The Role of Lexical Properties and Cohesive Devices in Text Integration and Their Effect on Human Ratings of Speaking Proficiency

Scott Crossley, Amanda Clevinger & Youjin Kim

To cite this article: Scott Crossley, Amanda Clevinger & Youjin Kim (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, DOI: 10.1080/15434303.2014.926905

To link to this article: http://dx.doi.org/10.1080/15434303.2014.926905

Published online: 18 Aug 2014.

Submit your article to this journal

Article views: 532

View related articles

View Crossmark data

Citing articles: 1 View citing articles
The Role of Lexical Properties and Cohesive Devices in Text Integration and Their Effect on Human Ratings of Speaking Proficiency

Scott Crossley, Amanda Clevinger, and YouJin Kim

Georgia State University, Atlanta, Georgia, USA

There has been a growing interest in the use of integrated tasks in the field of second language testing to enhance the authenticity of language tests. However, the role of text integration in test takers’ performance has not been widely investigated. The purpose of the current study is to examine the effects of text-based relational (i.e., cohesion) and propositional-specific (i.e., lexical) features on subsequent source text integration in spoken responses. In addition, the current study investigates the effects of word integration on human ratings of speaking performance. Sixty test-takers’ speaking samples from the listen/speak section of the TOEFL-iBT were analyzed in terms of textual integration at the lexical and cohesion level and how this integration predicted human scores of speaking proficiency. The results indicate that the properties of the source text are strongly predicative of which words test takers will integrate into their response. Moreover, it was found that text integration is an important factor that affects human ratings of speaking proficiency. The findings of the study are discussed in terms of item difficulty and construct representation.

INTRODUCTION

Non-native speakers (NNS) of English attending universities where English is the language of instruction need to have a high level of spoken English proficiency to accomplish a variety of academic tasks. Specifically, their target language use domain requires them to read academic texts, listen and comprehend academic lectures, and integrate this listening and reading into oral reports and class discussions (Douglas, 1997). Thus, integrating content from reading and listening samples (i.e., source text) are primary components of spoken academic success. Such integrated tasks authentically resemble the type of tasks that are integral to academic contexts and best represent the interdependent relationship between input and output in academic situations (Cumming, Kantor, Powers, Santos, & Taylor, 2000; Cumming, Grant, Mulcahy-Ernt, & Powers, 2005; Cumming, Kantor, Baba, Eouanzoui, Erdosy, & James, 2006; Lewkowicz, 1997).
An important component of students’ ability to integrate information from source text into spoken responses is recall ability (i.e., the ability to retrieve information and events that occurred in the past). Recall of items can either be based on learner characteristics or based on specific linguistic properties of a text. As a learner characteristic, the ability to recall items from discourse generally reflects working memory skills, which is the ability to temporarily store and manipulate information (Baddeley, 2003). Working memory capacity is an important component of recall that reflects the efficiency and quality of language processing in real time (Miyake & Friedman, 1998). As a linguistic property of text, research suggests that there are two types of information that affect the efficiency of encoding and the subsequent recall of discourse: relational information and proposition-specific information (McDaniel, Einstein, Dunay, & Cobb, 1986). Relational information refers to the organizational aspects of a text and how propositions are embedded in the text (i.e., text cohesion). Proposition-specific information pertains to the lexical items (i.e., words) that comprise a proposition and the relationship between words within a proposition. Both types of information are identified as important components of second language (L2) processing with L2 learners often having a difficult time identifying key ideas (i.e., proposition-specific information) and perceiving relationships among ideas (i.e., relational information; Powers, 1986).

Our goal in this study is to examine text-based relational and propositional-specific relationships in a listening source text and how these relationships influence subsequent source text integration in L2 learners’ integrated spoken responses. Furthermore, we examine the effects of word integration on expert ratings of speaking proficiency. Thus, this study addresses two research questions: 1) Which words from a listening source text are integrated into a spoken response and can these words be predicted based on relational and propositional properties in the source text? and 2) Can these relational and propositional properties predict human ratings of speaking proficiency? To address these research questions, we analyze a small corpus of integrated speaking samples selected from the listen/speak section of the TOEFL (Test of English as a Foreign Language) iBT public use dataset.

We first investigate which content words (defined as nouns, verbs, auxiliary verbs, adjectives, and adverbs) from the listening source text are most likely integrated or not integrated by the test taker into a spoken response and whether or not these words can be classified based on relational and propositional-specific properties. We next use relational and propositional-specific features to predict human scores of speaking proficiency. Such an approach allows us to examine if the lexical properties of the source text affect text integration and, more importantly, if these properties positively or negatively influence human judgments of speaking proficiency. If such relationships exist, then aspects of human judgments of speaking proficiency may not be learner specific (i.e., independent measures of learner language proficiency), but rather influenced by properties of the text. That is to say, the stimuli found in the source text may affect the elicitation of test-takers’ spoken responses, which might in turn influence human judgments (Lee, 2006). Such a finding would have important implications for test development, particularly related to item difficulty and construct representation.
LITERATURE REVIEW

Text Properties and Recall

Relational aspects of a text are subsumed under theories of text cohesion, which explain how linguistic features of text can help to organize and embed information. There is a variety of linguistic features related to text cohesion, including connectives, anaphoric references, and word overlap. Text cohesion refers to the presence of these linguistic cues that allow the reader to make connections between the ideas in the text. Cohesion is often confused with coherence, which refers to the understanding that the reader extracts from the text. This understanding may be derived from cohesion features but may also interact with prior knowledge and reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996; O’Reilly & McNamara, 2007).

Among various linguistic features related to text cohesion, connectives (e.g., and, but, also) are probably the most commonly discussed. Connectives play an important role in the creation of cohesive links between ideas and clauses (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994) and provide clues about text organization (van de Kopple, 1985) that promote greater text comprehension. Anaphoric reference (i.e., the resolution of pronominal antecedents) is also an important indicator of text cohesion (Halliday & Hasan, 1976). Accurate anaphoric resolution can lead to greater text comprehension and quicker text processing times (Clifton & Ferreira, 1987).

