Speaking Proficiency and Text Integration: The Impact of Linguistic Differences and Strategy Use

Chloe Ramirez[1], Daniel Lee[2], Emily Robinson[3]

[1] Department of Public Health, University of California, Los Angeles, USA

[2] School of Social Work, University of Sydney, Australia

[3] Department of Psychology, University of Chicago, USA

Introduction

In academic contexts, text integration skills (i.e., integrating material from reading or listening input into speaking or writing tasks) are presumed to be critical elements of academic success for second language (L2) learners of English. This basic notion is premised on the idea that academic settings require students to both read academic texts and listen to academic lectures while integrating information from both sources into oral and written reports as well as class discussions (Douglas, 1997). Integrated writing and speaking tasks that combine these skills best represent the demands placed on students in academic contexts, and such tasks have become common in a number of standardized testing situations designed to measure students' readiness for academic contexts (Cumming, Grant, Mulcahy-Ernt, & Powers, 2005; Cumming, Kantor, Baba, Erdosy, Eouanzoui, & James, 2006;).

Students' success at recalling and integrating previous information can be based on diverse learner characteristics (e.g., working memory), strategy use (e.g., note-taking strategies), and on linguistic properties of a text (e.g., word repetition or word frequency). Working memory capacity (WMC) denotes the ability to temporarily store and

manipulate information simultaneously (Baddeley, 2003) and it is an important component of recall that might impact the quality and efficiency of real time language processing (Miyake & Friedman, 1998). Previous studies have also shown that note-taking strategies can positively affect lecture summarization (Carrell, 2007). In terms of the linguistic properties of text, two types of information that affect the efficiency of encoding of discourse and its subsequent recall have been noted in previous research: proposition-specific information and relational information (McDaniel, Einstein, Dunay, & Cobb, 1986). Proposition-specific information refers to lexical items (i.e., words) that are found within a proposition (e.g., a sentence, clause, or idea) and the semantic relationships between these words. Relational information pertains to organizational elements with a text and how propositions are embedded (i.e., text cohesion). Both proposition-specific and relational information are important factors in L2 processing because L2 learners often have difficulty identifying relationships among ideas (i.e., relational information) and detecting key ideas (i.e., proposition specific information; Powers, 1986).

The purpose of the current study is to examine how learner characteristics (e.g., working memory, language proficiency, and gender) and the linguistic properties of listening source texts (e.g., the cohesive and lexical properties of source texts) influence source text integration in standardized language assessment test focused on integrated speaking tasks. Further, we assess associations between learner characteristics and linguistic properties in the source texts with expert ratings of speaking proficiency.

Test Takers' Individual Characteristics

In the current study, we examined a variety of test takers' individual characteristics including proficiency level as measured by the Test of English as a Foreign Language (TOEFL) Institutional Testing Program (ITP), first language, gender, and working memory. Language proficiency has been one of the most widely addressed individual characteristic, and researchers often investigate proficiency as a mediating variable of test performance. For instance, Appel and Wood (2016) reported that high level learners were less dependent on reading sources during integrated writing tasks. Barkaoui found that overall English language proficiency significantly contributed to TOEFL iBT writing scores (2013) and that participants' writing performance was mediated by task types but not proficiency (2015). Lastly, Hill and Liu (2012) reported that that language proficiency interacted with background knowledge in TOEFL iBT reading tasks. Overall, previous L2 assessment research has suggested that learner proficiency along with other variables such as background knowledge and task types may be associated with test takers' language performance.

Gender and age are other individual characteristics of test takers and L2 learners that have been examined. As an example, Breland, Lee, Najarian, and Muraki (2004) examined gender effects on TOEFL CBT writing and found that gender was a significant predictor of writing success, with females tending to obtain higher scores than males. Multiple studies have demonstrated that younger learners develop proficiency in a L2 faster than older learners (DeKeyser 2000; McDonald 2000).

Another individual characteristic of interest in test takers is WMC, which refers to "the ability to maintain information in an active and readily accessible state, while concurrently and selectively processing new information" (Conway, Jarrold, Kane,

Miyake, & Towese, 2007, p. 3).” Over the last two decades, WMC has been increasingly investigated and findings suggest it is an important cognitive factor that affects L2 learning and processing (Wen, Mota, & McNeil, 2015). For instance, Linck, Osthus, Koeth, and Bunting (2013) conducted a meta-analysis that included 79 studies and 3,707 participants that focused on associations between working memory and a range of learning outcomes such as L2 comprehension. The results suggested that working memory is an important component of L2 processing and proficiency outcomes. In contrast, Kormos and Trebits (2011) reported a more limited role for WMC in the oral production of L2 learners such that WMC might only affect L2 syntactic production. Recent studies also do not provide a strong evidence for a strong relationship between WCM and L2 listening comprehension even when using multiple WMC measures (Andringa, Olsthoorn, Beuningen, Schoonen, & Julstijn, 2012; Vandergrift & Baker, 2015). Research has relationships between WMC, L2 performance, and L2 language proficiency level. For instance, Kormos and Sáfár (2008) reported that phonological short-term memory capacity was mediated by proficiency level. Overall, although WMC has been suggested as an important individual characteristic, its role might not be consistent across different L2 tasks that involve different types of processing.

The last individual characteristic we consider is L2 learners’ note-taking strategies. Previous early L2 research suggests an association between students’ note-taking strategies and listening comprehension performance as measured by multiple choice tests (Dunkel, 1988). For instance, Dunkel (1988) reported that total number of words and information units in test-takers notes were significantly associated with test performance. Cushing (1993) reported that test-takers’ academic status and listening

comprehension proficiency positively affected the quality and content of notes. More recently, Carrell (2007) found that note-taking and test performance are moderately related. In sum, previous research suggests that students' note-taking strategies vary and that the quality and quantity of note-taking might be associated with language performance.

Text Properties and Recall

In the current study, the linguistic properties of a text are operationalized in terms of two types of information (i.e., relational information and proposition-specific information). Relational aspects in texts are most commonly related to text cohesion, while proposition-specific information is related to lexical elements. A variety of linguistic features such as connectives, anaphoric references, and word overlap have been used to measure text cohesion (Crossley, Kyle, & McNamara, 2017). These cohesion features provide readers with explicit text markers meant to signal connections between ideas in a text that can help develop a coherent model of the text. However, cohesion is different from text coherence. Coherence refers to the understanding that the reader extracts from the text and, while it can often develop with the help of cohesion features (e.g., connectives and word overlap), it can also develop because of prior knowledge and/or reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996).

