CaMLA Working Papers

2015–04

Variability in the MELAB Speaking Task: Investigating Linguistic Characteristics of Test-Taker Performances in Relation to Rater Severity and Score

Geoffrey T. LaFlair
Northern Arizona University
United States

Shelley Staples
Purdue University
United States

Jesse Egbert
Brigham Young University
United States

Variability in the MELAB Speaking Task: Investigating Linguistic Characteristics of Test-Taker Performances in Relation to Rater Severity and Score

Authors

Geoffrey T. LaFlair
Northern Arizona University
Applied Linguistics

Shelley Staples
Purdue University
Second Language Studies

Jesse Egbert
Brigham Young University
Linguistics and English Language

Table of Contents

About the Authors

Geoffrey LaFlair is a PhD candidate in applied linguistics at Northern Arizona university. His primary research interests lie in the areas of language assessment and statistical methods. His work has been published in *Applied Linguistics*.

Shelley Staples is assistant professor of second language studies at Purdue University. Her research focuses on corpus analyses of specialized spoken and written registers, particularly for applications to language teaching and assessment. Her work has recently been published in journals such as *Applied Linguistics*, *English for Specific Purposes*, and *Journal of English for Academic Purposes*.

Jesse Egbert is assistant professor in the Department of Linguistics and English Language at Brigham Young University. His primary areas of interest are corpus linguistics, register variation, English grammar, and statistical methods. He has published research articles in a variety of peer reviewed journals, including *International Journal of Corpus Linguistics*, *Applied Linguistics*, *Linguistics and Education*, and *Journal of English for Academic Purposes*.

Abstract

The overall goal of this study was to examine the extent to which variability across test-taker performances is captured by score and affected by variability in rater severity. First, a Rasch analysis examined rater severity and rater use of the MELAB speaking scale. Second, the linguistic characteristics of test-taker performances were investigated in terms of their relationship with assigned scores and their relationship to rater severity. The results of the Rasch analyses indicated a wide range of rater severity and underuse of the lower end of the scale. The results of the linguistic analyses showed significant correlations with features of speech, interaction, and language and test-taker score. However, no significant correlations were found between linguistic features of test takers' performances and rater severity. The results of these analyses provide evidence that the linguistic features typical of conversation occur more frequently as performance increases in the MELAB. Additionally, they provide partial evidence that the linguistic features of test-taker language elicited by the MELAB speaking task do not vary across raters.

Introduction

The MELAB speaking assessment is a variant of an oral proficiency interview (OPI). It is similar in structure to the more widely known and researched ACTFL OPI. However, unlike the Official/Certified ACTFL OPI, the scores for MELAB speaking performances are awarded by a single rater. The goal of the MELAB speaking assessment is to provide test takers with an opportunity to demonstrate their conversational language abilities. To date, rater behavior has not been investigated for the MELAB speaking test. Additionally, little research has been conducted on linguistic features that test takers produce. Thus, it is important to systematically review the extent to which raters are interchangeable (e.g., in their severity) and the extent to which features of spoken language elicited by the MELAB speaking test relate to test takers' speaking scores and raters' variability. Both are part of a more general question concerning the degree to which the MELAB speaking assessment varies across raters.

Reliability of ratings is important in all performance-based assessments. Much of the reliability research on OPIs has shown acceptable-to-high rates of traditional measures of inter-rater reliability (Henning, 1992; Thompson, 1995). However, high reliability (Pearson's *r*) can be achieved by raters who differ in difficulty provided they rank order the test takers in a similar fashion. Estimates of rater reliability do not show differences in rater severity or the extent to which the raters are using the scales on the rubric.

Research in performance-based assessment has consistently shown that raters will vary in their degree of severity regardless of training (Bonk and Ockey, 2003; Eckes, 2005; McNamara, 1996; Myford & Wolfe, 2003; Stahl & Lunz, 1996). However, intra-rater reliability remains fairly stable across administrations of performance assessments and can become more consistent with training (Lunz, Wright, & Linacre, 1990; Stahl & Lunz, 1996; Weigle, 1998). Lunz, Wright and Linacre (1990) and Stahl and Lunz (1996) have argued that it may be more beneficial to focus on improving intra-rater reliability and adjusting scores for rater severity using the Rasch model than to attempt to train raters until they are all interchangeable regarding their severity. Rasch analysis can also give insights into how the raters use the rating scale. This analysis can show whether the scale is working as it is intended. It evaluates the raters' use of the scales on the rubric to ensure that with each increase in score that is awarded to the test takers, there is an accompanying increase in ability level. Additionally, it evaluates the extent to which raters use all of the parts of the scale (Linacre, 2012).

Most of the research on test-taker language in OPIs has focused on comparisons with conversation. It has been analyzed largely through the lens of conversation analysis (Brown, 2003, 2005; Johnson & Taylor, 1998; Lazaraton, 1992; van Lier, 1989). These studies have emphasized that interactional elements in OPIs such as turn taking, topic control, and question-response patterns are different from conversation, and that raters vary in their use of these features. While such discussions are

important for understanding the relationship of OPIs to the domain of target language use (TLU), they do not determine the extent to which test-taker language differs in relation to raters' variability in scoring, or whether the variability within test-taker language is related to test-taker score. In addition, this research has primarily employed qualitative methods.

Most quantitative analyses of relationships between test-taker score and test-taker language have focused on independent and integrated monologic tasks. These studies have focused on a small range of features such as number of words, type-token ratio, rate of speech, and number of error-free clausal units (Douglas, 1994; Iwashita, Brown, McNamara, & O'Hagan, 2008; Jamieson & Poonpon, 2013). This research has provided interesting insights into the relationship between linguistic features that are of interest in scoring situations. However, few studies have specifically examined lexico-grammatical features that are known to be related to conversation. The selection of a wider range of linguistic variables based on empirical evidence can help to improve our understanding of what differentiates language learners linguistically at different levels. It allows for the inclusion of features that characterize interaction and language use in spoken discourse, but may not be specified in the rating criteria or the rater training.

Linguistic elements of conversation (e.g., pronouns, contractions, stance features) have long been a part of corpus linguistic analyses (e.g., Biber, 1988; Biber, Johansson, Leech, Conrad, & Finegan, 1999). These studies have revealed the importance of particular linguistic features used in spoken interactive discourse. Along with quantifying the use of interactional features such as turns, discourse markers, backchannels, and questions, these studies have also highlighted the importance of such variables as pronouns, contractions, and stance devices (e.g., adverbials and modals, TO complement clauses such as *want to*, conditionals, and causatives). Through comparisons with written discourse, corpus linguistics has also shown that conversation has fewer features such as nouns (particularly nouns that modify other nouns), nominalizations, attributive adjectives, noun + OF phrases (e.g., *source of water*), and relative clauses (Biber, 1988; Biber et al., 1999). In addition, corpus linguistic studies have shown that grammatical complexity differs across speech and writing, with speech containing more clausal features (e.g., adverbial dependent clauses) and writing containing more phrasal features (e.g., longer noun phrases) (e.g., Biber, Gray, & Poonpon, 2011). These methods have been used widely to describe domains of language use. However, they have less frequently been used in examinations of the elicited language of spoken productive tasks.

Two recent studies that have addressed this issue are Kang (2013) and Biber, Gray, and Staples (2014). Kang (2013) used corpus analysis along with other methods to examine differences across the Common European Framework of Reference (CEFR) proficiency levels. Along with the more standard measures of spoken assessment (e.g., fluency, error-free T-units), other features such as pronouns, nouns, modals, and vocabulary frequency levels were included in the analysis. Key differences were found across CEFR score levels, with higher scoring test takers using fewer first- and third-person pronouns and nouns as well as more second-person pronouns and modals.

Biber et al. (2014) investigated the language used in TOEFL iBT tasks by identifying lexico-grammatical features across different modes, tasks, and score levels, comparing speech and writing to identify the distinctive features of spoken (monologic) tasks in the TOEFL iBT. Features such as adverbs and finite adverbial clauses were used more frequently in spoken tasks, and desire verb + *to* clauses (e.g., *want to*) and clausal *and* features were used less frequently in higher scored spoken responses. Nouns, nominalizations, nouns + OF phrases, attributive adjectives, premodifying nouns, and finite relative clauses were all used less frequently in spoken responses than written responses. While Kang (2013) and Biber et al. (2014) provide important insight into the language used in speaking tasks, both of these studies were based on monologic tasks. None have examined test-taker speech in an OPI context, which focuses on a particular type of dialogic speaking task, namely between the test taker and the rater.

