. The files could then be randomized and then rated so that after the rating is complete, the results could be calculated and organized by retrieving the information stored in the database for each sound file.

Additionally, the larger number of students in the experimental group who have been in the United States for over one year may serve as a source of inequality between the two groups. However, an argument against this can be seen when looking at the spread of scores in figure 6. Both the experimental group and the control group do not have a disproportionate amount of high performing participants in the 13 to 18 score range. This suggests that being the United States between one to two years did not provide an advantage over students who had spent less time in the United States.

Different language backgrounds of the participants of the two groups also imposes a limitation on this study. A similar number of Chinese speakers composed a majority for all groups. Spanish, Russian, and Korean speakers were similarly distributed between the two groups. Generally, these language backgrounds do not provide specific advantages in stress and intonation which transfer into English with the exception of Korean which shares some similar characteristics of intonation (Swan & Smith, 2001). Dravidian languages and South Asian languages were also noted as not providing specific advantages in stress and intonation usage which transfers into English. There are two varieties of Portuguese, European and Brazilian. The European variety is a stress-timed language like English, but the Brazilian variety is a syllable-timed language like Spanish (Swan & Smith 2001). Native language transfer would be likely in the case of European Portuguese but not so for Brazilian Portuguese. There were no Portuguese speakers in the control group, but there were two Brazilian Portuguese speakers in

the Experimental group. Although there are some differences in the native language backgrounds represented in the two groups, neither the experimental group nor the control group seem to benefit from native language transfer as few speakers of native languages other than Chinese participated in this study.

A warm-up effect occurring in this study is also possible. The first target in this study proved to be the most difficult. However, this may be in part due to the participants warming up to the testing conditions and performing better as the test progresses. This would be a result of each task being given to participants in the same order in each interview in which the task containing target one was always presented first. It is, however, the opinion of this researcher that target one has been consistently difficult for participants in the oral part of the EPT. This section occurs after the midpoint of the complete oral component of the EPT. However, target one is the first target in the beginning of a new section of the test, so it is also possible that performance on target one may be subject to a warm-up effect in the operational oral component of the EPT.

Research Question 1

1. What effects, if any, does an interlocutor's performance have on a test taker's response in constructed dialogue tasks?

The test seemed easier when the experimental conditions were applied as indicated by the trend in lower mean target difficulty compared to the mean target difficulty under control conditions when mean difficulty is calculated from the number of correctly produced targets divided by the number of targets resulting in a mean difficulty of one as having a lower mean difficulty (easier) than targets which have a mean difficulty of zero (difficult). This same trend

appears in overall mean difficulty based on target performance in which the experimental conditions produced a mean difficulty of .50 and the control conditions produced a mean difficulty of .41. In the molecular view of the linguistic focus, the experimental conditions yielded an overall mean difficulty score of .43 compared to .35 under control conditions. The difference in overall mean difficulty becomes less noticeable at the molecular level mean task difficulty in which there was a mean difficulty of .18 in the experimental conditions and a mean difficulty of .13 in the control conditions. The slight differences in mean difficulty for each task as shown in figure 4 and figure 5 support this. The differences in mean difficulty between the control group and the experimental group for each linguistic focus as shown in figure 2 and figure 3 indicate that the experimental conditions most noticeably produce an effect in examinee performance on certain linguistic foci like narrowing questions, choice questions and responses to choice questions and more specifically suprasegmental targets like 7, 10, 11, 13, 14, and 17 as illustrated in figure 1.

The trend of the experimental group having overall lower mean difficulty than the control group appears to a much lesser extent in the groups which were not required to take phase II of the oral part of the EPT. At the target level, the experimental group had a mean difficulty of .55 and the control group had a mean difficulty of .53. The atomic level linguistic focus mean difficulty score for the experimental group was .56 and .54 for the control group. The molecular view identified a mean difficulty of .43 for both the experimental group and the control group. The experimental group also had lower mean difficulty at the atomic level of tasks in with a mean difficulty of .47 compared to the control group which had a mean difficulty of .46. At the molecular level the mean difficulty was .17 for the experimental group and .19 for the control group indicating that the control group was able to correctly produce all the targets in a task than

the experimental group. However, the differences in mean difficulty for the targets, linguistic foci, and tasks appear as .02 differences or less which does not indicate a strong advantage for the experimental group over the control group. Indeed the control group had the same mean difficulty as the experimental group in molecular linguistic focus mean difficulty and a lower mean difficulty at the molecular level of tasks.

The trend of the experimental group having overall lower mean difficulty than the control group appears to a much greater extent in the groups which were required to take phase II of the oral part of the EPT. This sample represents examinees who would actually take phase II in the university's oral component of the EPT. The experimental groups target level, atomic level linguistic focus, and atomic level task mean difficulties were .46, .51, and .40 respectively. The control groups target level, atomic level linguistic focus, and atomic level task mean difficulties were .27, .32, and .20 respectively. The molecular level linguistic focus mean difficulty was .44 for the experimental group and .24 for the control group. The molecular level task mean difficulty was .19 for the experimental group and .05 for the control group.

The overall mean difficulty for participants in the experimental group who were not required to take phase II and the overall mean difficulty for the participants in the experimental group who were required to take phase II support the ability of phase I to discriminate between students who have a higher pronunciation ability and students who would benefit from taking the university's pronunciation course. Similar support can be seen from the overall mean difficulty for participants in the control group who were not required to take phase II and the overall mean difficulty for participants in control group who were required to take phase II. In the cases of both the experimental groups and the control groups, participants which were found not to be

required to take phase II had lower mean difficulty than participants which were found to be required to take phase II.

The higher overall higher performance of the experimental group seemed be a result of participants which were required to take phase II. The groups which were not required to take phase II did not show a strong difference in mean difficulty from the control group to the experimental group. This suggests that the experimental conditions benefitted the sample of participants who actually took phase II of the university's oral part of the EPT.

This trend in lower mean difficulty under experimental conditions for participants who experienced experimental conditions based on features of spontaneous discourse speech (Blaauw 1994; Shriberg et al., 1998) gives evidence that an interlocutor's performance can affect an examinee's performance on scripted dialogues which assess primary phrase stress and intonation usage. This most noticeably affects participants who are at a lower level of pronunciation ability as identified by the large difference in overall performance in the experimental group and the control group which were required to take phase II of the university's oral component of the EPT.

In analyzing individual targets from phase II required groups and phase II not required groups, the mean difficulty appears very similar in both the experimental and control groups in some cases. In other cases, it seems that experimental group was helped by interlocutor performances indicated by sharp increases in the number of participants that produced targets correctly. However, there are also instances in which the control group performs better than the experimental group indicated by sharp rises in the number of participants producing targets correctly in the control group and sharp decreases in the number of participants producing targets correctly in the experimental group.

The actual benefit of the experimental conditions for the phase II required group, phase II not required group, and both groups together is different in each case. The experimental group which was required to take phase II displayed the greatest benefit from the experimental conditions compared to the control group which was required to take phase II. The experimental group which was not required to take phase II displayed the least benefit from the experimental conditions compared to the control group which was required to take phase II. Lastly the experimental groups from the phase II required and phase II not required groups combined compared to the control groups from the phase II required and phase II not required groups illustrate an amount of benefit from the experimental conditions that falls between that which is seen in the phase II required experimental group and phase II not required experimental group. Since phase I of the oral section of the EPT is designed to separate students who are at a higher level of pronunciation ability and the experimental conditions seemed to affect the participants of this study who required to move on to phase II, it seems as if the effects of an interlocutor's performance on the response of the participants has diminishing returns at higher levels of examinee pronunciation ability.

On individual targets, there are indeed cases in which the experimental conditions seemed to hinder the performance of the experimental group compared to the control group. However, the effects of interlocutor performance seem to have benefited the experimental group required to take phase II in almost all cases except two targets in which the control group performed better. This is important because the participants in this group reflect the performance of students who would actually take this part of the test according to the current operational version of the EPT. Additionally, judging the actual benefit of the performance based on the harm of individual targets may not be a valid assessment. It is important to be reminded of the

purpose of the oral section of the EPT, which is to determine whether students should be placed into a pronunciation service course, recommended to take a pronunciation service course, or not required to a pronunciation service course. The content in the oral section of the EPT reflects the content taught in the pronunciation service course. In that class, students learn to produce each linguistic focus. They are not simply taught parts of the pronunciation required to correctly produce a linguistic focus. Therefore, looking at the effects of the interlocutor's performance at a fine-grained level may not be useful in interpreting these results for the purposes of application to the oral section of the EPT. Instead, analysis should focus on whether or not participants who experience experimental conditions can produce a more target-like linguistic focus which can be more clearly seen at the atomic level of linguistic focus mean difficulty.

Unfortunately, the effects of an interlocutor's performance cannot be fully understood from a single exploratory study. Future research with a greater number of participants and results identifying mean difficulty across different language groups may shed further light on this issue.

The question as to why this phenomenon occurs still remains. The linguistic foci consisting of *contrasts among parallel phrases*, *true tag* question, and *choice question* used in this study do not rely on the interlocutor's script to determine the correct placement of primary phrase stress and intonation. In other words, by simply speaking these naturally the targets can be produced correctly without a scripted dialogue. Therefore, if the participants in the control group perceived their script as text isolated from discourse, it would be expected to see similar mean difficulty between the control group and the experimental group on these linguistic foci. However, this is not the case. A simpler explanation for the phenomenon of higher performance in the experimental group might come from looking at test taker anxiety. Phase II of the oral

part of the EPT requires participants to read aloud. This creates two avenues for anxiety in using foreign language. A general trend has been established between high levels of anxiety and low performance on oral exams (Hewitt and Stephenson, 2012; Phillips, 1992). Anxiety in speaking may be combined with anxiety in reading in a foreign language as well (Saito, Horwitz, and Garza, 1999). Therefore, high amounts of participant anxiety may have affected the performance of participants in the present study.