While many text features are related to cohesion, connectives such as *and, but,* or *also* are probably the most common cohesive devices reported in linguistic research. Connectives can help create cohesive links between ideas and clauses at the sentence level (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994). These links can help develop greater text organization (van de Kopple, 1985) and thus promote increased text

comprehension. However, there is some indication that connectives are not linked to text coherence, especially for advanced readers (Crossley & McNamara, 2010, 2011). Another common cohesive device that is used to link sentences is lexical overlap (i.e., overlap between words; Halliday & Hasan, 1976). Previous research has shown that lexical overlap can improve text readability and text processing (Crossley, Greenfield, & McNamara, 2008; Rashotte & Torgesen, 1985). However, similar to the use of connectives, lexical overlap at the sentence level has not been shown to be linked to text coherence (Crossley & McNamara, 2010, 2011). As compared to links between sentence level text segments (known as local cohesion), global cohesion devices that link larger segments of text together (e.g., at the paragraph level) have shown links with text coherence. These cohesive devices include lexical overlap between paragraphs (Crossley & McNamara, 2011; Crossley, Kyle, & McNamara, 2017; Foltz, 2007) and causal relations among text segments (Graesser, McNamara, Louwerse, & Cai, 2004).

Unlike relational information, proposition specific features refer to lexical elements within propositions and how words may be easier to recall because of their lexical properties. For instance, research has shown that concrete words have advantages in recall and comprehension tasks as compared to abstract words (Gee, Nelson, & Krawczyk, 1999; Paivio, 1991). Other lexical properties that influence recall include word imageability (Paivio, 1968), word polysemy (i.e., the number of senses per word, Davies & Widdowson, 1974), and word associations (Nelson, McEvoy, & Schreiber, 1990). Additionally, word recall can also be influenced by word familiarity and frequency. Word familiarity has demonstrated strong effects on word identification and recall (Paivio, 1991), although it is not as strong of a predictor as word imageability

(Boles, 1983; Paivio & O'Neill, 1970). High frequency words are named more rapidly (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004) and recognized quicker (Kirsner, 1994) than lower frequency words.

Text Integration

To be successful, language users have to integrate four language skills (i.e., speaking, listening, writing, and reading) in real-world contexts. As a result, integrating language skills is an important pedagogical component in the L2 classroom. Teaching learners how to integrate language skills can help students interact more naturally in an authentic environment (Oxford, 2001) by requiring students to receive, transmit, and demonstrate their knowledge as well as organize and regulate that knowledge for communicative purposes (Butler, Eignor, Jones, McNamara, & Suomi, 2000). From a testing perspective, integrating language skills is simplified by asking test-takers to discuss and include key propositions and terms found in listening and/or reading materials in their spoken or written responses. Standardized tests such as the Test of English as a Foreign Language (TOEFL) include integrated tasks because they represent an important authentic academic skill that affords test-takers the opportunity to manipulate and control language data that may not rely on their prior knowledge (Hamp-Lyons & Kroll, 1996; Wallace, 1997). Integrated tasks allow test-takers to produce contextually appropriate language (Hamp-Lyons & Kroll, 1996), identify and extract relevant information from the source text(s), and synthesize and organize this information into their responses (Feak & Dobson, 1996). In short, integrated tasks encourage test-takers to produce more authentic language (Plakans & Gebril, 2012).

To date, studies examining text integration have focused mainly on integrated writing tasks which require test takers to write using source texts. These studies have generally investigated the differences between integrated and independent writing in terms of linguistic features or have examined how linguistic features are predictive of human ratings of integrated writing. For instance, Guo et al. (2013) found integrated essays, as compared to independent essays, focused more on organizational cues, used a more detached style of informational writing, and contained more context-independent lexical items. Cumming et al. (2005, 2006) reported that higher-rated integrated essays generally contained more words, more words per T-unit, and a greater diversity of words.

Few studies have focused on text integration in speaking tasks. Barkaoui, Brooks, Swain and Lapkin (2012) investigated the strategic behaviors test-takers used during integrated speaking tasks. However, they failed to find clear relationships between strategy use and integrated speaking scores. A more recent study by Crossley, Clevinger and Kim (2014) examined the linguistic properties of source material on recall and human ratings of speaking proficiency in a small corpus of TOEFL speaking responses. Their findings demonstrated that the relational and propositional properties of words in the source texts were significant predictors of text integration. Specifically, they found that the average incidence of word occurrence in the source text, the frequency of integrated words in the source text (as measured by an external reference corpus), and the integration of words found in positive connective clauses in the source text predicted whether a word was integrated into a test-taker response or not with over 98% accuracy. They also found that the incidence of integrated words from the source text predicted 51% of score variance in speaking proficiency ratings.

Current Study

The findings reported by Crossley et al. (2014) indicated that linguistic properties in the source texts could strongly influence text integration in test-taker responses. Because the human ratings of integrated speaking proficiency appeared to be influenced by different levels of text integration, Crossley et al. concluded that the relational and proposition-specific elements of a text should be controlled during test development. For instance, if a source text was low in relational and proposition specific elements, it might lead to less information recall which could influence human judgments of quality. However, the study by Crossley et al. (2014) included several limitations. First, the study was a pilot study that focused on a small number of test-taker responses (N = 60). In addition, the study did not take into consideration learner characteristics such as WMC, language proficiency, gender, and age. Furthermore, although integrated TOEFL speaking tasks allow students to take notes, students' note-taking strategies were not examined. To date, the extent to which test takers' individual characteristics mediate such relationships has not been systematically examined.

In the current study, we conduct a partial replication of Crossley et al. (2014) by examining if the relational (i.e., cohesive) and proposition-specific (i.e., lexical) properties of words in source texts found in the integrated speaking section of the TOEFL-iBT are predictive of their integration into a spoken response within a relatively large test-taker population. However, unlike Crossley et al. (2014), we assess whether a number of individual differences (e.g., working memory, gender, age, note-taking strategies, and language proficiency) and the lexical and cohesion properties of integrated words are predictive of speaking response quality while controlling for random factors

such as participants and task. We focused on TOEFL integrated listen/speak responses referencing academic genres as found in the TOEFL-iBT. The listen/speak integrated tasks ask test-takers to first listen to a spoken source text, such as an academic lecture or a conversation in an academic context. The test-taker then provides a spoken response to a question based on the listening prompts, and their answer is recorded for later assessment. These answers generally include relationships between the examples in the source text and also the task topic. Expert raters then score these speech samples using a standardized rubric that assesses delivery, language use, and topic development.

The current study is guided by the following three research questions (RQs):

1. Do the relational and propositional properties of words in source texts predict their rate of integration into spoken responses?
2. Which individual characteristics of test-takers are predictive of human ratings of speaking quality?
3. Can relational and propositional properties in spoken responses along with individual characteristics predict human ratings of speaking proficiency?

Method

Participants

The study included 280 participants who were enrolled in Intensive English Programs (IEP) in the Atlanta, Georgia area at the time of data collection. Participants were recruited from intermediate and advanced English classes to ensure they had appropriate language skills to take the integrated listen/speak section of TOEFL-iBT. The participants spoke a number of different first languages. The first languages most strongly represented in the data were Arabic (22%), Portuguese (22%), Spanish (18%),

and Chinese (10%). In terms of their gender distribution, 47% of the participants were male and 53% were female. The average age of the participants was 24 years. Of the 280 participants, full data was only retrievable for 263 of the participants. Four participants were missing working memory scores because of technical problems. Six participants were missing institutional TOEFL scores because they failed to take the tests. Another six participants were missing speaking scores either because of technical difficulties or because the participants did not complete the question. One participant did not fill out the demographic survey.

