# APPLICATIONS OF TEXT ANALYSIS TOOLS FOR SPOKEN RESPONSE GRADING

**Scott Crossley, Georgia State University**
**Danielle McNamara, Arizona State University**

This study explores the potential for automated indices related to speech delivery, language use, and topic development to model human judgments of TOEFL speaking proficiency in second language (L2) speech samples. For this study, 244 transcribed TOEFL speech samples taken from 244 L2 learners were analyzed using automated indices taken from Coh-Metrix, CPIDR, and LIWC. A stepwise linear regression was used to explain the variance in human judgments of independent speaking ability and overall speaking proficiency. Automated indices related to word type counts, causal cohesion, and lexical diversity predicted 52% of the variance in human ratings for the independent speech samples. Automated indices related to word type counts and word frequency predicted 61% of the variance of the human scores of overall speaking proficiency. These analyses demonstrate that, even in the absence of indices related to pronunciation and prosody (e.g., phonological accuracy, intonation, and stress), automated indices related to vocabulary size, causality, and word frequency can predict a significant amount of the variance in human ratings of speaking proficiency. These findings have important implications for understanding the construct of speaking proficiency and for the development of automatic scoring techniques.

**Key words:** Language Testing, Speaking Proficiency, Computational Linguistics, Corpus Linguistics, Machine Learning

**APA Citation:** Crossley, S., & McNamara, D. Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. Retrieved from http://llt.msu.edu/issues/june2013/crossleymcnamara.pdf

**Received:** August 27, 2012; **Accepted:** February 5, 2013; **Published:** June 1, 2013



## INTRODUCTION

Our interest in this study is to better understand the underlying linguistic features that are predictive of communicative competence in second language (L2) learners. Communicative competence refers to the ability of language learners to organize language appropriately using grammatical and textual skills as well as develop pragmatic competence at the illocutionary and sociolinguistic level (Bachman, 1990; Canale & Swain, 1980). Which linguistic features are most predictive of communicative competence is the subject of some debate (Iwashita, Brown, T. McNamara, & O'Hagan, 2008), but the debate lends itself to empirical analyses approached from a variety of inter-related disciplines such as corpus linguistics, cognitive linguistics, and natural language processing.

The goal of this study is to use these disciplines to investigate communicative competence through the quantitative analysis of human ratings of speaking proficiency using independent speaking tasks as found in the Test of English as a Foreign Language (TOEFL) internet-Based Test (iBT). Our purpose is to examine which linguistic features in speech samples are most predictive of human ratings of speaking proficiency. We view human judgments of speaking proficiency as one of many potential gold standards of communicative competence (Iwashita et al., 2008), with the understanding that such judgments have significant consequences for L2 learners attempting to study abroad or to obtain jobs that require a certain level of communicative proficiency.

In this study, we focus on higher-level linguistic features related to speech delivery (i.e., number of words or ideas), language use (i.e., vocabulary and grammar) and topic development (i.e., idea density, coherence, and content relevance) and their relation to communicative competence as compared to lower level features found in speech delivery (i.e., fluency, intonation, rhythm, and pronunciation). We focus on higher-level linguistic features because automated speech recognition (ASR) tools are currently unreliable, especially in less predictable contexts such as independent speech samples and for L2 learners in general (Xi, Higgins, Zechner, & Williamson, 2008). Instead of relying on ASR tools, we use human transcribers to transcribe a selection of speech samples and then use computational tools to automatically examine the linguistic features in the speech samples. Thus, like Zechner, Higgins, Xi, and Williamson (2009), we see our study as additive because it provides information about speaking proficiency that goes beyond speech delivery. Our assumption is that, over time, technological advances will afford accurate automated speech recognition allowing full automation of speech analysis on the multiple dimensions that underlie speaking proficiency. However, until that point arrives, we must depend on human transcription to accurately analyze most of the linguistic features of spoken text.

We have two primary objectives in this study. The first is to better understand the construct of speaking proficiency through an analysis of human judgments of speaking proficiency. The second is to investigate the potential for automated indices related to aspects of proficiency (e.g., speech delivery, language use, and topic development) to advance automated scoring algorithms. Understanding the principal aspects underlying the construct of speaking proficiency will afford a better understanding of how humans evaluate communicative competence in relationship to its development in a cross-sectional corpus of L2 learners. Investigating the predictive ability of automated indices of human judgments will also promote the development of automated scoring techniques (AST). Dependable ASTs may eventually allow L2 learners without access to native or high-proficiency English speakers to receive natural feedback and pedagogical assessment regarding their speaking performance, thus making language assessment and materials available to a wider range of learners. ASTs also have the advantage of reduced costs (compared to human assessors), speed, flexibility, and reliability (Higgins, Xi, Zechner, & Williamson, 2011).

**Speaking Proficiency**

Exact definitions of speaking proficiency are not readily available or agreed upon, making generalizations about the construct difficult (Shin, 2005). However, two sets of standardized guidelines, though criticized, prove informative. These two guidelines come from the speaking proficiency rubrics published by the American Council on the Teaching of Foreign Languages (ACTFL, 1999) and TOEFL (Educational Testing Services, 2004). These two rubrics are not comparable at an assessment level because the ACTFL rubric assesses proficiency level by level whereas the TOEFL rubric is used to assess participants regardless of proficiency level. However, they do provide general standards for assessing speaking proficiency.

The ACTFL guidelines describe novice, intermediate, and advanced speakers at the low, mid, and high levels as well as superior speakers (ten levels of speaking proficiency in total). According to the ACTFL guidelines, speaking proficiency corresponds to accuracy and fluency in conversation on a variety of topics. The linguistic features related to proficiency include the concreteness, abstractness, and complexity of perspectives, the coherence of narrative, topical knowledge, argument structure, the ability to engage in extended discourse, syntactic and lexical complexity, discourse strategies (i.e., turn-taking), and phonology (i.e., pitch, stress, and tone). According to the ACTFL guidelines (1999), superior speakers communicate accurately and fluently and are able to participate in a number of conversational settings revolving around a variety of topics. From a communicative perspective, superior speakers can easily explain complex matters and provide coherent and extended narrations. Features of superior speakers include interactive and discourse strategies that are characterized by complex syntactic, lexical, and intonational devices. In general, superior speakers make few errors and rarely make errors that

interfere with communication; nevertheless, their discourse may be influenced by the language patterns in their L1. In contrast, novice low speakers are described as having no real functional ability and are potentially unintelligible because of pronunciation problems. Novice low speakers are generally limited to exchanging greetings, giving identities, and naming familiar objects in the immediate environment. These limitations stop novice low speakers from participating in actual conversational exchanges.

The TOEFL speaking rubric for independent speech is less complex than the ACTFL rubric reporting on only four levels of proficiency (1 through 4) along with a level 0 in which the speaker makes no attempt to respond or responds to an unrelated topic. Each level has a general description that addresses the completeness, intelligibility, and coherence of the speech sample along with three additional characteristics (delivery, language use, and topic development). Delivery focuses on pace, pronunciation, and intonation. Language use highlights the use of grammar and vocabulary in the speech sample while topic development focuses on the relationship between the ideas in the sample. Speakers who receive a score of 4 on the TOEFL-iBT independent rubric are characterized as fulfilling the demands of the task with only minor lapses in completeness. The response is evaluated as highly intelligible and exhibits sustained discourse that presents a well-paced flow of words with only minor difficulties in pronunciation or intonation patterns. In reference to language use, highly rated responses demonstrate effective use of grammar and vocabulary and exhibit fairly high degrees of automaticity. Speakers at this level also have well developed topics that are coherent and demonstrate clear relationships between ideas. Speakers who receive a score of 1 provide a response that is characterized as containing limited content and/or coherence with minimal connections to the task and speech that is largely unintelligible. The response has consistent problems in pronunciation, stress, and intonation with choppy delivery that contains frequent pauses and hesitations. Grammatically and lexically, the response is severely limited and may contain practiced or formulaic expression. The response also fails to express relevant content and lacks substance beyond the basic ideas of the prompt.

