Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment

Geoffrey T. LaFlair

University of Hawai‘i at Mānoa , USA

Shelley Staples

University of Arizona, USA

Abstract

Investigations of the validity of a number of high-stakes language assessments are conducted using an argument-based approach, which requires evidence for inferences that are critical to score interpretation (Chapelle, Enright, & Jamieson, 2008b; Kane, 2013). The current study investigates the extrapolation inference for a high-stakes test of spoken English, the Michigan English Language Assessment Battery (MELAB) speaking task. This inference requires evidence that supports the inferential step from observations of what test takers can do on an assessment to what they can do in the target domain (Chapelle et al., 2008b; Kane, 2013). Typically, the extrapolation inference has been supported by evidence from a criterion measure of language ability. This study proposes an additional empirical method, namely corpus-based register analysis (Biber & Conrad, 2009), which provides a quantitative framework for examining the linguistic relationship between performance assessments and the domains to which their scores are extrapolated. This approach extends Bachman and Palmer's (2010) focus on the target language use (TLU) domain analysis in their study of assessment use arguments by providing a quantitative approach for the study of language. We first explain the connections between corpus-based register analysis and TLU analysis. Second, an investigation of the MELAB speaking task compares the language of test-taker responses to the language of academic, professional, and conversational spoken registers, or TLU domains. Additionally, the language features at different performance levels within the MELAB speaking task are investigated to determine the relationship between test takers' scores and their language use in the task. Following previous studies using corpus-based register analysis, we conduct a multi-dimensional (MD) analysis for our investigation. The comparison of the language features from the MELAB with the language of TLU domains revealed that support for the extrapolation inference varies across dimensions of language use.

Keywords: Corpus linguistics, domain analysis, multi-dimensional analysis, performance assessment, register analysis, validity argument

Introduction

In this article, we demonstrate the use of corpus-based register analysis for evaluating evidence for the validity of the interpretations of test scores, particularly the extrapolation from test scores to real-world situations. We begin with an overview of an argument-based approach to validity. This is followed by a comparison of two analytic frameworks that are crucial to analyzing target domain language and language assessment tasks: the situational analysis component of corpus-based register analysis (Biber & Conrad, 2009) and TLU domain analysis (Bachman & Palmer, 1996, 2010). Within an argument-based approach to validity, TLU domain analysis initially takes place during the creation of the test, but it can be used as well to investigate the extrapolation inference, which requires a post-hoc evaluation of the relationship between the test tasks and the target domain, after the test has been developed. Finally, we report the results of a study which applies corpus-based register analysis to the investigation of the extrapolation inference in a validity argument for the Michigan English Language Assessment Battery (MELAB) speaking task. To do so, we compare test takers' performance on the MELAB oral proficiency interview to the language of spoken registers that represent the TLU domains: office hour interactions, service encounters, study groups, conversation, and nurse-patient interactions. In our comparison, we investigate two underlying assumptions: 1) that the linguistic features elicited by the MELAB is similar to the language in TLU domains, and 2) that the frequency of use of these linguistic features elicited by the MELAB approximates their frequency of use in the TLU domain as scores on the MELAB increase.

Concepts of validity arguments

One approach that has evolved out of validity research is the argument-based approach (Kane, 2013). Under this approach, the focus of a number of current validity studies is twofold: (1) the development of an interpretation and use argument (IUA), which lays out the claims about test score interpretation and use (Kane, 2013); and (2) the development of a validity argument, which is an evaluation of the IUA (Chapelle et al., 2008b; Kane, 1992, 2013). Analyses that were traditionally conducted to investigate construct, content, and criterion validity still exist in the argument-based approach. However, instead of being conceptualized as different types of validity, these traditional analyses are used to support various inferences that form an IUA. An IUA may vary from one test to another depending on the test's proposed interpretations and uses. However, tests with high-stakes decisions and more ambitious claims require more evidence to support the chain of inferences in their IUAs (Kane, 2013).

To use an argument-based approach to validity research, researchers need to identify the inferences that are critical to score interpretation and use (Chapelle et al., 2008b; Kane, 2013), because these form the inferential steps from the observed performance on the test to expected performance in the target domain. For example, Kane (1992, 2013) identified a minimum of three possible inferences–scoring, generalization, and extrapolation–which are made when interpreting and using test scores. Chapelle et al. (2008b) expanded on Kane's three inferences and identified six inferences that were made in one high-stakes language test: a domain definition inference, an evaluation (i.e., scoring) inference, a generalization inference, an explanation inference, an extrapolation inference, and a utilization inference. A common metaphor for these inferences is that they are bridges that link the various components in the interpretation and use of an assessment. For example, the extrapolation inference links the language of test performances to the

expected language performance in the target domain; like bridges, these inferences need support.

The logical structure typically used in a validity argument is Toulmin's (1958, 2003) argument structure (Chapelle et al., 2008b; Kane, 2013; Mislevy, Steinberg, & Almond, 2003). When Toulmin's framework is applied to language testing, inferences provide a means of making a *claim*, or conclusion, about a test taker's language abilities on the basis of *grounds* for the claim (e.g., data or observations). The inference depends on a *warrant*, which is an established procedure, a general rule, or general principle for making claims based on the grounds. The warrant requires *backing* in the form of scientific theories, bodies of knowledge, or precedents. Inferences are subject to *rebuttals* which weaken the strength of the link between the claim and its grounds (Chapelle et al., 2008b; Kane, 2013).

One assumption underlying the extrapolation inference in language testing is that specific contextual features affect both language test performance and language use in the target domain of interest to test users (Bachman, 1990; Biber & Conrad, 2009; Canale & Swain, 1980; Chapelle, Enright, & Jamieson, 2008a; Hymes, 1974). Accounting for the effect of context on language use is important in the TOEFL validity argument. Chapelle et al. (2008b) maintained that task-based perspectives to test development should be included as dual grounds alongside competency-based perspectives. The former interprets test scores in light of contextual features of language use situations. The latter interprets test scores in regards to constructs of language ability.

The analysis that we are proposing fits into the task-based perspective: the language elicited by test tasks and the language used in target domains can be characterized by features of their contexts. It is an analysis of what Kane (2013) calls *observable attributes*–or tendencies to perform or behave in some way. These observable

attributes are defined by their target domains. For example, if speaking in academic settings is considered an observable attribute of test takers' language ability, then it is defined by the types of linguistic (e.g., relative clauses, modals) and extra-linguistic characteristics (e.g., features of participants, setting, and communicative purposes) of office hours, study groups, and service encounters in academic settings, which have been shown to influence the types of linguistic features that are used by speakers (Biber, 2006). Thus a task that can simulate similar situational characteristics of the target domain should elicit language that is similar to the language of the target domain, and research showing that it does so can serve as support for the extrapolation inference in the validity argument for the test.