The overall goal of this study is to examine the extent to which variability across test-taker performances is captured by score and affected by variability in rater severity. First, it examines rater severity and rater use of the scale. If raters vary in their ranges of severity, then it calls into question the extent to which the raters are administering the same task, with task operationalized as the test-taker performance. Ineffective use of the scale may indicate that the descriptors in the bands of the rubric do not help the raters identify meaningful differences in test-taker language across levels. The linguistic aspect of this analysis focuses on the relationship of test-taker language to score and rater

severity. Systematic relationships between linguistic features and score would indicate that performances at a particular score level display similar characteristics. Examining the linguistic features that are characteristic of higher scoring performances allows us to determine if higher scores are related to greater use of conversation-like features. Finally, systematic relationships between test-taker language and rater severity would indicate that raters are eliciting different language depending on their severity. If there are not any systematic relationships between rater severity and linguistic characteristics, it could partially support the argument that test-taker performances are the same (linguistically) from one rater to the next. The results of the four research questions below will be used as evidence to establish the extent to which MELAB speaking tasks are similar across raters.

Research Questions

1. Are raters closely aligned in terms of severity?
2. Are raters using the full MELAB speaking rating scale?
3. Is the MELAB speaking score related to linguistic features in test-taker performances? If so, are higher scorers using more linguistic features that are characteristic of conversation?
4. Is rater severity related to linguistic features in test-taker performances?

Methods

This section is divided into four subsections which describe the methods used in the major stages of this study. In the first section we describe the MELAB speaking assessment, with an emphasis on the MELAB speaking dataset used in this study. In the second section we describe the design and construction of the MELAB OPI corpus. In the third section we describe the methods for the various linguistic analyses we performed on the corpus. The final section contains details about the re-rating of the interviews, the Rasch analysis of the speaking scores, and the statistical analyses performed.

MELAB Speaking Dataset

The data for this study come from archived MELAB speaking tasks from the 2013 calendar year. Before the speaking samples were selected, certified MELAB speaking raters were recruited to participate in the study. A total of 22 certified raters from Armenia, Canada, India, Jordan, Portugal, and the U.S volunteered to participate in the study. The raters comprised both native and nonnative speakers of English.

Next, a random sample of 98 speaking samples was selected for consideration in the study from the total number of archived MELAB speaking tasks that had been administered by the 22 raters. The number of samples that were randomly selected from each rater was proportional to their contribution to the total number of tasks that were available for sampling. For the summary statistics and the Rasch analysis the MELAB speaking scale was transformed from the 4 point scale with pluses and minuses to a ten point scale (0–9) (see Table 1). The total number of tasks per rater used in this study ranged from three to seven (see Table 2). The descriptive statistics for these 98 tasks show an average score of 6.73 and a standard deviation of 1.62, which indicates that most of our sample consists of test takers who were awarded scores at the higher end of the scale for the MELAB speaking task. Additionally, as indicated by the descriptive statistics for the Writing, Listening, and GCVR sections in Appendix A and the MELAB 2013 test report, the test takers in this sample received higher mean scores on other parts of the MELAB than the 2013 test taking population.

Table 1: Traditional and Transformed Score Scales for MELAB Speaking Task

Traditional	Transformed
4+	9
4-	8
3+	7
3	6
3-	5
2+	4
2	3
2-	2
1+	1
1	0

The MELAB speaking assessment is a variety of oral proficiency interview (OPI). It consists of a live-rated interview between a certified MELAB speaking rater and a test taker. The task is holistically rated on a ten-point scale (see http://www.cambridgemichigan.org/test-takers/tests/melab/). The descriptors for each of

Table 2: Descriptive Statistics for Raters Included in this Study

Rater	# of tasks	Average original rating	SD of original ratings
1	6	6.57	1.27
2	6	6.67	1.63
3	5	5.20	2.59
4	3	8.00	1.00
5	4	7.50	1.91
6	4	4.25	0.96
7	4	6.00	0.82
8	5	7.80	1.79
9	2	6.00	1.41
10	5	6.20	1.92
11	5	7.20	1.30
12	7	7.86	1.21
13	2	6.50	0.71
14	3	5.00	1.00
15	2	6.50	0.71
16	3	7.33	0.58
17	3	6.67	0.58
18	6	8.33	1.03
19	7	6.29	1.60
20	5	5.80	0.84
21	7	7.86	0.90
22	4	6.50	1.29
Total	98	6.73	1.62

Table 3: Salient Features of MELAB Speaking Tasks

Salient Feature	Subcategories	Examples
Speech	Fluency	rate of speech, pausing/hesitation, prosody
	Intelligibility	accent, articulation, delivery
Interaction	Conversational development	interactional facility, topic development
	Conversational comprehension	mutual comprehension (test taker comprehensibility and rater speech adjustment)
Language	Grammar Vocabulary	lexical range, use of lexical fillers, utterance length, utterance complexity, syntactic control, morphology

the four scales describe the test taker's communicative ability in three broad domains of salient features (see Table 3). The first of these salient features is *Speech*. In the rubric, this is described as the test taker's ability to communicate fluently and intelligibly with regard to pace, accentedness, and degree of similarity to native-like pronunciation. The next salient feature is *Interaction*. This category contributes to the holistic score regarding the test taker's ability to maintain/initiate and understand/be understood in conversation. The third salient feature that raters consider when awarding a score is *Language*. This category includes lexical range, complexity and length of utterances, and control of syntax and morphology.

Corpus

The MELAB OPI corpus comprises transcriptions of five-minute segments of 98 MELAB OPIs. While this limitation to five minutes was less than ideal, it was necessary due to budget constraints. A qualitative comparison of several of these five minute segments to the complete interview did not reveal any major differences in the content or linguistic nature of the interviews. These segments are composed of the first five minutes of the audio file after the beginning of the interview. In some of the interviews, the first one or two minutes of the interaction consisted only of formalities, such as the examiner ensuring that he or she had properly recorded the test taker's information. In these cases, the transcription began after these formalities were resolved. Each of the 98 transcribed texts was split into two texts, one for the speech of the MELAB rater and one for the speech of the MELAB test taker. This resulted in a total of 196 texts. Each of these 196 files contains the following information: unique exam ID, testing center ID, test-taker ID, test date, test-taker age, test-taker gender, test-taker L1, reason for test, original rater score, and second rater score. Each interaction was transcribed by a trained transcriber using a standard set of transcription conventions that were modified from those used for the TOEFL 2000 Spoken and Written Language (T2K-SWAL) corpus (see Biber, 2006). After transcription, each text was manually examined, and all inconsistencies were standardized. Details regarding the length of these texts and the overall MELAB OPI corpus are contained in Table 4.

Identification of Linguistic Features

Before beginning our study, we identified a variety of linguistic features that we were interested in exploring based on previous findings on spoken conversational language, primarily from corpus linguistics (Biber, 1988, 2006; Biber et al., 1999; Csomay, 2005; Friginal, 2009; Staples, 2015; White, 1994). The linguistic variables can be grouped according to the organizational framework provided in Table 3: speech features, interaction features, and language features. While many of the features in the categories of speech and interaction are likely familiar from previous studies of OPIs, for the category of language we included features known to characterize conversational discourse. Features such as pronouns and contractions are commonly discussed as indicating a high degree of interactivity. Expressions of stance (e.g., adverbials, modals, and complement clauses) are another characteristic of conversation. Speakers can express their stance (personal feelings, attitudes, and evaluations) more overtly/explicitly by using features such as first person pronouns and stance verbs (e.g., *I want to*), that overtly mark the agent (the speaker). Alternatively, speakers can express stance more implicitly, by using adverbials (*certainly, actually*), in which the speaker does not need to be overtly identified as the agent of the stance (Biber et al., 1999, p. 864–865). We also included features of grammatical complexity based on previous research showing that speech contains more clausal features (e.g., adverbial dependent clauses) in comparison with writing (e.g., Biber et al., 2011). Finally, vocabulary use was examined—conversation is characterized by many more high frequency words than writing (Biber et al., 1999). We also included several features that are more characteristic of writing (e.g., nouns, nominalizations) to provide a measure of

Table 4: Design of the MELAB OPI Corpus

Sub-Corpus	Texts	Mean words per text	Total words
MELAB Rater	98	305.9	29,980
MELAB Test Taker	98	465.2	45,588
Total	196	386	75,568

contrast. The complete list of features we measured, along with examples, are included in Table 5.