**Analyses of Speaking Proficiency**

Most investigations of speaking proficiency have examined the relationships between the linguistic features contained within speaking samples (either through primary trait scores or feature counts) and human ratings of holistic speaking proficiency as reported by guidelines such as the ACTFL and TOEFL rubrics discussed above.

An early example of this approach is Adams' 1980 study in which he examined connections between the holistic speaking scores on the Foreign Service Institute (FSI) Oral Interview Test of Speaking and primary trait scores of speaking proficiency (i.e., human judgments of accent, comprehension, vocabulary, fluency, and grammar). Adams found that the strongest predictors of holistic speaking scores were analytic judgments of grammar and vocabulary. Analytic ratings evaluating accent and fluency failed to discriminate the holistic scores at lower levels of proficiency. Adams concluded that grammar and vocabulary were the most important indicators of speaking proficiency across all levels. As levels increased, other factors such as pronunciation, sociolinguistic factors, and fluency also became important predictors of proficiency.

Bejar (1985) examined inter-rater reliability among a number of raters of the Test of Spoken English (TSE). His primary aim was not to analyze the predictive strength of primary trait scores of general speaking proficiency (defined as comprehensibility), but rather to reduce scoring costs by examining the potential to use one rater instead of two. However, in the process of investigated reliability between raters, Bejar also conducted a series of correlational analyses comparing primary trait scores (grammar, fluency, and pronunciation) to scores of speaking comprehensibility for 560 TSE examinees. Bejar found that the strongest correlations with scores of comprehensibility were for pronunciation followed by fluency and grammar.

In a later study, T. McNamara (1990) examined speaking proficiency using primary trait scores and

holistic human ratings retrieved from the Speaking sub-test of the Occupational English Test (OET) administered to health professionals. McNamara used the primary trait scores reported by the human raters (grammar and expression, intelligibility, comprehension, appropriateness, and fluency) to model the given holistic speaking score reflecting overall communicative effectiveness. Like Adams (1980), McNamara found that ratings of grammar and expression were the strongest predictors of the holistic scores for speaking proficiency.

More recently, Iwashita, Brown, T. McNamara, and O'Hagan (2008) examined 200 speech samples taken from five tasks in the TOEFL-iBT (two independent speaking tasks and three integrated speaking tasks). The 200 speech samples were transcribed and then analyzed using a variety of methods to investigate the linguistic features in the samples (e.g., grammatical accuracy, grammatical complexity, vocabulary, pronunciation, fluency). In collecting scores of grammatical accuracy, Iwashita et al. depended on primary trait scores for a variety of grammatical features (i.e., verb tense, third person singular, plural markers, preposition, articles) and global accuracy of use (i.e., error free T-units as a percentage of total T-units). Primary trait scores were also collected for grammatical complexity (i.e., T-unit complexity ratio, dependent clause ratio, verb phrase ratio, mean length of utterances), fluency (i.e., filled pauses, repairs, mean length of runs), and pronunciation (i.e., intonation, rhythm, pronunciation of words and syllables). When collecting scores for vocabulary use, Iwashita et al. used VocabProfile (Cobb, 2002) to collect measurements of high and low frequency words along with type and token counts. Iwashita et al. reported that many linguistic features within the speaking samples varied according to proficiency including grammatical accuracy (all measures), grammatical complexity (verb phrase complexity, mean length of utterance), vocabulary (type and token counts), pronunciation (syllables), and fluency (speech rate, unfilled pauses, total pause time). However, not all measures were strong predictors of level (with the exception of target like syllables, speech rate, and the number of words).

Overall, these studies generally support the notion that grammatical, lexical, and pronunciation measures are all strong predictors of human judgments of speaking proficiency. However, the patterns are not shared across studies. For instance, in two studies (Bejar, 1985; Iwashita et al., 2008) pronunciation and fluency were significant predictors of speaking proficiency, but in other studies (Adams, 1980; T. McNamara, 1990), measures of pronunciation and fluency were not strong predictors of holistic scores of speaking proficiency. In addition, these studies may not be directly comparable because the FSI test proceeds through a series of levels whereas the other tests assess proficiency based on a single prompt. However, when taken holistically, the findings are informative as general indicators of human judgments of speaking proficiency.

**Automatic Scoring Techniques**

More recent studies addressing the role that linguistic features play in predicting human scores of speaking proficiency focus on automatic scoring techniques (AST). The purpose behind AST is to test automated indices that demonstrate construct validity with features of speaking proficiency in an effort to predict human ratings. If accurate predictions of human ratings are possible, AST can offer the benefits of reducing the costs of scoring, scoring times, and scoring errors. Success using AST has been reported in complex tasks such as essay scoring (Burstein et al., 1998; Crossley & D. McNamara, 2012; D. McNamara, Crossley, & McCarthy, 2010) and tightly controlled speaking tasks that involve either the production of factual information (Leacock & Chodorow, 2003) and or use predictable contexts (Bernstein, 1999; Bernstein, van Moere, & Cheng, 2010).

Most research in AST for speaking proficiency comes from Educational Testing Services, whose interests lie in providing quick and accurate feedback to TOEFL-iBT takers and users of TOEFL Practice Online (TPO). Likely the most advanced automatic scoring tool for speaking proficiency currently available is *SpeechRater*, but even this tool fails to measure a variety of features important in judging speaking proficiency (e.g., intonation, grammatical complexity, lexical sophistication, relevancy of topic) and fails

to reach similar levels of agreement as human raters (Higgins et al., 2011). A major limitation to *SpeechRater* is its ability to automatically recognize and transcribe speech. As reported by Xi et al. (2008), *SpeechRater* reports a word accuracy of only 53%, using speech recognition software developed by Multimodal Technologies Inc.

Recent studies using *SpeechRater* (Higgins et al., 2011; Xi et al., 2008) have examined the potential for *SpeechRater* to predict human ratings of speaking proficiency in both TOEFL and TPO speaking samples. Generally these studies report on three, weighted rating schemes for the holistic ratings (equal weighting, weightings devised by a content advisory committee, and a least squared approach/empirical weighting) and the capacity for five automated features related to speech pronunciation, fluency, vocabulary, and grammar to model these scores. The five features reported in these studies and their description can be found in Table 1. It should be noted that the algorithms underlying many of these features are either ill defined or not defined in the literature.

Table 1. *Features Reported for SpeechRater by Higgins et al. (2011)*

| Index | Feature | Description | Extraction method |
|---|---|---|---|
| Amscore | Pronunciation | Compares the pronunciation of non-native speech to a reference pronunciation model | Not reported |
| Wpsec | Fluency | Speech articulation rate | Words per second |
| Tpsecutt | Fluency and vocabulary | Unique words normalized by speech duration | Types of words divided by the total length of speech |
| Wdpchk | Fluency and vocabulary | Average length of speech chunks | Not reported |
| Lmscore | Grammar | Compares the language of non-native speech to a reference language model and models the probabilities of word sequences | Not reported |

Using the weighted human evaluations and looking at single scores and combination of scores (two items, three items, and six items), Higgins et al. (2011) reported moderate scoring accuracies for the sets of six combined items with correlations of between .674 and .729 for the human ratings of the TOEFL samples and correlations between .509 and .574 for the human ratings of the TPO samples. Correlations using fewer than six items were lower. The best model developed using *SpeechRater* was able to explain a little over half the variance in human ratings of speaking proficiency ($r$ = .729, $r^2$ = .531) for the human ratings of the TOEFL samples.