Target language use domain analysis and corpus-based register analysis

In order to provide such linguistically based support for the extrapolation inference, a corpus-based methodology can be used. We introduce the use of a corpus-based methodology by showing the relationship between TLU domain analyses from language testing (Bachman, 1990) and corpus-based register analysis (Biber & Conrad, 2009). Both of these analyses are based on theories of communicative language competence (Canale & Swain, 1980; Hymes, 1972). First, Bachman (1990) laid out two frameworks: a framework for describing language abilities and a framework for describing the characteristics of test tasks and the TLU domain. Both of these frameworks adopted the perspective that communicative competence in a language includes knowledge of how context can govern the use of language. Bachman (1990) argued that the context of the TLU domain is important to consider in language test development:

> One way to conceive of a language test is as a means for controlling the context in which language performance takes place. From this perspective, the

> characteristics of the test method [including the task] can be seen as analogous to the features that characterize the context of situation, or speech event [of the TLU domain]. (p. 111)

In other words, the tasks on a language test can be viewed as an approximation, or a simulation, of the tasks in the target domain. The extent to which the characteristics of TLU and test tasks overlap could affect the extent to which linguistic features overlap. Bachman and Palmer's (1996; 2010) TLU analysis framework offers a method for identifying the characteristics of target domains that may affect language use so that test tasks can be evaluated and compared to the target domain. This method includes examining the features of the setting, the scoring rubric, the language input of the task, the expected response, and the relationship between the input and the expected response. In the development of the TOEFL validity argument, understanding the contextual features of the TLU domain and simulating them in assessment tasks was integral to investigating the evidence for the domain description inference of the IUA (Chapelle et al., 2008a). While Bachman and Palmer's framework provides a thorough method for developing test tasks so that the language they elicit is relevant to the target domain, it does not provide a robust, quantitative approach to examine the language of the responses beyond the use of analytic rubrics.

Corpus-based register analysis shares several similarities with TLU analysis in its approach to characterizing language use situations along with a quantitative framework for examining the linguistic characteristics of the language use situation. Register, as defined in Biber and Conrad's (2009) framework, is a language variety characterized by its situation of use. A register analysis contains three components: a situational analysis that identifies characteristics such as the speaker's role and setting; a linguistic analysis; and a functional interpretation of the linguistic features in the situational context. More

specifically, situational features can include the speaker's role in a communicative event, the setting of the event, the purpose for communicating, and the personal relationship between participants. All of these situational characteristics impact the linguistic forms used by speakers due to the functional needs of the communicative event. Biber and Conrad's (2009) framework for situational analysis is based on earlier work by Biber (1994) that draws from Hymes' (1974) SPEAKING[1] framework.

A major advantage of corpus-based register analysis is that it generally utilizes multi-dimensional (MD) analysis, a quantitative method of linguistic analysis that allows for a consideration of co-occurring language features that contribute to functional language use and that can be interpreted as related to the situational characteristics of tasks. Thus, corpus-based register analysis integrates many of the characteristics of Bachman and Palmer's TLU analysis into a statistical procedure (factor analysis) that allows for quantifiable comparisons of linguistic and functional language use across test tasks and TLU domains. The first column in Table 1 shows the set of characteristics that are considered in a TLU analysis when developing test tasks (Bachman & Palmer, 2010). The second column shows characteristics that are included in corpus-based register analysis (Biber & Conrad, 2009). As can be seen from the table, both approaches are concerned with similar situational characteristics; however, they are organized differently. For example, in a situational analysis *topic* is a characteristic of the register while in a TLU analysis *topic* is part of the characteristics of the input and the response.

Table 1. Characteristics included in TLU analyses (Bachman & Palmer, 2010) and corpus-based register analysis (Biber & Conrad, 2009).

TLU Characteristics	Potential Register Characteristics
• Characteristics of the setting (e.g., participants) • Characteristics of the rubric (e.g., time constraints) • Characteristics of the input (e.g., format, language, topic) • Characteristics of the response (e.g., format, language, topic) • Relationship between input and response (e.g. reactivity, scope)	• Participants (e.g., number of participants) • Relations among participants (e.g., interactiveness, social roles, power and asymmetry) • Channel (e.g., mode, medium) • Production circumstances (e.g., real time, planned, scripted) • Setting (e.g. private, public, sharing same time and space) • Communicative purposes (e.g., general, specific, expressions of stance) • Topic (e.g., general, specific, academic)

Although the features in Table 1 are not exhaustive, the similarities between the two sets of characteristics illustrate the potential for the use of corpus-based register analysis as a tool for evaluating inferences that are made when interpreting and using a test. Additionally, if a productive task is supported with evidence of a thorough TLU domain analysis, then it is plausible that the language produced by the test takers will be similar to the language of TLU domains, especially at higher score levels. Corpus-based register analysis can be used to evaluate this proposition. In other words, analyses can be conducted in the development stages to ensure adequate representation of the domain and consistent design of test tasks (i.e., analyses used for support of a domain definition inference). This can be followed by empirical analyses in the appraisal stages of validation to investigate if the test "controls the context" to the extent that test takers' production is similar to real-world production (i.e., analyses used for support of an extrapolation inference). The investigation conducted in the present study examines evidence for the

extrapolation inference because it occurs after the design stages of the MELAB OPI. The goal of this study is to appraise, or evaluate, the extent to which test-taker language in the MELAB OPI is similar to language used in the academic, professional, and conversational domains.

Using corpus-based register analysis to investigate productive assessments

Investigating the linguistic features of productive assessments is certainly not new. Previous studies have utilized corpus-based methods to conduct research on productive assessments by examining the relationship between specific linguistic features of test-taker responses and rubric score bands (Biber, Gray, & Staples, 2014; Jamieson & Poonpon, 2013; Kang, 2013; LaFlair, Staples, & Egbert, 2015; Yan & Staples, 2017), rater perceptions of test-taker performance across rubric score bands (Brown, Iwashita, & McNamara, 2005), production in real-life situations (Brooks & Swain, 2014; Weigle & Friginal, 2015), or features of the task (Kyle, Crossley, & McNamara, 2016). Table 2 highlights six studies on spoken language elicited by test tasks. The columns from left to right indicate the study, the number of linguistic features included at the outset of the analysis in each study, the final number of linguistic features that were retained after the statistical analyses in the study, a summary of the research design of the study, and examples of the retained features. The retained features represent the significant subset of the larger number that were included in regression analyses (LaFlair et al., 2015; Jamieson & Poonpon, 2013), ANOVA/Friedman analyses (Brooks & Swain, 2016; Brown et al., 2005; Kang, 2013), and discriminant function (DF) analyses (Kyle et al., 2016). The comparison of the initial number of linguistic features with the subset of significant features shows a large disparity between the two numbers. For example, Kyle et al. (2016) started with 202 linguistic features, with the goal of using DF analysis to classify spoken performance correctly into task types (i.e., independent and integrated) based on the

linguistic features in the performances. They conducted two studies using this method, and in total nine variables were used by the DF analysis to classify the performances into task types. The consideration of linguistic features individually does reduce a large number of linguistic features down to a smaller set of linguistic features. However, it ignores the co-occurrence patterns among the individual features that vary across task types as well as the functional aspects of these co-occurring features. Furthermore, a large number of features are lost in the analyses and the features that are kept after the statistical analysis may be difficult to interpret with respect to their communicative functions.