Baran-Łucarz (2011) found that pronunciation ability level of Polish learners of English was correlated with anxiety measured by the Foreign Language Classroom Anxiety Scale (FLCAS). Thus, learners with a higher level of pronunciation measured through reading passages aloud were less anxious according to the FLCAS, which might offer some explanation as to why there was less difference overall in mean difficulty displayed by the two groups not required to take phase II. A generalizability study on group oral tests of Japanese learners of English found that there was a person-by-occasion effect on examinee performance indicating that differences in examinee interactions can affect raters' perceptions of the ability of examinees (Van Moere, 2006). This supports the notion that interaction between interlocutor and examinee can affect an examinee's performance but does not necessarily point toward test anxiety as the source. However, it highlights interaction as a source of performance variance. Extending this finding, it may explain that participants in this study simply found the experimental conditions based on spontaneous speech more calming and could therefore perform better compared to the control conditions which were based on read speech.

An alternate hypothesis to anxiety might be unique effects of interlocutor animation. Many studies have focused on anxiety. However, findings in the literature on anxiety do not seem to be able to explain the phenomenon in this study completely. Therefore, an alternative

hypothesis might be that the effects of interlocutor animation may be a unique and unexplored test facet.

Interlocutor animation would encompass the extent to which an interlocutor speaks with animation. In this study, the experimental conditions with performances based on spontaneous speech represent a more animated interlocutor performance compared to the control conditions based on read speech which represent less animated interlocutor performance.

In this study, the students who experienced more animated speech tended to perform better on a pronunciation placement test assessing primary phrase stress and intonation. An interlocutor animation effect could then be said to allow learners at a lower level of pronunciation ability to perform better, but the exact nature of why animated interlocutor performance might affect pronunciation is not yet known as in some cases it seems that animated performance had a negative impact on the experimental group on some targets. This may be a result of a native language group effect to animated performance.

Research Question 2

2. How, if at all, do the findings for the first research question affect the oral section of the EPT?

Currently no guidelines for reading the scripted dialogue section of the oral part of the EPT are given to raters. As a result, the interview experience may vary during the constructed dialogue section of the oral part of the EPT. By investigating the first research question, it was found that an interlocutor's performance while reading a scripted dialogue may affect the examinee's performance.

Bachman provides detailed checklists to mitigate such issues (1988). Such a checklist would not be necessary for the oral section of the EPT. Many facets in the oral section of the

EPT are already standardized through the current procedure. For example, the locations of the ratings take place in university classrooms where the rater and the examinee can sit face-to-face in close proximity, the raters participate in standardized training together annually, and Phase II of the oral section of the EPT requires all examinees to read the same test material.

Although this test is already tightly controlled, Phase II of the oral part of the EPT which contains the constructed dialogue section is a part of this test in which there is potential for interlocutor performance variation within the interlocutor's performance of scripted lines which prompt the examinee's response. As previously discussed in this study, the rater's performance seems to affect the examinee response. However, modifying the interlocutor performance may not be the panacea to the issues in this section of the test as even after the examinees are presented with a performance which contains features used in the experimental performance in this study, they may still respond with poor primary phrase stress and intonation production.

Implementation of dynamic assessment in the oral section of the EPT may offer an different way to produce a similar effect as what arose out of manipulating the interlocutor performance in this study. In phase II of the oral part of the EPT the instructions require the examinee to read each line two times "smoothly and naturally." However, as Lado (1964) pointed out, reading pronunciation tests may not always produce the most natural speech. Utilizing dynamic assessment provides the rater/interlocutor with the opportunity to supply feedback on the performance to the examinee between first and second readings. Lado (1964) suggests that the specific pronunciation targets should not be identified, but by providing feedback on performance only, the rater can guide the examinee away from simply reading the script to speaking the part without making the targets known the examinee. This implementation of dynamic assessment is similar to Antón's work on the Spanish diagnostic speaking test where

students receive suggestions for improvement after they fall short of the tasks expectations in their first attempt (2009). Incorporating this method of assessment could be used throughout phase II of the oral part of the EPT, but it would be most beneficial in sections testing the use of suprasegmentals as Lado notes that "[reading aloud] is not a serious limitation on testing sound segments" (1964, p. 84). Additionally, this research did not look into other parts of phase II other than the constructed dialogue section.

This study scratches the surface of the depth of this issue. It contrasted the performance of the experimental conditions which contained features identified in discourse description studies with the control conditions which differed from the experimental conditions in that they did not contain these features beyond what is used naturally by reading aloud. This study found that under experimental conditions participants who were required to take phase II of the oral section of the EPT performed better, but because of the small sample size of 10 participants in the experimental group and 12 in the control group, it is difficult to generalize to the entire population of students required to take phase II. Additionally, similar effects may be produced through the use of dynamic assessment as simply modifying the instructions of the task may not address the issue of participant anxiety.

In limiting the scope of this research, this study did not identify how raters actually perform in the constructed dialogue section of the oral part of the EPT. As a result, it is unknown how the rater's perform naturally. This study recommends future research with larger samples to better understand the factors affecting student performance on constructed dialogues assessing primary phrase stress and intonation usage.

Future Research

Currently trained rater's having participated in the annual recalibration session are considered reliable as the operational oral section of the EPT follows a single rater model. Future research on the oral section of the EPT might look at rater reliability by comparing the rating of multiple raters. There are three possible models for such research. The first requires multiple raters in a room and one interlocutor. Thus, all the rating is completed at the same time. This model is not consistent with the operational oral part of the EPT as it has multiple raters in the testing room and the role of rater/interlocutor is split. Another possible model would audio record oral interviews and raters other than those who conducted the audio recorded interviews to listen to the recording in separate rooms one time. The scoring of the raters could then be compared. A third model would require multiple ratings of an examinee. This would maintain the rater as interlocutor role, but examinee performance may be affected by warming up to the first rater. The second model with multiple raters scoring audio recorded interviews would be the most practical. The ratings could also be compared to the original rating to determine if there is any advantages to recorded ratings.

Introducing dynamic assessment into phase II of the oral part of the EPT to encourage the examinee to produce a more natural reading performance and promote the rater as a source of encouragement for participants may help to improve examinee performance. Lado (1964) identified reading aloud as detrimental to testing some phonological features as there is a difference in what is acceptable for reading and what is acceptable for speaking. Using dynamic assessment, it would be possible to identify whether feedback between first and second readings could prompt the examinee to speak naturally rather than read aloud. Giving feedback to encourage participants to relax and speak naturally may also reduce some level of anxiety as a

form of feedback through interaction as support from group members was found in a post-test survey to reduce test anxiety in a study looking at how individual tasks and group tasks affected students' performance on a test (Sun, 2011). Performance ratings of examinees who experience dynamic assessment in this way could be compared to their performance without this treatment to determine its effect on performance. Additionally, comparing results from post-test surveys would aid in determining the reactions of examinees to such an assessment method.

This study did not analyze outliers. Deeper insights might be gained from evaluating the response patterns of participants from each group who produced unusual response patterns. Their interviews could be reviewed again more closely to see what they are doing that causes them to produce targets correctly or incorrectly. To further understand how individuals are affected by interlocutor performance, future research might also include case studies which include interviews, acquiring more detailed background information about participants, and further testing. This is especially important as each language learner is an individual with his or her own experiences with English pronunciation. Case studies may attempt to identify common learner variables among participants that seem facilitate or hinder target-like responses as a result of variations in interlocutor performance.

Conclusions

The descriptive statistics in this study indicate that interlocutor performance effects examinee performance in constructed dialogue tasks assessing English primary phrase stress and intonation usage. Examinees who were required to take phase II of the university's oral part of the EPT seemed to perform better when the interlocutor uses which used experimental conditions rather than control conditions. Groups who were not required to take phase II performed

similarly under experimental and control conditions. It is hypothesized that participants who were required to take phase II were more affected by the experimental conditions as they simply helped to reduce the anxiety of the participants. However, this may also be the result of an interlocutor animation effect as the current research anxiety cannot fully explain the effects found in this study. Research into dynamic assessment may prove to be a suitable alternative to producing similar effects as found this study as requiring a focus on rater/interlocutor performance may overburden the focus of raters.

References

Ayers, G. (1994). Discourse functions of pitch range in spontaneous and read speech. In J. Venditti (Ed.), *Ohio State University working papers in linguistics* (Vol. 44, pp. 1-49). Columbus, Ohio: The Ohio State University Department of Linguistics.

Antón, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*(3), 576-599.

Antón, M. (2012). Dynamic assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 106-119). New York, NY: Routledge.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164.

Baran-Łucarz, M. (2011). The relationship between language anxiety and the actual and perceived levels of foreign language pronunciation. *Studies in Second Language Learning and Teaching, 1*(4), 491-514.

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perception difference between read and spontaneous speech . *Speech Communication, 4*(14), 359-375.

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals, 36*(4), 498-506.

Davidson, F. (1996). *Principles of statistical data handling*. Thousand Oaks, CA: Sage Publications, Inc.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of langauge testing* (pp. 197-207). New York, NY: Routledge.