Overall, AST for assessing speaking proficiency have demonstrated moderate results in predicting human judgments of speaking quality. While disappointing, the limitations of AST approaches for speaking proficiency rest in the low accuracy of speech recognition technologies. Once these technologies reach the reliability of human transcribers, the predictive capability of AST should increase substantially because the technology to analyze language use and topic development is available. This technology and its predictive ability are the focus of this study.

## METHOD

The goal of this paper is to determine the degree to which human judgments of speaking proficiency can be predicted using automated indices related to delivery (i.e., number of words or ideas), language use (i.e., vocabulary, grammar), and topic development (i.e., idea density, coherence, content relevance), the

three major constructs of speech as found in the TOEFL-iBT scoring rubric. However, we only focus tangentially on the first construct of speech delivery found in the TOEFL-iBT scoring rubric. That is to say, we focus on the number of words or ideas (i.e., the flow of ideas), but not on the pronunciation and prosody of the speech sample because accurate and reliable methods for assessing pronunciation and prosody are unavailable. In addition, to take advantage of current state of the art automated indices related to language use and topic development, our analysis assumes the existence of a speech recognition system that matches human accuracy in speech transcription.

To examine if language features beyond phonology and prosody affect human judgments of speaking proficiency, we analyzed a transcribed corpus of scored spontaneous speech samples taken from the TOEFL-iBT public use dataset using automated linguistic features taken from various computational tools such as Coh-Metrix (Graesser, D. McNamara, Louwerse, & Cai, 2004; D. McNamara & Graesser, 2012), the Computerized Propositional Idea Density Rater (CPIDR; Brown, Snodgrass, Kemper, Herman, & Covington, 2008) and Linguistic Inquiry and Word Count (LIWC, Pennebaker, Francis, & Booth, 2001). To predict human judgments of speaking proficiency, we divided the corpus of speech samples into training and test sets. We then conducted correlations and computed linear regression models comparing the expert ratings of spoken proficiency (both proficiency in spontaneous speech and overall spoken proficiency) and the scores reported by the computational tools using the training set only. The results of this analysis were later extended using the regression model to the held back, independent test set data, and finally to the complete corpus.

**TOEFL-iBT Public Use Dataset (Speech Samples)**

The TOEFL-iBT public use dataset comprises data collected from TOEFL-iBT participants from around the world. The public use dataset contains three separate datasets: item level scores, speech samples, and writing samples. The speech sample dataset includes speech samples from 480 examinees on six speaking tasks stratified by quartiles (240 participants taken from two test forms). The six speaking tasks include two independent speaking tasks and four integrated speaking tasks. These tasks represent the speaking content and speaking expectations of academic situations (Cumming, Grant, Mulcahy-Ernt, & Powers, 2004). The independent tasks ask participants to speak on familiar topics and draw upon their own ideas, opinions, and experiences (i.e., spontaneous speech). These speaking tasks last 45 seconds. The integrated tasks require participants to read and/or listen to material and then respond to a prompt that requires the participant to use information from the reading and listening material. These speaking tasks last 60 seconds. The TOEFL-iBT public use dataset includes human scores of speaking proficiency for each task and a combined overall speaking proficiency score.

In this study, we are primarily interested in the independent speaking tasks because they better reflect naturalistic speech in that they are more spontaneous, less context dependent, and participants are not allowed to take notes. They also provide participants the opportunity to explore topics of interest, speak about real life situations, and use their own past experiences to demonstrate their basic communicative skills (Cumming et al., 2004). Such considerations afford a more accurate and valid measurement of the construct of interest (i.e., speaking proficiency; Bernstein et al., 2010; Higgins et al., 2011). In addition, we consider overall speaking proficiency scores as reported for the TOELF-iBT data. These scores combine the score from both independent and integrated speaking tasks.

As mentioned earlier, our focus is on predicting how textual features related to language use and topic development explain the variance in human scores of speaking proficiency. Our approach is to assume the existence of a speech recognition system that is as accurate as human transcribers. Such an approach affords us the opportunity to look at the lexical features, grammatical properties, and the idea density of the speech samples produced by the TOEFL participants. Thus, a trained transcriber transcribed each of the independent speech samples from 125 participants randomly selected from test form 1 and 125 participants randomly selected from test form 2 ($N$ = 500). The transcriber only transcribed the speaker's

words and did not transcribe metalinguistic data (e.g., pauses, breaths, grunts) or filler words (e.g., ummm, ahhhh). Other disfluencies that were linguistic in nature (e.g., false starts, word repetition, repairs) were retained. If a word was not transcribable, that word was annotated with an underscore. Periods were added to the samples at the end of idea units. A second transcriber then reviewed the transcripts for accuracy. Descriptive information for the transcribed samples including means and standard deviations are located in Table 2.

Table 2. *Descriptive Statistics for the Initial Transcribed TOEFL-iBT Spoken Corpus*

| Form/ Item | n | Prompt | Mean total score (*SD*) | Number of words (*SD*) | Mean Score (*SD*) |
|---|---|---|---|---|---|
| 1/1 | 125 | Students work hard but they also need to relax. What do you think is the best way for a student to relax after working hard? Explain why. | 15.13 (4.57) | 80.54 (22.54) | 2.49 (.82) |
| 1/2 | 125 | Some people think it is alright to stay up late at night and sleep late in the morning. Others think it is better to go to bed early at night and wake up early. Which view do you agree with? Explain why. | | 92.29 (25.09) | 2.64 (.87) |
| 2/1 | 125 | Talk about the most important gift you have ever received. Describe the gift and explain why it was significant. | 15.42 (4.21) | 87.52 (23.50) | 2.63 (.79) |
| 2/2 | 125 | Do you think your life is easier or more difficult than your grandparents' lives? Use examples and details to explain your answer. | | 86.91 (24.67) | 2.64 (.83) |

An immediate problem with the transcribed speech samples was their length. Many of the automated indices computed for this analysis need a minimum of 100 words in order to report reliable values. The 100-word cut off is contingent on a sample providing the necessary lexical coverage to compute indices such as word concreteness, word frequency, lexical diversity, and word familiarity. Below a threshold of 100-words, such indices may be unreliable because the sample may not provide enough linguistic representation (i.e., not enough words to generalize about lexical, syntactic, and cohesion elements in the text). Considering these restrictions, we combined the independent speech samples produced by each participant in order to meet the text length requirements of the automated indices. After combining the two independent speech samples for each participant, six samples still fell beneath the 100-word threshold. These samples were removed from the corpus leaving us with a final corpus of 244 transcribed speech samples: 122 speech samples from test form 1 and 122 speech samples from test form 2 (see Table 3 for descriptive statistics). A limitation to such an approach is that it does not provide us a method to assess potential prompt-based differences.

Table 3. *Descriptive Statistics for Final Combined TOEFL-iBT Corpus Used in Study*

| Form | n | Number of words (*SD*) | Mean item score (*SD*) | Mean total score (*SD*) |
|---|---|---|---|---|
| 1 | 122 | 174.87 (41.93) | 5.20 (1.50) | 15.30 (4.50) |
| 2 | 122 | 177.46 (43.52) | 5.34 (1.41) | 15.63 (4.10) |

### Survey Instrument

The rubric used by the raters in this study was developed specifically for the TOEFL speaking proficiency test. The rubric provides a holistic score of speaking proficiency and is based on a 0–4 Likert scale with a score of 4 representing sustained coherent discourse and a score of 1 representing a response that is limited in content and/or coherence and is largely unintelligible. A score of 0 represents a response in which the speaker makes no attempt to respond or the response is unrelated to the topic. Raters are asked to consider three criteria when providing a holistic score: delivery (i.e., pronunciation), language use (i.e., grammar and vocabulary), and topic development (i.e., content and coherence).