In the category of speech, we investigated the test takers' rate of speech and hesitation markers (also called filled pauses). Rate of speech was operationalized as syllables per second. Audio files were first segmented by speaker (rater or test taker), as were the transcripts. The number of syllables for each utterance was identified by using the freeware website http://www.readability-score.com/. The site was chosen as the most reliable freeware site after multiple tests comparing hand counts and automatic counts. The number of syllables was then divided by the length of the speaker's utterances (including intra-turn pauses), in seconds. Hesitation markers were operationalized as the forms *um*, *uh*. A computer program was written to determine the frequency of such forms in each interaction for each speaker (focusing on test-taker speech). The forms were then normed per 100 words.

Within the category of interaction, we examined the following features: discourse markers, backchannels, number of questions, number of turns, and turn length (in words). Computer programs were written to identify all of these features. The coding principles for discourse

Table 5: Linguistic Features Included in this Study

Category	Features
Speech	• Rate of speech (syllables per second) • Hesitation markers (e.g., *um*, *uh*)
Interaction	• Discourse markers (e.g., *now*, *well*) • Backchannels (e.g., *yeah*, *uh huh*) • Number of questions • Number of turns • Turn length (measured in words)
Language	• Contractions • Pronouns (1st, 2nd, and 3rd person) • Vocabulary (1–500 most frequent) • Vocabulary (501–3,000 most frequent) • Vocabulary (not among 3,000 most frequent) • Certainty adverbials (e.g., *certainly*, *of course*) • Likelihood adverbials (e.g., *maybe*, *possibly*) • Possibility modals (e.g., *could*, *might*) • Linking adverbials (e.g., *also*, *besides*) • Finite adverbial clauses (e.g., *When you came back* did you like show your pictures to your family back where you're from?) • Because clauses (e.g., Now I'm happy *because I take the lesson driver and I can drive.*) • If clauses (e.g., *If I have long break* I prefer to go to other state.) • WH-relative clauses (e.g., I want to work in hotels *which will be in five stars.*) • THAT-relative clauses (e.g., What was your favorite thing at Disney World *that you saw*?) • TO verb complement clauses (e.g., I want to *study mechanical engineering.*) • THAT verb complement clauses (e.g., Do you think *that you'll continue to live in the same place*?) • Attributive adjectives (e.g., *good* job, *new* friends) • Premodifying nouns (e.g., *sales* job, *language* class) • Nouns • Nominalizations (e.g., *admission*, *education*) • Noun + OF (e.g., source *of water*)

markers were based on previous research, including an identification of the forms commonly used as discourse markers (e.g., *well, now*), an identification of the form's position in the utterance, and an examination of the lexical items in the immediate environment around the form.

Backchannels were more easily identified by their form (e.g., *yeah*), the lack of words following the form before the next turn, and the lack of a question in the turn preceding the form (to distinguish the backchannel from a minimal response). We ensured that the frequency counts were accurate by manually counting features from sample texts and comparing results from the automated counts. The following dialogue shows the coding of discourse markers <dm>, backchannels <bc>, hesitation markers <hes> and questions <qu>:

Rater:	What uh <hes> which schools are you applying for, are you interested in going? <qu>
Test Taker:	Now? <qu>
Rater:	Yeah. [minimal response; not coded as backchannel or discourse marker]
Test Taker:	Yeah, <dm> uh <hes> now I uh <hes> I am in Eastern Michigan University.
Rater:	Okay. <bc>
Test Taker:	I st- now I study just language.
Rater:	Uh huh. <bc>

The frequency of discourse markers, backchannels, and number of questions were calculated for each speaker in each interaction and normed to 100 words. The number of turns was not normed and the number of words per turn was calculated by determining the number of words (excluding backchannels) and dividing that by the number of turns.

Most of the features in the category of language were annotated automatically using the Biber tagger and counted using a program (also developed by Biber) called TagCount (Biber, 1988, 2006). The Biber tagger is a rule-based and probability-based tagger that has been widely used in corpus research since the late-1980s. TagCount is a program that automatically calculates the normed rates of occurrence (per 1,000 words) for more than 150 linguistic features in corpus texts. However, in this study we investigate the use of only a small subset of these linguistic variables. We used a program in conjunction with the Biber tagger called Complexity, developed by Gray (2011), which measures a range of syntactic features associated with phrasal compression and clausal elaboration (see Gray, 2011). We transformed all rates of occurrence for these linguistic features to per 100 words.

After an analysis of the accuracy of the tags in our corpus, we deemed it necessary to correct the tags for several features. Some of these features were corrected automatically using Perl scripts. Other features were corrected manually in the corpus using an interactive fix-tagging program developed for use in Biber and Gray (2013).

Vocabulary use was examined in terms of word frequency. We measured word frequency using an online tool called WordandPhrase (http://www.wordandphrase.info) (Davies, 2011–). This tool uses a large list of the most frequent words in English to report the percent of a text that is composed of words in various frequency bands, 1–500 most frequent words, 501–3,000 most frequent words, and words that are not among the 3,000 most frequent words.

Analysis

The analysis of rater severity was conducted using the Rasch one-parameter model. To achieve a link among raters for the Rasch analysis, each speaking sample was rated a second time by CaMLA certified raters who agreed to participate in this study. The original design used a computer program to assign second ratings to ensure that the second rater was different than the first. In total, 19 of the original 22 participants completed their assigned ratings. The 12 ratings that were assigned to participants who did not complete the study were reassigned to the other participants based on their availability. The analysis was conducted using Facets (Linacre, 2012). Because raters were the focus of the analysis, test takers were centered at zero and the rater facet was allowed to float.

The rates of occurrence for each of the features listed in Table 5 were calculated for the MELAB test takers. These features were used to measure (a) relationships between MELAB speaking score and test-taker language use, and (b) relationships between language production and MELAB rater severity. To examine the relationship between speaking score and test-taker language use,

correlations were conducted between the normed rates of occurrence for the linguistic features listed in Table 5 and the initial MELAB score received by the test taker. In addition, a multiple regression was conducted to determine the weight of the individual linguistic features and combined predictive power. To investigate relationships between the test-taker language production and MELAB rater severity, we conducted correlations between the normed rates of occurrence for the linguistic features and a measure of rater severity.

Results

This study had four research questions. The first two research questions investigated the range of rater severity in a sample of raters from the MELAB speaking tasks as well as their use of the MELAB speaking task rubric categories. The third research question examined the occurrence of individual linguistic features produced by test takers in the MELAB speaking task and their relationship to speaking score. The fourth research question investigated the relationship between the linguistic features and rater severity.

To answer the first two research questions we conducted a Rasch analysis using Facets (Linacre, 2012). Following criteria from Linacre (2012), mean-square fit statistics between 0.50 and 1.50 were considered productive for measurement. Overall, the raters fit the Rasch model (see Appendix A). Raters 10 and 13 showed considerable underfit and overfit, respectively. It was decided that both raters would be included in the remaining analyses for two reasons. The first is related to the effect that the over- and underfitting measures have on measurement. While these patterns of rater fit are outside of the desired range, Linacre (2012) argues that fit statistics of less than two and greater than 0 are not degrading. Although the infit measure for Rater 13 is 2.01, its standard error is 0.69. Thus Rater 13's true measure may be well within the acceptable limits or well outside. The second reason is related to one of the foci of this study—rater comparability in the MELAB speaking task. Because the goal of the study is to examine what variation in rater leniency/severity exists, excluding raters would limit the ability to assess variation across raters in the sample.