Materials

Background survey. A background survey was created to collect the following information: age, gender, the highest educational degree, other foreign language learning experience, time spent in the US, time spent studying English, grade point average (GPA) in the IEP, and previous TOEFL scores. The survey was conducted on-line using Qualtrics.

Working memory tests. In the current study, complex WMC was measured using two different working memory tests which were administered using E-Prime 2.0: an aural running span test and a listening span test. Because the current study used the TOEFL integrated speaking tests, which used listening prompts, the listening span test was developed based on the original reading span test (Daneman & Carpenter, 1980; Kim, Payant, & Pearson, 2015). The listening span test was similar to that used in previous SLA studies (Mackey, Adams, Stafford, & Winke, 2010; Mackey & Sachs, 2011). The test consisted of 72 sentences with the sequences ranging from three to six spans, and the order of each sequence was randomly presented. For each sentence,

participants were asked to judge plausibility (i.e., whether its content is possible in the real world by pressing either “yes” or “no” on the computer keypad). After they answered the plausibility question, they heard a letter (e.g., “P”), and at the end of each span, they were asked to recall all of the letters they heard in the correct order. The listening span test was piloted with 10 native speakers of English and 3 non-native speakers of English in order to verify the accuracy of the expected judgments. We scored the listening span test using a partial-credit scoring rather than all-or-nothing scoring following Conway et al., (2005). One point was given for each correctly recalled letter, and, thus, the possible total score was 72.

In order to provide a working memory test which is not overly dependent on L2 proficiency, we also used an aural running span test (Broadway & Engle, 2010). Broadway and Engle (2010) tested the validity of the running span test, and found that it is predictive of higher order cognition. Since then a growing number of second language studies have used the running span test (e.g., Kim, Payant & Pearson, 2015). In this test, participants heard a series of letters and were asked to recall the last *n* items from lists that are *m* + *n* items long. The number of letters to recall was pre-determined; however, participants were not informed of the total number of letters that they would hear in the series. For instance, participants would see the message “remember the last 4 letters” on the monitor, but they were not informed a priori of the total number of letters to be presented aurally in any given sequence. The span of letters ranged from three to six, and there were six sets letters in each span. In total, participants were asked to recall a total of 108 letter items. Based on Broadway and Engle (2010), participants received one point

for each correctly recalled item in correct serial position. Thus the possible total score of the running span test was 108.

Institutional TOEFL. Participants completed an institutional TOEFL exam, which utilizes retired items from the paper-based TOEFL. The institutional TOEFL includes three sections: Listening comprehension (k=50, 30-40 minutes), Structure and written expression (k=40, 40 minutes), and Reading comprehension (k=50, 50 minutes). The three sections take approximately two hours to complete in total.

TOEFL iBT speaking tasks. Participants also completed two non-operational research versions of the integrated listen/speak TOEFL iBT speaking tasks. Each version consists of two speaking tasks which are based on two types of listening sources: (1) listening to a conversation in an academic context; and (2) listening to a lecture. For each question, students were given 20 seconds to prepare for their response and 60 seconds to respond to the prompt. Participants were allowed to take notes during the tests, but they were not required. The two conversational listening sources included in this study including a discussion between two professors about a student missing class because she was on the swimming team (swimming topic) and a conversation between two students about note-taking in class (note-taking topic). The two lecture sources included a lecture on reciprocity from an anthropology class (reciprocity topic) and a lecture about fungus from a botany class (botany topic).

Procedure

All participants attended two data collection sessions. They completed the institutional TOEFL on Day 1 and then completed the background survey, the two working memory tests, and the two integrated listen/speak tasks from the TOEFL iBT

speaking test (listening to a conversation vs. listening to a lecture) on Day 2. On average, participants spend approximately two hours in the lab on the first day, and one hour and 20 minutes in the lab on the second day. The order of the data collection for the two speaking tasks on day two was counter-balanced and randomly assigned to participants.

Transcription

Each spoken response was transcribed by a trained transcriber. The transcriber ignored filler words (e.g., umm, ahh) but did include other disfluency features such as word repetition and repairs. Periods were inserted at the end of each idea unit. All transcriptions were independently checked for accuracy by a second trained transcriber. The same trained transcriber transferred all the notes written by the test-takers into an electronic format. The vast majority of all notes were lexical in nature (i.e., the notes consisted of words and not symbols or abbreviations).

Note-taking

To assess student note-taking, we calculated the number of word lemmas (i.e., word roots) shared between the source text and the notes taken by each participant. We calculated two different note-taking features for the number of lemma tokens (i.e., all words) and types (i.e., unique words) shared between the notes and the source text.

Human Ratings

Two expert TOEFL raters scored each speaking response. The raters used the TOEFL-iBT integrated speaking task rubric, which provides a holistic score (see http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf). The score is based on a 0-4 scale with a score of 4 representing the highest score. Three criteria formed the basis of ratings: delivery (i.e., pronunciation and prosody), language use (i.e., grammar

and vocabulary), and topic development (i.e., content and coherence). Text integration is not addressed in the rubric but the rubric notes task fulfillment, which requires text integration.

Inter-rater reliability for the human scores reported a Cohen's Kappa of .697 and a Pearson's correlation of r = .714. If the two scores differed by less than two points, the average of the raters' scores was included in the dataset. If the scores between the two raters differed by more than one point, a third rater scored the sample, and the final score was the average of the two closest scores (cf. Bejar, 1985; Carrell, 2007; Sawaki, Stricker, & Oranje, 2008).

Language Feature Variables

A variety of cohesion and syntactic values were calculated to assess if word lemmas were integrated from the source text (i.e., the listening samples) into the test-taker speaking responses. We consider these source internal variables because each word in the source text was assigned a cohesion or syntactic value based on features found in the source texts. These features included the number of repetitions of the word within the source (cohesion), if the word was in the subject or object position in a clause (syntax), or if the word was coordinated in a phrase or a clause (syntax). After source internal values were assigned, they were matched to the words produced by the test-takers in their spoken responses in order to examine features for words that were not integrated (i.e., found in the source text, but not in the test-taker response) and words that were integrated (i.e., found in the source text and the test-taker responses). A different procedure was conducted for lexical features. For lexical features, words in each test-taker's response were separated into .txt files that contained either integrated or non-integrated words.

These files were then run through the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) in order to calculate a number of lexical features (see below for discussion of these features). We considered these features to be response internal.

The source and response internal features were used to predict which words were integrated into spoken responses (i.e., RQ 1). The source and response internal features were also used to predict human ratings of speaking proficiency in conjunction with individual characteristics and topic (RQ 2).