### Human Ratings

Two expert TOEFL raters scored each speaking sample in the corpus. The use of a holistic score affords greater efficiency in scoring and likely lowers the cognitive burden on the raters (Xi, 2007). While inter-rater reliability scores are not provided for the TOEFL-iBT scores in the public use dataset, reported weighted kappa values for similarly double scored TOEFL speaking samples generally range from .77 for one score and up to .93 for three scores (Xi et al., 2008). For our analysis, the scores from the two independent samples were combined to create an independent speaking score. The scores from the two independent samples and the four integrated samples were combined to form an overall speaking proficiency score.

### Variable Selection

A variety of indices were collected from the computational tools Coh-Metrix, CPIDR, and LIWC. We selected these tools because they each report on language features that are theoretically important for speaking proficiency and the indices they report are dissimilar enough to warrant inclusion. In addition, the tools are easy to access. For instance, Coh-Metrix is freely available for use on-line (http://www.cohmetrix.memphis.edu/) and CPIDR is available for free download (http://www.ai.uga.edu/caspr/). LIWC is available for a small fee at http://www.liwc.net/. We selected a sub-set of indices from each tool based on theoretical correlates to the TOEFL scoring subsections of the speaking rubric (i.e., delivery, language use, and topic development). That is to say, we selected indices a-priori that should not only correlate highly with human judgments of quality, but should also provide broad coverage of concepts associated with communicative competence (cf. Zechner et al., 2009). These indices and their links to the speaking rubric are discussed briefly below and summarized in Table 4. For a full description of Coh-Metrix refer to D. McNamara and Graesser (2012) and Graesser et al. (2004). For a full description of CPIDR see Brown et al. (2008) and for LIWC consult Pennebaker et al. (2001).

#### *Delivery*

Delivery is described in the TOEFL-iBT speaking rubric as the well-paced flow or fluidity of expression. For high-scored samples, the speech is generally clear, but may contain lapses in pronunciation or intonation patterns. Low-scored samples have consistent difficulties in pronunciation, stress, and intonation and delivery is choppy, fragmented, or telegraphic. The speech is also characterized by frequent pauses or hesitations. To assess the delivery of the participants, we used CPIDR to count the number of ideas and the number of words in the sample. The number of ideas and words relates to the flow of the speech in that more words and ideas would indicate fewer pauses and hesitations and thus more fluid speech. We do not consider other aspects of delivery (i.e., phonological and prosodic properties).

**Ideas.** CPIDR measures the number of ideas in text by counting parts-of-speech and, using a set of readjustment rules, the number of ideas. The basic premise behind CPIDR is that every verb, adjective, adverb, conjunction, and preposition roughly equates to an idea. The readjustment rules in CPIDR condense complicated verb phrases and other phrases into single idea units. CPIDR reports two idea

Table 4. *Selected Variables for each TOEFL Subsection*

| Delivery subsection | | Language use subsection | | Topic development subsection | |
|---|---|---|---|---|---|
| Feature | Tool | Feature | Tool | Feature | Tool |
| Number of ideas | CPIDR | Tense | LIWC | Lexical overlap | Coh-Metrix |
| Idea density | CPIDR | Modals | Coh-Metrix | LSA Givenness | Coh-Metrix |
| Number of words | CPIDR | Word type count | Coh-Metrix | MED | Coh-Metrix |
| | | Word frequency | Coh-Metrix | Causality | Coh-Metrix |
| | | Lexical diversity | Coh-Metrix | Connectives | Coh-Metrix |
| | | Polysemy | Coh-Metrix | Logical operators | Coh-Metrix |
| | | Hypernymy | Coh-Metrix | Key words | Coh-Metrix |
| | | Meaningfulness | Coh-Metrix | Topic adherence | Coh-Metrix |
| | | Familiarity | Coh-Metrix | | |
| | | Concreteness | Coh-Metrix | | |
| | | Imageability | Coh-Metrix | | |

indices: the number of ideas and the idea density (calculated by dividing the number of ideas by the number of words). CPIDR also has a speech mode that rejects repetitions common in natural speech samples.

**Number of Words.** CPIDR also calculates the number of words in a text. Examining the number of words produced by a learner is a common approach for measuring L2 fluency (Norris & Ortega, 2009; Reid, 1990).

***Language Use***

Language use, as described by the TOEFL-iBT speaking rubric, comprises both grammar and vocabulary. High-scored speaking samples demonstrate the effective and generally automatic use of grammar and vocabulary. Low-scored speaking samples demonstrate a severely limited range of grammar and vocabulary that restricts or prevents the expression of ideas. To assess the grammar use of the participants, we used LIWC to compute the use of past, present, and future time and Coh-Metrix to calculate the use of modals of possibility. To assess participants' vocabulary use, we collected values for indices related to breadth of lexical knowledge (e.g., type counts, lexical diversity, and word frequency), depth of lexical knowledge (e.g., word hypernymy, word polysemy, and word meaningfulness) and core lexical items (e.g., word familiarity, imageability, and concreteness) from Coh-Metrix. All of the selected indices are discussed below.

**Past, Present, and Future Time**. LIWC computes tense and aspect in word use using counts for highly frequent words in the past (e.g., *ate*, *did*, *was*), present (e.g., *eat*, *do*, *is*), and the future (e.g., *might*, *should*, *will*).

**Modals of Possibility.** Coh-Metrix reports a count for the number of modals of possibility (e.g., *may*, *might*, *could*) and obligation (e.g., *must*, *need to*, *have to*) found in a text. The modal count does not include future particles (e.g., *will* and *won't*).

**Breadth of Knowledge Indices**. Breadth of knowledge indices relate to the number of words a learner knows and are important indicators of vocabulary use (Crossley, Salsbury, McNamara, & Jarvis, 2010, 2011). Coh-Metrix measures the number of word types produced, word frequency, and lexical diversity. A word type count measures the number of unique words produced. The primary frequency counts in

Coh-Metrix come from CELEX (Baayen, Piepenbrock, & Gulikers, 1995), the database from the Centre for Lexical Information, which consists of frequencies taken from the early 1991 version of the COBUILD corpus, a 17.9 million-word corpus. Coh-Metrix reports frequency indices for both the CELEX spoken and written subcorpora as well the combined corpus. The indices reported are for raw values, proportion values, or logarithmic values. Coh-Metrix measures lexical diversity using a variety of indices that demonstrate small text length effects (McCarthy & Jarvis, 2010). These indices include the Measure of Textual, Lexical Diversity (MTLD; McCarthy & Jarvis, 2010), *D* (Malvern, Richards, Chipere, & Duran, 2004), and *M* (Maas, 1972).[1]

**Depth of Knowledge Indices**. Depth of knowledge indices relate to how well a learner knows a word. These indices are also important indicators of word knowledge (Crossley, Salsbury, & D. McNamara, 2009, 2010). Coh-Metrix calculates depth of knowledge using WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) and the MRC Psycholinguistic Database (Coltheart, 1981). Using WordNet, Coh-Metrix measures word polysemy (the number of senses words have) and word hypernymy (the depth of a word in a conceptual, taxonomic hierarchy).[2] These indices also related to word ambiguity and specificity respectively. Using the MRC Psycholinguistic Database, Coh-Metrix reports on word meaningfulness (i.e., the number of associations a word has according to human raters; Gilhooly & Logie, 1980; Paivio, 1965; Toglia & Battig, 1978).