Finocchiaro, M., & Sako, S. (1983). *Foreign language testing: A practical approach*. New York: Regents Publishing Company, Inc.

Fulcher, G. (2003). *Testing second language speaking*. London ; New York: Longman.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Hewitt, E., & Stephenson, J. (2012). Foreign language anxiety and oral exam performance: A replication of Phillips's MLJ study. *The Modern Language Journal*, *96*(2), 170-189.

International Student and Scholar Services. (2012). *Statistics*. Retrieved from http://www.isss.illinois.edu/about/statistics.html

Lado, R. (1964). *Language testing: The construction and use of foreign language tests*. New York: McGraw-Hill Book Company.

Madsen, H. (1983). *Techniques in testing*. New York: Oxford University Press.

Phillips, E. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal*, *76*(1), 14-26.

Saito, Y., Horwitz, E. K., & Garza, T. J. (1999). Foreign language reading anxiety. *Modern Language Journal*, *83*(2), 202-218.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R.,...van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech?. *Language and Speech*, *41*(3-4), 443-492.

Stansfield, C. W., & Kenyon, D. M. (1992a). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*(3), 347-364.

Stansfield, C. W., & Kenyon, D. M. (1992b). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*(2), 129-141.

Sun, Y. (2011). The Influence of the Social Interactional Context on Test Performance: A Sociocultural View. *Canadian Journal Of Applied Linguistics*, *14*(1), 194-221.

Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems*. (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411-440.

Appendix A: IRB Submission

IRB Exempt Form

Project Title: The Effects of Interlocutors on Student Performance on Constructed Dialogue Tasks Assessing Primary Phrase Stress Production

1.1 Responsible Project Investigator. The RPI must be a non-visiting member of UIUC faculty or staff who will serve as project supervisor at UIUC. Students, interns, post-doctoral researchers, and visiting faculty from other campuses may not serve as RPI, but should be listed as investigators, if applicable.

Last Name: Davidson	First Name: Fred	Academic Degrees: Ph.D
Dept. or Unit: Linguistics	Office Address:	Mail Code:
Street Address:	City:	Zip Code:
Phone:	Fax:	E-mail:
UIUC Status (please mark one): Non-visiting member of X Faculty ☐ Staff		

1.2 Investigators. Please list: All investigators who are different from the RPI, including those from other institutions. Include all persons who will be directly responsible for the project's design or implementation, the consent process, data collection, data analysis, or follow-up.

Last Name: Boyd	First Name: Ryan	Academic Degrees: BA
Dept. or Unit: Linguistics	Office Address:	Mail Code:
Street Address:	City:	Zip Code:
Phone:	Fax:	E-mail:
UIUC Affiliation (please mark one): ☐ Faculty ☐ Staff X Student ☐ Visiting Scholar ☐ Non-UIUC Affiliate of (Institution)		

1.3 Please review the 2 categories of exemption listed below and indicate the category or categories that apply to your research. (Note: Exemptions do NOT apply for prisoners, or for research that <u>specifically targets</u> persons who are cognitively impaired or persons who are economically or educationally disadvantaged.)

☐ **1**. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as: research on regular and special education instructional

strategies, *or* research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

X 2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures (paper or web-based questionnaires, reading of words and phrases [examples included with application], oral description of pictures [examples need to be included with application], retellings of stories or other facts of common knowledge [e.g., Little Red Riding Hood], interview procedures (guided interviews [questionnaire or list of topics included with application]; guided interactions between two or more speakers, free conversations [with the subjects' stated right to have the entirety or any part of their conversation be excluded]), or observation of public behavior (e.g., recordings from publicly-broadcasted internet, radio, and television, publicly available corpora, etc.), ***unless****:*

information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; ***and*** any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

Note: This category does not apply to the following types of research involving children: surveys, interviews, and observations of public behavior when the investigator is a participant in the activities being observed.

Note for researchers collecting audio/video data: Researchers who collect any such data need to specify in the application and in the consent form the duration for which the data will be kept and who will have access to it. The subjects must also be told explicitly that they can ask to have any recording be deleted and thereby excluded from the study. In addition, the consent form needs to have separate consent for a) playing the data at conferences, b) publishing the data (if is to be published, the consent form needs to explain how the data will be published). Data can only be published for not-for profit purposes.

If the proposed research does not qualify in any of these categories, please complete the IRB 1 application form at www.irb.uiuc.edu.

2. Research Summary. In layman's language, please summarize the objectives and significance of the research.

The objectives of the research are to:
1) Identify the effects of interlocutor performance on test taker response in constructed dialogue tasks

2) Determine whether performance training can improve the validity of constructed dialogue tasks designed to test primary phrase stress usage in American English

Significance of the research

1) The research will contribute to the body of knowledge on improving the validity of testing second language speaking
2 a) The research seeks to improve the validity of the Oral English Placement Test

3. Participants. Describe who will participate in this research and how these persons will be recruited.

The participants will be incoming University of Illinois at Urbana-Champaign students who are required to take the Oral English Placement Test and are at least 18 years of age. The participants must not have previously taken the Oral English Placement Test.

Participants will be recruited via an advertisement on the English Placement Test webpage and email from the English Placement Test research assistant to students registered for the Oral English Placement.

The recruitment email is attached.

An incentive will be advertised. The incentive is that students will be able to experience the testing conditions of the Oral English Placement Test by participating in this study prior to taking the actual Oral English Placement Test.

4. Research Procedures. Specifically describe what the participants will do and where the activities will take place. Outline the approximate dates and durations for specific activities, including the total number of treatments, visits, or meetings required and the total time commitment.

Please include a copy of each of your measures as attachments.

Through the recruiting process, participants will be directed to contact the investigator and request one of the designated meeting times.

The participants will meet with the test administrator on the main campus of University of Illinois. They will be asked to sign a consent form. Each consent form will be given a number. Participants will then be asked to complete a questionnaire about the language background. The questionnaire will be given a number corresponding to the number on the consent form.

The language questionnaire is attached.

Participants will meet individually with the test administrator (interlocutor). A participant will engage in a constructed dialogue task. The interlocutor will read his designated line in the dialogue and the participant will read his designated response line for each item. Each participant will meet for one session lasting approximately 20 minutes. Each session will be audio recorded.

5. Data Collection. Please explain how confidentiality will be maintained during and after data collection. If appropriate, address confidentiality of data collected via e-mail, web interfaces, computer servers and other networked information.

The test administrator will ask the student to present a student ID card and compare the ID card with the participant schedule to verify the participant's identity for test scheduling purposes. It is possible that someone may recognize the voice of a participant in a recording. Therefore, the following confidentiality measures will be taken to protect participant data.

Confidentiality will be maintained in that personally identifiable information of the participants will not be recorded during audio recorded parts of the session (i.e. only the constructed dialogue task will be recorded in which participants and interlocutors read a script).

Scoring sheets, audio recordings, and language questionnaires will be given a number to correspond to the numbered consent forms. Recordings will be stored in password-protected, secure server for two years, and backed up in a locked case to which only the investigator has the key and stored in a locked room.

6. Consent Process Describe when and where voluntary consent will be obtained, how often, by whom, and from whom. Attach copies of all consent forms (as well as assent forms for those under age 18 if any).

Consent forms will be distributed to each participant by the test the test administrator prior to the start of each participant's session. Participant consent forms are attached.

7. Dissemination of Results. What is (are) the proposed form(s) of dissemination (e.g., journal article, thesis, academic paper, conference presentation, sharing within the industry or profession, etc.)?
The proposed forms of dissemination are:
1) Thesis
2) Conference presentation
3) Journal article
4) Sharing within the profession

8. Individually identifiable information. Will any individually identifiable information, including images of subjects, be published, shared, or otherwise disseminated? Please mark the appropriate box below.

X Yes

☐ No

Note: If yes, subjects must provide explicit consent or assent for such dissemination. Provide appropriate options on the relevant consent documents.

Consent Form

AGREEMENT TO PARTICIPATE IN RESEARCH (participants)

Responsible Project Investigators: Fred Davidson

Investigator: Ryan Boyd

Departments of Linguistics

University of Illinois at Urbana-Champaign

707 S. Mathews Ave., Urbana, IL 61801, USA

Purpose of this Language Test Research

The purpose of this study is to learn how a test giver can affect a test takers performance on a speaking test. You must be an international student required to take the Oral English Placement Test. You must not have previously taken the Oral English Placement Test. You must be 18 years of age or older.

What You Will Be Expected to Do

If you agree to participate in this research, you will complete: (1) a language background questionnaire in which you will provide information about your language learning experience (approx. 5 min.); (2) an interview in which you will be required to read lines in a dialogue with a test giver (approx. 15 min.). The interview will be audio recorded.

Your Rights to Confidentiality

The obtained data will be treated with absolute confidentiality. You will be given a number to conceal your actual identity. No information will be released that could reveal your identity. All the data will be stored in a secure location and only the responsible project investigators and the investigator will have continuous access to them.

Your Rights to Withdraw at Any Time

Your participation in this research is voluntary. You may withdraw from it or discontinue participation at any time, and you may require that your data be destroyed, without any consequences. The decision to participate, decline, or withdraw from participation will have no effect with your future relations with the University of Illinois nor will it affect your score on any test administered by the University of Illinois.

Benefits and Possible Risks

The benefits to you are that you can experience the Oral English Placement Test before taking the real test. Your participation also benefits the field of language testing and can help improve the Oral English Placement Test. To our knowledge, there are no risks or discomforts involved in this research beyond those found in everyday life.