TAALES. TAALES is a computational tool that is freely available, user-friendly, works on most computer operating systems (Linux, Mac, Windows), allows for batch processing of text files, and incorporates over 250 classic and recently developed indices of lexical sophistication. These indices measure word frequency, lexical range, n-gram frequency and proportion, academic words and phrases, word information, lexical and phrasal sophistication, bigram and trigram strength of association, contextual distinctiveness, word neighbor information, lexical decision times, age of exposure, and semantic lexical relations (hypernymy and polysemy). Each of these are discuss briefly below. For more detailed accounts of TAALES please see Kyle & Crossley (2015).

Word frequency indices. TAALES calculates a number of word frequency indices with frequency counts retrieved from the SUBTLexus database (Brysbaert & New, 2009). the British National Corpus (BNC; 2007) and the five genres found in the Corpus of Contemporary American English (COCA; academic, fiction, magazine, news, and spoken texts; Davies, 2010). TAALES calculates scores for all words (AW), content words (CW), and function words (FW).

Range indices. In addition to frequency information, TAALES computes range indices which calculate how many texts within a corpus a word appears (i.e., specificity). Range indices were computed from the spoken (574 texts) and written (3,083 texts) subsets of the BNC, SUBTLEXus (8,388 texts), the five genres found in COCA (190,000 texts in the complete corpus).

N-gram frequency and proportion indices. TAALES calculates bigram and trigram frequencies and proportion scores (i.e., the proportion of n-grams in a text that are common in a reference corpus) from both the written (80 million words) and spoken subcorpora (10 million words) of the BNC and from the five genres represented in COCA (440 million words).

N-gram association measures. TAALES calculates five association measures for each bigram and trigram found in the reference corpora: Mutual Information (MI), Mutual Information Squared (MI^2), t-score, ΔP, and collexeme score (Gries, 2013). MI, MI^2, and t-score are bidirectional measures of association between constituent words in an n-gram. While MI and, to a lesser extent, MI^2 tend to highlight n-grams composed of low-frequency words, t-score tends to favor n-grams composed of high-frequency words. ΔP is a directional association measure and calculates the probability of the second word in a bigram given the occurrence of the first word in it. The collexeme association measure calculates the strength of association between lexemes.

Contextual distinctiveness. TAALES calculates several indices related to contextual distinctiveness approach which measure the diversity of contexts in which a word is encountered (Brysbaert & New, 2009; McDonald & Shillcock, 2001). These indices come from The Edinburgh Associative Thesaurus (EAT) index based on

empirical free association data collected by Kiss, Armstrong, Milroy, & Piper (1973), the University of South Florida (USF) (Nelson, McEvoy, & Schreiber, 1998) stimuli count index based on a written free association task, semantic diversity (SemD) based on a computationally-derived latent semantic analysis (LSA) measure (Hoffman, Ralph & Rogers, 2013), and relative entropy index calculated by McDonald and Shillcock (2001) for 8,000 English lexemes as they occurred in the spoken BNC.

Word recognition norms. TAALES reports on lexical decision (LD) and word naming (WN) behavioral norms obtained from The English Lexicon Project (ELP), a large publicly available psycholinguistic dataset (Balota et al., 2007). The ELP includes LD and WN task response latencies and accuracies collected from 816 native English-speaking subjects. Latencies (i.e., response times) and accuracies were calculated in response to 40,481 real words (and an additional 40,481 nonwords for the LD task).

Word neighborhood information. TAALES reports on the word neighborhood information found in ELP. These indices are based on orthographic, phonographic, and phonological neighborhood information for 40,481 words that report word neighborhood size and frequency indices. All neighborhood frequency values are based on the 131 million-word Hyperspace Analogue to Language (HAL) corpus frequency norms (Lund & Burgess, 1996).

Age of exposure. TAALES reports on age of exposure indices that calculate a comprehensive model of word complexity, Age of Exposure, which replicates the learning curve of lexical concepts based on their associations with other words (Dascalu, McNamara, Crossley, & Trausan-Matu, 2016). Hypothetically, AOE indices model the

word learning process as a function of language experience with language based on a large-scale corpus.

Word information indices. Word information in TAALES originate from the MRC Psycholinguistic Database (Coltheart, 1981), Brysbaert, Warriner, & Kuperman (2013), and Kuperman, Stadthagen-Gonzales, & Brysbaert (2012). Word information scores are computed for word age of acquisition, concreteness, familiarity, imageability, and meaningfulness.

Statistical Analyses

In order to address our three research questions, a number of statistical analyses were conducted. Prior to all analyses, we first checked for multicollinearity between all the linguistic variables in the analysis, which was operationalized as any two variables demonstrating a strong correlation ($r > .700$). We next conducted correlations between the variables and the speaking scores for each task for each participant to ensure that variables entered into the model demonstrated a significant and meaningful linear relation with the dependent variable ($p < .001$, $r > .100$). We selected a cut-off of $p < .001$ to correct for any Type I errors. For research question 1, we first conducted an initial Multivariate Analysis of Variance (MANOVA) to select the linguistic variables that demonstrated the strongest differences between the integrated and unintegrated words. We then entered the significant MANOVA variables that did not demonstrate multicollinearity into a discriminant function analysis (DFA) on the entire set of speaking samples to provide confirmatory evidence for the strength of these variables in classifying the words as integrated or unintegrated. The model reported by this DFA was then used to predict group membership of the speaking samples using leave-one-out-

cross-validation (LOOCV). The LOOCV procedure allows testing of the accuracy of the model on an independent data set. The DFA analysis can provide evidence that source internal variables are predictive of which words test-takers will integrate into their responses.

Our second statistical analysis was to determine if the linguistic features and individual differences (e.g., working memory and institutional TOEFL sub-scores) could be used to predict the human ratings for the individual integrated speaking tasks while accounting for both pooled and individual variance among participants as opposed to one pooled group by including subjects as random effects (i.e., assigning a unique intercept for each participant). We used R (R Core Team, 2015) for our statistical analysis and the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015) to construct linear mixed effects models (LME). We also used the package *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2015) to analyze the LME output and derive p-values for individual fixed effects. Final model selection and interpretation was based on *t* and *p* values for fixed effects, post-hoc comparisons of categorical variables, and visual inspection of residuals distribution. To obtain a measure of effect sizes, we computed correlations between fitted and predicted residual values, resulting in an R^2 value[1]. Prior to running an LME model, we examined correlations between the linguistic features and the individual characteristics and the speaking scores in order to select variables for inclusion in the LMEs that reported at least a small effect size ($r > .100$) and that were not multicollinear

[1] We used R^2_{GLMM} to present the variance explained in our model. Historically, using R^2 in mixed-effects models has been problematic because R^2 algorithms may report decreased or increased R^2 in larger models. R^2_{GLMM} calculates marginal and conditional R^2 that are less susceptible to these problems. Marginal effects are concerned with the variance explained by fixed factors while conditional effects concern the variance explained by both fixed and random factors (Nakagawa & Schielzeth, 2012).