**Core Lexical Items**. Core lexical items are more likely to be basic category words (Brown, 1958; Murphy, 2004). Basic category words are words that are generally learned first and are characteristic of emerging lexicons. One method to measure core lexical items is through measuring word familiarity, concreteness, and imageability scores as found in the MRC Psycholinguistic database. These scores measure lexical constructs such as spoken word exposure (familiarity), word abstractness (concreteness), and the evocation of mental and sensory images (imageability).

### *Topic Development*

Another key element of speaking proficiency as defined by the TOEFL rubric is topic development. Highly rated speech samples have well developed responses that are coherent and relationships between ideas that are clear. Low-rated speech samples lack substance beyond expressing basic ideas and potentially contain heavy repetition. To investigate topic development, we used Coh-Metrix to assess the cohesive devices in the sample, key word use, and prompt adherence.

**Lexical and Semantic Coreferentiality.** Coh-Metrix reports on a variety of indices of cohesion. These include four forms of lexical co-reference between sentences: noun overlap between sentences, argument overlap between sentences, stem overlap between sentences, and content word overlap between sentences. Coh-Metrix measures semantic coreferentiality and given/new information using Latent Semantic Analysis (LSA; Hempelmann, et al., 2005), which is a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts (Landauer, D. McNamara, Dennis, & Kintsch, 2007).

**Structural Cohesion.** Coh-Metrix computes the Minimal Edit Distance (MED) for a sample by measuring differences in the sentential positioning of content words. A high MED value indicates that content words are located in different places within sentences across the text suggesting lower structural cohesion.

**Causality.** LIWC calculates causality using key word counts for words associated with causality, which is important for developing situational cohesion (Zwaan, Langston, & Graesser, 1995). These include words such as *cause*, *consequence*, *how*, *produce*, *purpose*, *therefore*, *thus*, and *since*. Coh-Metrix measures causality by calculating the number of causal verbs, the number of causal particles, and the ratio of causal particles to causal verb (Dufty, Hempelmann et al., 2005). The causal verb count in Coh-Metrix is calculated using the number of main causal verbs (e.g., *kill*, *throw*, and *pour*) identified through

WordNet (Fellbaum, 1998; Miller et al., 1990). The causal particle count is calculated using a pre-defined set of causal particles such as *because, consequence of,* and *as a result.*

**Connectives and Logical Operators.** Coh-Metrix also calculates the density of connectives and logical operators. The first dimension of connectives contrasts positive versus negative connectives, whereas the second dimension is associated with particular classes of cohesion as identified by Halliday and Hasan (1976) and Louwerse (2001) such as positive additive (*also*, *moreover*), negative additive (*however*, *but*), positive temporal (*after*, *before*), negative temporal (*until*), and causal (*because*, *so*) measures. The logical operators measured in Coh-Metrix include variants of *or*, *and*, *not*, and *if-then* combinations.

**Key Words.** Coh-Metrix reports two indices of key word use that can be used to measure the topic development in text. Key word lists are first extracted from a corpus of samples for a specific prompt using an algorithm that calculates the number of essays in the sample corpus in which the word appears and the frequency distribution of the word. Using this algorithm, the number and proportion of key words that occur in each sample are calculated.

**Topic Adherence**. Coh-Metrix calculates the relevance of the ideas contained in the text as compared to the topic by computing the semantic similarity between the prompt and the participant responses using LSA. Reponses that more specifically address the prompt will report higher LSA values.

### Statistical Analysis

We conducted two analyses. The first analysis assessed links between our selected indices, the transcribed speech samples, and human scores for the independent speech samples. The second analysis assessed links between our selected indices, the transcribed speech samples, and the combined human scores for all six tasks (i.e., the overall speaking proficiency resulting from the independent and integrated task scores). We wanted to test our results on an independent corpus, so we divided the speaking samples into a training set ($n$ = 163) and a test set ($n$ = 81). The purpose of the samples in the training set was to identify which of the variables taken from CPIDR, LIWC, and Coh-Metrix best correlated with the human scores assigned to each speech sample. These variables were later used to predict the human scores for the samples in the training set using a linear regression model. We selected a multiple regression analysis for its transparency and flexibility and because it provides a reliable weighting system. After the initial regression analysis on the speech samples in the training set, the speech samples in the test set were analyzed using the models reported in the training set to calculate the predictability of the variables in an independent corpus (Whitten & Frank, 2005).

In order to allow for a more reliable interpretation of the multiple regressions, we ensured that there were at least 20 times more cases (speech samples) than variables (the automated indices) in our final analysis. We used Pearson correlations to select the variables for the multiple regressions, selecting only those variables that demonstrated significant correlations with the human ratings while not demonstrating multicollinearity with other variables.

## RESULTS

### Independent Speaking Proficiency Scores

#### *Pearson Correlations Training Set*

We selected the indices that (a) demonstrated the highest Pearson correlation when compared to the human ratings for the independent speech samples and (b) did not demonstrate multicollinearity with other indices. The 14 selected variables and their measures along with their $r$ values and $p$ values are presented in Table 5 sorted by the strength of the correlation.

Table 5. *Variables Selected for Independent Score Regression Analysis Based on Pearson Correlation Strength*

| Variable | Measure | *r* value | *p* value |
|---|---|---|---|
| Word type count | Vocabulary breadth | .75 | < .001 |
| D | Lexical diversity | .44 | < .001 |
| Key type proportion | Topic development | -.37 | < .001 |
| Word meaningfulness all word | MRC Database | -.36 | < .001 |
| Key type count | Topic development | .34 | < .001 |
| Word familiarity content words | MRC Database | -.32 | < .001 |
| Present tense | Grammar | -.27 | < .001 |
| Word imageability content words | MRC Database | -.26 | < .001 |
| Incidence of causal verbs | Casuality | -.26 | < .001 |
| Past tense | Grammar | .24 | < .010 |
| CELEX content word frequency | Frequency | -.23 | < .010 |
| Ratio of causal particles to causal verbs | Causality | .17 | < .050 |
| Minimal edit distance content words | Cohesion | .16 | < .050 |
| Word polysemy | Vocabulary depth | -.16 | < .050 |

*Note.* ($n$ = 163)

### *Multiple Regression Training Set*

A stepwise linear regression analysis was conducted for the 14 variables. These 14 variables were regressed onto the raters' evaluations for the 163 transcribed speech samples in the training set. These variables were checked for multicollinearity using both variance inflation factors (VIF) values and tolerance. All VIF values and tolerance levels were at about 1, indicating that the model data did not suffer from multicollinearity (Field, 2005).

The linear regression using the 14 variables yielded a significant model, $F(3, 159) = 73.471$, $p < .001$, $r = .762$, $r^2 = .581$. Three variables were significant predictors in the regression: word type count, ratio of casual particles to verbs, and *D* (lexical diversity). The remaining 11 variables were not significant predictors and were left out of the subsequent model. The regression model is presented in Table 6. The

results from the linear regression demonstrate that the combination of the three variables accounts for 58% of the variance in the human evaluations of independent speaking proficiency for the 163 speech samples examined in the training set.

### *Test Set Model*

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how the model would function on an independent data set (the 81 transcribed speech samples held back in the test set). The model produced an estimated value for each speech sample in the test set. We then conducted a Pearson Correlation between the estimated score and the actual score. We used this correlation along with its $r^2$ to demonstrate the strength of the model on an independent data set. The model for the test set yielded $r = .721$, $r^2 = .520$. The results from the test set model demonstrate that the combination of the three variables accounted for 52% of the variance in the evaluation of the 81 speech samples comprising the test set.