The analysis of infit and outfit mean-square statistics also provided an indication of how the raters are using the scale—or a way of diagnosing rater scale usage (Linacre, 2012). Infit and outfit statistics that fall below 0.50 are considered "muted" and indicate the use of a narrow range of values on the rating scale (also called overfit). Fit statistics above 1.50 are termed "noisy" (or under fit), and these indicate overuse of extreme categories on the rating scale Linacre (2012)—or the unpredictability of a rater's score. Three raters over fit the model indicating overuse of a small set of scale

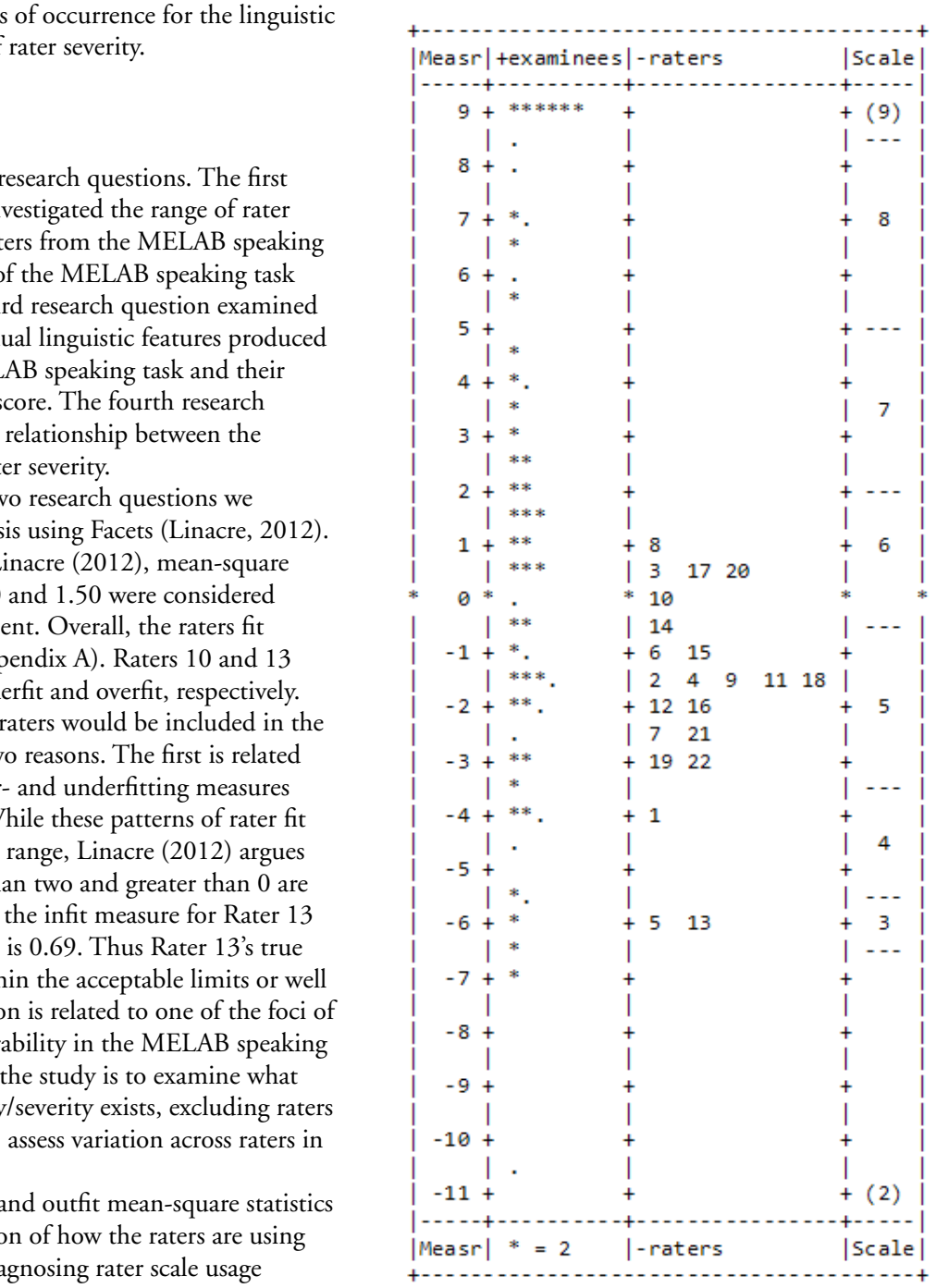

Figure 1: Facets Map of Estimated Test-taker Ability, Rater Severity, and Scale Use.

categories. Two raters underfit the model, indicating the overuse of categories at the ends of the scale. The two underfitting raters were estimated to be the two most lenient raters.

To answer the rater severity question, Table 6.0 from the Facets output was examined (see Figure 1). Figure 1 has four columns of information. The first column represents the logit scale on which the test-taker ability and rater severity can be evaluated. The second column titled *examinees* shows the estimated range of test-taker ability from our sample. Because test takers were not the focus of this study, they are centered at zero and the results from test takers will not be discussed. The next two columns are the columns of interest for Research Question 1 and Research Question 2.

The *raters* column shows the range of rater severity. Raters falling near the top of Figure 1 are more severe, and raters at the bottom of Figure 1 are more lenient. As can be seen, the range of rater severity was 6.69 logits, which is the equivalent of two full score points on the transformed MELAB scale and a full score point on the traditional MELAB scale. This indicates that the raters in this sample varied considerably in their range of severity. There were three especially lenient raters, who had a severity estimate below -3.50 logits.

The *scale* column also provides information to indicate the extent to which the raters used the full scale of the MELAB speaking rubric. This column shows the highest and lowest categories that the raters used. In addition, it shows the likelihood of a score to be given an ability estimate. For example, test takers who are estimated to be at an ability level between about 6–7 logits would likely receive an 8 (MELAB 4-). Another more precise visualization of these results can be found in Figure 2 below, which shows the ability level in logits on the x-axis and the probability of receiving a score on the y-axis. The possible score points are represented by the numbered lines in the plot. The points on the x-axis at which the numbers peak show the ability level of the test takers that will most likely receive that score. It can be seen from the plot that test takers whose ability estimates are 4 and 7 logits are most likely to receive scores of 7 (MELAB 3+) and 8 (MELAB 4-) respectively. In addition, all of the score bands have a distinct peak above an ability level except for 3 (MELAB 2). This is an indication that this sample of raters was unlikely to award this score to the test takers.