(r > .700). We conducted two stepwise LMEs. The first LME examined the associations of individual characteristics (e.g., working memory, age, and institutional TOEFL scores) and topic on the speaking scores. This model included subjects as random effects. Descriptive statistics for the continuous scaled individual characteristics used in this analysis are reported in Table 1. The second LME model was conducted to examine the associations of these individual characteristics along with topic and linguistic features on speaking scores.

[Insert Table 1 about here]

Results

Classifying Integrated and Unintegrated Words

MANOVA. Prior to conducting the MANOVA, all assumptions for the MANOVA were checked and met. The MANOVA used the integrated and unintegrated words from each text as the independent variables and the linguistic indices as the dependent variables. Seventeen indices were selected from the MANOVA for the DFA based on their effect sizes. Selected indices did not theoretically overlap with each other (see Table 2 for descriptive statistics for these indices). The MANOVA results demonstrated that words integrated into test-takers spoken responses from the source text were more frequent, had lower age of acquisition, had a greater range, had more orthographic and phonological neighbors, had more free associations, were repeated more often in the source text (i.e., the occurrence of word in source text index), occurred more often in the source text in clausal coordinations and as objects of prepositions, had greater age of exposure, had greater character bigram frequency, and were named more quickly

than unintegrated words. Conversely, the words not integrated into test-takers spoken responses from the source text were less meaningful and less concrete.

[Insert Table 2 about here]

Discriminant function analysis. We conducted a stepwise discriminant function analysis (DFA) to confirm that the indices selected in the MANOVA indeed discriminated between integrated and unintegrated words. A DFA generates a discriminant function, which is then used in an algorithm to predict group membership (i.e., whether the words were integrated or unintegrated). For the DFA, we used the 17 indices from MANOVA analysis. The stepwise DFA retained 11 of these indices as significant predictors of whether a word was integrated in the test-takers' response or unintegrated (see Table 2 for details on whether the variable was retained in the DFA) and removed the remaining six variables as non-significant predictors based on their predictive strength.

The results demonstrate that the DFA using these eleven indices correctly allocated 1049 of the 1052 word lists as being integrated or unintegrated, $\chi 2$ (1) n=1052) = 1040.068, $p < .001$, for an accuracy of 99.7% (chance level for this analysis is 50%). The Kappa value for this analysis was .994, which suggests almost perfect agreement between the predicted classification of the word lists and their actual classification. The results from the LOOCV were identical to the initial DFA (see Table 3 for the confusion matrix for this analysis). The results indicate that the 11 variables can predict with almost perfect accuracy if a word is integrated or unintegrated from the source text.

[Insert Table 3 about here]

Predicting Human Ratings of Speaking Proficiency

Pearson correlations. After controlling for multicollinearity, *p* values, and effect sizes, we were left with 31 variables. These variables related to key words, and Institutional TOEFL reading, listening, and structure subscores, cohesion, syntactic, and lexical sophistication scores taken from the integrated words, note-taking, and working memory (see Table 4 for Pearson correlation results). For our baseline model that answered RQ 2, we included all individual characteristics that showed at least a small effect size ($r > .100$) along with topic and gender. In order avoid overfitting the full LME model, which addressed RQ 3, we only selected the linguistic indices that demonstrated at least a medium effect size ($r > .300$) with speaking scores and all individual characteristics that showed at least a small effect size ($r > .100$) along with topic and gender. Thus, we included the five linguistic features that showed the highest correlations in the model along with the three TOEFL subscore variables, one note-taking variable, one working memory variable (listening span score), and two categorical variables (*gender* and *topic*).

[Insert Table 4 about here]

Linear mixed effects models. A baseline stepwise LME model considering participants' individual characteristics and topic revealed significant effects for note taking, TOEFL listening and structure scores, and topic. The model indicated that students who included more word types from the source into their notes scored higher. In addition, students with higher TOEFL listening and structure scored higher as did students who responded to the "note-taking" topic (i.e., a conversation task). The model reported a marginal R^2 of .361 and a conditional R^2 of .719. Table 5 displays the

coefficients, standard error, *t* values, and *p* values for each of the fixed effects. Inspection of residuals suggested the model was not influenced by homoscedasticity.

[Insert Table 5 here]

A full model including the nested baseline model and linguistic features revealed significant effects for two linguistic features, *Number of shared words between response and source* and *Occurrence of shared words (noun in object position*) between response and source, TOEFL listening and structure scores, and topic. Results indicated that students who had a greater number of words integrated from the source into their response received higher speaking scores. However, if the students integrated words from the source texts that were in the object position, they received lower scores. As in the baseline model, students with higher TOEFL listening and structure scored higher. In terms of topic, students who responded to the "note taking" topic scored higher than students who wrote on the fungus and reciprocity topic but not the swimming topic (i.e., students scored higher on the conversation tasks than the lecture tasks). Contrasts indicated that student who wrote on swimming topic scored higher than on the fungus and reciprocity topics. The model reported a marginal R^2 of .588 and a conditional R^2 of .754. Table 6 displays the coefficients, standard error, *t* values, and *p* values for each of the fixed effects. A log likelihood comparison found a significant difference between the baseline and full models, ($\chi 2(2) = 193.210$, $p < .001$), suggesting that the inclusion of linguistic features contributed to a significantly better model fit. Inspection of residuals suggested the model was not influenced by homoscedasticity.

[Insert Table 6 here]

Discussion

Integrating content from surrounding language is an important indicator of academic success and, in order to better assess the potential for academic success in test-takers, standardized tests now reflect this reality. An important element of integrating content is the ability to recall information from previously exposed discourse. Recall can be aided by individual characteristics such as working memory or language proficiency, strategy use such as note-taking, or based on the linguistic properties of the preceding discourse. The purpose of this study was to examine if linguistic features in source texts could explain word recall and integration for items administered in the listen/speak section of the TOEFL-iBT and to what extent individual characteristics such as working memory and proficiency level and/or linguistic features could predict human judgments of speaking proficiency.

The results provide evidence that words integrated into spoken responses from the source text had word properties that would afford their recall. Twelve linguistic indices related to lexical items (i.e., propositional-specific information), text cohesion (relational information), and syntactic features predicted to an almost perfect accuracy (99.7%) whether words from the source text would be integrated into test-takers' spoken responses. The majority of these variables were lexical in nature and demonstrated that words in the source text that were more frequent, had more associations, were named more quickly, contained more frequent character bigrams, and had more phonographic neighbors were more likely to be integrated into the response. Two cohesion variables were also significant predictors in the DFA indicating that words that were repeated more often in the source texts and words that were found in coordinated phrases were more likely integrated into test-takers responses. Lastly, one syntactic feature (nouns that were

objects of a preposition) was a predictor in the DFA indicating that nouns used in descriptive phrases were more likely integrated into the spoken response.