Table 6. Linear Regression Analysis to Predict Independent Speaking Scores: Training Set

| Entry | Variable Added | *r* | $r^2$ | ß | *SE* | *B* |
|---|---|---|---|---|---|---|
| Entry 1 | Word type count | .75 | .56 | .070 | .006 | .838 |
| Entry 2 | Ratio of causal particles to causal verbs | .76 | .57 | .117 | .057 | .107 |
| Entry 3 | D | .76 | .58 | -.015 | .007 | -.144 |

*Notes.* Estimated Constant Term is -0.089; *ß* is unstandardized Beta; *SE* is standard error; *B* is standardized Beta

## Overall Speaking Proficiency Scores

### *Pearson Correlations Training Set*

We selected the indices that demonstrated the highest Pearson correlation when compared to the overall human ratings for the speech samples and did not demonstrate multicollinearity with other indices. The 13 selected variables and their measures along with their *r* values and *p* values are presented in Table 7, sorted by the strength of the correlation. In general, the variables selected for this analysis and the independent scores analysis were the same as the indices selected in the independent speaking proficiency score analysis. The only major differences were that *word polysemy* demonstrated a significant correlation with independent ratings but not the overall scores and *incidence of modals* correlated significantly with overall scores, but not independent scores.

Table 7. *Variables Selected for Overall Score Regression Analysis Based on Pearson Correlation Strength*

| Variable | Measure | *r* | *p* |
|---|---|---|---|
| Word type count | Vocabulary breadth | .82 | < .001 |
| D | Lexical diversity | .52 | < .001 |
| Key type proportion | Topic development | -.43 | < .001 |
| Word meaningfulness all word | MRC Database | -.36 | < .001 |
| Word familiarity | MRC Database | -.34 | < .001 |
| Key type count | Topic development | .31 | < .001 |
| Word imageabilty content words | MRC Database | -.28 | < .001 |
| Present tense | Grammar | -.27 | < .001 |
| Number of causal verbs and particles | Causality | -.25 | < .001 |
| CELEX content word frequency | Frequency | -.24 | < .010 |
| Past tense | Grammar | .23 | < .010 |
| Minimal edit distance content words | Cohesion | .22 | < .010 |
| Modals | Rhetoric | -.16 | < .050 |

*Notes.* (*n* = 163)

### *Multiple Regression Training Set*

As with the independent scores analysis, a linear regression analysis was conducted using the 13 variables. These variables were regressed onto the raters' evaluations of overall speaking proficiency for the participant samples in the training set. These variables were checked for multicollinearity using both variance inflation factors (VIF) values and tolerance. All VIF values and tolerance levels were at about 1,

indicating that the model data did not suffer from multicollinearity (Field, 2005).

The linear regression using the 13 variables yielded a significant model, $F(2, 160) = 164.251$, $p < .001$, $r = .820$, $r^2 = .672$. Two variables were significant predictors in the regression: word type count and word frequency. The remaining 11 variables were not significant predictors and were left out of the subsequent model. The regression model is presented in Table 8. The results from the linear regression demonstrate that the combination of the two variables in the independent speech samples found in the training set accounts for 67% of the variance in the human evaluations of overall speaking proficiency.

Table 8. *Linear Regression Analysis to Predict Overall Speaking Scores: Training Set*

| Entry | Variable Added | *r* | $r^2$ | ß | *SE* | *B* |
|---|---|---|---|---|---|---|
| Entry 1 | Type count | .82 | .66 | .195 | .011 | .797 |
| Entry 2 | CELEX content word frequency | .82 | .67 | -1.480 | .720 | -.095 |

*Notes:* Estimated Constant Term is 1.232; *ß* is unstandardized Beta; *B* is standardized Beta; *SE* is standard error.

### *Test Set Model*

As in the first analysis, we used the *B* weights and the constant from the training set multiple regression analysis to estimate how the model functioned on an independent data set (the 81 speech samples held back in the test set). The model for the test set yielded $r = .779$, $r^2 = .607$. The results from the test set model demonstrate that the combination of the five variables in the speech samples found in the test set accounted for 61% of the variance in the human evaluations of overall speaking proficiency.

## DISCUSSION

This study has provided evidence for the predictive capacity of automated indices related to language use, and topic development to assign scores to speech samples in a manner similar to the scores assigned by human raters. For instance, indices related to breadth of vocabulary (i.e., word type count and lexical diversity) and cohesion (i.e., causality) predicted 52% of the variance in human ratings for the independent speech samples in the test set. For human judgments of overall speaking proficiency (including scores taken from both the independent and the integrated speaking sample), automated indices predicted 61% of the variance of the human scores in the test set using indices related to breadth of vocabulary (word type count and word frequency). These two analyses demonstrate that even in the absence of indices related to delivery (e.g., flow, phonological accuracy, intonation, and stress), automated indices related to breadth of vocabulary knowledge and cohesion can predict a significant amount of the variance in human ratings of speaking proficiency. These findings have important implications for understanding the construct of speaking proficiency (and, by proxy, communicative competence) and for the development of automatic scoring techniques. Below, we discuss the features that informed our models of speaking proficiency and discuss our model's success as compared to that of previous models. We also discuss how the findings from this study inform our understanding of speaking proficiency, and how the findings can be used to enhance automatic scoring techniques.

### Word Type Counts

By far, the strongest predictor for both independent scores and overall scores of speaking proficiency was the number of word types produced by the speaker. For the independent score analysis (training set), the number of word types explained almost 56% of the variance in the human ratings of proficiency. For the overall score analysis (training set), the number of word types explained 66% of the variance in the human ratings of proficiency. Such trends indicate that speakers who produce a greater number of unique words are judged to have greater speaking proficiency. The word type count indicates that more proficient

speakers can access and retrieve a greater number of words than low proficient speakers, thus indicated a larger vocabulary. A post-hoc analysis also demonstrates that the word type count index strongly correlates with the number of words in the text ($r$ = .856) indicating that the index may also tap into issues of delivery (i.e., the flow of speech) as well as depth of vocabulary.

**Casuality**

Causality was also a significant predictor of speaking proficiency with speech samples that contained a higher ratio of causal particles to verbs being judged as more proficient (explaining 1% of the variance in the human scores). This finding demonstrates that it is not merely the number of causal verbs used in a sample, but more importantly, the use of causal particles such as *because* and *as a result*. A greater number of these particles in relation to the overall number of causal verbs should develop greater causal relations between ideas.

**Lexical Diversity**

Another significant predictor of speaking proficiency is the level of lexical diversity used by the speaker. For the independent scores (training set), samples that contained greater lexical diversity were scored more highly (explaining 1% of the variance in the human scores). Thus, it is not only the number of unique types that the speaker produces, but the number of these types in relation to the overall number of tokens found in the sample.

**Word Frequency**

Word frequency was also a significant predictor of speaking proficiency. In relation to overall judgments of speaking proficiency (as in human judgments of writing ability, McNamara et al., 2010; Crossley & McNamara, 2012), samples that contained more infrequent words were scored higher (explaining 1% of the variance in the human scores). The ability to produce more infrequent words likely relates to greater vocabulary size.