Dissemination

This research may be disseminated in conferences, and it may be published in conference proceedings and journal articles. Your name will be kept confidential, but it is possible that someone may recognize your voice from the recording. Do you allow the RPI and the investigator to play your recording during conferences as an anonymous participant?

Yes ☐ No ☐

Your Rights to Ask Questions at Any Time

You may ask questions about the research at any time by emailing the responsible project investigator at (email redacted). If you have any questions about your rights as a participant in this study, please contact the University of Illinois Institutional Review Board at (phone number redacted) (you may call collect) or via email at (email redacted).

Giving Consent to Participate

By signing the consent form, you certify that you are 18 years of age or older, that you have read and understand the above, that you have been given satisfactory answers to any questions about the research, and that you have been advised that you are free to withdraw your consent and to discontinue participation in the research at any time, without any prejudice.

Participant: I have read and understand the above information, and voluntarily agree to participate in this research.

_____

Name (printed)

_____ _____

Signature Date

Please keep one copy for your records and return the other copy to the researcher.

Recruitment Email

This email is sent from the EPT research assistant.

[Subject] English Placement Test (EPT) Practice opportunity.

Are you nervous about taking the EPT this semester? Good news, there is an opportunity for you to practice a section of the Oral English Placement Test.

By participating in research study on Phase II of the oral English Placement Test, you will be able to experience the testing conditions before you take the real test.

Your participation in this study is voluntary and will not affect your EPT score.

You must meet the following conditions to participate.

-You must be an international student required to take the Oral English Placement Test

-You must not have previously taken the Oral English Placement Test

-You must be 18 years of age or older

Contact Ryan Boyd at: (email redacted) to schedule an appointment.

Language Background Questionnaire

A. General Information

1. Sex: ☐ F ☐ M
2. Age: _____
3. Do you have vision or hearing problems? _____
4. University year: 1 2 3 4 5 6 7 8 9

   ☐ Undergrad ☐ Graduate ☐ Post Doctoral ☐ Others

5. If you checked "others"specify:

   _____

6. Major: _____

B. Known Languages and Uses

1. Native language: _____ Dialect: _____
2. Mother's native language: _____ Dialect: _____
3. Father's native language:_____ Dialect: _____
4. Language(s) spoken at home during childhood: _____
5. Country of residence during childhood: _____
6. Did you ever attend a school which taught in a language other than your native language?

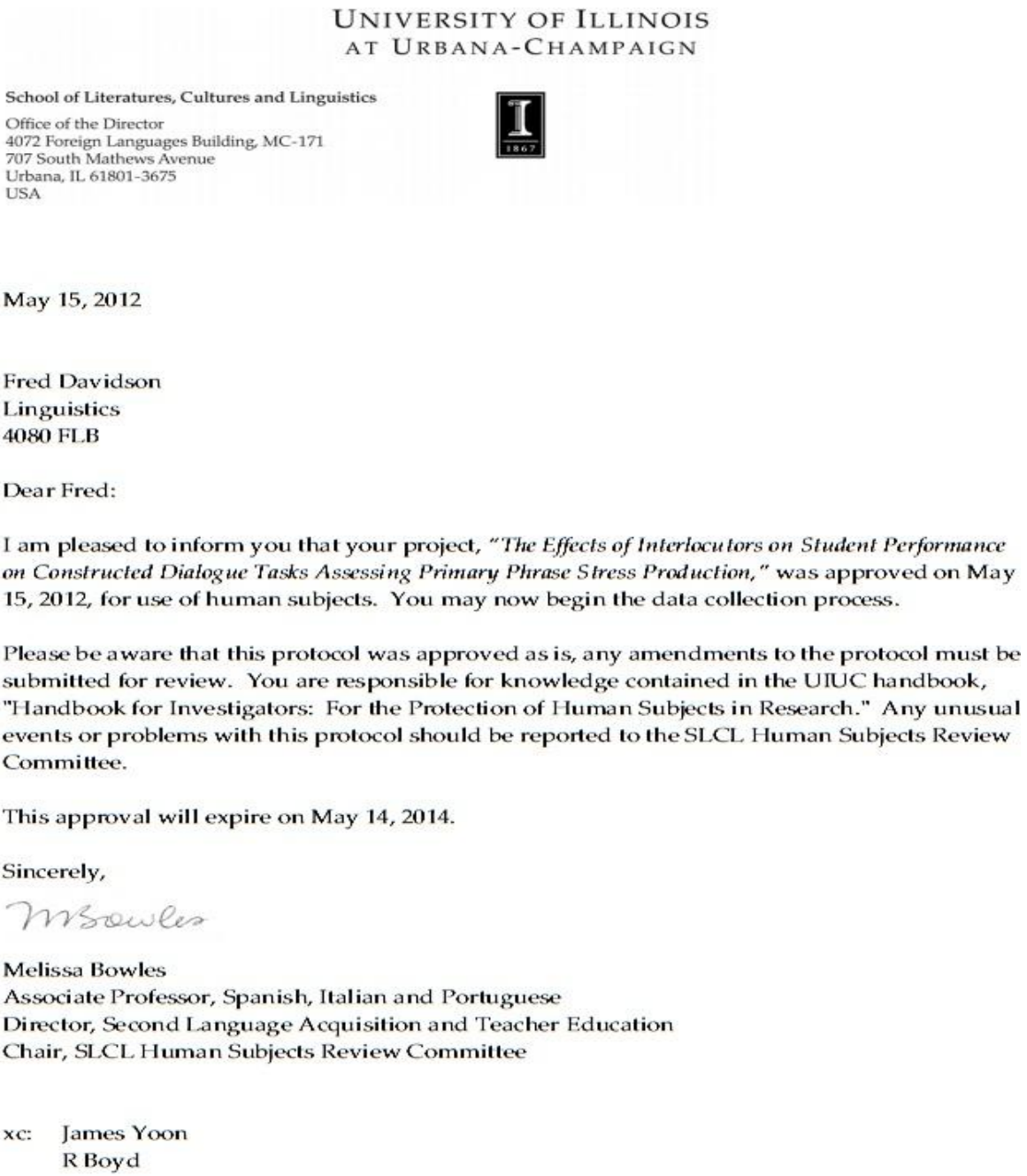
   If so when and in which language?

   _____

7. Other language(s) that you know and proficiency levels

Language	Reading	Writing	Speaking	Listening
English	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native
_ _ _ _ _ _ _	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native
_ _ _ _ _ _ _	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native
_ _ _ _ _ _ _	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native	☐ Beginner ☐ Intermediate ☐ Advanced ☐ Near-native

Figures



UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

School of Literatures, Cultures and Linguistics
Office of the Director
4072 Foreign Languages Building, MC-171
707 South Mathews Avenue
Urbana, IL 61801-3675
USA

May 15, 2012

Fred Davidson
Linguistics
4080 FLB

Dear Fred:

I am pleased to inform you that your project, *"The Effects of Interlocutors on Student Performance on Constructed Dialogue Tasks Assessing Primary Phrase Stress Production,"* was approved on May 15, 2012, for use of human subjects. You may now begin the data collection process.

Please be aware that this protocol was approved as is, any amendments to the protocol must be submitted for review. You are responsible for knowledge contained in the UIUC handbook, "Handbook for Investigators: For the Protection of Human Subjects in Research." Any unusual events or problems with this protocol should be reported to the SLCL Human Subjects Review Committee.

This approval will expire on May 14, 2014.

Sincerely,

MBowles

Melissa Bowles
Associate Professor, Spanish, Italian and Portuguese
Director, Second Language Acquisition and Teacher Education
Chair, SLCL Human Subjects Review Committee

xc: James Yoon
R Boyd

Figure 17: IRB approval letter. This figure shows that the research met IRB regulations.

Appendix B: Test Specification

General Description

The rater and examinee will read their respective turn(s) of each dialogue twice. The rater is expected to read the context of the dialogue. The prompts are printed in the oral component of the EPT interview sheet. The prompts consist of seven dialogues which may be found in examinees' daily lives. Each dialogue asks the examinee to take one or two turns within the length of each dialogue.

Prompt Attributes V1

(Redacted to maintain test security)

Prompt Attributes V2

(Redacted to maintain test security)

Prompt Attributes V3

(Redacted to maintain test security)

Variables

This test is designed to determine if the way in which the interlocutor performs his lines can affect student test performance. The two types of interlocutor performance are an "in character" performance referred to as experimental conditions and a "reading aloud" performance referred to as control conditions. In order to obtain a similar amount of data for each variable, approximately 50% of participants should be tested with the "in character" performance and the other 50% should be tested with the "reading aloud" performance.

The "in character" performance requires the interlocutor to read the scripted dialogue in a way such that the interlocutor assumes the role of a person actually in the situation presented in

the task. As a result the lines should be read with more contextualized emotion, emotion that emerges from the rater adopting the role presented in the task context. Characteristics of such emotion may include dramatic pitch changes, extended pauses, and meaningful word stress assignment. Because each dialogue and context is unique, the same experimental conditions cannot be applied to each task. The focus should be on making the "in character" performance distinct from the "reading aloud" performance.

The "reading aloud" performance requires that the interlocutor read the scripted dialogue lines without assuming the role of a person actually in the situation presented in the task. Instead the rater maintains the identity of a rater and reads the script aloud without contextualized emotion. As a result, there should be no special emphasis other than those resulting from the basic phonological rules of English used in the United States in order to make this performance distinct from the experimental conditions.