This study also focused on predicting human judgments of speaking responses in terms of individual characteristics, topic, and linguistic features related to both source and response internal variables. A baseline model using only individual characteristics and topic included four variables as significant predictors of human ratings. These included note-taking, TOEFL ITP listening scores and structure scores, and topic. The note-taking variable indicated that students who included more word types (i.e., individual words) from the source text in their notes received a higher score. In terms of topic, lecture tasks led to lower scores than the note-taking conversation as did swimming conversation. The note-taking conversation led to higher scores when compared to the swimming conversation likely because the topic was more common (note-taking as compared to swimming) as was the context (two students talking as compared to two professors).

No demographic variables were significant predictors of speaking proficiency in the LME model. In addition, no working memory test scores were significant predictors in the LME model even though a correlation demonstrated a weak relationship between the listening span and speaking scores (r = .154, see Table 2), while the correlation between the running span and speaking scores was not significant (r = .030). The descriptive statistics for the working memory scores reported in Table 1 do not indicate a ceiling effect and show a relatively robust range and variance scores, suggesting that our study included participants who had a range of working memory capacity. The findings of the study are, to some extent, in line with previous L2 listening testing literature which showed a lack of evidence for the significant relationship between WMC and L2 listening

(Andringa et al., 2012). In addition, the participants in the current study participated in listen/speak tasks which allowed them to take notes and use them while speaking. Such task characteristics likely reduce the need to rely on working memory during oral responses (i.e., using strategies to overcome cognitive differences).

When linguistic variables were incorporated into the model, the model significantly outperformed the baseline model and included two linguistic features. The first feature indicated that responses that received higher scores included a greater number of shared words between the response and the source text suggesting that the degree of text integration was the most important factor that predicted human scores. Additionally, responses received a lower score if the responses included a greater number of nouns from the response text that were located in the object position. The latter finding likely indicates that test-takers that focused on ancillary information in the source text (i.e., not the main subjects of the source text) received lower scores. The LME models also indicated that test-takers with better listening and structure scores scored higher on the speaking section of the TOEFL-iBT. Lastly, topic was an important predictor. Specifically, test-takers received lower scores on the lecture tasks than the conversation tasks. Unlike the baseline LME, note-taking was not a significant predictor of human scores when linguistic features were included.

In combination, the findings from the DFA and LME models indicate that linguistic elements in the source text (i.e., cohesive and syntactic features) and lexical properties of word strongly predict which words test-takers integrate into their spoken responses. The findings demonstrate that words that are repeated words more often in the source text and nouns that either coordinated of found as object as preposition in the

source text are more likely to be integrated into test-takers' responses. In addition, words in the source text that are more frequent, have more associations, are named more quickly, have more common characters and have more phonographic neighbors are more likely integrated into test-takers' responses. These findings suggest that properties of the source text along with properties of the words within the source text assist in text recall and may aid test-takers in noticing and integrating key words and/or concepts into their responses. The findings also show that integration of words from the source text is a significant predictor of human judgments of speaking proficiency although nouns in the object position are not. These linguistic features are still important predictors of speaking proficiency even when individual characteristics such as language proficiency, working memory skills, age, and gender along with topic and strategies such as note-taking are included in the model.

As noted by Crossley et al. (2014), these findings have important inferences for the difficulty of test items because listening samples that contain less sophisticated words that are easier to recall and contain greater cohesion between these words appear to lead to better recall of key words from the source text. The integration of these words by test-takers into their spoken response may lead to higher ratings of speaking proficiency indicating that source texts containing words with greater recall properties (i.e., words that are more frequent words and have greater associations) and discourse structures that lead to greater recall (i.e., key words, words in cohesive structures, and words that are objects of prepositions) may positively influence test-taker scores when compared to source texts with lower lexical, cohesion, and syntactic recall properties.

Thus, test designers need to carefully consider lexical and cohesive properties between test items to ensure balance among items across different versions of their tests. When developing speaking assessment tests, developers should consider that linguistic properties of source texts strongly influence text integration, which in turn can impact human ratings of integrated speaking proficiency. If a text contains relational, propositional, and syntactic features that do not lead to recall of items, human ratings of speaking proficiency may decrease. On the other hand, if a source text contains relational, propositional, and syntactic feature that do increase recall, ratings of speaking proficiency may increase. As a result, if test has two forms or multiple versions of test are administered with different source texts that differ in the amount of relational, propositional, and syntactic features, one form or test may prime greater recall of source text words/concepts resulting in increased speaking proficiency scores when compared to the other. While not easy to measure, natural language processing tools like TAALES would prove helpful in assessing the properties of words within source texts. For instance, if multiple forms of a test are developed, TAALES could be used to measure differences in the lexical properties of each form (i.e., differences in word frequency, words' phonological and orthographic neighbors, and word meaningfulness) to ensure balance across forms. This could provide a level of certainty that each form would lead to similar integration of words from the source text. Additionally, test developers could identify key words in source texts and ensure that each form included a similar number of key terms.

Conclusion

The current study shows that the relational, propositional, and syntactic properties of source texts are almost perfect predictors of text integration and that lexical integration from the source text into the spoken response (especially nouns) acts as a strong predictor of human ratings of speaking proficiency that goes beyond individual differences such as working memory and listening skills, test-taking strategies such as note-taking, and topic. Overall the findings indicate that the properties of the source text can predict which words will be included in the response as well as predict human ratings of speaking proficiency. The finding that properties in the input appear to have an effect on the elicitation of spoken responses (Lee, 2006) raises concerns about integrated speaking assessments which may inadvertently place greater weight on recall ability than other elements of speaking proficiency such as language use, delivery, and topic development. Future studies would benefit from the inclusion of multiple source texts that are controlled such that they differ in their frequency and type of relational and propositional properties. Such studies could better examine the relationship between linguistics properties in the source text and speaking proficiency score and provide direct support for our interpretation of the findings from this study.

Overall, this study in conjunction with Crossley et al. (2014) provides strong evidence that linguistic features in the source text can influence text recall and text integration. However, these results cannot be generalized to other types of sources beyond the listen/speak tasks in the TOEFL-iBT. Unlike Crossley et al. (2014), the current study did control for several test-taker variables such as proficiency, age, gender, and working memory. In addition, this study examined a wider range of linguistic features taken from a number of contemporary natural language processing tools.

Together, these additions provide additional strength to the argument that lexical, cohesion, and syntactic features in the source text can influence text recall and text integration and that this integration is a predictor of test performance.

Acknowledgments: This research was funded by the Educational Testing Service (ETS) under a Committee of Examiners and the Test of English as a Foreign Language research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researchers. TOEFL® test material is reprinted by permission of Educational Testing Service, the copyright owner.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823. doi: 10.1111/j.1467-9280.2006.01787.x

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning, 62*(S2), 49–78.

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly, 13,* 55-71. doi:10.1080/15434303.2015.1126718

Baddeley, A. (2003). Working memory and language: an overview. *Journal of Communication Disorders*, *36* (3), 189 – 208.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47-89). New York: Academic Press.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*(2), 283-316. doi: 10.1037/0096-3445.133.2.283

Balota, D. A, Yap, M.J,, Cortese, M.J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. doi: 10.3758/BF03193014.