Lastly, lexical overlap (i.e., overlap of words and stems between and within sentences) also aids in text comprehension by facilitating meaning construction that can improve text readability and text processing (Crossley, Greenfield, & McNamara, 2008; Douglas, 1981; Kintsch & van Dijk, 1978; Rashotte & Torgesen, 1985).

Proposition-specific features refer to words within a proposition and how these words are recalled based on their lexical properties. Concrete words (i.e., non-abstract words), for instance, have been shown to have advantages in recall and comprehension tasks compared to abstract words (Gee, Nelson, & Krawczyk, 1999; Paivio, 1991). Similarly, word imageability, which refers to how easily one can construct a mental image of a word in one’s mind, is a strong predictor of word recall (Paivio, 1968). Previous research has found strong correlations between word concreteness, word imageability, and recall mainly because research participants are more likely to generate images for concrete words (Marschark, 1985; Marschark & Hunt, 1989; Marschark, Richman, Yuille, & Hunt, 1987; Paivio, 1971, 1986) than abstract words, and words that arouse stronger images are easier to recall than those that do not. Word polysemy (i.e., the number of senses a word has) is also of interest because words with more senses exhibit a greater degree of ambiguity and are, therefore, likely more difficult to process (Davies & Widdowson, 1974).

In addition, word recall also results from familiarity and frequency effects. For instance, word familiarity, which is related to word exposure, is a strong predictor of recall, but not as strong a predictor as imagery (Boles, 1983; Paivio & O’Neill, 1970). Word familiarity has also been consistently shown to aid in word identification, which in turn aids in recall (Paivio, 1991). A number of studies have also demonstrated that high-frequency words are recognized quicker (Kirsner, 1994) and named more rapidly (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Chambers, 1973; Frederiksen & Kroll, 1976) than lower-frequency words and that reading samples with more frequent words lead to greater text readability (Crossley et al., 2008). In reference to relationships between words, word meaningfulness (i.e., the number of lexical associations a word contains) has been shown to affect memory performance (Nelson

Text Integration

From a pedagogical perspective, the practice of integrating language skills (i.e., speaking, listening, writing, and reading) in the L2 classroom has long been a favored instructional approach found in both content-based language instruction and task-based language instruction. Such integration treats language skills as fundamentally interactive compared to being discrete, segregated skills. Integration in a classroom assists language learners in learning to interact naturally with language in an authentic environment (Oxford, 2001). Interacting effectively includes receiving, transmitting, and demonstrating knowledge as well as organizing and regulating that knowledge (Butler, Eignor, Jones, McNamara, & Suomi, 2000). In practical terms, integration requires the inclusion of key words and propositions taken from reading and/or listening materials in subsequent language production. In response to the growing interest in adopting integrated tasks in language instruction, recent versions of standardized tests such as the TOEFL-iBT also require test takers to integrate reading or listening source material into written or spoken responses. Such integrated tasks represent a vital academic skill that affords contextual language use (Hamp-Lyons & Kroll, 1996), allows test takers to demonstrate their ability to manipulate information that goes beyond their prior knowledge (Hamp-Lyons & Kroll, 1996; Wallace, 1997) and encourages authentic language use (Plakans & Gebril, 2012). Thus, integrating source text information tests a students’ ability to identify and extract relevant information in the source text(s) and organize and synthesize information (or understanding of this information) in their response (Cumming et al., 2000; Feak & Dobson, 1996).

The majority of research investigating text integration has focused on integrated writing (i.e., writing based on information found in source texts). These studies have mainly focused on analyzing differences in linguistic features between integrated and independent writing tasks (i.e., writing from prompt only; Guo, 2011) or investigating how linguistic features are predictive of human ratings of integrated writing (Cumming et al., 2005, 2006). For instance, Guo (2011) found that independent essays were characterized as argumentative, reliant on verbs to provide examples from previous experiences, involved and interactional, and structurally more complex. Integrated essays, on the other hand, were characterized as more focused on organizational cues, using a more detached style of writing that was more informational, and containing lexical items that were more context independent. In terms of predicting human judgments of integrated writing, Cumming et al. (2005, 2006) found that higher-rated essays tended to contain more words, greater diversity of words as evidenced by type-token ratios, and more words per T-unit.

In sum, despite a growing interest in recall and text integration in L2 testing contexts, there have been few to no studies that examine the extent to which linguistic characteristics of a source text (i.e., relational and proposition-specific properties of words) impact test-takers’ responses. This is especially true of spoken responses based on listening source texts. More importantly, how source text word integration impacts human ratings of speaking samples has not been systematically investigated. If the lexical and cohesive properties of words in the source text promote their integration into a spoken response, then some aspects of spoken responses may be, to a degree, test-taker independent (i.e., text-based). If these same linguistic aspects are predictive of human
judgments of speaking quality, then we would have evidence supporting the complexity of interactions between textual features, test-taker output, and human judgments of quality. Notably, this evidence would provide a specific link between the relational and propositional elements found in source texts and test-taker responses and rater judgments. The current study addresses these possibilities.

METHODS

Our purpose in this study is twofold. First, we examine if relational (i.e., cohesive) and proposition-specific (i.e., lexical) properties of words found in source texts aid in those words’ recall and eventual integration into a response. Second, we examine if the lexical and cohesive properties of integrated words are predictive of human judgments of speaking proficiency. Our domain of interest is specifically narrowed to TOEFL integrated listen/speak responses referencing academic genres as found in the TOEFL-iBT public use dataset. The listen/speak tasks require that L2 test takers listen to a spoken source text (e.g., an excerpt from a lecture) and then summarize the lecture in speech by developing relationships between the examples in the source text and the overall topic. Expert raters score these speech samples using a standardized rubric that assesses delivery, language use, and topic development.