Table 2. Numbers of Individual Linguistic Features Included in Statistical Tests in Corpus-based Studies of Oral Assessment Data

Study	Initial number of features	Final number of features	Summary of research design	Examples of features related to score/proficiency level, context, or task type
Brooks & Swain (2014)	24	14	Investigated differences in test takers' use of linguistic features across three contexts (test, classroom, out-of-classroom); linguistic features were dependent variables in Friedman tests	Less grammatical complexity, more grammatical inaccuracies, more speech organizers in test contexts than non-test contexts; more connectives, more passive verbs, more nominalizations, more words from the first 1000 band, more words from the second 1000 band, more off-list words, more total content words in test and in-class contexts than out-of-class contexts
Brown et al. (2005) (RQ 4)	30	18	Compare test takers' mean use of linguistic features across score levels; linguistic features were dependent variables in ANOVAs	Higher speech rate, more word tokens, more word types, target-like pronunciation of syllables, number of clauses, more t-units, better global accuracy , lower type–token ratio, fewer unfilled pauses
Jamieson & Poonpon (2013)	19	12	Examine the relationship between linguistic features and score level; linguistic features were predictor variables in a multiple regression	Longer mean length of run, more syllables per second, increase in overall pitch range, fewer silent pauses, more error-free C-units, higher word count, more prepositional phrases, more passives, more adjectives, more key ideas, more conjunctions, extent of introduction framing as scores increase
Kang (2013)	65	36	Compare test takers' mean use of linguistic features across	Higher speech rate, shorter/fewer pauses, increase in phonation time ratio, more error-

Study	Initial number of features	Final number of features	Summary of research design	Examples of features related to score/proficiency level, context, or task type
			proficiency levels; linguistic features were dependent variables in ANOVAs	free t-units, more clauses, more complex t-units, better grammatical accuracy, more word types, more tokens, more words from the first 1000 band, more academic words, modals, nominalizations, articles, prepositions in higher proficiency levels
Kyle et al. (2016)	202	9	Classify test taker responses into their task types; linguistic features were predictor variables in a discriminant function analysis	Type–token ratio, personal pronouns, motion prepositions, range of content words, mental verbs, spoken bi-gram frequency, givenness, meaningfulness, insight words were effective in predicting task type
LaFlair et al. (2015)	28	5	Compare test takers' mean use of linguistic features across score levels; linguistic features were predictor variables in a multiple regression	More syllables per second, fewer hesitation markers, more likelihood adverbs, fewer first-person pronouns, more certainty adverbs as scores increased

The study by Brooks and Swain (2014) is of particular interest because they interpreted their results as having a bearing on the extrapolation inference of the IUA in the TOEFL validity argument. They found that the language produced in the speaking task was more prone to error, more grammatically and lexically complex, and more formal than language used in out-of-class and in-class situations. They attributed this result in part to differences in situational characteristics between the test task and the target domain and concluded that this exposes a "weak link" in the IUA (Interpretation/Use Argument) for the TOEFL iBT.

These studies reflect strengths and weaknesses in using individual linguistic features as the basis for analysis of test performances. One strength is that the wide range of linguistic features included in these studies is a part of the multi-faceted construct of spoken English. A weakness is that lexical and grammatical units of analysis are analyzed as if their occurrences are independent. However, all linguistic features are correlated to some extent. When language is separated into such fine-grained features, it can be difficult to discern and interpret patterns of variation both within and across studies (Biber et al., 2014). Furthermore, it is difficult to understand the role that these individual linguistic features play in communicative functions of language.

Corpus-based register analysis that includes multi-dimensional (MD) analysis can account for the co-occurrence of linguistic features and provide insight into the use of linguistic features for communicative purposes. Biber et al. (2014) importantly show that dimensions of language use in TOEFL iBT spoken (and written) tasks are better predictors of score level than individual linguistic features. MD analysis has also been used to show that performances from TOEFL iBT independent writing tasks are different than (e.g., including more narrative features and more features of personal

opinions) disciplinary writing in university settings, which has important bearing on the current study's focus on the extrapolation inference (Weigle & Friginal, 2015). The advantage of MD analysis is that each dimension typically accounts for a number of linguistic features. This reduces the number of predictors in an analysis (i.e., holistic dimensions instead of individual linguistic features) while retaining a large number of linguistic features. Furthermore, it shifts the focus from finding individually statistically significant features to identifying trends in co-occurring patterns of language use. Additionally, the interpretations of dimensions allow for insights into how test takers use specific linguistic features in combination for various communicative purposes. As a result, this method allows for an evaluation of one type of support for the extrapolation inference of the validity argument by examining the use of linguistic features for communicative purposes across language elicited by a test (in this study, the MELAB) and its target domains. This study answers two research questions:

1. To what extent are linguistic features of dimensions of language use elicited by the MELAB similar to language observed in target domains?
2. To what extent are linguistic features of dimensions of language use elicited by the MELAB similar to language used in the target domain as scores increase?

Method

This study uses a corpus-based register approach, which involves quantitative linguistic analysis (using multi-dimensional analysis) as well as a situational analysis, which qualitatively examines the situational characteristics of the registers in this study (MELAB OPI, conversation, academic and professional interactive registers). Here, we first describe the corpora used in the study, followed by the situational and multi-dimensional analysis.

The MELAB OPI

The MELAB OPI is designed to measure intermediate to advanced speaking ability in academic, professional, and social domains. It is accepted by over 800 institutions in the United States and Canada; most of these are educational institutions but many are organizations involved in the certification of medical professionals such as nursing boards, of which 13 US state boards were listed as accepting organizations (Cambridge Michigan Language Assessments, 2016). The National Council of State Boards of Nursing (NCSB) conducted a standard setting study on the MELAB in 2012 in order to establish a passing English language proficiency standard for entry-level nurses and provide their members with another option for testing English language proficiency (Qian, Woo, & Banerjee, 2014). The MELAB OPI consists of an interview between one test taker and one examiner. Although the interview is live scored, it is also recorded, allowing us to transcribe test data for corpus creation.

Corpora

The MELAB OPI corpus (LaFlair et al., 2015; Staples et al., 2017) was created in 2014 and includes a random sample of 98 OPIs selected from MELAB OPI administrations during 2013. The first five minutes of these 98 MELAB speaking assessment samples were transcribed to build the corpus. After transcription, the MELAB OPI corpus was divided into two speaker groups, to make it possible to analyze the examiner and test-taker discourse separately (see LaFlair et al., 2015 for more information about the corpus and the test). The test-taker half of the MELAB OPI is composed of performances that received ratings between 2 and 4 on the MELAB rubric (note that + and – scores can be given). As is indicated in Table 3, the majority of the performances were awarded 3− or higher.

The MELAB OPI corpus was compared to five registers in three reference corpora, each of which represents a register in the TLU domain. These three reference corpora are the US Nurse/Patient (UNSP) corpus, the T2K-SWAL corpus of spoken language in academic settings, and the American Conversation sub-corpus of the Longman Corpus of Spoken and Written English (Longman corpus). The UNSP is composed of interactions between standardized patients (actors) and nurses (Staples, 2015). Standardized patients are actors trained to interact with health care providers in the same way, and are often used in assessment contexts. The T2K-SWAL is composed of spoken interactions from office hours (professors and students), study groups, and service encounters (customers and servers) in US university settings (Biber, 2006). The Longman corpus comprises natural conversations between US speakers (Biber, Johansson, Leech, Conrad, & Finegan, 1999). Information about the design of the reference corpora can be found in Table 4.