Barkaoui, K. (2013). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*, 241-259.

Barkaoui, K. (2015). Test takers' writing activities during the TOEFL iBT writing tasks: A stimulated recall study. *ETS Research Report* (RR-15-04, TOEFLiBT-25).

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2012). Test-Takers' Strategic Behaviors in Independent and Integrated Speaking Tasks. *Applied Linguistics*, *34*(3), 304-324. http://dx.doi.org/10.1093/applin/ams046

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bejar, I. (1985). *The preliminary study of raters for the test of spoken English*. (Monograph Series No. MS-18). Princeton, NJ: ETS.

Boles, D. B. (1983). Dissociated imageability, concreteness, and familiarity in lateralized word recognition. *Memory & Cognition, 11*, 511–519.

Breland, H., Lee, Y, Najarian, M., & Muraki, E. (2004). An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups. *TOEFL researcher reports report 76*. ETS.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*(2), 563-570.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41* (4), 977-990. doi: 10.3758/brm.41.4.977

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*. doi:10.3758/s13428-013-0403-5

Butler, F. A., Eignor, D., Jones, S., McNamara, T. & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: a working paper.* TOEFL Monograph Series (MS-20). Princeton, NJ: Educational Testing Service.

Carrell, P. L. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks.* (TOEFL Monograph Series No. MS-35). Princeton, NJ: ETS.

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

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769-786.

Crismore, A., Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students. *Written Communication 10*, 39–71.

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly, 11* (3), 250-270.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42* (3), 475-493.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion. *Behavior Research Methods.*

Crossley, S. A. & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Cumming, A., Grant, L., Mulcahy-Ernt, P. & Powers, D. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL.* TOEFL Monograph Series, No. 26. Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype tasks for the new TOEFL.* TOEFL Monograph Series, Report No. 30. Princeton, NJ: Educational Testing Service.

Cushing, S. T. (1993. April). L2 proficiency, academic status, and lecture note content. Paper presented at TESOL, Atlanta, GA.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading *Journal of Verbal Learning and Verbal Behavior, 19*(4), 4500-4466.

Dascalu, M., McNamara, D.S., Crossley, S.A., & Trausan-Matu, S. (in press). Age of Exposure: A Model of Word Learning. *Proceedings of the 30th Association for the Advancement of Artificial Intelligence (AAAI) Conference*.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447-464.

Davies, A., & Widdowson, H. (1974). Reading and writing. In J. P. Allen & S. P. Corder (Eds.), *Techniques in applied linguistics* (pp. 155–201). Oxford: Oxford University Press.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition* 22.4, 499–533.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations*. TOEFL Monograph Series. Educational Testing Service: Princeton, NJ.

Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL Quarterly*, *22*, 259-281.

Feak, C. & Dobson, B. (1996). Building on the impromptu: a source-based academic writing assessment. *College ESL, 6* (1), 73-84.

Foltz, P. W. (2007) Discourse coherence and LSA. In T. K Landauer, W. Kintsch, D. McNamara & S. Dennis. (Eds.) *LSA: A Road to Meaning*. Lawrence Erlbaum Publishing.

Gee, N. R., Nelson, D. L., & Krawczyk, D. (1999). Is the concreteness effect a result of underlying network interconnectivity? *Journal of Memory and Language, 40*, 479–497.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). *Coh-Metrix*: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36, (2), 193-202. doi: 10.3758/BF03195564

Gries, S.T. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics, 18*(1), 137-165.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18* (3), 218-238.

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

Hamp-Lyons, L. & Kroll, B. (1996). Issues in ESL writing assessment: an overview. *College ESL, 6* (1), 52-72.

Hill, Y., Liu, L. (2012). Is there any interaction between background knowledge and language proficiency that affects TOEFL iBT reading performance. RR-12-22, TOEFLibT-18, ETS Research Report.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods*, *45*(3), 718-730.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition. 37*, 549 – 581.

Kirsner, K. (1994). Implicit processes in second language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283–312). San Diego, CA: Academic Press.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973) An associative thesaurus of English and its computer analysis. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), *The Computer and Literary Studies*. Edinburgh: University Press.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), Second language task complexity: researching the cognition hypothesis of language learning and performance (pp. 267-289). Amsterdam: John Benjamins. Kucera H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzales, H. and Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods*, *44*(4), 978-990. doi:10.3758/s13428-012-0210-4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). *lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)*. R package version 2.0-11

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49* (4), 757-786.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23* (2) 131–166 10.1191/0265532206lt325oa

Linck, J. A., Osthus, O., Koeth, J. T., & Bunting, M. F. (2013). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review. 21*, 861-883.

Longo, B. (1994). Current research in technical communication: the role of metadiscourse in persuasion. *Technical Communication, 41*, 348–352.

Lund, K., Burgess, C. & Atchley, R. A. (1995). Semantic and associative priming in a high-dimensional semantic space. *Cognitive Science Proceedings (LEA)*, 660-665.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60*(3), 501-533. doi: 10.1111/j.1467-9922.2010.00565.x

Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning 62*(3), 704-740. doi: 10.1111/j.1467-9922.2011.00649.x

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E., (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language, 25*, 645–656.

McDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics* 21.3, 395–423.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44*(3), 295-322.

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.

Miyake, A. & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy (Ed.), *Foreign language learning* (pp. 339 – 364). Mahwah, NJ: Lawrence Erlbaum Associates.

Nakagawa, S., & Schielzeth, H. (2013), A general and simple method for obtaining *R*2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133–142. doi:10.1111/j.2041-210x.2012.00261.x

Nelson, D. L. & Friedrich, M. A. (1980). Encoding and cuing sounds and senses. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 717–731.

Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. (1990). Encoding context and retrieval conditions as determinants of the effects of natural category size. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 31–41.

Oxford, R. (2001). Integrated skills in the ESL/EFL classroom. *ESL Magazine*, *6* (1).

Paivio, A. (1968). A factor analytic study of word attributes and verbal learning. *Journal of Verbal Learning and Verbal Behavior, 7*, 41–49.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, & Winston.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45* (3), 255–287.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*, 18–34.

Powers, D. E. (1986). Academic demands related to listening skills. *Language Testing, 3* (1), 1-38.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rashotte, C.A. & Torgesen, J.K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20*, 180-188.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). Factor Structure of the TOEFL Internet-Based Test (iBT): Exploration in a Field Trial Sample. (TOEFL Monograph Series No. RR-08-09). Princeton, NJ: ETS.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning, 65*, 390-416

van de Kopple, W. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36*, 82–95.

Wallace, C. (1997). IELTS: Global implications of curriculum and materials design. *ELT Journal, 51*, 370-373.

Wen, Z., & Mota, M., & McNeil A. (Eds.) (2015). *Working memory in second language acquisition and processing*, Multilingual Matters. Bristol: UK.

Williams, J. N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 427-441). New York: Routledge.