TOEFL-iBT Public Use Speech Sample Dataset

The TOEFL-iBT public use dataset comprises data collected from TOEFL-iBT participants from around the world in the years 2006 and 2007. The public use dataset is composed of three separate datasets: item level scores, speech samples, and writing samples. The speech sample dataset includes speaking responses from 480 examinees on six speaking tasks stratified by quartiles (240 participants taken from two test forms). The six speaking tasks include two independent and four integrated speaking tasks that, overall, represent the general topics and expectations of academic situations (Cumming et al., 2004). The four integrated tasks are further subdivided into read/listen/speak and listen/speak tasks. The read/listen/speak tasks prompt the test taker to read one passage and listen to another passage and then either summarize opinions from the passages or combine and convey important information from the passages. The listen/speak tasks are of two topics: a campus situation or an academic course. Of interest to this study is the latter, which contains a listening passage of about 250 words excerpted from an academic lecture. Test takers are expected to listen to the passage, prepare a response in 20 seconds, and then summarize the lecture in a 60-second recording. Test takers are allowed to take notes during the process. The TOEFL-iBT public use dataset includes human scores of speaking proficiency for each task and a combined overall speaking proficiency score that is a combination of scores from all six speaking tasks (i.e., both the independent and integrated speaking tasks).

Corpus Used in the Current Study

For this analysis, we randomly selected 60 listen/speak responses from 60 different TOEFL-iBT participants. We balanced the responses across the two forms contained in the TOEFL-iBT
public use dataset so that the corpus contained 30 responses from each form. A trained transcriber transcribed each of the independent speech samples from the 60 test takers. 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 1.

Human Ratings

Two expert TOEFL raters scored each speaking sample in the corpus. The rubric used by the raters in this study was developed specifically for the TOEFL-iBT integrated speaking tasks (see http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf). The rubric provides a holistic score of integrated speaking proficiency and is based on a 0–4 scale with a score of 4 representing adherence to the task; clear, fluid, and sustained speech; good control of basic and complex structure; coherent expression; effective word choice; and a clear progression of ideas that conveys relevant information required by the task. The rubric does not specifically address text integration but notes that the response should fulfill the demands of the task and that pace may vary as the listener attempts to recall information. 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 and prosody), language use (i.e., grammar and vocabulary), and topic development (i.e., content and coherence).

While interrater reliability data are not provided for the TOEFL-iBT scores in the public use dataset, reported weighted Kappas for similarly double scored TOEFL speaking samples generally range from .77 for scores on a single task and up to .93 for scores summed across three tasks (Xi, Higgens, Zechner, & Williamson, 2008). The final score for a given response was the average of the raters’ scores if the two scores differed by less than two points. Generally, a third rater scores the sample if the scores between the first two raters differ by more than one point. In such a case, the final score is the average of the two closest raters (cf. Bejar, 1985; Carrell, 2007; Sawaki, Stricker, & Oranje, 2008). In the current study, we used the reported final scores in the public use dataset.

**Table 1**

Descriptive Statistics for the Transcribed TOEFL-iBT Listen/Speak Corpus

<table>
<thead>
<tr>
<th>Form Item</th>
<th>n</th>
<th>Lecture area (topic)</th>
<th>Number of words (standard deviation)</th>
<th>Average score on integrated task (standard deviation)</th>
<th>Average combined score (standard deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>30 Anthropology class (reciprocity)</td>
<td>93.200 (18.260)</td>
<td>2.133 (0.730)</td>
<td>13.433 (3.380)</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>30 Botany class (fungus)</td>
<td>96.333 (17.293)</td>
<td>2.167 (0.647)</td>
<td>14.300 (3.395)</td>
</tr>
</tbody>
</table>
Variable Calculation

We calculated a variety of lexical and cohesion values to assess which words were integrated from the source text (i.e., the listening samples) into the test-taker speaking responses. For these source internal variables, each content word (i.e., nouns, verbs, adjectives, and adverbs) in the source text was assigned a value based on either outside databases or internal counts (e.g., the number of repetitions of the word with the source). Once values had been assigned to each word, these values were matched to the words produced in the test-taker response to examine the properties of the words that were integrated (i.e., found in the source text and the test-taker responses) and those that were unintegrated (i.e., found in the source text but not in the test-taker response). For each cohesion variable we calculated average incidence scores.\footnote{We did not use the raw scores because these scores correlated highly with the number of words integrated or not integrated. Because there were fewer words integrated, the raw scores were biased.} The average incidence scores consisted of the average number of occurrences in the source text for a feature that was integrated into the response text (i.e., the average incidence of integrated words in the source text). For each lexical variable, average scores were computed for all the content words in the response that were integrated and for all the words in the source text that were unintegrated (i.e., the average word concreteness score for both the integrated and unintegrated content words). In total, we calculated 10 variables (see Table 2 for descriptive information for these variables).

We also calculated lexical and cohesion scores to predict the human ratings of speaking proficiency. The cohesion scores included incidence scores for the number of tokens integrated into the response, incidence scores for the number of types integrated and not integrated into the response, and incidence scores for the number of times these types occurred in the source text (i.e., a token count). We also calculated average incidence scores for the types integrated and not

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dimension</th>
<th>Calculation</th>
<th>Derived from</th>
<th>Number of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word frequency</td>
<td>Lexical</td>
<td>Average</td>
<td>SUBTLEXus/Source text</td>
<td>2</td>
</tr>
<tr>
<td>Word concreteness</td>
<td>Lexical</td>
<td>Average</td>
<td>MRC</td>
<td>1</td>
</tr>
<tr>
<td>Word familiarity</td>
<td>Lexical</td>
<td>Average</td>
<td>MRC</td>
<td>1</td>
</tr>
<tr>
<td>Word imageability</td>
<td>Lexical</td>
<td>Average</td>
<td>MRC</td>
<td>1</td>
</tr>
<tr>
<td>Word meaningfulness</td>
<td>Lexical</td>
<td>Average</td>
<td>MRC</td>
<td>1</td>
</tr>
<tr>
<td>Incidence of word occurrence in source text</td>
<td>Cohesion</td>
<td>Average</td>
<td>Source text</td>
<td>1</td>
</tr>
<tr>
<td>Use of word in pronominal referent in source text</td>
<td>Cohesion</td>
<td>Average</td>
<td>Source text</td>
<td>1</td>
</tr>
<tr>
<td>Use of word in positive connector clause in source text</td>
<td>Cohesion</td>
<td>Average</td>
<td>Source text</td>
<td>1</td>
</tr>
<tr>
<td>Use of word in negative connector clause in source text</td>
<td>Cohesion</td>
<td>Average</td>
<td>Source text</td>
<td>1</td>
</tr>
</tbody>
</table>