Item 1

Guiding Language V1

This item tests… [1]

Guiding Language V2

This item tests primary phrase stress use in contradiction with non-modal auxiliaries [1].

Guiding Language V3

The experimental item tests primary phrase stress use in contradiction with non-modal auxiliaries [1]. Under experimental conditions the examiner script should be spoken at a high pitch from the beginning and transition to a low pitch by the end of the script.

Experimental V1

I am confused about your food preference.

E. I thought you didn't like the taste of fish.

I. Actually, I [1] don't like the taste of fish.

Experimental V2

I am confused about your job status.

E. I thought you were being promoted.

I. Actually, I [1] have been promoted.

Operational Task

(Redacted to maintain test security)

Item 5

Guiding Language V1

This item tests contrast among three parallel phrases requiring two primary stresses per phrase [2-7], … [8], and stress and intonation of one yes/no question [9-10].

Guiding Language V2

This item tests contrast among three parallel phrases requiring two primary stresses per phrase [2-7], stress of a contrast in a parallel phrase using one true-question tag [8], and the intonation of a confirmation tag [9].

Guiding Language V3

The experimental item tests contrast among three parallel phrases requiring two primary stresses per phrase [2-7], and the stress of a contrast in a parallel phrase using one true-question tag [8]. Under experimental conditions the examiner script requires vowel lengthening on *why*.

Operational Task

(Redacted to maintain test security)

Experimental V1

Family members are preparing dinner.

E. Why are you chopping so many vegetables?

I. I'm looking forward to a delicious meal. I've got [2] lettuce for [3] salad, [4] peppers for [5] seasoning, and [6] carrots for making it [7] healthy. It's not [8] that many, [9] is [10] (rising intonation) it?

Experimental V2

Family members are preparing dinner.

E. Why are you chopping so many vegetables?

I. I'm looking forward to a delicious meal. I've got [2] lettuce for [3] salad, [4] peppers for [5] seasoning, and [6] carrots for making it [7] healthy. It's not that many, [8] is [9] (rising intonation) it?

Guiding Language V3

Family members are preparing dinner.

E. Why are you chopping so many vegetables?

I. I'm looking forward to a delicious meal. I've got [2] lettuce for [3] salad, [4] peppers for [5] seasoning, and [6] carrots for making it [7] healthy. It's not that many, [8] is it?

Item 6

Guiding Language V1

This item tests the stress and intonation of one repetition question [11-12] and the stress and intonation of one narrowing question [13-14].

Guiding Language V2

This item tests the stress and intonation of one repetition question [10-11] and the stress and intonation of one narrowing question [12-13].

Guiding Language V3

The experimental item tests the stress and intonation of one repetition question [9-10] and the stress and intonation of one narrowing question [11-12]. Under experimental conditions the examiner should speak *three* and *cheeseburgers* loud and slow with a pause between the two words in both of parts of the examiner's script.

Operational Task

(Redacted to maintain test security)

Experimental V1

Office workers after fall break

E. I ate an entire turkey last week.

I. [11] Wh<u>a</u>t did you [12] (rising intonation) eat?

E. An entire turkey, one of my brothers prepared it.

I. [13] Wh<u>i</u>ch [14] (falling intonation) brother?

<u>Experimental V2</u>

Office workers chatting after the weekend

E. I ate three cheeseburgers yesterday.

I. [10] Wh<u>a</u>t did you [11] (rising intonation) eat?

E. Three cheeseburgers, one of my brothers made them.

I. [12] Wh<u>i</u>ch [13] (falling intonation) brother?

<u>Experimental V3</u>

Office workers chatting after the weekend

E. I ate three cheeseburgers yesterday.

I. [9] Wh<u>a</u>t did you [10] (rising intonation) eat?

E. Three cheeseburgers. One of my brothers made them.

I. [11] Wh<u>i</u>ch [12] (falling intonation) brother?

<u>Item 7</u>

<u>Guiding Language V1</u>

This item tests the stress and intonation of one choice question [15-18], stress on an answer to a choice question [19], and stress on new information on a non-final content word [20].

Guiding Language V2

This item tests the stress and intonation of one choice question [14-17], stress on an answer to a choice question [18], and stress on new information on a non-final content word [19].

Guiding Language V3

The experimental item tests the stress and intonation of one choice question [13-16], stress on an answer to a choice question [17], and stress on new information on a non-final content word [18]. Under experimental conditions *let's watch a movie tonight* should be spoken quickly and include vowel lengthening on the last syllable, *night*. The examiner's second scripted line should be spoken with a lengthening on the first *I*.

Operational Task

(Redacted to maintain test security)

Experimental V1

Two friends are planning a get-together

E. Let's watch a movie tonight.

I. Alright, do you want to watch a [15] scary [16] (rising intonation) movie or [17] funny [18] (falling intonation) movie?

E. I don't mind. I can watch either. What about you?

I. [19] Funny movies are the best because those movies are more [20] entertaining than scary ones.

Experimental V2

Two friends are planning a get-together

E. Let's watch a movie tonight.

I. Alright, do you want to watch a [14] sc<u>a</u>ry [15] (rising intonation) movie or [16] f<u>u</u>nny [17] (falling intonation) movie?

E. I don't mind. I can watch either. What about you?

I. [18] F<u>u</u>nny movies are the best. That's because the funny ones are more [19] entert<u>ai</u>ning the than scary ones.

Experimental V3

Two friends are planning a get-together

E. Let's watch a movie tonight.

I. Alright, do you want to watch a [13] sc<u>a</u>ry [14] (rising intonation) movie or [15] f<u>u</u>nny [16] (falling intonation) movie?

E. I don't care. I can watch either. What about you?

I. [17] F<u>u</u>nny movies are the best. That's because the funny ones are more [18] entert<u>ai</u>ning the than scary ones.

Appendix C: Experimental Test

<u>Rater Instructions</u>: Each of the following items is part of a conversation. I will read the first part. You will read the second part. Make your responses smooth and natural. Let's read each conversation twice.

1. *I am confused about your job status.*

I say. I thought you were being promoted.

You say. Actually, I [1] h<u>a</u>ve been promoted.

2. *Family members are preparing dinner.*

I say. Why are you chopping so many vegetables?

You say. I'm looking forward to a delicious meal. I've got [2] l<u>e</u>ttuce for [3] s<u>a</u>lad, [4] p<u>e</u>ppers for [5] s<u>ea</u>soning, and [6] c<u>a</u>rrots for making it [7] h<u>ea</u>lthy. It's not that many, [8] <u>i</u>s it?

3. *Office workers chatting after the weekend*

I say. I ate three cheeseburgers yesterday.

You say. [9] Wh<u>a</u>t did you [10] <u>ea</u>t?

I say. Three cheeseburgers. One of my brothers made them.

You say. [11] Wh<u>i</u>ch [12] brother?

4. *Two friends are planning a get-together*

I say. Let's watch a movie tonight.

You say. Alright, do you want to watch a [13] sc<u>a</u>ry [14] movie or [15] f<u>u</u>nny [16] m<u>o</u>vie?

I say. I don't care. I can watch either. What about you?

You say. [17] F<u>u</u>nny movies are the best. That's because the funny ones are more [18] entert<u>ai</u>ning than the scary ones.

<u>Participant Instructions</u>: Each of the following items is part of a conversation. I will read the first part. You will read the second part. Make your responses smooth and natural. Let's read each conversation twice.

1. *I am confused about your job status.*

Examiner: I thought you were being promoted.

You say: Actually, I have been promoted.

2. *Family members are preparing dinner.*

Examiner: Why are you chopping so many vegetables?

You say: I'm looking forward to a delicious meal. I've got lettuce for salad, peppers for seasoning, and carrots for making it healthy. It's not that many, is it?

3. *Office workers chatting after the weekend*

Examiner: I ate three cheeseburgers yesterday.

You say: What did you eat?

Examiner: Three cheeseburgers. One of my brothers made them.

You say: Which brother?

4. *Two friends are planning a get-together*

Examiner: Let's watch a movie tonight.

You say: Alright, do you want to watch a scary movie or funny movie?

Examiner: I don't care. I can watch either. What about you?

You say: Funny movies are the best. That's because the funny ones are more entertaining than the scary ones.

Appendix D: Linguistic Foci in Context

1. Task: *I am confused about your job status.*

I say. I thought you were being promoted.

You say. Actually, I [1] have been promoted. 1. Contradiction: Target

2. Contrasts among parallel phrases: Targets 2-7

2. Task: *Family members are preparing dinner.*

I say. Why are you chopping so many vegetables?

You say. I'm looking forward to a delicious meal. I've got [2] lettuce for [3] salad, [4] peppers for [5] seasoning, and [6] carrots for making it [7] healthy. It's not that many, [8] is it?

3. True tag question: Target 8

3. Task: *Office workers chatting after the weekend*

I say. I ate three cheeseburgers yesterday.

You say. [9] What did you [10] eat? 4. Repetition question: Targets 9 and 10

I say. Three cheeseburgers. One of my brothers made them.

You say. [11] Which [12] brother? 5. Narrowing question: Targets 11 and 12

6. Choice question: Targets 13-16

4. Task: *Two friends are planning a get-together*

I say. Let's watch a movie tonight. 7. Answer to a choice question: Target 17

You say. Alright, do you want to watch a [13] scary [14] movie or [15] funny [16] movie?

I say. I don't care. I can watch either. What about you? 8. New information: Target 18

You say. [17] Funny movies are the best. That's because the funny ones are more [18] entertaining than the scary ones.