**Additional Indices**

A variety of indices did correlate significantly with human judgments, but were not kept in the regression models for either the independent or overall scores of speaking proficiency. These indices, while not predictive in the final models, do provide supplementary evidence as to additional features of the speech samples that may influence human judgments of quality. Of particular interest are those indices that correlated with both independent and overall scores and did not strongly overlap conceptually with those indices reported by the models. For instance, key type indices were strongly correlated with human scores of both independent and combined speaking proficiency. These indices indicated that the proportion of key types was negatively correlated with speaking proficiency, but that key type counts were positively correlated. This likely indicates that speakers that focused on summarizing the prompt were judged to be less proficient (i.e., they repeated a greater proportion of the words in the prompt) than those speakers that focused on smaller elements of the prompt (i.e., the overall count of words repeated from the prompt was high, but the proportion was low). Speaking samples that were score highly also contained more lexical sophistication in relation to the word meaningfulness, familiarity, and imageability. The correlations demonstrate that for both independent and overall scores, samples that contained less meaningful, familiar, and imageable words were scored higher. Thus, test takers that produced more lexically sophisticated words were judged to have greater speaking proficiency. In addition, significant correlations between human scores and past and present tense indices indicate that speakers that use more verbs in the past tense than in the present tense are judged to be more proficient speakers. Lastly, multiple edit distance indices demonstrated significant correlations with both independent and overall speaking proficiency such that speech samples that were scored higher had greater multiple edit distance suggesting lower structural cohesion. Alternatively, this index may be tapping into the presence of structural diversity (in a similar way that lexical diversity indices can tap into both lexical cohesion and lexical

sophistication; McCarthy & Jarvis, 2010). From this perspective, the positive correlations make more intuitive sense.

Of interest, also, are those selected indices that have theoretical overlap with the construct of speaking proficiency but did not demonstrate significant correlations with human judgments. Lexically, indices that measure word hypernymy did not significantly correlate with human judgments indicating that word specificity may not be important elements of speaking proficiency in the selected samples. More surprisingly, most cohesion indices did not correlate with judgments of speaking proficiency (with the exception of causal particles, discussed above, and the minimal edit distance indices reported in Tables 5 and 7). Cohesive devices (e.g., connectives, logical operators, lexical and semantic overlap) are important textual elements that connect text segments and help develop textual coherence (D. McNamara, Kintsch, Songer, & Kintsch, 1996). According to the TOEFL scoring rubric, text cohesion is an important element of speaking proficiency and is found under the subscale of topic development. According to the topic development subscale, high proficiency samples are "well developed and coherent" and "the relationships between ideas are clear." However, coherence, at least as a factor of human judgments, appears to be unrelated to the use of cohesive devices. Such a finding is likely the result of the background knowledge of the expert raters, who are capable of generating appropriate inferences to bridge conceptual gaps in texts. Such raters generally benefit from texts low in cohesion because the texts induce them to generate inferences and build coherent models of text. In contrast, texts with explicit cohesion likely produce less coherent mental representation on the part of expert raters (Crossley & D. McNamara, 2010, 2011; D. McNamara, 2001).

**Comparison to Past Studies**

One problem with comparing our models to previous studies is that the majority of studies focusing on speaking proficiency analyzed differences in text features by proficiency level and generally depended on primary trait scores (Adams, 1980; Bejar, 1985; Iwashita et al., 2008; T. McNamara, 1990) or had narrow speaking domains (e.g., answering questions or reading a passage: Balogh, Bernstein, Cheng, & Townshend, 2007; Bernstein, 1999; Bernstein, DeJong, Pisoni, & Townshend, 2000). With the exception of Higgins et al. (2011), most previous studies did not model human ratings of speaking proficiency or use automated indices (Iwashita et al. did report on some automated lexical features). Like many of the earlier studies that examined text differences as a function of level (i.e., Adams, 1980; Iwashita et al, 2008; T. McNamara, 1990), many of our predictors were lexical in nature (type count, lexical diversity, word frequency). However, none of our grammatical indices were significant predictors in our final model though some did demonstrate significant correlations.

In comparison to the models reported by Higgins et al. (2011), we find that our models predict a greater, but not significant amount of the variance in the human ratings for both two item and six item scores. For instance, Higgins et al.'s model for two item scores predicted 47% of the variance in the human scores as compared to the 52% reported by our model. For the total scores of the six items, Higgins et al.'s model predicted 53% of the variance in the human scores as compared to the 61% reported by our model. Of interest in light of the findings of our current study is that Higgins et al. did not report significant correlations between their word type count index and human judgments of speaking proficiency. Two word type count indices did demonstrate significant correlations (*unique words normalized by total word duration* and *unique words normalized by speech duration*) of which one was included in their final model (*unique words normalized by speech duration*). However, even this index demonstrated much lower correlations with human scores ($r$ = .408) when compared to the word type count index used in the current study ($r$ = .746 for independent scores and $r$ = .815 for combined scores).

Of course, direct comparisons between the two approaches are problematic because Higgins et al.'s models were built using automatically transcribed data, where the word error rates reported by the ASR were around 50%, whereas our approach assumed the existence of an accurate speech recognition system.

Comparisons are also difficult because our models did not include indices related to pronunciation. In addition, Higgins et al. did not report on the variance explained by each of their indices in their regression model and instead only reported on the variance of the combined indices, making comparisons difficult. Lastly, many of the indices used by Higgins et al. were undefined, making direct comparisons between the computed indices problematic.

**Speaking Proficiency**

Standardized guidelines (e.g., ACTFL and TOEFL) generally equate speaking proficiency to the ability to communicate accurately and fluently on a variety of topics. Elements of this accuracy and fluency include syntactic complexity, lexical sophistication, text coherence, topical knowledge, accurate pronunciation and use of prosodic features, and use of appropriate discourse strategies. Speaking proficiency can also be described as a sub-skill of communicative competence (Iwashita et al., 2008), which identifies language organization (i.e., grammatical and textual skills) and pragmatic competence (i.e., illocutionary and sociolinguistic skills) as primary elements of communicative success (Bachman, 1990; Canale & Swain, 1980).

Our analysis focuses solely on language organization and demonstrates that the number of unique words produced by the test taker and the lexical diversity of the sample (i.e., breadth of knowledge indices), casual cohesion, and the sophistication of the words (i.e., word frequency) used in the sample are primary predictors of speaking proficiency. An important component of our study is that the tested features adhere to the relations hypothesized for the construct they represent (i.e., speaking proficiency or communicative competence; Shin, 2005). Given this, we have confidence that our models have not only predictive validity, but also face validity. The results of our study demonstrate that, to some degree, speaking proficiency is related to the size of a speaker's vocabulary and the speaker's ability to make links between ideas.

**Automated Scoring Techniques (AST)**

From a practical perspective, the results of this study promote the possibility of more accurately assessing speaking proficiency automatically contingent upon the development of more reliable automatic speech recognition tools. Not only that, the index that is most predictive in this study is an index based on a simple type count. Such an index is not computationally expensive, appears reliable in both training and test sets and across human scores (i.e., for both independent and overall scores), and likely taps into not only vocabulary size, but also number of words produced. An AST based on such a simple technique and combined with other relatively simple and computationally light indices such as lexical diversity and word frequency would prove beneficial in standardized tests (such as the TOEFL), for classroom assessments, for program and administrative assessment (such as those needed in intensive English programs), and to provide direct feedback to students (as in systems such as TPO).

Unlike human raters, ASTs can score items quickly, reliably, cost-effectively, and in a manner which can be linked to the construct of interest (in this case speaking proficiency). Unlike ASTs, human raters also may not be completely objective in their ratings and are candidates for fatigue and ordering effects (Hoyt, 2000; Murphy & Anhalt, 1992). Perhaps the greatest advantage of ASTs would be in cost and time. Human raters require training, time to score, and monitoring, all of which are costly and time consuming (Higgins et al., 2011). Reductions in cost and time would allow most second language learners to receive feedback on their speaking proficiency regardless of location (i.e., in the absence of a native speaking or advance speaker population) and promote the assessments of language skills that are traditionally more difficult to evaluation. That is to say, ASTs for speaking proficiency would allow the quick and effective evaluation of speech along with more traditional assessments of grammar, vocabulary, and writing skills.