MRC = Medical Research Council Database.
integrated into the response (i.e., what was the average occurrence of these types in the source text). Lastly, we calculated average incidence scores for how often integrated and not integrated words occurred as pronominal referents and how often these words were used in positive and negative connective clauses. Lexically, we calculated both source internal and source external variables (i.e., variables for the words found in the test-taker response that were not found in the source text). The source internal variables calculated were the average scores for the integrated and not integrated words in reference to their word frequency, word concreteness, word familiarity, word imageability, and word meaningfulness. We also computed source external features for the content words produced in the test-taker responses that were not found in the source texts. The variables calculated were the number of content words in the response not found in the source text as well as average scores for word frequency, word concreteness, word familiarity, word imageability, word meaningfulness, word hypernymy, and word polysemy for the content words in the response that were not in the source. While these features tell nothing about text integration, they do provide information about the importance of language used outside of the source text in predicting human judgments of speaking proficiency. In total, we calculated 31 variables for predicting human judgments of integrated speaking proficiency. Table 3 contains descriptive information for these variables. The variables and their calculations are discussed in greater detail below.

Source Internal Variables

We computed a variety of lexical and cohesion features that were source internal. Our lexical variables were word frequency, word concreteness, word familiarity, word imageability, and word meaningfulness. Our cohesion features were word repetition, the use of pronominal referents, and words used in positive and negative connector clauses.

Cohesion indices. For each content word in the source text, we calculated how often the word was repeated in the source text, how often a pronominal referent was used to refer to the word in the source text, and how many times the word in the source text was found in a positive connective constituent (e.g., a constituent that is connected with and or also) or a negative connective constituent (e.g., a constituent that is connected with but). For these cohesion features we computed raw incidence and average incidence scores. Cohesion cues such as those computed provide explicit information about words in the text and their relationship to one another (McNamara et al., 1996; O’Reilly & McNamara, 2007). Increased text cohesion can both improve and facilitate text comprehension (Gernsbacher, 1990), especially for low-skilled students (McNamara et al., 1996) and L2 learners (Crossley et al., 2008; Freedle & Kostin, 1993).

Lexical indices. For each content word in the source text, we calculated the average word frequency in the source text (see below) and word frequency as found in an external reference corpus (the SUBTLEXus corpus; Brysbaert & New, 2009). We also calculated average scores for word property features, such as familiarity, imageability, concreteness, and meaningfulness, by using the MRC Psycholinguistic Database (see below; Coltheart, 1981).

Internal word frequency was calculated by computing the frequency of occurrence of each content word in the source text and normalizing the frequency by text length. External word frequency indices were calculated by using the SUBTLEXus corpus. Frequency values in the SUBTLEXus corpus are based on subtitles taken from 2,046 U.S. films from the years
1900–1990, 3,218 U.S. films from the years 1990–2007, and 4,575 U.S. television series. The total size of the SUBTLEXus corpus is 51 million words, with 16.1 million words from television shows, 14.3 million words from films before 1990, and 20.6 million words from films after 1990. We selected the SUBTLEXus corpus because the corpus better reflects the frequency of words in spoken language than more commonly available corpora of written language. Word frequency indices have proven predictive of lexical development (Laufer & Nation, 1995), text readability (Crossley et al., 2008; Freedle & Kostin, 1993) and L2 lexical proficiency (Crossley, Salsbury, McNamara, & Jarvis, 2011a, 2011b).

We selected four lexical properties from the MRC Database: concreteness, familiarity, imageability, and meaningfulness. For each property, we calculated the average score of all the content words in the source text. Word concreteness measures how concrete or abstract a word is. Word imageability measures how easy it is to construct a mental image of a word, and word familiarity measures how commonly a word is experienced. Lastly, word meaningfulness relates to how many lexical associations a word has. While similar, these properties do differ, depending