Appendix E: Figures

Control Group (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Job search	Contradiction	Target 1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
Preparing dinner	Contrasts among parallel phrases	2	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1
		3	0	1	0	0	1	1	0	1	1	0	1	1	0	0	0	1	0	0	0	1	1	0	1
		4	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1
		5	0	1	0	0	1	1	0	1	1	0	1	1	0	0	0	1	0	0	0	1	1	0	1
		6	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1
		7	0	1	0	1	1	1	0	1	1	0	0	1	0	0	0	1	0	0	0	1	1	0	1
		8	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0
Office workers chatting	Repetition question	9	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0
		10	0	0	0	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
	Narrowing question	11	0	1	0	0	0	0	0	1	1	0	1	0	1	0	1	0	1	1	1	1	0	1	1
		12	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	1	0	1	1
Planning a get-together	Choice question	13	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		14	1	1	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
		15	0	1	0	0	0	1	0	1	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0
		16	1	1	1	0	0	0	1	0	1	0	1	1	0	0	1	1	0	0	0	0	1	0	0
	Answer to a choice question	17	0	1	0	1	1	1	0	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1
	New information	18	0	1	1	0	0	1	1	1	0	0	1	0	0	0	1	1	1	0	1	0	1	0	1
		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
		SUM	2	15	2	4	11	13	3	17	15	3	10	7	3	2	4	12	3	4	3	12	12	4	10
		MEAN	0.11	0.83	0.11	0.22	0.61	0.72	0.17	0.94	0.83	0.17	0.56	0.39	0.17	0.11	0.22	0.67	0.17	0.22	0.17	0.67	0.67	0.22	0.56

Figure 18. Control group participant target ratings.

Control Group (Atom)					
Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.13	Contradiction	0.13	Job search	0.13
2	0.35				
3	0.48				
4	0.39				
5	0.48				
6	0.39	Contrasts among parallel			
7	0.48	phrases	0.43	Preparing	
8	0.65	True tag questions	0.65	dinner	0.46
9	0.30				
10	0.26	Repetition question	0.28		
11	0.52			Office workers	
12	0.48	Narrowing question	0.50	chatting	0.39
13	0.26				
14	0.30				
15	0.30				
16	0.43	Choice question	0.33		
17	0.70	Answer to a choice question	0.70	Planning a get-together	0.42
18	0.52	New information	0.52		

Figure 19. Control group atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	MEAN
	Contradiction	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0.13
Control Group Linguistic Focus (Molecule)	Contrasts among parallel phrases	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0.35
	True tag questions	0	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0.65
	Repetition questions	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09
	Narrowing questions	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	1	1	0.26
	Choice questions	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09
	Answer to a choice question	0	1	0	1	1	1	0	1	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1	0.70
	New information	0	1	1	0	0	1	1	1	0	0	1	0	0	0	1	1	1	0	1	0	1	0	1	0.52
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
	SUM	0	5	1	1	3	4	1	7	5	1	4	1	2	2	2	4	2	3	2	2	5	3	4	
	MEAN	0.00	0.63	0.13	0.13	0.38	0.50	0.13	0.88	0.63	0.13	0.50	0.13	0.25	0.25	0.25	0.50	0.25	0.38	0.25	0.25	0.63	0.38	0.50	

Figure 20. Control group participant molecular level linguistic focus ratings and mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	MEAN
	Job search	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0.13
Control Group Tasks (Molecule)	Preparing dinner	0	1	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0.30
	Office workers chatting	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04
	Planning a get-together	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
	SUM	0	2	0	0	1	1	0	3	1	0	0	0	0	0	0	1	0	1	0	0	2	0	0	
	MEAN	0.00	0.50	0.00	0.00	0.25	0.25	0.00	0.75	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.25	0.00	0.00	0.50	0.00	0.00	

Figure 21. Control group participant molecular level task ratings and mean difficulty.

Experimental Group (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Job search	Contradiction	Target 1	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0
		2	1	1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0
		3	1	1	0	1	0	0	1	0	1	0	0	1	1	0	1	0	0	1	0	0	0	1	1	0
		4	1	1	0	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0
	Contrasts	5	1	1	0	1	0	0	1	0	1	0	0	1	1	1	1	0	0	1	0	0	0	1	1	0
	among parallel	6	1	1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	1	0	1	0	1	1	0
	phrases	7	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	1	1	1	1	0	0	1	0
Preparing dinner	True tag questions	8	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0
	Repetition	9	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0
	question	10	1	1	1	1	0	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	1	1	1	0
Office workers	Narrowing	11	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	0	0	0	0	1	1
chatting	question	12	0	1	1	1	1	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	1	0	0	1
		13	1	1	1	1	1	0	0	0	0	1	0	0	1	1	1	1	0	1	1	0	0	0	0	1
		14	1	1	1	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0	1	0	0	1	1	0
	Choice	15	1	1	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0
	question	16	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0
	choice question	17	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
Planning a get-together	New information	18	1	1	0	1	1	1	0	0	0	1	0	1	0	1	1	1	0	1	1	1	0	1	1	0
		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
		SUM	16	17	8	17	9	6	9	2	9	9	6	6	14	10	13	8	4	16	7	4	4	8	11	4
		MEAN	0.89	0.94	0.44	0.94	0.50	0.33	0.50	0.11	0.50	0.50	0.33	0.33	0.78	0.56	0.72	0.44	0.22	0.89	0.39	0.22	0.22	0.44	0.61	0.22

Figure 22. Experimental group participant target ratings.

Experimental Group (Atom)					
Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.17	Contradiction	0.17	Job search	0.17
2	0.38				
3	0.46				
4	0.42				
5	0.50				
6	0.42	Contrasts among parallel			
7	0.75	phrases	0.49		
8	0.58	True tag questions	0.58	Preparing dinner	0.50
9	0.29				
10	0.50	Repetition question	0.40		
11	0.71				
12	0.58	Narrowing question	0.65	Office workers chatting	0.52
13	0.54				
14	0.58				
15	0.33				
16	0.29	Choice question	0.44		
17	0.92	Answer to a choice question	0.92		
18	0.63	New information	0.63	Planning a get-together	0.55

Figure 23. Experimental group atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	MEAN
	Contradiction	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0.17
Experimental Group Linguistic Focus (Molecule)	Contrasts among parallel phrases	1	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0.29
	True tag questions	1	1	1	1	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0	0.58
	Repetition questions	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.17
	Narrowing questions	0	1	1	1	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0	0	1	0.46
	Choice questions	1	1	0	1	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0.25
	Answer to a choice question	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0.92
	New information	1	1	0	1	1	1	0	0	0	1	0	1	0	1	1	1	0	1	1	1	0	1	1	0	0.63
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
	SUM	6	7	3	7	4	3	1	1	2	4	4	3	4	4	4	4	1	6	4	1	2	3	3	2	
	MEAN	0.75	0.88	0.38	0.88	0.50	0.38	0.13	0.13	0.25	0.50	0.50	0.38	0.50	0.50	0.50	0.50	0.13	0.75	0.50	0.13	0.25	0.38	0.38	0.25	

Figure 24. Experimental group participant molecular level linguistic focus scores and mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	MEAN
	Job search	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0.17
Experimental Group Tasks (Molecule)	Preparing dinner	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0.21
	Office workers chatting	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.13
	Planning a get-together	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0.21
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
	SUM	2	3	0	3	1	0	0	0	0	0	1	0	1	1	1	1	0	2	1	0	0	0	0	0	
	MEAN	0.50	0.75	0.00	0.75	0.25	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.25	0.25	0.25	0.25	0.00	0.50	0.25	0.00	0.00	0.00	0.00	0.00	

Figure 25. Experimental group participant molecular level task ratings and mean difficulty.

Control Group not Required to Take Phase II (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13
Job search	Contradiction	Target 1	0	0	1	0	0	0	0	0	1	0	1	0	0
		2	1	1	1	1	0	0	0	0	0	0	1	0	1
		3	1	1	1	1	0	1	1	0	0	1	1	0	1
		4	1	1	1	1	0	0	0	0	0	1	1	0	1
		5	1	1	1	1	0	1	1	0	0	1	1	0	1
	Contrasts among	6	1	1	1	1	0	0	0	0	0	1	1	0	1
	parallel phrases	7	1	1	1	1	0	0	1	0	0	1	1	0	1
Preparing dinner	True tag questions	8	1	1	1	1	1	1	1	1	1	0	1	1	0
	Repetition	9	1	0	1	1	0	1	0	0	0	1	0	0	0
	question	10	0	1	1	1	1	0	1	0	0	0	1	0	0
Office workers chatting	Narrowing	11	1	0	1	1	0	1	0	1	1	1	0	1	1
	question	12	0	1	1	0	1	1	1	1	0	1	0	1	1
		13	1	0	1	1	0	0	0	0	0	1	0	0	0
		14	1	1	1	1	0	0	0	0	0	1	0	0	0
		15	1	0	1	1	0	1	0	0	0	1	0	0	0
	Choice question	16	1	0	0	1	0	1	1	0	0	0	1	0	0
	Answer to a choice question	17	1	1	1	1	0	1	0	0	1	1	1	1	1
Planning a get-together	New information	18	1	0	1	0	0	1	0	0	0	0	1	0	1
		Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13
		SUM	15	11	17	15	3	10	7	3	4	12	12	4	10
		MEAN	0.83	0.61	0.94	0.83	0.17	0.56	0.39	0.17	0.22	0.67	0.67	0.22	0.56

Figure 26. Control group not required to take phase II participant target scores.