## CONCLUSION

We used a variety of automated tools to assess the linguistic features of transcribed speech samples in order to predict human ratings of speaking proficiency. The linguistic features measured were linguistic items that strongly overlap with features theorized to represent the construct of speaking proficiency (Xi et al., 2008). Thus, our confidence in our findings and their interpretation rests not only on the predictive ability of our models, but also on the face validity of our selected indices and the content validity of our speech samples and the assigned human ratings.

Our analysis presupposed the existence of a reliable speech recognition system (ASR), making immediate application of our models problematic. However, the theoretical value of the models and how they help interpret and better understand human judgments of speaking proficiency remain. We predict that future developments of ASR systems will afford greater automation of the models presented in this study as well as provide the means to develop needed automated indices related to the phonological and prosodic properties of speech samples (through the increased accuracy of such AST systems) in addition to indices of accentedness (Munro & Derwing, 2001; Munro, Derwing, & Sato, 2006) and speaking rate (Munro & Derwing, 1998), both of which may influence expert raters. Between that unrealized future and the present, additional studies could improve upon the current methods and approaches. For instance, knowing the word length constraints for most lexical resources, longer speech samples along with their respective human assessments should be collected and analyzed. Longer speech samples would not only allow for more reliable automatic assessments, but would also afford opportunities to control for potential prompt-based differences. Future studies should also consider the use of a variety of speaking proficiency rubrics other than the TOEFL rubric alone. The TOEFL rubric, while validated in multiple studies, lacks the depth of other speaking rubrics (e.g., the ACTFL rubric) and was developed with academic registers in mind. Thus, it may not completely represent speaking proficiency. Lastly, the development of automated indices that measure grammatical and syntactic complexity, incidences of errors, and coherence as a property of expert raters (as compared to cohesion indices that measure textual properties) would likely advance the predictive ability of our automated models and our understanding of how expert raters assign scores of speaking proficiency. Such knowledge would promote a better understanding of L2 communicative competence and second language acquisition.

## NOTES

1. Maas, as reported by Coh-Metrix, is reverse-scaled so that lower numbers indicated greater lexical diversity.

2. A lower hypernymy score relates to less specific words, while a higher score relates to more specific words.

## ACKNOWLEDGMENTS

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would like to also thank the anonymous reviewers and the editorial staff at *Language Learning & Technology*. In addition, the authors are indebted to Michael Laspina for his assistance in transcribing the speaking samples used in this paper.

## ABOUT THE AUTHORS

Scott Crossley is an Assistant Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, and second language acquisition. He has published articles in second language lexical acquisition, multi-dimensional analysis, discourse processing, speech act classification, cognitive science, and text linguistics.

Danielle McNamara is a Professor at Arizona State University. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.

## REFERENCES

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Firth (ed.), *Measuring spoken proficiency* (pp. 1–6). Washington, DC: Georgetown University Press.

American Council on the Teaching of Foreign Languages. (1999). *ACTFL proficiency guidelines–speaking: Revised 1999.* Hastings-on-Hudson, NY: ACTFL Materials Center.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (Eds.) (1995). *The CELEX Lexical Database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium.

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

Balogh, J., Bernstein, J., Cheng, J., & Townshend, B. (2007). Automated evaluation of reading accuracy: assessing machine scores. *Proceedings of The International Speech Communication Association Special Interest Group on Speech and Language Technology in Education* (SLaTE), Farmington, PA.

Bejar, I. I. (1985). *A preliminary study of raters for the test of spoken English* (ETS RR-85-5). Princeton, NJ: Educational Testing Service.

Bernstein, J. (1999). *PhonePass Testing: Structure and Construct.* Menlo Park, CA: Ordinate Corporation.

Bernstein, J., DeJong, J., Pisoni, D., & Townshend, B. (2000). Two experiments in automated scoring of spoken language proficiency. *Proceedings of InSTILL (Integrating Speech Technology in Language Learning*, Dundee, Scotland.

Bernstein, J., van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27* (3), 355–377. DOI: 10.1177/0265532210364404

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods, 40*(2), 540–545.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65*, 14–21.

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: a prototype automated scoring system for GMAT analytical writing assessment* (ETS RR-98-15). Princeton, NJ: Educational Testing Service.

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

Cobb, T. (2002). The Web Vocabulary Profiler. Retrievable http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33*, 497–505.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the 32nd annual conference of the Cognitive Science Society.*

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading, 35*(2), 115–135.

Crossley, S. A, Salsbury, T., & McNamara, D. S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning. 59*(2), 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*(3), 573–605.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2010). Predicting lexical proficiency in language learners using computational indices. *Language Testing, 28*(4), 561–580.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182–193.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107–145.

Dufty, D., Hempelmann, C., Graesser, A., Cai, C., & McNamara, D.S. (2005). An algorithm for detecting causal and intentional information in text. *Presentation at the 15th Annual Meeting of the Society for Text and Discourse*, Amsterdam, Netherlands.

Educational Testing Services (2004). Independent Speaking Scoring Rubrics. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Field, A. (2005). *Discovering statistics using SPSS*. London, UK: Sage Publications.

Gilhooly K. J., & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, 12, 395–427.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193–202.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London, UK: Longman.

Hempelmann, C.F., Dufty, D., McCarthy, P.M., Graesser, A.C., Cai, Z., & McNamara, D.S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 941–946). Mahwah, NJ: Erlbaum.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*(2), 282–306. DOI: 10.1016/j.csl.2010.06.001

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods 5*, 64–86.

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

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Leacock, C., & Chodorow, M. (2003). C-rater: scoring of short-answer questions. *Computers and the Humanities 37*(4), 389–405.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12*, 291–315.

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik, 8,* 73–79.

Malvern, D. D. Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, UK: Palgrave Macmillan. DOI: 10.1057/9780230511804

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*, 381–392. DOI: 10.3758/BRM.42.2.381

McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51–62.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27*(1), 57–86.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P.M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1–43.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52–75.

Miller, G. A, Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Five papers on WordNet. *Cognitive Science Laboratory*, Princeton University, No. 43.

Munro, M., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*, 159–182.

Munro, M., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition, 23,* 451–468.

Munro, M. J., Derwing, T. M., & Sato, K. (2006). Salient accents, covert attitudes: Consciousness-raising for pre-service second language teachers. *Prospect 21*(1), 67–79.

Murphy, G. L. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, K. R., Anhalt, R. L. (1992). Is halo error a property of the raters, ratees, or the specific behaviors observed? *Journal of Applied Psychology, 72*, 494–500.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 4*, 32–38.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC): LIWC2001*. Mahwah, NJ: Lawrence Erlbaum Associates.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–210). Cambridge, UK: Cambridge University Press.

Toglia, M.P., & Battig, W.R. (1978). *Handbook of semantic word norms*. New York, NY: Erlbaum.

Shin, S. K.. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, 22(1) 31–57 DOI: 10.1191/0265532205lt296oa

Whitten, I. A. & Frank, E. (2005). *Data Mining*. San Francisco, CA: Elsevier. DOI: 10.1093/bioinformatics/bth261

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated Scoring of Spontaneous Speech Using SpeechRater (SM) v1.0*. Educational Testing Service, Research Report RR-08-62, Princeton, NJ.

Xi, X. (2007). Evaluating analytic scoring the TOEFL Academic Speaking Test (TAST) for operational use. *Language Testing, 24*(2) 251–286. DOI: 10.1177/0265532207076365

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51*(10), 883–895. DOI: 10.1016/j.specom.2009.04.009

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*, 292–297.