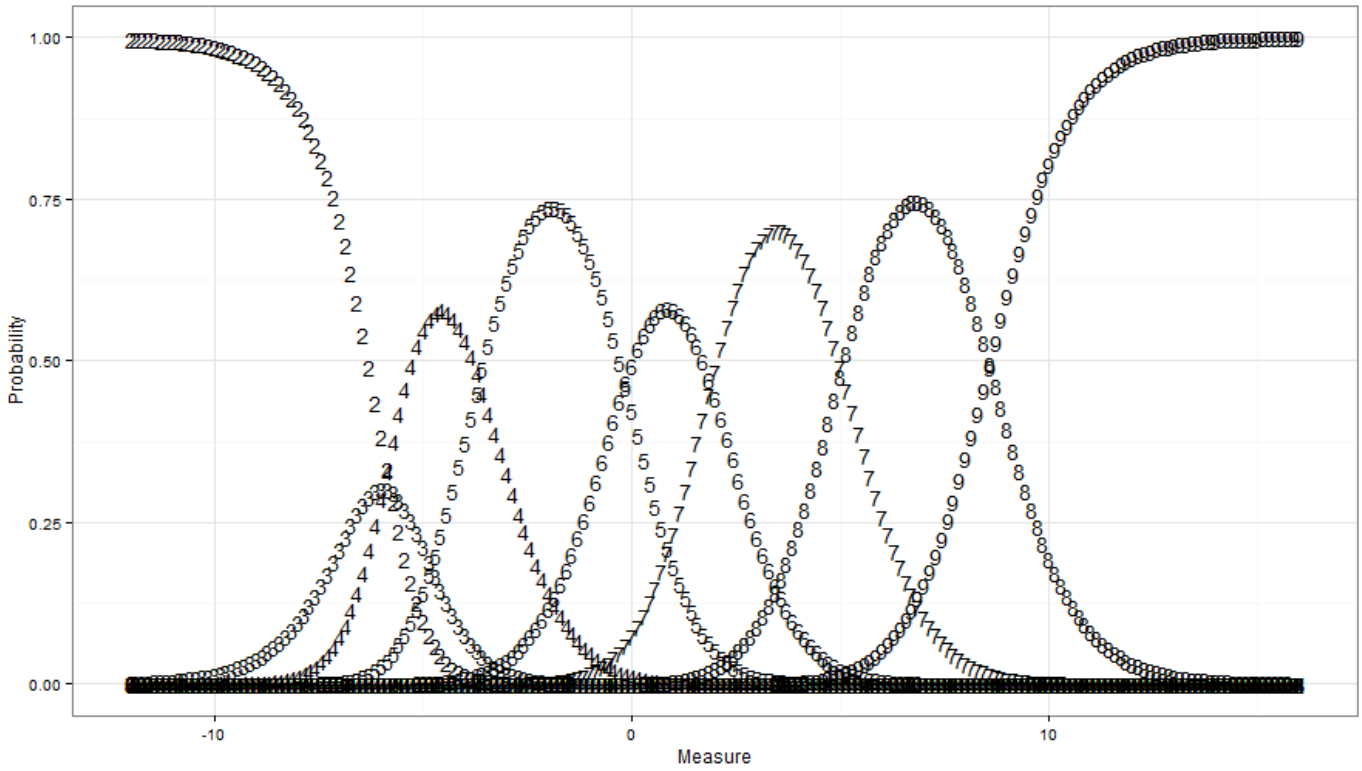

Figure 2: Scale Category Probability Curves

Table 6: MELAB Scale Category Statistics

Category Counts				Quality Control		Rasch-Andrich Thresholds	
Score	**Used**	**%**	**Cum. %**	**Measure**	**Outfit Mnsq**	**Measure**	**S.E.**
2 (2-)[1]	4	2%	2%	-6.21	0.60		
3 (2)	4	2%	5%	-5.24	0.90	-5.76	0.67
4 (2+)	14	8%	13%	-3.85	1.20	-5.95	0.55
5 (3-)	33	19%	31%	-1.53	0.90	-3.65	0.39
6 (3)	37	21%	52%	0.98	0.70	-0.16	0.30
7 (3+)	48	27%	80%	3.35	0.70	1.91	0.27
8 (4-)	26	15%	94%	6.28	1.10	5.03	0.32
9 (4)	10	6%	100%	8.20	1.70	8.58	0.46

[1] MELAB scale conversions are in parentheses.

The corresponding table for Figure 2 (Table 6) shows the same pattern. The first group of columns in the table (*Category Counts*) provides frequency information about the use of the rubric. *Category Counts* indicates how often the scale category was used (*Used*), the percent of ratings that the counts represent (*%*), and the cumulative percentage (*Cum. %*). It can be seen that the lower end of the MELAB scale is used much less frequently than the higher end. In the sample for this study, there was not a single score awarded from the lowest possible category. The second group of columns (*Quality Control*) contains a column for the average ability *Measure* associated with each scale category and its outfit mean-square (*Outfit Mnsq*) value. The average measure for each rating scale category increased as the category increased. This indicates that the scale is functioning as it was intended to (a higher score = more ability). The *Outfit Mnsq* column shows the fit of each category. As with the rater analysis, the fit for scale categories should fall between 0.50 and 1.50. The highest category exceeded 1.50, indicating that there was some unpredictability in the use of the highest category. The third group of columns (*Rasch-Andrich Thresholds*) provides information about the ability level at which the probability of being awarded a higher score was equal to the probability of receiving a lower score. *Rasch-Andrich Thresholds* has two subcolumns. The *Measure* column shows the ability level at which a test taker has an equal probability of falling into one of the adjacent categories. This is also reflected in Figure 2 where the curves intersect. For example a test taker with an ability of -5.95 would be as likely to receive a 4 (MELAB 2+) as a 3 (MELAB 2). Here it is also important to see an ordered increase in measure with each increase in the scale. None of the categories show disorder, however the *Rasch-Andrich Measure* values for a 4 (MELAB 2+) and a 3 (MELAB 2) were within each other's standard error of measure. This suggests that the two scores may not have been applied as distinct score points. Given this overlap in the *Rasch-Andrich Measure*, and the overlap of the scale probability curves in Figure 2, it appears that for the lower range of ability levels, the plus/minus distinction could be problematic when rating lower ability level examinees. However, the scales that represent higher abilities were distinct and ordered. As most of the examinees were assigned scores in the higher ability range, it appears that the scale does a good job of targeting those examinees (Bonk & Ockey, 2003).

The third and fourth research questions focused on test-taker speech. First, the linguistic characteristics of test-taker performances were identified. These characteristics were then investigated for relationships with MELAB speaking scores and rater severity. Descriptive statistics for the test-taker group, including minimums, maximums, means, and standard deviations for the thirty linguistic features included in this study can be seen in Table 7. The table shows that test takers are using many features of conversation (e.g., questions, contractions, and pronouns). However, they are also using features more characteristic of writing, such as nouns. Importantly, the *Min* and *Max* columns show that there was a wide range in variability in some of the linguistic features such as hesitation markers, contractions, first-person pronouns. This variability is explored in more depth below in relation to score level.

Table 7: Descriptive Statistics of Linguistic Features used by Test Takers Normed to 100 Words (except where indicated)

Linguistic Feature	Min	Max	Mean	SD
Rate of speech[1]	2.05	5.29	3.29	0.63
Hesitation markers	0.24	17.63	6.40	4.14
Discourse markers	1.02	13.18	4.41	1.84
Backchannels	0.00	7.27	0.99	1.39
Number of questions	0.00	3.76	0.45	0.56
Number of turns[2]	14.00	72.00	36.24	12.73
Turn length[3]	4.37	31.61	13.39	6.02
Contractions	0.44	6.37	3.47	1.41
1st person pronouns	2.20	13.88	9.32	2.45
2nd person pronouns	0.00	6.24	1.37	1.35
3rd person pronouns	0.00	6.82	1.63	1.38
Vocabulary (1–500 most frequent)[4]	56.0	83.0	71.0	6.0
Vocabulary (501–3,000 most frequent)[4]	4.0	13.0	8.0	2.0
Vocabulary (Not among 3,000 most frequent)[4]	10.0	36.0	21.0	5.0
Certainty adverbials	0.00	4.20	0.70	0.69
Likelihood adverbials	0.00	2.01	0.24	0.32
Possibility modals	0.00	2.50	0.60	0.55
Linking adverbials	0.00	4.88	1.86	1.07
Finite adverbials	0.00	3.38	1.03	0.66
Because	0.00	2.19	0.67	0.47
If	0.00	2.17	0.29	0.42
WH-relative clauses	0.00	1.36	0.10	0.20
THAT-relative clauses	0.00	2.70	0.49	0.49
TO verb complement clauses	0.00	1.19	0.16	0.25
THAT verb complement clauses	0.00	0.74	0.13	0.18
Attributive adjectives	0.26	4.84	1.98	0.85
Pre-modifying nouns	0.00	4.27	1.13	0.78
Nouns	10.22	22.61	15.46	2.29
Nominalizations	0.54	6.03	2.57	1.18
Noun + OF	0.00	2.30	0.63	0.53

[1] Speech rate was measured as syllables per second
[2] Number of turns was not normed
[3] Turn length was measured as number of words per turn
[4] Vocabulary for each of the three frequency levels was measured as percentage of total vocabulary within the particular frequency level

The third research question asked whether MELAB speaking score is related to linguistic features in test-taker performances. In order to answer this question, we measured the relationship between the MELAB speaking rating scale and the linguistic features listed above. The results of these correlations can be seen in Table 8 below. While there were many statistically significant relationships at the $p < 0.05$ level, we will focus on interpreting the correlations with absolute values that are greater than |0.25|. Although correlations at or near |0.25| are relatively weak, we wanted to be as inclusive as possible in our interpretation of the results due to the exploratory nature of this study.

The strongest correlations were between speaking score and the linguistic variables in the speech category. Rate of speech and hesitation markers showed correlations of 0.69 and -0.52, respectively. This shows that higher scoring test takers typically speak faster and hesitate less in their speech. In other words, the results show that speaking score is closely related to fluency.

The results for the category of interaction showed one significant correlation: a moderate positive relationship between speaking score and number of turns ($r = 0.33$). This finding suggests that higher scoring test takers had a greater level of interactivity in their performances.

In the category of language, significant correlations over |0.25| were found between speaking score and vocabulary, contractions, first person pronouns, certainty adverbials, and TO verb complement clauses. A positive relationship was found between speaking score and the use of highly frequent vocabulary ($r = 0.31$), and a corresponding negative relationship between lower frequency words and speaking score ($r = -0.30$). This suggests that successful test takers rely on higher frequency words that are likely to be familiar to a native speaker, and that they tend to avoid words that are more technical and specialized.

There was a positive relationship between speaking score and two types of adverbials—certainty adverbials ($r = 0.40$) and likelihood adverbials ($r = 0.25$). Certainty and likelihood adverbials are frequent in conversation (Biber et al., 1999). In fact there is some evidence showing that certainty adverbials are used more frequently in conversation than in other spoken contexts (Staples & Biber, 2014).