Table 3. Overview of the Test-taker discourse in the MELAB corpus.

Score band	Texts	Mean words/text	Total words
2	3	404.67	1214
2+	5	410.40	2052
3−	16	375.12	6002
3	17	419.41	7130
3+	26	469.31	12,202
4−	12	532.25	6387
4	19	557.95	10,601
Total	98	465.22	45,588

Table 4. Overview of reference corpora.

Corpus	Texts	Mean words/text	Total words
Nurse (UNSP)	50	925.64	46,282
Patient (UNSP)	50	362.70	18,135
Customer (T2K-SWAL)	21	1707.33	35,854
Server (T2K-SWAL)	21	2508.19	52,672
Professor (T2K-SWAL)	11	2934.36	32,278
Student (T2K-SWAL)	11	1508.09	16,589
Study Groups (T2K-SWAL)	23	6262.87	144,046
Conversation (Longman)	709	5656.58	4,010,518
Total	896	4862.02	4,356,374

Situational analysis

We conducted a situational analysis of both test taking and TLU registers using the framework from Biber and Conrad (2009, p. 40). This framework, as discussed above, allows researchers to qualitatively examine differences across such situational characteristics as the topics and communicative purposes as well as number of participants and relationships among them (e.g., degree of power/asymmetry). As such, it aligns with TLU analysis (Bachman, 1990; Bachman & Palmer, 2010). The analysis of the situational context took place both before and after the linguistic analysis, and involved reading previous research on these registers (e.g., Biber, 2006; Staples, 2015), discussion of the situational characteristics of the registers by the researchers, as well as qualitative examination of transcripts. The situational analysis is provided here to foreground our interpretations of the quantitative linguistic analysis found in the results and discussion.

All of the registers contain a number of similar situational characteristics: there are at least two participants who take turns interacting to create the discourse. They share the same physical and temporal setting, and the discourse is produced in real time.

Key differences across the situational contexts include the topics and communicative purposes of the interaction and the social roles and relationships between participants (including degree of asymmetry). Below, we discuss these differences, particularly with respect to differences between the MELAB OPI and the target domains.

The MELAB OPI is characterized by a restricted range of topics, including the test-taker's academic and professional interests and experience. They may also include more personal topics, such as family, friends and adjusting to life in a new country. The overall purpose of the MELAB is to provide test takers with an opportunity to demonstrate their spoken language abilities. The test takers' goals include gaining entrance to a university or professional program.

Study groups are even more restricted in terms of topic and purpose than the MELAB, with personal topics limited to occasional comments and goals focused on conveying and gathering information, as well as recalling content and instructions from classes. Office hours tend to focus on student questions about course content, advising concerns, and future plans. Nurse–patient interaction focuses on assessing the patient's current state of health and addressing the patient's health concerns. Professors and nurses provide information to students and patients, respectively, and aim to gather information from their interlocutors in order to provide advice or to assess the patient's condition. Service encounters have both interpersonal and transactional purposes, especially in the context of an academic campus. Many of the service workers are fellow students, so students use the encounters to chat with friends and acquaintances. Finally, face-to-face conversation has the broadest range of topics and purposes; speakers often discuss recent and distant past events in the form of narratives, and the purpose of interacting is much more social and interpersonal than in the other registers.

In terms of social roles and relationships among participants, the MELAB is different from the target registers in that the participants have no prior knowledge of each other and do not intend to build a relationship, so there is less focus on interpersonal and social purposes. Instead, there is a marked asymmetry between the two participants, with examiners playing a gatekeeping role that may impact the test takers' future academic and career plans. In face-to-face conversation, the roles of the participants may vary, but there is no expected asymmetry between the participants. This lack of asymmetry can to a large extent also characterize study groups. In both registers, the participants know each other to some extent.

Office hours and nurse–patient interactions are both characterized by a great deal of asymmetry. However, in both situations there is also a desire to mitigate this asymmetry. Professors will generally know their students already; in the nurse–patient interactions included in this study, the nurses have an interest in building a relationship with the patients.

These brief descriptions of the situational characteristics of the registers under analysis in this study provide an overview of the different factors that may lead to linguistic variation. In addition, they help point to possible interpretations of those linguistic differences owing to the functions of language in these different situational contexts.

Multi-dimensional analysis

In conducting our MD analysis, we followed the framework provided by Biber and Conrad (2009). After performing our initial situational analysis, we reviewed previous research to select appropriate linguistic features for the linguistic analysis, including those features identified from previous research on spoken assessment (e.g., Biber et al., 2016; Jamieson & Poonpon, 2013; Kang, 2013; LaFlair et al., 2015) as well

as features identified in the spoken registers we compared to the MELAB (e.g., Biber, 2006; Biber et al., 1999; Staples, 2015). The final set of 41 linguistic features can be found in the Appendix. These features were then analyzed using the Biber tagger and Tagcount, two programs that identify and count specific linguistic features (Biber, 2006). Measures were taken to insure tagger accuracy for the MELAB corpus, including running post-tagging scripts to improve the accuracy of the tagger and manually checking all occurrences of *that* in the files, which was identified as a problematic feature based on previous research (Biber & Gray, 2013). All of the other corpora had already undergone extensive tag checking and fixing as part of previous analyses.

We then performed a factor analysis on the normed rates of occurrence of each of the 41 features, using the statistical software program R (R Core Team, 2016; Revelle, 2016; Wickham, 2009). We used principal axis factoring and a Promax rotation. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) was .70, acceptable for continuing with the factor analysis.

The scree plot of eigenvalues revealed a definitive break between the fifth and sixth factors, so a five-factor solution was chosen. Together, these factors accounted for 35% of the variance of the linguistic features in the corpus, which is slightly below average for MD analyses (Egbert & Staples, forthcoming). Variables were only included in the analysis if they met a minimal factor loading threshold of +/−.30. Based on this criterion, 36 of the original 41 linguistic variables were retained. Each variable was only included on the factor where it loaded the strongest. The MD analysis resulted in five dimension scores for each text and the dimensions were functionally interpreted as follows:

Dimension 1: Oral Narrative

Dimension 2: Suggestions and Future Possibilities

Dimension 3: Listener-centered vs. Speaker-centered Discourse

Dimension 4: Informational Elaboration

Dimension 5: Stance

To demonstrate how MD analysis can be used to investigate the extrapolation inference, the presentation and discussion of the results will be limited to three of the five dimensions, Dimension 1, Dimension 2, and Dimension 4. They were selected because they exemplify results of the MD analysis that have bearing on the extrapolation inference. For readers interested in seeing the full results of this method, the descriptive statistics and correlational results for all five dimensions can be found in Tables A2 and A3 in the Appendix. Table 5 shows the three dimensions and the co-occurring linguistic features for each that were identified by the factor analysis. Of the dimensions reported, one is typified by both positive loading features and negative loading features. For example, positive loading features on Dimension 1 include features that are associated with recounting events such as the past tense and third-person pronouns; negative loading features include stance verbs followed by a *to* complement clause (e.g., *I* ***want to study engineering***), which are not typically found in oral narratives. Other dimensions are typified by positive loading features only. For example, Dimension 2 is largely marked by the presence of the present tense and modals.