<table>
<thead>
<tr>
<th>Variable</th>
<th>Dimension</th>
<th>Source</th>
<th>Integrated/Unintegrated</th>
<th>Source</th>
<th>Calculation</th>
<th>Derived from</th>
<th>Number of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word frequency</td>
<td>Lexical</td>
<td>Yes</td>
<td>Both</td>
<td>Yes</td>
<td>Average</td>
<td>SUBTLEXus</td>
<td>3</td>
</tr>
<tr>
<td>Word concreteness</td>
<td>Lexical</td>
<td>Yes</td>
<td>Both</td>
<td>Yes</td>
<td>Average</td>
<td>MRC</td>
<td>3</td>
</tr>
<tr>
<td>Word familiarity</td>
<td>Lexical</td>
<td>Yes</td>
<td>Both</td>
<td>Yes</td>
<td>Average</td>
<td>MRC</td>
<td>3</td>
</tr>
<tr>
<td>Word imageability</td>
<td>Lexical</td>
<td>Yes</td>
<td>Both</td>
<td>Yes</td>
<td>Average</td>
<td>MRC</td>
<td>3</td>
</tr>
<tr>
<td>Word meaningfulness</td>
<td>Lexical</td>
<td>Yes</td>
<td>Both</td>
<td>Yes</td>
<td>Average</td>
<td>MRC</td>
<td>3</td>
</tr>
<tr>
<td>Word hypernymy</td>
<td>Lexical</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Average</td>
<td>WordNet</td>
<td>1</td>
</tr>
<tr>
<td>Word Polysemy</td>
<td>Lexical</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Average</td>
<td>WordNet</td>
<td>1</td>
</tr>
<tr>
<td>Incidence in the source</td>
<td>Cohesion</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Raw</td>
<td>Source text</td>
<td>1</td>
</tr>
<tr>
<td>text of integrated tokens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence of type</td>
<td>Cohesion</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>Raw</td>
<td>Source text</td>
<td>2</td>
</tr>
<tr>
<td>occurrence in source text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence of token</td>
<td>Cohesion</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>Average/Raw</td>
<td>Source text</td>
<td>4</td>
</tr>
<tr>
<td>occurrence in source text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of word as pronominal</td>
<td>Cohesion</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>Average</td>
<td>Source text</td>
<td>2</td>
</tr>
<tr>
<td>referent in source text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of word in positive</td>
<td>Cohesion</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>Average</td>
<td>Source text</td>
<td>2</td>
</tr>
<tr>
<td>connector clause in source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of word in negative</td>
<td>Cohesion</td>
<td>Yes</td>
<td>Both</td>
<td>No</td>
<td>Average</td>
<td>Source text</td>
<td>2</td>
</tr>
<tr>
<td>connector clause in source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of words in the</td>
<td>Lexical</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Average</td>
<td>Test-taker responses</td>
<td>1</td>
</tr>
<tr>
<td>response not found in source</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We selected four lexical properties from the MRC Database: concreteness, familiarity, imageability, and meaningfulness. For each property, we calculated the average score of all the content words in the source text. Word concreteness measures how concrete or abstract a word is. Word imageability measures how easy it is to construct a mental image of a word, and word familiarity measures how commonly a word is experienced. Lastly, word meaningfulness relates to how many lexical associations a word has. While similar, these properties do differ, depending
on the word. For example, *chassis* is a highly concrete word but is not strongly imageable or familiar. *Concert* is both an imageable and familiar word but is not highly concrete. *Result* is a familiar word, but it is not highly imageable or concrete. *Person* is a relatively familiar word but not strongly meaningful. The word property indices reported by the MRC Database have been demonstrated to be strongly predictive of human judgments of lexical competence (Crossley, Salsbury, McNamara, & Jarvis, 2011a, 2011b), human judgments of second language (L2) writing quality (Crossley & McNamara, 2012), the development of L2 lexical proficiency (Salsbury, Crossley, & McNamara, 2011), and text difficulty (Crossley, Allen, & McNamara, 2012; Freedle & Kostin, 1993). For each index, higher values equate to greater lexical properties (i.e., more concrete, familiar, imageable, and meaningful words).

**Source external variables**

We also selected a variety of features that were source external (i.e., features of the words produced by the test takers in their speaking responses that were not found in the listening source text). These features were lexical in nature, focused only on content words, and included the number of unique words produced and their word familiarity, imageability, concreteness, and meaningfulness scores as computed from the MRC Psycholinguistic Database along with SUBTLEXus word frequency counts. For each of these features, we computed an average score. In addition, we computed two indices related to word ambiguity (polysemy) and word specificity (hypernymy)\(^2\) using the WordNet database (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1993). The ambiguity index (i.e., polysemy) measured the number of senses a word contained on average, whereas the specificity index (i.e., hypernymy) measured the average number of levels a word has in a conceptual, taxonomic hierarchy.

**Statistical Analyses**

To answer our research questions, we conducted a number of statistical analyses. The first analysis was to determine differences between content words from the source text that were integrated versus those that were unintegrated (i.e., research question 1). For this analysis, we conducted an initial Multivariate Analysis of Variance (MANOVA) to select the source internal variables from Table 2 that demonstrated the strongest differences between the integrated and unintegrated words. We then used those variables from the MANOVA that did not demonstrate multicollinearity in a discriminant function analysis to provide confirmatory evidence for the strength of these variables in classifying the words as integrated or unintegrated. This analysis is meant to provide evidence that source internal variables can predict what words will be integrated into test-takers responses.

Our second statistical analysis was to determine if the source internal and source external variables found in Table 3 could be used to predict the human ratings for the integrated speaking task and the combined overall ratings of speaking proficiency, which consisted of the two independent speaking scores and the four integrated speaking scores for each test taker (i.e., research

---

\(^2\)Word hypernymy is related to word concreteness, because the lower a word is located on a hypernymic hierarchy the more likely the word will be concrete (Rosch, 1973; Tversky & Hemenway, 1984).
question 2). In this analysis, we first computed Pearson Product-Moment correlations to select variables for a regression analyses. We then used those variables that did not demonstrate multicollinearity, were not highly correlated to the test-takers’ response lengths, and demonstrated a significant \( r \) value (i.e., \( p < .050 \)) for inclusion in our regression analyses. The regression analysis did not examine interaction between the variables. These analyses are meant to provide evidence for the strength of source internal and source external variables in predicting human judgments of speaking proficiency.

RESULTS

Classifying Integrated and Unintegrated Words

The first research question asked which words from a source text are integrated into test-takers’ spoken response and whether these words can be predicted on the basis of relational and propositional properties in the source text.

**MANOVA**

A MANOVA was conducted by using the lexical and cohesion indices taken from the source internal variables as the dependent variables and the integrated and unintegrated words from each text as the independent variables. Thus, we had 2 independent variables (the lists of integrated and unintegrated words for each of the 60 test takers) and 10 dependent variables (the average scores for the selected lexical and cohesion features for the words in the integrated and unintegrated lists). All assumptions for the MANOVA were met, and all indices except the \textit{average word meaningfulness} index demonstrated significant differences between integrated and unintegrated words in the test-takers’ spoken responses. We next checked for multicollinearity (defined as \( r > .900; \) Tabachnick & Fidell, 2001) between variables. If two variables demonstrated multicollinearity, the variable with the lower-effect size (\( \eta^2 \)) was excluded from the analysis, and the variable with the larger-effect size was included. After checks for multicollinearity, we were left with 7 indices (see Table 4 for descriptive statistics for these indices). The results show that words integrated into test-takers spoken responses from the source text are repeated more often in the source text, are more frequent in the source text, are found more often in positive and negative connective clauses, are less frequent in a reference corpus, but are more familiar and imageable.