Speaking score was positively correlated with contractions ($r = 0.27$), another linguistic feature commonly found in conversation (Biber et al., 1999). Additionally, higher scores were often characterized by

Table 8: Results of Bivariate Correlations Between Melab Speaking Scores and Thirty Linguistic Features

Linguistic Feature	Correlation with Speaking Score (*r*)
Rate of speech	0.69*
Hesitation markers	-0.52*
Discourse markers	0.16
Backchannels	0.08
Number of questions	-0.17
Number of turns	0.33*
Turn length	0.06
Contractions	0.27*
1st person pronouns	-0.28*
2nd person pronouns	0.17
3rd person pronouns	0.03
Vocabulary (1–500 most frequent)	0.31*
Vocabulary (501–3,000 most frequent)	-0.07
Vocabulary (not among 3,000 most frequent)	-0.30*
Certainty adverbials	0.40*
Likelihood adverbials	0.25*
Possibility modals	-0.02
Linking adverbials	0.24*
Finite adverbials	0.00
Because	0.00
If	0.00
WH-relative clauses	0.12
THAT-relative clauses	0.23*
TO verb complement clauses	-0.25*
THAT verb complement clauses	0.00
Attributive adjectives	-0.03
Pre-modifying nouns	-0.03
Nouns	-0.21*
Nominalizations	0.13
Noun + OF	0.11

* The correlation was significant at $p < 0.05$.

the use of fewer first person pronouns (r = -0.28). This finding suggests that test takers achieve higher scores on the MELAB speaking task when their speech is less speaker-focused.

Somewhat surprisingly, the results showed that only one of the linguistic variables associated with grammatical complexity was related to speaking score. TO complement clauses had a small negative correlation with speaking score (r = -0.25). These clause structures were often found in combination with first person pronouns (e.g., *I want to . . .*). This pattern is explored further in the discussion below.

As a final step toward answering the third research question, we performed a multiple regression analysis to determine the extent to which MELAB speaking score can be predicted by multiple linguistic predictor variables. In this regression analysis, we entered all linguistic variables that correlated at or above |0.25| as independent variables. A stepwise multiple regression revealed that five of the linguistic variables explained 58.3% of the variance (R^2 = 0.58, $F(5, 92)$ = 25.77, $p < 0.001$). Together, these results show that test takers are likely to receive higher scores when they use the linguistic features below. The regression results support the findings from the correlation analyses reported above, providing further evidence that features related to fluency (syllables per second, hesitation markers) and features of spoken discourse (adverbials, pronouns) are good predictors of MELAB OPI scores.

- More syllables per second (β = 0.47, t = 5.37, $p < 0.001$)
- Fewer hesitation markers (β = -0.20, t = -2.44, p = 0.02)
- More likelihood adverbials (β = 0.15, t = 2.18, p = 0.03)
- Fewer first person pronouns (β = -0.18, t = -2.50, p = 0.01)
- More certainty adverbials (β = 0.18, t = 2.42, p = 0.02)

The final research question asked how rater severity is related to the linguistic features in test-taker performances. To answer this question, correlations were conducted between the normed rates of occurrence for the test-taker speech and the rater severity measure. Table 9 shows correlations between the relationships between the linguistic features and rater severity.

Table 9: Correlations Between Linguistic Features and Rater Severity

Linguistic Feature	Correlation with Rater Severity (r)
Rate of speech	0.19
Hesitation markers	0.00
Discourse markers	0.04
Backchannels	-0.14
Number of questions	-0.10
Number of turns	0.05
Turn length	0.06
Contractions	0.01
1st person pronouns	-0.10
2nd person pronouns	0.07
3rd person pronouns	-0.06
Vocabulary (1–500 most frequent)	0.02
Vocabulary (501–3,000 most frequent)	-0.12
Vocabulary (not among 3,000 most frequent)	0.02
Certainty adverbials	0.04
Likelihood adverbials	-0.03
Possibility modals	-0.10
Linking adverbials	0.08
Finite adverbials	-0.02
Because	0.09
If	-0.12
WH-relative clauses	-0.13
THAT-relative clauses	0.04
TO verb complement clauses	0.02
THAT verb complement clauses	-0.07
Attributive adjectives	-0.07
Pre-modifying nouns	-0.07
Nouns	-0.07
Nominalizations	0.00
Noun + OF	-0.08

As can be seen, there were no significant correlations between any of the linguistic features used by test takers and the rater severity measure. In addition, all of the correlation magnitudes were less than |0.20|, with most being under |0.10|. These findings suggest no relationships between the linguistic features of test-taker performances and rater severity.

Discussion

The results of the Rasch analysis of rater severity and scale usage showed two trends. The first trend was that there was a wide range of variability regarding the severity measures of the raters. The raters varied in their severity by 6.69 logits, which amounted to two full score points on the transformed MELAB speaking scale and a difference of a full point on the traditional MELAB speaking scale. The resulting large range of rater severity measures mirrors the results of other Rasch analyses of other interactive speaking tasks (Bonk & Ockey, 2003; Eckes, 2005). However, there were no significant correlations between rater severity and the linguistic features used by test takers. This suggests that rater severity is not related to any particular linguistic features. It also provides some evidence that the MELAB speaking task is similar across test takers, at least from the perspective of test-taker speech.

Additionally, the presence of raters who overfit and underfit the Rasch model is indicative of raters who are not using the scale consistently (Weigle, 1998). The Rasch analysis also showed that the full MELAB rating scale is not used. In our semi-random sample of 98 tasks the bottom three bands (1, 1+, and 2-) were never used. This could be due to the sample of test takers being composed of people who scored higher on the full MELAB. It is unlikely that test takers with higher levels of ability in other skills would receive low speaking scores. Furthermore, receiving a MELAB score of 2 from this group of raters was as likely as receiving a 2- or a 2+. The upper ends of the scale are functioning as should be expected. In the upper ends there appears to be logical increases in test-taker ability as the scale increases at higher levels.

Given the use of the MELAB speaking test in high stakes situations (e.g., licensure and university admission), it would be worthwhile to conduct regular training sessions. The goal of these sessions could be to decrease the range of rater severity measures as well as to increase intra-rater reliability. The latter seems like a more achievable goal (Stahl & Lunz, 1996; Weigle, 1998). Furthermore, it may be possible to establish MELAB speaking raters' measure of severity and then make final score adjustments based on their estimates. Regular training may also improve the use of the scales at the lower end of the rating scale. Additionally, it is worth considering whether the lower levels of the scale need revision or elimination. When the analysis of the scales are considered in the context of the elicited language, the scale could be revised to highlight those features that are typical in lower proficiency performances (e.g., higher occurrence of personal pronouns and speaker focused language) or not present in lower proficiency performances (e.g., likelihood and certainty adverbials, rates of speech).

Related to this, a number of the linguistic features used by test takers showed significant correlations with test-taker score. In particular, the fluency features of rate of speech (r = 0.69) and hesitation markers (r = -0.52) showed the highest levels of correlations with test-taker score. Fluency has been identified as one of the most important indicators of test-taker proficiency (Ginther, Dimova, & Yang, 2010; Götz, 2013; Kang, Rubin, & Pickering, 2010; Kormos & Denes, 2004; Riggenbach, 1991).

Speech rate has been a consistent measure of fluency in numerous studies (Ginther et al., 2010; Götz, 2013; Kang et al., 2010; Kormos & Denes, 2004; Riggenbach, 1991). In this study, the mean speech rate (measured as syllables per second) was 3.28 sps. This rate is above the mean range reported by Levelt (1987) for L1 speakers (2–3 sps), but below the means reported by Kowal, Wiese, and O'Connell (1980) (3.5 sps) and Ginther, Dimova, and Yang (2010) (3.6 sps). The correlation between speech rate and test-taker score is consistent with previous results for other spoken tasks, both OPIs (Young, 1995), and other types of standardized tests (Iwashita et al., 2008; Kang, 2013).

The number of hesitation markers was also related to test-taker score (r = -0.52), with fewer hesitation markers used by test takers receiving higher scores. Examples 1 and 2 below show, respectively, the use of hesitation markers by lower and higher scoring test takers. Hesitation markers are bolded in the examples.