Table 5. Overview of Staples et al. (2017) Dimensions 1, 2, 4, and their linguistic features.

Dimension	Positive features	Negative features
1. Oral Narrative	Past tense, Third-person pronouns, *That* deletion, Word count, Predicative adjectives, Communication verbs + *that* complement clauses, Certainty verbs + *that* complement clauses, Communication verbs, Type–token ratio, Subordinate clauses (other than causative or conditional)	Stance verb + *to* clause
2. Suggestions and Future Possibilities	Present-tense verbs, Prediction modals, Conditional clauses, Possibility modals, Contractions, Necessity modals, Causative verbs	NA
4. Informational Elaboration	Word length, Prepositions, Nominalizations, Attributive Adjectives, *That* relative clauses, Amplifiers, *Wh* relative clauses	NA

Results

The goal of this study was to examine evidence for the extrapolation inference for the MELAB OPI. Here, we present results from three of the five dimensions identified above to answer both our research questions. Within our discussion of each dimension, we answer the first question, *To what extent are linguistic features elicited by the MELAB similar to language observed in target domains?* by providing a comparison between the distributions (means and standard deviations) of dimension scores from the MELAB corpus and the TLU registers, represented by the reference corpora (nurse-patient interaction, service encounters, office hours, study groups, and conversation). To answer the second research question, *To what extent are linguistic features elicited by the MELAB similar to language used in the target domain as scores increase?* we examine

the trend of the distributions across score levels. We also report correlational analyses to determine the magnitude of the linear relationship between MELAB score and dimension score such that higher level test takers use more of the features associated with the TLU registers. For each of the three dimensions, we provide excerpts from the MELAB corpus and the reference corpora to further illustrate our findings.

Dimension 1: Oral Narrative

Dimension 1 is composed of both positive features and a negative feature. Positive scores on this dimension indicate more use of oral narrative linguistic features such as the past tense, third-person pronouns, and *that* deletion. Negative scores indicate more use of stance verbs followed by *to* clauses. Figure 1 shows the scores of the MELAB corpus and the reference corpora on Dimension 1: Oral Narrative. In the plot, the corpora are on the *x*-axis and the dimension scores are on the *y*-axis. The points represent each observation (individual points representing the dimension score of each of the recorded, transcribed interactions) within the corpora, the mean dimension scores of the interactions are indicated by the middle horizontal bar, and the standard deviation of the dimension scores are represented by the upper and lower horizontal bars. Speakers in the reference corpora tended to use the features of this dimension at roughly similar mean rates to each other and at higher rates than the test takers. Among the reference corpora, patients and interlocutors in conversation used these features at the highest mean rates. These higher rates could be an effect of similar communicative purposes (i.e., describing past events). Thus, to answer research question 1, we can see that across the MELAB scores, the use of oral narrative is much lower than what we find in the TLU domains, particularly conversation.

The excerpts below are examples of Oral Narrative from conversation and the MELAB corpus. In each of these excerpts the past tense is in bold, third-person

pronouns are capitalized, and desire + *to* clauses (not typical of oral narration) are underlined. In comparing Excerpts 1 and 2, it is evident that the excerpt from conversation contains more features of Oral Narrative than the excerpt from the MELAB.

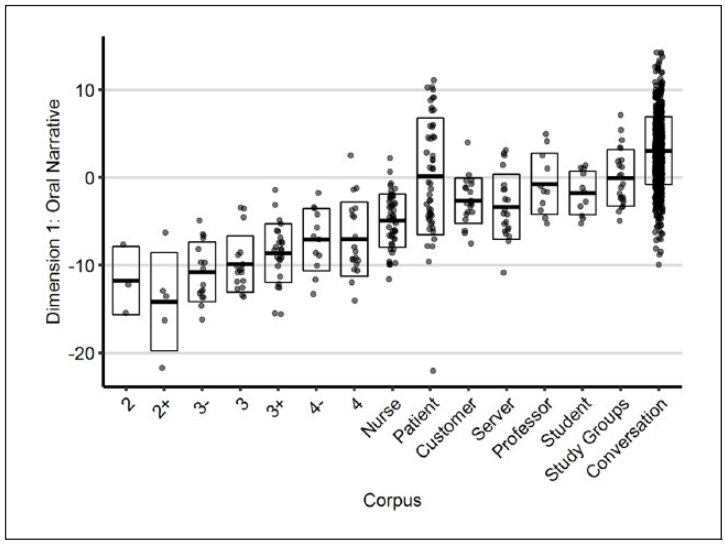

Figure 1. Distributions of MELAB (2-4) and Reference Corpora on Dimension 1.

Excerpt 1: Conversation, File 139201; Dimension 1 Score = +13.97

Speaker A: I never look at this, I, I, IT **was** two, three weeks old, all THEY had **was** you know the front page so xxx check this out, some guy's on cocaine, the last one man Juan Jones Breckland County, **pleaded** guilty to second degree burglary, HIS sentencing is scheduled today.

Excerpt 2: MELAB, File 4_A_6C.txt; Test-taker score 2+, Dimension 1 Score = −21.73

Test taker: 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.

Examiner: Uh huh.

Test taker: 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.

Figure 1 also addresses research question 2. It shows that the higher scoring test takers on the MELAB used more positive features of Dimension 1 than lower scoring test takers. The relationship between performance score and dimension scores was positive, moderate, and significant (r = 0.44). For this dimension, the gradual increase in the use of positive features as test score increases shows gradual steps toward approximating the use of Oral Narrative in the target domains. Additionally, when we compare Excerpt 2 with Excerpt 3, it is clear that Excerpt 3, which was awarded a score of 4, contains more positively loading features and fewer negatively loading features of Dimension 1 than the lower scoring performance (Test-taker score of 2+). This illustrates that the test takers who received higher scores on the MELAB demonstrated more use of Oral Narrative features than those who received lower scores.

Excerpt 3: MELAB, File 9_B_21C.txt; Test-taker score 4, Dimension 1 Score = +2.50

Test taker: And then I **applied** to University <unclear> as well.

. . .

Test taker: quite late because uh uh I **thought** I would fall under the exception that THEY have IT's like um the exception is uh that if you are studying in an English language school system before coming to Canada then you might be you know uh accepted

Dimension 4: Informational Elaboration

Positive scores on Dimension 4: Informational Elaboration represent more use of features such as attributive adjectives, prepositional phrases, relative clauses, and

nominalizations. These linguistic features were used in the reference corpora at differing mean rates, indicating variability in the rates at which target domains use these features. Figure 2 shows a split in the reference corpora's mean use of features, with professors and interlocutors in study groups using these features more. These higher rates of use can be explained by the need for professors and interlocutors in study groups to share information. We can also see that the dispersion of Dimension 4 scores for the MELAB in general is more closely aligned with registers of academic discourse (study groups and office hours) as well as the discourse of nurses. It is less aligned with the discourse of patients and that found in service encounters (customers and servers). Thus, to answer research question 1, we can see similarities between the MELAB and many of the TLU registers, but particularly office hours and study groups, two registers that require more detailed discussion of information.