**Discriminant function analysis**

To provide confirmatory evidence that the indices above do discriminate between integrated and unintegrated words, we conducted a stepwise discriminant function analysis (DFA). 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 six strongest indices from the MANOVA analysis based on the strength of their reported effect size.
(see Table 4). We selected only the top six variables to have a minimum of 10 events per predictor variable (EPV). Such a ratio is standard to control for overfitting in similar models (Concato, Peduzzi, Holford, & Feinstein, 1995).

The stepwise DFA using these six variables retained three variables as significant predictors of whether a word was integrated in the test-takers’ response or unintegrated (Average incidence of word occurrence in source text, Average word frequency in source text, and Average word use in positive connective clause) and removed the remaining three variables as non-significant predictors based on their predictive strength.

The results show that the DFA using these three indices correctly allocated 118 of the 120 word lists as being integrated or unintegrated, \( \chi^2 (1, n=120) = 112.258, p < .001 \), for an accuracy of 98.3% (chance level for this analysis is 50%). The reported Kappa value for this analysis was .967, indicating almost perfect agreement between the predicted classification of the word lists and their actual classification.

**Leave-one-out-cross-validation**

We also tested the six indices from the DFA by using a leave-one-out-cross-validation (LOOCV) analysis. In this analysis, we chose a fixed number of folds that equaled the number of observations (i.e., the 120 word lists). In LOOCV, one observation in turn is left out for testing and the remaining instances are used as the training set (i.e., the 119 remaining word lists). We assessed the accuracy of the DFA model by testing its ability to predict the classification of the omitted instance. Such an approach affords us the opportunity to test the model generated by the DFA on an independent dataset (i.e., on data that is not used to train the model). If similar results are found between the entire set and the LOOCV set, our level of confidence in the model increases, supporting the extension of the analysis to external datasets.

The results from the LOOCV were identical to the initial DFA, correctly allocating 118 of the 120 word lists as being integrated or unintegrated, reporting an accuracy of 98.3% (see Table 5 for the confusion matrix for this analysis). These results provide evidence that three variables related to word repetition in the source text, word frequency in the source text, and use in a positive...
TABLE 5
Predicted Word Type (integrated or unintegrated) from DFA

<table>
<thead>
<tr>
<th>Actual word type</th>
<th>Predicted word type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total set</td>
<td>Integrated</td>
</tr>
<tr>
<td>Integrated</td>
<td>58</td>
</tr>
<tr>
<td>Unintegrated</td>
<td>0</td>
</tr>
<tr>
<td>LOOCV Set</td>
<td>Integrated</td>
</tr>
<tr>
<td>Integrated</td>
<td>58</td>
</tr>
<tr>
<td>Unintegrated</td>
<td>0</td>
</tr>
</tbody>
</table>

connective clause in the source text can predict with almost perfect accuracy whether the lists contained integrated or unintegrated words.

Predicting Human Ratings of Speaking Proficiency

For our second research question, we were interested in how the lexical and cohesion properties of the words in the source text and the lexical properties of the words in the participants’ responses that were not found in the source text could predict the averaged human assessments of speaking quality for the listen/speak integrated task and the combined speaking scores (i.e., general speaking proficiency scores based on the six speaking tasks in the TOEFL). Thus, in these analyses we examined both source internal and source external variables. Because we are analyzing the test-takers’ spoken responses, we were able to compute word type scores for the word repetition variables. These scores allow us to analyze how many word types were integrated or not integrated into the spoken response as compared to the number of tokens integrated.

**Listen/speak integrated speaking proficiency scores**

**Pearson correlations training set.** To select the variables for our regression analysis, we first computed Pearson product-moment correlations to assess both multicollinearity between the variables and the effect sizes of the variables. We set three thresholds for inclusion. The first threshold was that no selected variables could strongly correlate with each other ($r > .900$; Tabachnick & Fidell, 2001). The second threshold was that no selected variable could strongly correlate ($r > .900$) with the length of the speech sample. The third threshold was that selected variables needed to demonstrate a significant correlation ($p < .050$) with the dependent variable (the averaged human scores).

After controlling for multicollinearity, text length considerations, and effect size, we were left with eight variables related to integrated types and tokens, unintegrated types and tokens, total number of word types in the response that were not in the text, and word hypernymy scores for the words in the responses that were not in the source text (see Table 6 for Pearson correlation results).

**Linear multiple regression training set.** We conducted a stepwise linear regression analysis using the six variables that showed the highest correlation with the human scores of integrated speaking (from Table 6). Like the DFA, we selected only the top six variables to have a minimum
of 10 events per predictor variable (EPV). Acceptable ratios for linear regression models range from 4 EPV (Freedman & Pee, 1989) to 15 EPV (Stevens, 2002). The six variables were regressed onto the human ratings for the 60 transcribed speech samples. During this process, we also checked for additional multicollinearity through variance inflation factors (VIF) and tolerance values. All VIF and tolerance values were at about 1, supporting the notion that the model data did not suffer from multicollinearity (Field, 2005).

The linear regression using the six variables yielded a significant model, \( F(1, 58) = 69.146, p < .001, r = .737, R^2 = .544 \). One variable was a significant predictor: Total incidence in the source text of integrated words. The results from the linear regression show that this one variable accounts for 54% of the variance in the human ratings of integrated speaking proficiency for the 60 speech samples (see Table 7 for additional information).

**Leave-one-out-cross-validation.** We also tested the model from the stepwise regression analysis using a LOOCV analysis. For the listen-speak item score, the LOOCV set yielded \( r = .716, R^2 = .513 \). The results from the LOOCV set model show that the one variable (Total incidence in the source text of integrated tokens) accounted for 51% of the variance in the human ratings of integrated speaking proficiency for the 60 speech samples in an independent sample.

This analysis indicates that if test takers integrate words with higher cohesion values (i.e., words that are repeated more often in the source text), they will receive a higher score.