Example 1

T: Yes, I love buildings, I love even the people, they of course I, I wasn't for a long time here, but in this period **uh** I've found them very **um**, I mean, friendly, and **uh**, very very I think <unclear> I think, **uh**, what can I say? **Um**. <MELAB score: 2>

Example 2

T: **Uh**, more so for maybe the benefits that they might receive later on. Such as financial, because it's not that rewarding, especially in the, today's industry. It takes quite a few years before you can make a subta- substantial amount of money and be stable financially. <MELAB score: 4>

The result for hesitation markers is consistent with Iwashita et al. (2008) who found that L2 English speakers with higher proficiency levels used fewer hesitation markers on the TOEFL exam. However, it should be noted that there has been an inconsistent relationship between hesitation markers and proficiency level (e.g., Kang et al., 2010; Kormos & Denes, 2004).

In the construct of interaction, number of turns was moderately positively correlated with test-taker score at r = 0.33. This result is consistent with Csomay (2005) who found that a greater number of turns was an indicator of higher interactivity in discourse. Turn taking has been an important unit of study in conversation analyses of OPIs (Brown, 2003, 2005; Johnson & Taylor, 1998; van Lier, 1989; Young, 1995). This study adds to that literature by showing that turns are an important feature in distinguishing score levels in the MELAB speaking task.

The other features we examined were under the construct of language. However, as we will discuss, we believe a few of these features could also be included under the constructs of speech or interaction. First person pronouns were low to moderately negatively correlated with MELAB score at r = -0.28. Examples 3 and 4 show the use of first person pronouns by lower and higher scoring test takers. First person pronouns are bolded in test-taker speech.

Example 3

R: Okay, so in the U.S., you plan to continue or maybe to change? Your major?

T: Uh, at first **I**, yes, **I** decide to change the ma- **my** major in <unclear> technology

R: Uh huh.

T: Because **I** love it. Of course, **I**'m much very <unclear> you know. What can **I** say? Um.

R: What do you mean?

T: **I** love this course, but **I**'m not very familiar with it, you know? **I**'m familiar, but in primary limits, not <unclear>. **I** want to continue **my** education in that course.

R: Okay, so you have definitely decided to go to the U.S., not some other European country for your education?

T: No, **I** decided to go to United States, yes. <MELAB Score: 2>

Example 4

R: Do you think you'll continue to work with geriatric patients or you'll change your focus?

T: **I** would love to work with geriatric patients but not the full-time **I** would prefer to work in a clinic

R: Hm.

T: because with the geriatric patients you go more for the maintenance and keeping them in comfort

R: Uh huh.

T: But in clinics you get the like definite results more improvement

R: Right. Yes.

T: So it motivates you to work more in my opinion.

R: Right? Okay yes absolutely yeah yeah yeah so so you actually see full recovery at clinics

T: yes

R: where you won't with geriatrics

T: yeah.

R: yeah yeah

T: You can see the immediate results

R: yes.

T: as compared to geriatric patients because in long-term care centers it's more of maintaining where they are right now <MELAB Score: 4>

As can be seen, the lower scoring test taker (Example 3) remained oriented towards her own wants, needs, and opinions whereas the higher scoring test taker (Example 4) shifted to an orientation with the field of study she was interested in pursuing. This suggests an ability to discuss aspects of her future career, which is beyond the "here and now" and is a possible sign of greater complexity (e.g., Robinson, 2001). This finding parallels Kang (2013) who found that first person pronouns were used more by lower scoring test takers on the Cambridge English Exam. Less frequent use of first person also suggests less speaker-centered orientation, which indicates more interactional involvement. It also matches the CEFR descriptors of speaking proficiency, which describe more frequent speakers as able to move beyond familiar topics to more academic and professional topics (Council of Europe, 2001).

We can also see that first person pronouns are used along with TO complement clauses (also negatively correlated with score and discussed in more detail below) in a number of cases in the above example (*decide/d to*, *want to*). Although the higher scorer also uses these structures in response to the rater's question (*I would love to*, *I would prefer to*), there are many other types of structures represented in this test taker's speech.

Contractions showed a low to moderate positive correlation with test-taker score on the MELAB (r = 0.27). Examples 5 and 6 show the patterns of full and contracted forms (in bold) used by low and high scoring test takers, respectively.

Example 5

R: Okay, Ali, what do you do for a living?

T: For a living, **I am** doing now, uh **I am** working at the rental office car in Amman. <MELAB score 2+>

Example 6

R: Okay just tell just tell me about yourself.

T: **I'm** basically a physiotherapist uh from back India. Um **I'm** working as a physiotherapy aide uh in Hastings Manor, Bellville Ontorio. Uh I live in Belville. <MELAB score 4->

Contractions have been identified as a distinctive feature of conversation (Biber, 1988; Biber et al., 1999). Use of contractions can be seen as an effort saving device (Biber et al., 1999, p. 1048) or a feature of economy (Finegan & Biber, 1994; White, 1994). As such, it may also be related to measures of fluency, such as syllables per second. However, further research is needed to explore this interpretation.

Higher frequency vocabulary was positively correlated with speaking score (r = 0.31), while lower frequency vocabulary was negatively correlated (r = -0.30). This indicates that successful test takers are more likely to use higher frequency words, which are more common in conversation when compared with writing (Biber et al., 1999). Similarly, the data showed that successful test takers are also less likely to use lower frequency vocabulary, or words that would be less commonly used in conversation (Biber et al., 1999). It

is also important to note that the 1–500 most frequent words includes both function and content words, which may indicate a greater use of function words by higher scoring test takers.

Example 7 below comes from a test taker who received a near-perfect speaking score of 4-. Eighty-three percent of the words used by this test taker in the interaction were among the 500 most frequent words in English. These words are bolded in the excerpt below. The nonbolded, small caps words are those not on that list. As can be seen, the overwhelming majority of the words are high frequency. These are words that the raters are much more likely to hear in natural conversations, decreasing the time necessary to process and comprehend them and possibly making the test taker's speech sound more natural and native-like. These vocabulary results parallel Kang's (2013) finding that score was positively correlated with the use of high frequency words on the Cambridge English Exam.

Example 7

T: **Every day** BASIS. **So it was really hard, but then I started working which you know let me** COMMUNICATE **with the other people over here. And then, I was able to understand the people. So I can understand, at least I can say** EIGHTY **percent.** EIGHTY **five percent. Some** SLOGAN **I don't know about that. That's why I can't understand. But I can say and say, Hey, I'm so** SORRY, **can you** REPEAT **that in other words so I can understand. Because it happens with the** PATIENTS **too.** <MELAB Score: 4->

In contrast, Example 8 comes from a test taker who received a speaking score of 3. Only 63% of the words used by this test taker were among the 500 most frequent words in English. These words are bolded in the excerpt below.

Example 8

T: **After that I have to** REGISTRATED **at the** UM JORDANIAN **law** BAR.

R: Uh huh.

T: UH **like** UH **as under** TRAINING LAWYER **for two years.**

R: Uh huh.

T: UH **after passing the** EXAMS UH

R: This is a training period?

T: **Yeah, we** UH **have two years** TRAINING **and we have** UM **to pass the** EXAMS **at the end.**

R: Uh huh.

T: **Yeah, our our** UH JORDANIAN **law** BAR **put the** EXAMINATION **for the** UNDER-TRAINING LAWYERS <MELAB Score: 3>

Example 8 demonstrates the type of lower frequency vocabulary that was characteristic of lower performance. Examples of these words are *Jordanian*, *lawyer*, and *examination*. While the lower frequency words are likely to be known and understood by the raters, they are not as commonly used, especially in conversational discourse. Therefore, they slow down processing time and may at times sound unnatural or stilted.

The results of this study also revealed positive relationships between two categories of stance adverbials (certainty and likelihood) and speaking score. There was a negative correlation between the use of TO verb complement clauses and speaking score. Expressions of stance are a characteristic of interactive discourse in which speakers use lexico-grammatical markers to convey their personal feelings and attitudes. Speakers can express their stance more overtly/explicitly by using features such as first person pronouns and stance verbs (e.g., *I want to*), that overtly mark the agent (the speaker). Alternatively, speakers can express stance more implicitly, by using adverbials (*certainly*, *actually*), in which the speaker does not need to be overtly identified as the agent of the stance (Biber et al., 1999, p. 864–865). The results from this study suggest that higher scoring test takers are more likely to express personal stance

more implicitly through the use of adverbial stance markers, while lower scoring test takers are more likely to express their stance more overtly through using TO complement clauses. As discussed above, in this study, TO complement clauses were often used with first person pronouns, another feature associated with the performances of lower scorers. Example 9 below is from a test taker who received a speaking score of 3-. This test taker used no stance adverbials, but relied heavily on TO complement clauses. Note the use of first person pronouns in conjunction with these clauses.