In the examples of Dimension 4 from a study group and the MELAB, adjectives are in italics, prepositional phrases are underlined, relative clauses are in bold, and nominalizations are capitalized. These examples highlight the similarities between the Informational Elaboration of the MELAB and language used in study groups.

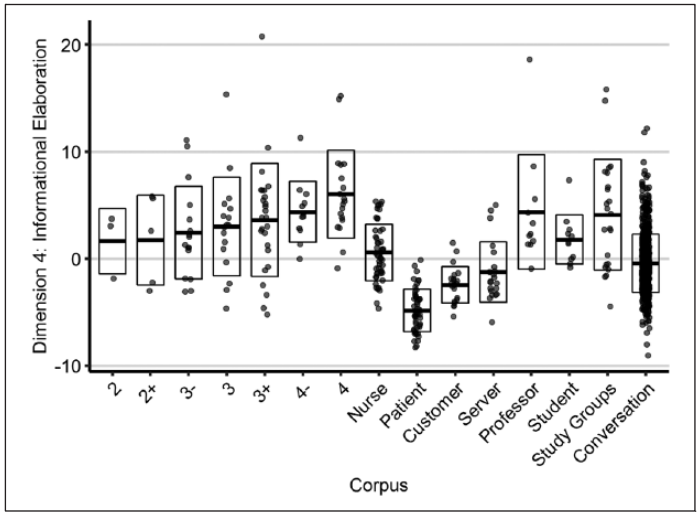

Figure 2. Distributions of MELAB (2-4) and Reference Corpora on Dimension 4.

Excerpt 4: Study Group, File Humhisgudpn037; Dimension 4 Score = +14.41

Speaker A: And the *communist* party of the country concerned should take that into account of course. And our *Chinese* friends had many *original* ideas **which they are implementing** in the course of *socialist* CONSTRUCTION in their country. They're giving birth to *new* ideas too **which take into CONSIDERATION** some *specific* conditions in China.

Excerpt 5: MELAB, File 9_D_8B; Test-taker score 4, Dimension 4 Score = +15.72

Test taker: But out of that has grown an interest really to to help people because uh SPONSORSHIP is not the only uh SOLUTION

Examiner: Uh huh.

Test taker: To many of the issues **that uh people who are in *refugee-like* SITUATIONS face**.

Examiner: Uh huh.

Test taker: So I get asked a lot of other questions **which have to do with other categories of IMMIGRATION**. And therefore I find that I need to expand my scope and also deepen my understanding of the *whole* IMMIGRATION uh area.

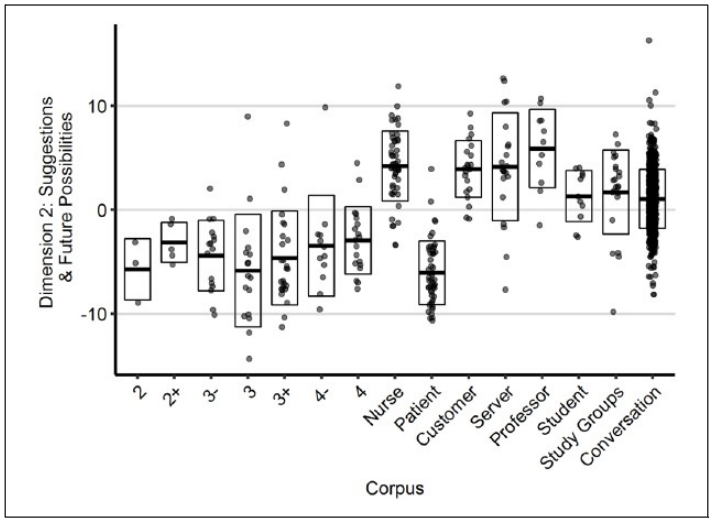

Figure 3. Distributions of MELAB (2-4) and Reference Corpora on Dimension 2.

However, when we turn to research question 2, we can also see from Figure 2 that there is a linear increase in test takers' use of these features that shows a trend away from nurses, patients, customers, servers, students, and conversation. The relationship between performance score and dimension score was positive, moderate, and significant ($r = 0.29$). Higher scoring test-takers tend to use the features of this dimension at slightly higher rates than professors and the interlocutors in study groups, highlighting an even stronger need for them to provide information during the interaction.

We can contrast the use of Dimension 4 features in Excerpt 5 above, from a higher scoring test taker, with that of Excerpt 6 below, from a lower scoring test taker. The speaker in Excerpt 6 still uses informational features at times but with less frequency than the speaker in Excerpt 5.

Excerpt 6: MELAB, File 4_B_14B.txt; Test-taker score 2+, Dimension 4 Score = −3.04

> Test taker: Because when you go to the bank, and you need to take uh maybe maybe some money you go there afternoon or <unclear> you can uh they can call you and uh you can do your business. So maybe you need uh sometimes you need uh sometimes <unclear> there are lots of pe-people, they will call you uh you need to go there you better go there uh <unclear> tomorrow or *next* day. Uh.

Dimension 2: Suggestions and Future Possibilities

Dimension 2: Suggestions and Future Possibilities was typified by greater use of linguistic features such as modals, conditionals, and the present tense. The reference corpora use these features at differing mean rates (see Figure 3). Nurses, customers, and servers tended to use these features more on average than students and interlocutors in study groups and conversation, who in turn used them at higher mean rates than patients. These differences in the use of linguistic features on this dimension is driven by communicative purpose. For example, nurses need to make suggestions and discuss future plans with their patients. It is clear that with the exception of patients, speakers in the reference corpora tend to use these linguistic features at higher rates than the MELAB OPI test takers. Thus, to answer research question number 1, there were few similarities between the MELAB OPI discourse and the discourse of the TLU registers.

The examples for Dimension 2 are from a nurse–patient interaction, an office hour interaction, and a MELAB performance. In the excerpts, modals are in bold, conditionals are underlined, and present tense is capitalized. Dimension 2 is typified by the use of these features to discuss plans and possibilities in the future, which is demonstrated by Excerpts 7 and 8. However, in the example from the test-taker production, it is clear that not many of these features are present, and the one that is present (i.e., present tense) is not used to discuss future possibilities.

Excerpt 7: UNSP, File ABN_46; Nurse Dimension 2 Score = +6.68

Nurse: We **can** always like discharge before you GO home. We **can** always provide you with documentation that for like outside counseling if you NEED. And I'**ll** make sure that the I'**ll** let our doctors KNOW.

Excerpt 8: Office Hours, File busbaoh_n156.txt, Student Dimension 2 Score = +3.47

Student: I **should** be done and **can** we go over two b? **Could** I have **could** I have used upcoming instead of forthcoming?

Professor: sure Student:

OK Professor:

<unclear>

Student: I just didn't KNOW if I **could** use upcoming so I just wanted forthcoming to say <unclear>

Excerpt 9: MELAB, File 6_B_19F.txt; Test-taker score 3, Dimension 2 Score = −3.67

Test taker: I uh ORDER conversation partner sometimes from my, my institute. They sometimes BRING one and TALK with him.

Examiner: Uh huh.

Test taker: And there was a station, asked people and they TRY to talk with him to practice English

Examiner: Uh huh.

Test taker: to improve myself, my English.