**Combined speaking proficiency scores**

**Pearson correlations training set.** We selected the same variables from the integrated speaking proficiency score analysis for this analysis (i.e., see Table 6). Pearson’s correlations

<table>
<thead>
<tr>
<th>Entry</th>
<th>Variable added</th>
<th>( r )</th>
<th>( R^2 )</th>
<th>( \hat{b} )</th>
<th>( B )</th>
<th>SE</th>
<th>( t )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry 1</td>
<td>Total incidence in the source text</td>
<td>0.737</td>
<td>0.544</td>
<td>0.039</td>
<td>0.737</td>
<td>0.005</td>
<td>8.315</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*Note. Constant = −0.093.*
TABLE 8
Correlations Between Selected Indices and Average Integrated Scores in Listen/Speak Corpus

<table>
<thead>
<tr>
<th>Index</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total incidence in the source text of integrated tokens</td>
<td>0.758</td>
<td>0.001</td>
</tr>
<tr>
<td>Average incidence of unintegrated word types in source text</td>
<td>-0.685</td>
<td>0.001</td>
</tr>
<tr>
<td>Total incidence of unintegrated word tokens from source text</td>
<td>-0.619</td>
<td>0.001</td>
</tr>
<tr>
<td>Incidence of integrated tokens from source text in the response text</td>
<td>0.583</td>
<td>0.001</td>
</tr>
<tr>
<td>Total number of types in the response that are not in the source text</td>
<td>0.531</td>
<td>0.001</td>
</tr>
<tr>
<td>Total incidence of integrated word types from source text</td>
<td>0.407</td>
<td>0.001</td>
</tr>
<tr>
<td>Average word hypernymy for words in response not from source text</td>
<td>-0.273</td>
<td>0.035</td>
</tr>
<tr>
<td>Average incidence of integrated word tokens in source text</td>
<td>-0.228</td>
<td>0.080</td>
</tr>
</tbody>
</table>

were conducted between these eight variables and the combined speaking proficiency scores. Seven of the eight variables showed a significant correlation with the human ratings, did not demonstrate multicollinearity, and were not strongly correlated with text length (see Table 8 for correlation results). The strongest six variables were used in a subsequent regression analysis.

**Multiple regression training set.** The same procedure was used for the combined human scores as was used for the listen-speak item scores. The linear regression using the six variables yielded a significant model, $F(1, 58) = 78.146, p < .001, r = .758, R^2 = .574$. The one predictive variable retained in this model was also the predictive variable in the regression model for the integrated scores: Total incidence in the source text of integrated words. This variable explained 57% of the variance in the human ratings that comprise the combined speaking proficiency scores (see Table 9 for details).

**Leave-one-out-cross-validation.** For the combined human score, the LOOCV set yielded $r = .740, R^2 = .547$. The results from the LOOCV set model show that the one variable accounted for 55% of the variance in the combined human ratings of speaking proficiency for the 60 speech samples.

Like the integrated analyses, this analysis shows that test takers who integrate words that repeat more often in the source text will receive a higher rating for their combined speaking abilities.

TABLE 9
Regression Analysis for Average Combined Scores in Listen/Speak Corpus: Training Set

<table>
<thead>
<tr>
<th>Entry</th>
<th>Variable added</th>
<th>r</th>
<th>$r^2$</th>
<th>$\beta$</th>
<th>$B$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry 1</td>
<td>Total incidence in the source text of integrated tokens</td>
<td>0.758</td>
<td>0.574</td>
<td>0.197</td>
<td>0.758</td>
<td>0.022</td>
<td>8.840</td>
<td>0.001</td>
</tr>
</tbody>
</table>

*Note. Constant = 2.464.*
DISCUSSION

The purpose of the current study was twofold. First, we examined if the lexical and cohesive properties of words found in a listening source text aid in the words’ recall and eventual integration into a speaking response. Second, we investigated if the lexical and cohesive properties of integrated and unintegrated words, along with words found in the response that were not in the source text, were predictive of human judgments of speaking proficiency.

The findings from this study indicate that words integrated into spoken responses from the source text were repeated more often in the source text than words that were unintegrated, were more frequent in the source text than all words, and were used more often in positive connective clauses. Three variables related to these features predicted to an almost perfect accuracy whether words from the source text would be integrated into a test-taker’s response. One variable related to the total incidence in the source text of integrated words accounted for 51% of the variance in the human ratings of integrated speaking proficiency in the cross-validated set. The same variable accounted for 55% of the variance in the combined speaking proficiency scores for the test takers in the cross-validated set.

The results from the discriminant function analysis reveal that the relational and propositional properties of the words in the source text can predict to an almost perfect accuracy whether words will be integrated into a test-taker’s response or not. Two relational variables (Average incidence of word occurrence in source text and Average word use in positive connective clause) with links to text cohesion along with one propositional variable (Average word frequency in source text) with links to lexical difficulty informed our model. These variables, all with strong connections to theories of word recall, show how text-based properties alone can predict which words from a source text are going to be integrated into a spoken response.

In reference to human judgments for both integrated and combined speaking proficiency, the strongest predictor of speaking quality was the total incidence in the source text of integrated words. This finding indicates that if a test taker fails to integrate word types that occur often in the source text, they will receive a lower proficiency score than those test takers who do integrate words that are repeated in the source text. In tandem, the findings from these two analyses indicate that stimuli in the source text (i.e., linguistic features related to cohesion) strongly predict what words test takers will include in their spoken responses. Thus, source texts that repeat important words/concepts more often (i.e., are more cohesive) aid test takers in spotting and reproducing key words/concepts in their responses. In addition, the integration of these words significantly influences human judgments of speaking proficiency, indicating that the reproduction of these key words/concepts by test takers is beneficial. These findings have important implications for item difficulty because listening samples that repeat key words will likely lead to better recall of those words by test takers, and this recall may lead to better evaluations of speaking proficiency by human raters, who likely take into account the dimension of topic development as found in the TOEFL scoring rubric (i.e., topic development in the response is likely best demonstrated through key word use). The findings also suggest that human raters may not be solely evaluating the representative construct (i.e., integrated speaking proficiency) but also recall ability. However, it should be noted that recall ability in an integrated speaking task may be a legitimate element of measuring text integration because, when asked to summarize, the ability to recall the source text should lead to better summarization.
Examples of low and high rater integrated speaking responses are presented below. Table 10 shows the values from these responses for the predictive variable from the integrated and combined score regression models.