Control Group not Required to take Phase II (Atom)

Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.23	Contradiction	0.23	Job search	0.23
2	0.46				
3	0.69				
4	0.54				
5	0.69				
6	0.54	Contrasts among parallel			
7	0.62	phrases	0.63		
8	0.85	True tag questions	0.85	Preparing dinner	0.63
9	0.38				
10	0.46	Repetition question	0.42		
11	0.69			Office workers	
12	0.69	Narrowing question	0.69	chatting	0.56
13	0.31				
14	0.38				
15	0.38				
16	0.38	Choice question	0.37		
17	0.77	Answer to a choice question	0.77		
18	0.38	New information	0.38	Planning a get-together	0.44

Figure 27. Control group not required to take phase II atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	MEAN
	Contradiction	0	0	1	0	0	0	0	0	1	0	1	0	0	0.23
Control Group Not Required to Take Phase II Linguistic Focus (Molecule)	Contrasts among parallel phrases	1	1	1	1	0	0	0	0	0	0	1	0	1	0.46
	True tag questions	1	1	1	1	1	1	1	1	1	0	1	1	0	0.85
	Repetition questions	0	0	1	1	0	0	0	0	0	0	0	0	0	0.15
	Narrowing questions	0	0	1	0	0	1	0	1	0	1	0	1	1	0.46
	Choice questions	1	0	0	1	0	0	0	0	0	0	0	0	0	0.15
	Answer to a choice question	1	1	1	1	0	1	0	0	1	1	1	1	1	0.77
	New information	1	0	1	0	0	1	0	0	0	0	1	0	1	0.38
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	
	SUM	5	3	7	5	1	4	1	2	3	2	5	3	4	
	MEAN	0.63	0.38	0.88	0.63	0.13	0.50	0.13	0.25	0.38	0.25	0.63	0.38	0.50	

Figure 28. Control group participant molecular level linguistic focus ratings and mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	MEAN
	Job search	0	0	1	0	0	0	0	0	1	0	1	0	0	0.23
Control Group Not Required to Take Phase II Linguistic Focus (Molecule)	Preparing dinner	1	1	1	1	0	0	0	0	0	0	1	0	0	0.38
	Office workers chatting	0	0	1	0	0	0	0	0	0	0	0	0	0	0.08
	Planning a get-together	1	0	0	0	0	0	0	0	0	0	0	0	0	0.08
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	13	
	SUM	2	1	3	1	0	0	0	0	1	0	2	0	0	
	MEAN	0.50	0.25	0.75	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.50	0.00	0.00	

Figure 29. Control group not required to take phase II participant molecular level task ratings and mean difficulty.

Control Group Required to take Phase II (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10
Job search	Contradiction	Target 1	0	0	0	0	0	0	0	0	0	0
		2	0	0	0	1	0	0	0	1	0	0
		3	0	0	0	1	0	0	0	1	0	0
		4	0	0	0	1	0	0	0	1	0	0
		5	0	0	0	1	0	0	0	1	0	0
	Contrasts among	6	0	0	0	1	0	0	0	1	0	0
Preparing	parallel phrases	7	0	0	1	1	0	0	0	1	0	0
dinner	True tag questions	8	0	0	0	1	0	1	1	1	0	0
		9	0	0	0	1	0	0	0	1	0	0
Office	Repetition question	10	0	0	0	0	0	0	0	0	0	0
workers		11	0	0	0	0	0	0	1	0	1	1
chatting	Narrowing question	12	0	0	0	1	1	0	0	0	0	0
		13	0	0	1	1	0	0	0	0	0	0
		14	1	0	1	0	0	0	0	0	0	0
		15	0	0	0	1	0	0	0	1	0	0
	Choice question	16	1	1	0	0	1	0	1	1	0	0
	Answer to a choice											
Planning a get-	question	17	0	0	1	1	0	1	0	1	1	1
together	New information	18	0	1	0	1	1	0	1	1	1	1
		Participant #	1	2	3	4	5	6	7	8	9	10
		SUM	2	2	4	13	3	2	4	12	3	3
		MEAN	0.11	0.11	0.22	0.72	0.17	0.11	0.22	0.67	0.17	0.17

Figure 30. Control group required to take phase II participant target ratings.

Control Group Required to take Phase II(Atom)

Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.00	Contradiction	0.00	Job search	0.00
2	0.20				
3	0.20				
4	0.20				
5	0.20				
6	0.20	Contrasts among parallel			
7	0.30	phrases	0.22	Preparing	
8	0.40	True tag questions	0.40	dinner	0.24
9	0.20				
10	0.00	Repetition question	0.10		
11	0.30			Office workers	
12	0.20	Narrowing question	0.25	chatting	0.18
13	0.20				
14	0.20				
15	0.20				
16	0.50	Choice question	0.28		
17	0.60	Answer to a choice question	0.60	Planning a get-together	0.40
18	0.70	New information	0.70		

Figure 31. Control group required to take phase II atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	MEAN
	Contradiction	0	0	0	0	0	0	0	0	0	0	0.00
Control Group Required to Take Phase II Linguistic Focus (Molecule)	Contrasts among parallel phrases	0	0	0	1	0	0	0	1	0	0	0.20
	True tag questions	0	0	0	1	0	1	1	1	0	0	0.40
	Repetition questions	0	0	0	0	0	0	0	0	0	0	0.00
	Narrowing questions	0	0	0	0	0	0	0	0	0	0	0.00
	Choice questions	0	0	0	0	0	0	0	0	0	0	0.00
	Answer to a choice question	0	0	1	1	0	1	0	1	1	1	0.60
	New information	0	1	0	1	1	0	1	1	1	1	0.70
	Participant #	1	2	3	4	5	6	7	8	9	10	
	SUM	0	1	1	4	1	2	2	4	2	2	
	MEAN	0.00	0.13	0.13	0.50	0.13	0.25	0.25	0.50	0.25	0.25	

Figure 32. Control group required to take phase II participant molecular level linguistic focus ratings and mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	MEAN
	Job search	0	0	0	0	0	0	0	0	0	0	0.00
Control Group Required to Take Phase II Tasks (Molecule)	Preparing dinner	0	0	0	1	0	0	0	1	0	0	0.20
	Office workers chatting	0	0	0	0	0	0	0	0	0	0	0.00
	Planning a get-together	0	0	0	0	0	0	0	0	0	0	0.00
	Participant #	1	2	3	4	5	6	7	8	9	10	
	SUM	0	0	0	1	0	0	0	1	0	0	
	MEAN	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.25	0.00	0.00	

Figure 33. Control group required to take phase II participant molecular level task ratings and mean difficulty.

Experimental Group not Required to Take Phase II (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10	11	12
Job search	Contradiction	Target 1	0	0	0	0	0	1	1	0	0	0	0	0
		2	0	1	1	1	0	0	1	0	1	0	0	1
		3	0	1	1	1	0	0	1	0	1	0	0	1
		4	0	1	1	1	0	0	1	0	1	0	0	1
		5	0	1	1	1	0	1	1	0	1	0	0	1
	Contrasts among	6	0	1	0	1	0	0	1	0	1	1	0	1
	parallel phrases	7	1	1	1	0	1	1	1	0	1	1	0	1
Preparing dinner	True tag questions	8	1	1	0	0	1	1	0	0	1	0	1	0
		9	0	1	1	0	0	0	0	0	1	0	0	0
	Repetition question	10	1	1	0	1	0	1	1	0	1	0	1	1
Office workers		11	1	1	1	1	1	1	0	1	1	0	0	1
chatting	Narrowing question	12	1	1	0	1	1	0	1	1	1	1	1	0
		13	1	1	0	0	1	1	1	1	1	0	0	0
		14	1	1	1	0	1	1	1	1	0	0	0	1
		15	0	1	0	0	1	0	0	1	1	0	0	0
	Choice question	16	0	1	0	0	0	0	0	1	1	0	0	0
	Answer to a choice													
Planning a get-	question	17	1	1	1	1	1	1	1	1	1	0	1	1
together	New information	18	0	1	0	0	1	1	1	1	1	1	0	1
		Participant #	1	2	3	4	5	6	7	8	9	10	11	12
		SUM	8	17	9	9	9	10	13	8	16	4	4	11
		MEAN	0.44	0.94	0.50	0.50	0.50	0.56	0.72	0.44	0.89	0.22	0.22	0.61

Figure 34. Experimental group required to take phase II participant target ratings.