To answer research question 2, MELAB OPI test takers used these features at similar mean rates across score level, as Figure 3 shows. The relationship between performance score and dimension score was positive, weak, and not significant (r = 0.14). There is not a clear pattern of use of these linguistic features across score levels on

this dimension, and they are underused in comparison to the reference corpora, indicating less need for discussing future possibilities and making suggestions during the interaction during the test.

Discussion

The purpose of the study was to investigate linguistic and functional evidence related to the extrapolation inference for the validity argument for the MELAB speaking test. This was accomplished first by conducting a situational analysis of the MELAB OPI and its TLU registers (nurse-patient interaction, service encounters, office hours, study groups, and conversation), which serves as a lens for interpreting the results of the linguistic analysis. Then we examined and compared the distributions of the dimension scores across the MELAB corpus and the reference corpora. This was followed by an analysis of the relationship between test takers' scores on the MELAB OPI and the dimension scores of their responses from the MD analysis. We also investigated whether higher scoring test takers used more of the features associated with the reference corpora (nurse-patient interaction, office hours, service encounters, study groups, and conversation).

The results of the situational analysis of the MELAB corpus and the TLU registers revealed key differences between the MELAB corpus and the TLU corpora in topic, participants' social roles and relationships, and communicative purposes. The difference in communicative purposes may have played a role in the extent to which test taker language approximated the target domains. The primary purpose for test takers to communicate on the test is to demonstrate language proficiency by answering questions and sharing professional and personal background. Similar to the context of the test, the primary purpose for communicating in study groups and office hours is to share information. Narrating and providing suggestions are not a primary communicative purpose in the test

task; however, these purposes are central to face-to-face conversation (narration) and nurse-patient interaction and office hours (providing suggestions). These situational differences were also reflected in the different patterns of use for linguistic features related to narration and providing suggestions.

The results of the comparison of the distributions of the MELAB and reference corpora across the three dimensions of the MD analysis show mixed support for the extrapolation inference within the validity argument for the MELAB. There were similarities in the mean dimension scores and standard deviations between the MELAB and many of the reference corpora with respect to Dimension 4: Informational Elaboration. However, there were differences between the MELAB corpus as a whole and the reference corpora with respect to Dimension 1: Oral Narrative and Dimension 2: Suggestions and Future Possibilities.

When we compared the distributions of the MELAB across score levels, we found that upper-score-level MELAB responses used more of the features of Oral Narrative, meaning that they began to approximate some the target domains represented by the reference corpora in their use of features for Dimension 1. Additionally, higher scoring test takers used Informational Elaboration features at similar rates to the professors and study groups. However, the responses to the MELAB OPI lack many of the linguistic features related to making suggestions and discussing future possibilities, regardless of MELAB score level. The results of the correlation analysis revealed moderate positive relationships between test takers' scores and their use of Oral Narrative features (Dimension 1) as well as their use of features related to Informational Elaboration (Dimension 4). There was not a discernible relationship between MELAB speaking test scores and Dimension 2.

The increasing (or decreasing) use of linguistic features as a test score increases can provide evidence for the extrapolation inference if the use of linguistic features at the

endpoint of the trend (i.e., the highest score on the rubric) approximates the use of the linguistic features in the reference corpora. Thus, the results of this analysis show relatively strong support, or backing, for extrapolating about high-scoring MELAB test takers' abilities to elaborate in study group sessions or as professors in office hours (e.g., if the test is used as a screening tool for international teaching assistants [ITAs]). Additionally, test users can be somewhat confident that incoming students who scored highly on the MELAB have the linguistic means to participate in discussions about course content in study groups. These results also show some backing for extrapolation about the ability of high scorers to have the linguistic means to narrate similarly to some of the target domains (e.g., nurses and servers) represented by the reference corpora.

Test users cannot be certain, however, about the test takers' abilities to talk about future events or to make suggestions. Test takers tend not to use these features in any of the scoring bands. This may limit test users' ability to extrapolate from performance on the task to performance as a nurse, professor, or ITA. Part of a nurse's job is to counsel, or make suggestions to, their patients, and professors (and potentially ITAs) use such language to help students solve problems. Since the test takers are not asked for advice or to make suggestions about future possibilities, then they seem not to have the opportunity to use these features in the MELAB. As a result, there is little evidence regarding the extent to which test takers can or cannot use these features in the TLU domain.

The findings of the present study and of those of Brooks and Swain (2014) illustrate that linguistic variation in test tasks are driven by their situational characteristics (e.g., communicative purpose). In addition, these findings underscore the importance of the role that the context of language use plays on actual language production, which has been highlighted as an important consideration in current test development frameworks (Bachman, 1990; Bachman & Palmer, 1996, 2010; Chapelle et al., 2008a).

This study also adds to the literature that examines the linguistic features of productive test tasks by illustrating the power of MD analysis as a tool. Rather than investigating individual features, which tends to result in few features being identified by the analysis as important (Brooks & Swain, 2014; Brown et al., 2005; Jamieson & Poonpon, 2013; Kang, 2013; Kyle et al., 2016; LaFlair et al., 2015), a large majority of the features that were initially selected for inclusion at the outset of the analysis were retained after the MD analysis was conducted (36 out of 41). This retention and grouping of co-occurring linguistic features reveals more interpretable patterns of language use, a more exhaustive comparison to language use in the target domain, and a more robust method for investigating the extrapolation inference.

The results are clearly limited by the small samples in the MELAB OPI corpus and the other reference corpora, with the exception of the conversation sub-corpus of the Longman Corpus of Spoken and Written English. Additionally, the present study clearly does not account for every linguistic variable that may represent the construct. For example, in previous studies (e.g., Brooks & Swain, 2014; Kang, 2013) grammatical accuracy and fluency variables are features of interest in the analysis of test-taker production, but they were not accounted for in the present study. It is possible that these features would play a role in one or more dimensions if they were identified in the test-taker corpus. Furthermore, conversation may be too broad of a domain to extrapolate to given its potentially wide range of contexts (e.g., informal social gatherings, family interaction). Future extrapolation studies would benefit from the inclusion of more features and a more nuanced comparison with conversation.

Conclusion

Current frameworks for investigating validity demand varied and robust evidence for the interpretations and uses of high-stakes language assessments. In this paper, we have

proposed a new method (corpus-based register analysis with MD analysis) for investigating evidence for the extrapolation inference. This method can be viewed as a linguistic parallel to traditional criterion validity studies. However, instead of investigating the relationship between test scores and criterion scores, we have proposed investigating the relationship between the uses of linguistic features that are found to co-occur through MD analyses as well as their functional interpretations. This method is supported by the similarities in the theoretical underpinnings between the TLU analysis framework and the corpus-based register analysis framework.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/ or publication of this article: This work was supported by CaMLA's Spaan Research Grant Program, 2014.

Note

1. A heuristic for organizing the contextual features of speech acts: S – Setting and Scene, P – Participants, E – Ends, A – Act Sequence, K – Key, I – Instrumentalities, N – Norms, G – Genre

References

Bachman, L. F. (1990). *Fundamental considerations in language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Biber, D. (1994). An analytical framework for register studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 31–56). Oxford, UK: Oxford University Press.