Example 9

T: Yeah, this is the first time. I uh I want pass exam MELAB. I **want to** go Canada and study <unclear> study there. Not only study study both study and work there. I **want to** study hotel management.

R: Uh huh.

T: I know in Armenia there is no universities where I can study hotel management and I **decided to** go there and study and have good work work experience.

R: Okay so you are interested in hotel management.

T: Yes, I **want to** become hotel manager.

R: Okay you want to become a hotel manager, very interesting. So do you have a bachelor's degree in Armenia, or you are applying for the bachelor's degree.

T: No, I um I study in I studied in French university in Armenia.

R: Uh huh.

T: But then I decide, change my, I **want to** change my job. I **want to** become hotel management. <MELAB Score: 3->

This more explicit stance marking seems to be more characteristic of the performance of lower scoring test takers. This finding interestingly parallels those of Biber et al. (2014), who found that lower scoring test takers on the TOEFL speaking tasks used more TO complement clauses with desire verbs (e.g., *want*).

Example 10 comes from a test taker who received a perfect score of 4. It contains several examples of the stance adverbials *really* and *actually*. These are being used as more implicit markers of the test taker's personal stance.

Example 10

T: Studying. Um, **actually**, this is my first MELAB and um, um, I **actually** got the notice from the university that they want to uh, they need an English proficiency from my side because it's **actually** my second <unclear> Europe high school. It's **really** late. I **actually**, I knew like late April, so I was just looking for something to get it **really** fast and to study it like easier than the TOEFL or the IELTS which are the most popular tests, right? So, yeah, part of my life is, part of my life in the last two weeks was this but **actually** the main part was my studying because grade twelve needs marks, marks, marks, marks. <MELAB Score: 4>

Conclusion

The results of this study show three important things about the MELAB speaking task. First, the raters vary widely in their severity, and the lower end of the MELAB rating scale is not being used to its full capacity. While this result should be tempered with the idea that the Rasch analysis was based on small samples, the next logical step in investigating the rater severity of MELAB raters would be to conduct a series of norming sessions using benchmarked samples of the MELAB. This would help to pin down the range of rater severity that exists in the ratings and would provide further insight into which raters may need further training to increase their inter- and intra-rater reliability. Furthermore, it would provide a more complete picture of how the rating scale is used. The sample in this study did not contain many performances that were awarded scores at the lower end

of the scale. A larger sample may contain more of these types of performances and would provide a more robust estimate of rater use of the speaking scale.

Second, linguistic features are systematically related to score level, which provides some evidence that test-taker performances (i.e., the MELAB task) are assessed in similar ways across raters. In addition, linguistic features typical of conversation increase as score increases. This provides some evidence for the generalizability of the task to conversation. However, this should be interpreted cautiously in light of the fact that the MELAB speaking task has yet to be subjected to explicit comparisons with conversation. In addition, previous corpus research has shown that test-taker scores are more likely to correlate with a constellation of linguistic features rather than individual features (Biber & Gray, 2013). A multi-dimensional analysis of the linguistic features used in the MELAB would provide further insight into the linguistic characteristics of this spoken task.

Taken together, despite the variability in rater scoring found in this study, there is evidence for similarity across tasks. This is somewhat supported by the correlational analysis of the linguistic features with proficiency. That certain linguistic features increase (or decrease) with an increase in proficiency indicates that the raters are using the scale effectively to rank order the test takers. The lack of a relationship between linguistic features and measures of severity indicates that raters are not systematically eliciting different language from test takers regardless of their estimated levels of leniency or severity. However, it is important to note that this study considered only half of the linguistic picture—that is, the test-taker discourse. In future research it will be important to consider the linguistic features in the rater discourse.

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: CUP.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Philiadelphia, PA: John Benjamins Publishing.

Biber, D. & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT ® Test: A lexico-grammatical analysis*. TOEFL iBT Research Report 19. Princeton, NJ: Educational Testing Service.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use the characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, doi:10.1093/applin/amu059

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London: Pearson.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1–25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. New York: Peter Lang.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Csomay, E. (2005). Linguistic variation within university classroom talk: A corpus-based perspective. *Linguistics and Education*, *15*(3), 243–274.

Davies, M. (2011–). *Word and phrase.info*. Available online at http://wordandphrase.info.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing*, *11*(2), 125–144.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performanceassessments: a many-facet Rasch analysis. *Language Assessment Quaterly*, *2*(3), 197–221.

Finegan, E. & Biber, D. (1994). Register and social dialect variation: An integrated approach. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 315–347). Oxford: OUP.

Friginal, E. (2009). *The language of outsourced call centers: a corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399.

Gray, B. (2011). *Exploring academic writing through corpus linguistics: When discipline tells only part of the story* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.

Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *System*, *20*(3), 365–372.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

Jamieson, J., & Poonpon, K. (2013). *Developing analytic rating guides for TOEFL iBT integrated speaking tasks*. TOEFL iBT Research Report 20. Princeton, NJ: Educational Testing Service.

Johnson, M., & Taylor, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In A. W. He & R. Young (Eds.). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Philadelphia, PA: John Benjamins.

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Cambridge English: Research Notes*, *52*, 40–48.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English. *Modern Language Journal*, *94*(4), 554–566.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.

Kowal, S., Wiese, R., & O'Connell, D. C. (1983). The use of time in storytelling. *Language and Speech*, *26*(4), 377–392.

Lazaraton, A. (1992). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, *13*(2), 151–172.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Linacre, J. M. (2012). *Facets computer program for many-facet Rasch measurement*, version 3.70.0. Beaverton, Oregon: Winsteps.com

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990) Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345.

McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Riggenbach H (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423–441.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57.

Stahl, J. A., & Lunz, M. E. (1996). Judge performance reports: Media and message. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice (Vol. 3)*. Norwood, NJ: Greenwood Publishing Group.

Staples, S. (2015). Examining the linguistic needs of internationally educated nurses: a corpus-based study of lexico-grammatical features in nurse-patient interactions. *English for Specific Purposes Journal*, *37*, 122–136.

Staples, S. & Biber, D. (2014). The expression of stance in nurse-patient interactions: An ESP perspective. In M. Gotti & D.S. Giannoni (Eds.), *Corpus analysis for descriptive and pedagogical purposes: ESP perspectives* (pp. 123–142). Bern: Peter Lang.

Thompson, I. (1995). A study of interrater reliability of the ACTFL oral proficiency interview in five European languages: Data from ESL, French, German, Russian, and Spanish. *Foreign Language Annals*, *28*(3), 407–422.

Weigle, S. (1998). Using Facets to model rater training effects. *Language Testing*, *15*(2), 263–287.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508.

White, M. (1994). *Language in job interviews: Differences relating to success and socioeconomic variables* (Unpublished dissertation). Northern Arizona University, Flagstaff, AZ.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning*, *45*(1), 3–42.

Appendix A: Rater Measurement Report

Table 1A: Facets' Rater Measurement Report

Rater	Count	Measure	S.E.	InfitMS	OutfitMS
13	11	-5.93	0.61	2.01	1.96
5	10	-5.79	0.72	1.54	1.39
1	10	-3.90	0.55	0.76	0.93
22	8	-3.16	0.64	0.74	0.87
19	11	-2.84	0.54	0.86	0.85
21	7	-2.59	0.76	0.44	0.52
7	8	-2.42	0.54	0.70	0.69
12	11	-2.09	0.61	0.88	0.97
16	7	-1.83	0.67	0.66	0.68
4	7	-1.61	0.65	1.34	1.35
18	10	-1.58	0.59	0.75	0.77
11	11	-1.48	0.55	0.51	0.49
2	10	-1.33	0.61	0.91	1.07
9	11	-1.25	0.53	0.57	0.64
15	10	-1.17	0.54	1.46	1.47
6	8	-1.01	0.51	1.11	0.91
14	7	-0.39	0.57	1.18	1.12
10	5	-0.18	0.81	0.22	0.22
3	9	0.39	0.51	0.98	0.94
17	9	0.53	0.52	0.94	1.07
20	9	0.67	0.53	0.68	0.60
8	11	0.76	0.54	0.89	0.92