Experimental Group not Required to take Phase II (Atom)

Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.17	Contradiction	0.17	Job search	0.17
2	0.50				
3	0.50				
4	0.50				
5	0.58				
6	0.50	Contrasts among parallel			
7	0.75	phrases	0.56	Preparing	
8	0.50	True tag questions	0.50	dinner	0.55
9	0.25				
10	0.67	Repetition question	0.46		
11	0.75			Office workers	
12	0.75	Narrowing question	0.75	chatting	0.60
13	0.58				
14	0.67				
15	0.33				
16	0.25	Choice question	0.46		
17	0.92	Answer to a choice question	0.92	Planning a get-together	
18	0.67	New information	0.67		0.57

Figure 35. Experimental group not required to take phase II atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	MEAN
Experimental Group Not Required to Take Phase II Linguistic Focus (Molecule)	Contradiction	0	0	0	0	0	1	1	0	0	0	0	0	0.17
	Contrasts among parallel phrases	0	1	0	0	0	0	1	0	1	0	0	1	0.33
	True tag questions	1	1	0	0	1	1	0	0	1	0	1	0	0.50
	Repetition questions	0	1	0	0	0	0	0	0	1	0	0	0	0.17
	Narrowing questions	1	1	0	1	1	0	0	1	1	0	0	0	0.50
	Choice questions	0	1	0	0	0	0	0	1	0	0	0	0	0.17
	Answer to a choice question	1	1	1	1	1	1	1	1	1	0	1	1	0.92
	New information	0	1	0	0	1	1	1	1	1	1	0	1	0.67
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	
	SUM	3	7	1	2	4	4	4	4	6	1	2	3	
	MEAN	0.38	0.88	0.13	0.25	0.50	0.50	0.50	0.50	0.75	0.13	0.25	0.38	

Figure 36. Experimental group not required to take phase II participant molecular level linguistic focus ratings and mean

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	MEAN
	Job search	0	0	0	0	0	1	1	0	0	0	0	0	0.17
Experimental Group Not Required to Take Phase II Tasks (Molecule)	Preparing dinner	0	1	0	0	0	0	0	0	1	0	0	0	0.17
	Office workers chatting	0	1	0	0	0	0	0	0	1	0	0	0	0.17
	Planning a get-together	0	1	0	0	0	0	0	1	0	0	0	0	0.17
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	
	SUM	0	3	0	0	0	1	1	1	2	0	0	0	
	MEAN	0.00	0.75	0.00	0.00	0.00	0.25	0.25	0.25	0.50	0.00	0.00	0.00	

Figure 37. Experimental group not required to take phase II participant molecular level task ratings and mean difficulty.

Experimental Group Required to Take Phase II (Atom)		Participant #	1	2	3	4	5	6	7	8	9	10	11	12
Job search	Contradiction	Target 1	0	0	0	0	0	1	0	0	0	1	0	0
		2	1	1	0	0	0	0	0	1	0	0	0	0
		3	1	1	0	0	0	0	1	1	0	0	1	0
		4	1	1	0	1	0	0	0	1	0	0	0	0
		5	1	1	0	0	0	0	1	1	0	0	1	0
	Contrasts among	6	1	1	0	0	0	0	0	1	0	0	1	0
Preparing	parallel phrases	7	1	1	1	1	0	1	1	1	1	1	0	0
dinner	True tag questions	8	1	1	0	1	0	1	1	1	0	1	1	0
		9	1	1	0	0	0	0	0	1	1	0	0	0
Office	Repetition question	10	1	1	0	0	1	0	0	0	0	0	1	0
workers	Narrowing	11	1	1	1	1	0	1	0	1	1	0	0	1
chatting	question	12	0	1	1	1	0	1	0	0	0	0	0	1
		13	1	1	1	0	0	0	0	1	0	1	0	1
		14	1	1	1	0	0	0	0	1	0	1	1	0
		15	1	1	1	0	0	0	0	1	0	0	0	0
	Choice question	16	1	1	1	0	0	0	0	1	0	0	0	0
Planning	Answer to a choice													
a get-	question	17	1	1	1	0	1	1	1	1	1	1	1	1
together	New information	18	1	1	1	1	0	0	1	0	0	1	1	0
		Participant #	1	2	3	4	5	6	7	8	9	10	11	12
		SUM	16	17	9	6	2	6	6	14	4	7	8	4
		MEAN	0.89	0.94	0.50	0.33	0.11	0.33	0.33	0.78	0.22	0.39	0.44	0.22

Figure 38. Experimental group required to take phase II participant target ratings.

Experimental Group Required to take Phase II(Atom)

Target #	MEAN	Linguistic Focus (Atom)	MEAN	Task # (Atom)	MEAN
1	0.17	Contradiction	0.17	Job search	0.17
2	0.25				
3	0.42				
4	0.33				
5	0.42				
6	0.33	Contrasts among parallel phrases	0.42	Preparing dinner	0.45
7	0.75				
8	0.67	True tag questions	0.67		
9	0.33				
10	0.33	Repetition question	0.33		
11	0.67			Office workers chatting	0.44
12	0.42	Narrowing question	0.54		
13	0.50				
14	0.50				
15	0.33				
16	0.33	Choice question	0.42		
17	0.92	Answer to a choice question	0.92	Planning a get-together	0.53
18	0.58	New information	0.58		

Figure 39. Experimental group required to take phase II atomic level target, linguistic focus, and task mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	MEAN
Experimental Group Required to Take Phase II Linguistic Focus (Molecule)	Contradiction	0	0	0	0	0	1	0	0	0	1	0	0	0.17
	Contrasts among parallel phrases	1	1	0	0	0	0	0	1	0	0	0	0	0.25
	True tag questions	1	1	0	1	0	1	1	1	0	1	1	0	0.67
	Repetition questions	1	1	0	0	0	0	0	0	0	0	0	0	0.17
	Narrowing questions	0	1	1	1	0	1	0	0	0	0	0	1	0.42
	Choice questions	1	1	1	0	0	0	0	1	0	0	0	0	0.33
	Answer to a choice question	1	1	1	0	1	1	1	1	1	1	1	1	0.92
	New information	1	1	1	1	0	0	1	0	0	1	1	0	0.58
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	
	SUM	6	7	4	3	1	4	3	4	1	4	3	2	
	MEAN	0.75	0.88	0.50	0.38	0.13	0.50	0.38	0.50	0.13	0.50	0.38	0.25	

Figure 40. Experimental group required to take phase II participant molecular level linguistic focus ratings and mean difficulty.

	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	MEAN
	Job search	0	0	0	0	0	1	0	0	0	1	0	0	0.17
Experimental Group Required to take Phase II Tasks (Molecule)	Preparing dinner	1	1	0	0	0	0	0	1	0	0	0	0	0.25
	Office workers chatting	0	1	0	0	0	0	0	0	0	0	0	0	0.08
	Planning a get-together	1	1	1	0	0	0	0	0	0	0	0	0	0.25
	Participant #	1	2	3	4	5	6	7	8	9	10	11	12	
	SUM	2	3	1	0	0	1	0	1	0	1	0	0	
	MEAN	0.50	0.75	0.25	0.00	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.00	

Figure 41. Experimental group required to take phase II participant molecular level task ratings and mean difficulty.

Target	
1	0.04
2	0.03
3	-0.02
4	0.03
5	0.02
6	0.03
7	0.27
8	-0.07
9	-0.01
10	0.24
11	0.19
12	0.11
13	0.28
14	0.28
15	0.03
16	-0.14
17	0.22
18	0.10

Figure 42. Target mean difficulty differences between experimental group and control group. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Linguistic Focus	Atom	Molecule
Contradiction	0.04	0.04
Contrasts among parallel phrases	0.06	-0.06
True tag questions	-0.07	-0.07
Repetition questions	0.11	0.08
Narrowing questions	0.15	0.20
Choice questions	0.11	0.16
Answer to a choice question	0.22	0.22
New information	0.10	0.10

Figure 43. Linguistic focus mean difficulty differences between experimental group and control group. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Task	Atom	Molecule
Job search	0.04	0.04
Preparing dinner	0.04	-0.09
Office workers chatting	0.13	0.09
Planning a get-together	0.13	0.17

Figure 44. Task mean difficulty differences between experimental group and control group. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Target #	
1	-0.06
2	0.04
3	-0.19
4	-0.04
5	-0.11
6	-0.04
7	0.13
8	-0.35
9	-0.13
10	0.21
11	0.06
12	0.06
13	0.28
14	0.28
15	-0.05
16	-0.13
17	0.15
18	0.28

Figure 45. Target mean difficulty differences between the experimental group and the control group which were not required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Linguistic Focus	Atom	Molecule
Contradiction	-0.06	-0.06
Contrasts among parallel phrases	-0.07	-0.13
True tag questions	-0.35	-0.35
Repetition questions	0.04	0.02
Narrowing questions	0.06	0.04
Choice questions	0.09	0.02
Answer to a choice question	0.15	0.15
New information	0.29	0.29

Figure 46. Linguistic focus mean difficulty differences between the experimental group and the control group which were not required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Task	Atom	Molecule
Job search	-0.06	-0.06
Preparing dinner	-0.08	-0.21
Office workers chatting	0.05	0.09
Planning a get-together	0.13	0.09

Figure 47. Task mean difficulty differences between the experimental group and the control group which were not required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Target #	
1	0.17
2	0.05
3	0.22
4	0.13
5	0.22
6	0.13
7	0.45
8	0.27
9	0.13
10	0.33
11	0.37
12	0.22
13	0.30
14	0.30
15	0.13
16	-0.17
17	0.32
18	-0.12

Figure 48. Target mean difficulty differences between the experimental group and the control group which were required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Linguistic Focus	Atom	Molecule
Contradiction	0.17	0.17
Contrasts among parallel phrases	0.20	0.05
True tag questions	0.27	0.27
Repetition questions	0.23	0.17
Narrowing questions	0.29	0.42
Choice questions	0.14	0.33
Answer to a choice question	0.32	0.32
New information	-0.12	-0.12

Figure 49. Linguistic focus mean difficulty differences between the experimental group and the control group which were required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.

Task	Atom	Molecule
Job search	0.17	0.17
Preparing dinner	0.21	0.05
Office workers chatting	0.26	0.08
Planning a get-together	0.13	0.25

Figure 50. Task mean difficulty differences between the experimental group and the control group which were required to take phase II. Note: Negative numbers represent a situation in which was easier for the control group than the experimental group.