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

Biber, D. (2008). Corpus-based analyses of discourse: Dimensions of variation in conversation. In Bhatia, V. K., Flowerdew, J. & Jones, R. H. *Advances in discourse studies* (pp. 100-114). New York: Routledge.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

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). *Grammar of spoken and written English*. London, UK: Pearson.

Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBT and real-life academic speaking activities. *Language Assessment Quarterly*, *11*(4), 353–373.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series*, *2005*(1), i–157.

Cambridge Michigan Language Assessments. (2016). *MELAB*. Retrieved from www.cambridgemichigan.org/institutions/products-services/tests/proficiency-certification/melab/

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008a). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008b). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1–25). New York: Routledge.

Egbert, J., & Staples, S. (forthcoming). Doing multi-dimensional analysis in SPSS, SAS and R. In T. Berber-Sardinha & M. Veirano (Eds.), *Multi-dimensional analysis*. London: Bloomsbury.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin Books.

Hymes, D. (1974). *Foundations in sociolinguistics*. Philadelphia, PA: University of Pennsylvania Press.

Jamieson, J. M., & Poonpon, K. (2013). Developing analytic rating guides for TOEFL iBT's integrated speaking tasks. *ETS Research Report Series*, *2013*(1), i–93.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels B1 to C2 and examinee L1 backgrounds. *Research Notes*, *52*, 40–48.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319–340.

LaFlair, G. T., Staples, S., & Egbert, J. (2015). Variability in the MELAB speaking task: Investigating linguistic characteristics of test-taker performance in relation to rater severity and score. *CaMLA Working Papers (2015–04)*.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

Qian, H., Woo, A., & Banerjee, J. (2014). *Setting an English language proficiency passing standard for entry-level nursing practice using the Michigan English Language Assessment Battery*. Chicago, IL: National Council of State Boards of Nursing.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org/.

Revelle, W. (2016). Psych: Procedures for psychological, psychometric, and personality research (Version 1.6.9) [Computer software]. Evanston, IL: Northwestern University. Retrieved from http://CRAN.R-project.org/package=psych

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Staples, S. (2015). *The discourse of nurse-patient interactions: Contrasting the communicative styles of us and international nurses* (Vol. 72). Philadelphia, PA: John Benjamins.

Staples, S., LaFlair, G. T., & Egbert, J. (2017). Comparing language use in oral proficiency interviews to target domains: Conversational, academic, and professional discourse. *Modern Language Journal*, *101*(1), 194-213.

Toulmin, S. E. ([1958] 2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Weigle, S. C., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes*, *18*, 25–39.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer. Retrieved from http://had.co.nz/ggplot2/book

Yan, X., & Staples, S. (2017). Investigating lexico-grammatical complexity as construct validity evidence for the ECPE writing tasks: A multidimensional analysis. *CaMLA Working Papers*.

Appendix

Table A1. Linguistic features included in the study.

Feature	Example
That deletion	I think *(that)* the distance is uh 300 kilometers.
Contractions	*can't, don't*
Present tense verbs	he *travels*
Second-person pronouns	*you, your, yours, yourself*
Emphatics	*just, a lot*
First-person pronouns	*I, me, my, mine, we, us, our*
Causative clauses	Now I'm happy *because I take the lesson driver and I can drive.*
Discourse particles	*now, well*
Hedges	*almost, more or less, kind of, sort of*
Amplifiers*	*greatly, totally, utterly, very*
Wh questions	*What* is your name?
Nouns	*test, book*
Prepositions	*to, of, for*
Attributive adjectives	*good* job, *new* friends
Past tense verbs	*saw, wondered*
Third-person pronouns	*he, she, him, her, them, they*
Nominalizations	*admission, education*
Possibility modals	*could, might*
Adverbs	*unfortunately, likely*
Prediction modals	*will, be going to*
Conditional clauses	*if I have a long break*
Necessity modals	*must, have to*

Feature	Example
Conjunctive adverbials*	*also, besides*
Other subordinate clauses	How did you know about the MELAB test *since it is virtually new* in Jordan?
Predicative adjectives	Oh yeah, that's *excellent.*
Wh relative clauses	I want to work in hotels *which will be in five stars*.
That relative clauses	What was your favorite thing at Disney World *that you saw*?
Premodifying nouns*	*sales* job
Communication verb + *that* complement clause	So you *said that* you're interested in […]
Certainty verb + *that* complement clause	I did not *know that it's such a cold city.*
Likelihood verb + *that* complement clause*	I really *think that only way to be able* […]
Certainty adverbials	*certainly, definitely, of course*
Likelihood adverbials	*perhaps, probably, maybe*
Stance verb + *to* complement clause	I want to *study mechanical engineering*.
Activity verbs*	*borrow, play, wait*
Communication verbs	*accuse, offer*
Mental verbs	*accept, imagine*
Causative verbs	*let, permit*
Type/token ratio	
Word length	
Word count	

* These features had factor loadings less than 0.30 and thus were dropped from the analysis.

Table A2. Descriptive Statistics for the sub-corpora/registers across five dimensions.

Corpus	Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
2	−11.77	3.92	−5.75	2.95	−8.13	1.54	1.63	3.06	−0.70	1.45
2+	−14.16	5.60	−3.15	1.92	−9.00	3.75	1.74	4.21	−0.94	3.21
3−	−10.77	3.39	−4.41	3.39	−7.86	3.85	2.43	4.33	−1.77	3.17
3	−9.87	3.21	−5.86	5.40	−6.98	1.65	2.99	4.60	−2.67	3.35
3+	−8.62	3.33	−4.66	4.52	−6.73	2.58	3.60	5.28	−0.04	5.24
4−	−7.08	3.54	−3.47	4.85	−8.19	1.85	4.39	2.85	2.08	5.19
4	−7.04	4.24	−2.94	3.24	−5.03	2.56	6.03	4.11	2.40	4.93
Nurse	−4.92	3.05	4.20	3.36	7.06	2.66	0.58	2.64	1.36	3.53
Patient	0.14	6.66	−6.06	3.06	−2.23	3.07	−4.86	2.00	3.02	5.02
Customer	−2.66	2.58	3.92	2.74	1.92	1.53	−2.45	1.71	−0.08	2.47
Server	−3.35	3.69	4.15	5.18	3.64	3.51	−1.25	2.83	−0.03	4.34
Professor	−0.74	3.48	5.89	3.80	1.63	2.44	4.38	5.36	2.00	3.16
Student	−1.76	2.48	1.31	2.47	−0.18	3.30	1.78	2.30	3.49	2.64
Study Groups	−0.05	3.22	1.70	4.04	−0.58	1.67	4.09	5.18	−0.72	2.70
Conversation	3.05	3.85	1.05	2.84	−0.39	1.41	−0.44	2.73	−0.35	3.04

Table A3. Correlation between the test-taker sub-corpora score levels and five dimensions.

Dimensions	Pearson's *r*
Dimension 1: Oral Narrative	0.44
Dimension 2: Suggestions and Future Possibilities	0.14
Dimension 3: Listener-centered vs. Speaker-centered Discourse	0.30
Dimension 4: Informational Elaboration	0.29
Dimension 5: Stance	0.32

Note: All relationships were significant except Dimension 2.