Participant 2007 2019 (low performing): Words in **bold** are integrated. Words in *italics* are not found in source text.

Fungus indirectly help **trees** because the fungus feeds and and that bring to be whole help **tree**. Whole **tree** is very **stable** and they will not **blow easily**. Another benefit is that **birds**, Animals are going to, **birds** they will **some bird** will live in their **tree** and animals are going to **leave** their **products**. It the **products** help the **trees** a lot.

Participant 2007 3266 (high performing): Words in **bold** are integrated. Words in *italics* are not found in source text.

The fungus within the **trees** eats up the **bad wood**, and this will **make** the **trunk empty**, and there are **two benefits**. Firstly, it will **make** the **tree stable** because the **trunk** will **be lighter**, and since the **trunk trunk** will **be lighter**, the **roots can anchor** the **tree**, therefore it will be **stable**. And secondly, since it is **hollow** it becomes a **great habitat for animals**, and animals inhabit within the **trees**, they will **produce wastes** which provides **nutrients** for the **tree**.

The strongest predictor of human scores of integrated speaking proficiency was the total incidence in the source text of integrated words. In the example above, the low-performing test taker failed to integrate words common in the source text such as **animals** (occurring 5 times in the source text), **hollow** (occurring 7 times in the source text), and **wood** (occurring 5 times in the source text), whereas the high-performing test taker integrated these words. This failure to integrate common words in the source text is indicative of poor performance in human ratings.

It is also interesting to note that none of the variables calculated by using outside databases or using words in the test-taker response that were not included in the source text were included in the final models reported in these analyses. Many of the lexical variables showed significant differences between integrated and unintegrated words (e.g., SUBTLEXUS word frequency and MRC counts of word familiarity and imageability), but these were not predictive in the DFA models. In addition, the number of words included in a test-taker’s response that were not in the source text, along with word hypernymy scores for words in the response that were not in the source text, both correlated significantly with human ratings of speaking proficiency, but neither were predictive in the subsequent regression models. This seems to indicate that source-specific variables are the most predictive of text integration and judgments of speaking proficiency.
The findings of the current study show that the relational and propositional properties of source texts are strong predictors of text integration and that a single relational property is a strong predictor of human ratings of speaking proficiency. The words integrated into a spoken response can be predicted on the basis of relational and propositional properties of texts, such as the average incidence of word occurrence in the source text, integrated word frequency in the source text, and the use of integrated words in the positive connective clauses in the source text. These indices have strong links to recall and indicate that words that are easier to recall are more likely to be integrated into a response. The relational properties of text are also strong predictors of the scores assigned to integrated responses and combined speaking proficiency. The findings from this study indicated that about 51% of the variance in integrated proficiency ratings in the cross-validated set can be predicted on the basis of one index that measures the incidence in the source text of integrated words. In addition, 55% of the variance in combined proficiency scores in the cross-validated set can be predicted by using the same variable. Overall, these findings suggest that the properties of the source text not only influence which words will be integrated into the response but may also influence human judgments of speaking proficiency. This raises potential concerns about the validity of integrated speaking assessments and the role that stimuli in the input have on eliciting spoken responses (Lee, 2006). However, this concern may be localized only to the listen/speak sections of the TOEFL iBT, which likely place a premium on recall ability. Such recall is likely not as important in integrated listen/speak tasks that ask for personal opinions along with summarization and in independent speaking tasks. However, additional studies are needed to assess this idea. Specifically, the relationship between characteristics in the source text and speaking proficiency scores should be directly investigated by manipulating multiple source texts so that they differ in the type and frequency of relational and propositional properties. Test-taker responses to these source texts should then be investigated to see if the manipulations in the source texts impact human judgments of integrated speaking proficiency and overall speaking proficiency (i.e., does the manipulation of integrated source texts influence the combined speaking scores from the six speaking tasks?).

The findings reported here have important implications for the development of standardized testing as well as instructional materials design. Knowing that text properties strongly influence text integration and knowing that text integration is an important element of human ratings of integrated speaking proficiency, controlling the relational and proposition-specific elements of a text becomes an important criterion when developing assessments, especially if there is more than one form being used and if the forms are meant to be comparable. If a text does not contain enough relational and proposition-specific elements, recall of items will be difficult and human judgments of quality may be lower. Conversely, if a source text contains too many relational and proposition-specific elements, recall will increase, and this may influence judgments of proficiency. If two forms of a text vary in the amount of relational and propositional elements, one form may lead to greater recall and higher proficiency scores than the other. In addition, the findings indicate that the repetition of key words could serve as a measure of content in test-takers’ responses. Thus, the approaches advocated here may prove useful in content analysis, but more empirical testing is necessary to support this contention.

While the findings from this study are strong and appear generalizable (as a result of the cross-validation methods used), we see this study as a first step in addressing the importance
of the source text on text recall and response integration. Because the current study analyzed listening text sources only, the results cannot be generalized to other types of sources like reading-based materials. Thus, future studies should look at larger test-taker samples and examine read/listen/speak tasks as well. Special care should also be taken to control for various test-taker variables such as working memory and note-taking strategies, which are influential in lecture summarization (Carrell, Dunkel, & Mollaun, 2002, 2004; Powers, 1986). Lastly, a large-scale analysis that includes source internal and a wider variety of source external variables is warranted to assess the relative importance of each type of variable on predicting speaking proficiency. Additional variables, such as synonymy or frequency calculations based on academic corpora such as MICASE and BASE, may explain additional variance in future models of integrated speaking quality. Such analyses would help to replicate the findings reported in this study and lead to a greater understanding of the role that source internal variables play in predicting text recall, its subsequent integration, and the effect of this integration on human judgments of integrated speaking quality.

ACKNOWLEDGMENTS

TOEFL® test material are reprinted by permission of Educational Testing Service, the copyright owner.

REFERENCES


O’Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. Discourse Processes, 43(2), 121–152.