Table 1
Descriptive statistics for individual differences

Variable	Mean	SD	Min	Max
Age	24.080	5.100	17	45
Time spent studying English (months)	54.293	56.114	0.5	360
Time spent in English speaking country (months)	7.298	9.054	0.5	60
Listen span partial score	44.717	8.354	10	60
Running span partial score	62.274	14.425	18	104
TOEFL listening score	50.875	5.313	37	68
TOEFL structure score	44.589	5.895	31	61
TOEFL reading score	48.597	7.204	31	66
TOEFL score total	480.217	53.282	330	630
TOEFL speaking score	2.146	0.705	1	4

Table 2
Descriptive statistics and MANOVA results for linguistic features.

Index	Integrated words mean (SD)	Unintegrated words mean (SD)	*F*	*p*	η 2	Retained in DFA
Frequency written all words (BNC)	0.421 (0.227)	-0.569 (0.151)	6953.182	< .001	0.869	Yes
Meaningfulness all words (MRC)	333.545 (45.356)	399.209 (10.723)	1044.108	< .001	0.499	Yes
Age of acquisition all words (Kuperman)	4.681 (0.665)	5.751 (0.386)	1018.489	< .001	0.492	No
Range all words (SUBTLEXus)	6819.979 (1095.866)	5221.998 (576.353)	876.106	< .001	0.455	No
Orthographic Neighbors	9.704 (1.91)	7.114 (0.922)	784.454	< .001	0.428	No
Free Association Stimuli (University of South Florida)	18.432 (7.568)	27.403 (4.063)	573.746	< .001	0.353	No
Word similarity (Latent Semantic Analysis)	0.166 (0.044)	0.134 (0.011)	270.344	< .001	0.205	Yes
Concreteness Brysbaert (all words)	2.41 (0.386)	2.696 (0.161)	245.668	< .001	0.19	No
Occurrence of word in source text	738.954 (950.317)	133.515 (88.505)	211.66	< .001	0.168	Yes
Phonographic neighbors (homophones included)	9.6 (1.347)	8.773 (0.255)	191.516	< .001	0.154	Yes
Noun (clausal coordinate) in source text	68.883 (89.937)	29.446 (29.415)	91.365	< .001	0.08	Yes
Noun (object of preposition) in source text	31.964 (46.575)	14.365 (10.263)	71.624	< .001	0.064	Yes
Free association tokens (EAT)	87.516 (12.324)	83.156 (4.042)	59.448	< .001	0.054	Yes
Word age of exposure	1.505 (0.986)	1.143 (0.525)	55.4	< .001	0.05	No
Character bigram frequency	3680.405 (599.501)	3495.413 (152.863)	47.028	< .001	0.043	Yes
Co-occurrence probability (McDonald)	0.679 (0.183)	0.738 (0.102)	40.512	< .001	0.037	Yes
Word naming response time (z-score)	-0.563 (0.08)	-0.546 (0.031)	20.338	< .001	0.019	Yes

Table 3
Confusion matrix for DFA integrated and unintegrated words

		Integrated words	Unintegrated words	
Whole set	Integrated words	523	3	526
	Unintegrated words	0	526	526

		Integrated words	Unintegrated words	
Cross-validated	Integrated words	523	3	526
	Unintegrated words	0	526	526

Table 4
Correlations between fixed factors and speaking scores

Variable	Type	*r*	*p*
Number of integrated words from response in sample	Key words	0.697	< .001
TOEFL listening score	TOEFL	0.569	< .001
TOEFL reading score	TOEFL	0.429	< .001
TOEFL structure score	TOEFL	0.422	< .001
Occurrence of shared word in source text	Cohesion	-0.317	< .001
Occurrence of shared noun (object position) in source text	Syntactic	-0.31	< .001
Word association (MI2) tri-grams (COCA news)	Lexical	0.306	< .001
Bi-gram frequency (BNC)	Lexical	-0.306	< .001
Number of word types from source text in notes	Note-taking	0.303	< .001
Bi-gram range (COCA academic)	Lexical	-0.293	< .001
Occurrence of shared word (clausal coordination) in source text	Syntactic	-0.278	< .001
Word hypernymy (noun)	Lexical	0.277	< .001
Age of acquisition Kuperman (content words)	Lexical	0.259	< .001
Bi-gram proportion (BNC)	Lexical	0.254	< .001
Word frequency all words (COCA academic)	Lexical	-0.247	< .001
Occurrence of shared word (phrasal coordination) in source text	Syntactic	-0.228	< .001
Word age of exposure	Lexical	0.227	< .001
Word naming response time (standard deviation)	Lexical	0.221	< .001
Range content words (SUBTLEXus)	Lexical	0.215	< .001
Polysemy (content words)	Lexical	0.203	< .001
Lexical decision time (standard deviation)	Lexical	0.2	< .001
Number of orthographic neighbors	Lexical	0.193	< .001
Character bigram frequency	Lexical	0.193	< .001

Number of phonographic neighbors (homophones excluded)	Lexical	0.185	< .001
Number of orthographic neighbors with lower frequency (mean number)	Lexical	0.182	< .001
Listen span partial score	Working memory	0.148	< .001
Word concreteness (Brysbaert)	Lexical	0.175	< .001
Imageability content words (SUBTLEXus)	Lexical	0.145	< .001
Free association types (EAT)	Lexical	0.141	< .001
Word similarity (Latent Semantic Analysis)	Lexical	0.122	< .010

Table 5
Baseline model for speaking proficiency scores

Fixed Effect	**Coefficient**	**Std. Error**	***t***	***p***
intercept	-1.564	0.312	-5.019	< 0.001
Number of word types from source text in notes	0.010	0.003	3.188	< 0.010
TOEFL listening score	0.062	0.007	8.542	< 0.001
TOEFL structure score	0.013	0.007	2.047	< 0.050
Topic: Swimming (baseline note-taking)	-0.146	0.068	-2.141	< 0.050
Topic: Fungus (baseline note-taking)	-0.252	0.047	-5.319	< 0.001
Topic: Reciprocity (baseline note-taking)	-0.237	0.070	-3.409	< 0.001

Table 6
Full model for speaking proficiency scores

Fixed Effect	**Coefficient**	**Std. Error**	***t***	***p***
Intercept	-0.926	0.237	-3.904	< 0.001
Number of integrated words from response in sample	0.014	0.001	14.398	< 0.001
Occurrence of shared noun (object position) in source text	-0.001	0.001	-2.286	< 0.010
TOEFL listening score	0.034	0.006	5.859	< 0.001
TOEFL structure score	0.014	0.005	2.939	< 0.010
Topic: Fungus (baseline note-taking)	-0.244	0.043	-5.686	< 0.001
Topic: Reciprocity (baseline note-taking)	-0.275	0.056	-4.891	< 0.001
Topic: Fungus (baseline swimming)	-0.140	0.055	-2.565	< 0.010
Topic: Reciprocity (baseline swimming)	-0.171	0.045	-3.835	< 0.001