Assessing Writing 18 (2013) 218–238

Contents lists available at SciVerse ScienceDirect

Assessing Writing

Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study

Liang Guo[a], Scott A. Crossley[b,*], Danielle S. McNamara[c]

[a] *Shanghai University of Finance and Economics, Shanghai, People's Republic of China*
[b] *Georgia State University, Atlanta, GA 30312, United States*
[c] *Arizona State University, Tempe, AZ 85287, United States*

ARTICLE INFO

Article history:
Received 21 June 2012
Received in revised form 16 May 2013
Accepted 16 May 2013
Available online 22 June 2013

Keywords:
Integrated writing tasks
Independent writing tasks
Computational linguistics
Writing assessment

ABSTRACT

This study explores whether linguistic features can predict second language writing proficiency in the Test of English as a Foreign Language (TOEFL iBT) integrated and independent writing tasks and, if so, whether there are differences and similarities in the two sets of predictive linguistic features. Linguistic features related to lexical sophistication, syntactic complexity, cohesion, and basic text information were investigated in relation to the writing scores for both integrated and independent samples. The results of this study show that linguistic features can be used to significantly predict essay scores in the integrated as well as the independent writing. When comparing across the two writing tasks, there are both similarities and differences in the two sets of predictive features. For instance, lexical sophistication was found to be a significant predictor for both tasks while features such as verbs in 3rd person singular form and semantic similarity were only significant predictors for the integrated task. These findings demonstrate that evaluation of the two writing tasks rely on similar and distinct features, and are at least partially assessed using different linguistic criteria. Implications of these findings for the assessment of second language (L2) writing are also discussed.

* Corresponding author.
E-mail addresses: guoguo18@gmail.com (L. Guo), scrossley@gsu.edu (S.A. Crossley).

1075-2935/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.asw.2013.05.002

In large scale testing situations, independent writing (i.e., timed, impromptu writing) has been widely used as a measure of second language (L2) academic writing ability. It is generally agreed that compared with indirect writing assessment (e.g., multiple-choice questions), independent writing tasks provide a more valid representation of underlying writing ability because they afford the assessment of writing performance beyond morphological and syntactic manipulation (Camp, 1993; Hamp-Lyons, 1991). Unlike indirect writing assessments, independent writing tasks prompt test-takers to produce an extended written argument built exclusively on their prior knowledge and/or experience. However, concerns have been raised about writing assessments that solely contain independent writing tasks because they risk decontextualizing the writing activity and obtaining an unrepresentative snapshot of the writer's abilities (Hamp-Lyons & Kroll, 1996; Horowitz, 1991). As a consequence, independent writing assessments may under-represent writing proficiency (Camp, 1993; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Weigle, 2002).

To address these concerns, integrated writing tasks (i.e., using reading and/or listening materials as stimuli for composing an essay) have been proposed as a promising alternative for standardized writing tests (Feak & Dobson, 1996; Jennings, Fox, Graves, & Shohamy, 1999; Plakans, 2008; Weigle, 2004). Like academic writing, which is often based on or stimulated by outside sources (Carson, 2001; Cumming et al., 2000; Feak & Dobson, 1996; Leki & Carson, 1997), integrated writing tasks prompt test-takers to respond to source text(s) presented in oral or written format. Thus, integrated writing tasks may more authentically resemble the type of writing that is integral to academic contexts in higher education (Cumming et al., 2000, 2005, 2006; Lewkowicz, 1997; Weigle, 2004).

The use of both independent and integrated writing as a means to assess academic writing skills has been adopted in the new version of the Test of English as a Foreign Language (TOEFL iBT). The underlying goal of including both writing tasks is to enhance the authenticity and validity of English as a Second Language (ESL) writing tests (Cumming et al., 2005, 2006; Huff et al., 2008). In terms of testing validity, the combined use of integrated writing tasks and independent writing tasks can diversify and improve overall measures of writing ability because no single task can be solely reliable to predict the writing ability of a test-taker (Cumming et al., 2005; White, 1994). The use of integrated writing tasks prompts test-takers to respond to source text(s), testing their ability to identify and extract relevant information in the source text(s) and organize and synthesize the information in the response they construct (Cumming et al., 2000; Feak & Dobson, 1996). Independent writing tasks, on the other hand, prompt test-takers to produce an extended written argument built exclusively on their prior knowledge and/or experience. Because of the source materials and the academic nature of integrated writing tasks, integrated essays are expected to be different from independent essays (Huff et al., 2008) and contain more sophisticated language forms (Cumming et al., 2005, 2006; Plakans & Gebril, 2012; Way, Joiner, & Seaman, 2000).

The purported advantages of combining independent and integrated essays into a single standardized test of writing proficiency, however, remain largely asserted. Since there has been little empirical evidence illustrating the differences in writing skills that integrated and independent writing tasks tap into, a clear understanding of how the concurrent use of integrated and independent writing tasks diversifies measurements of writing proficiency is elusive. The few studies that have examined integrated and independent task differences have not compared the construct coverage of the two tasks and have instead mainly focused on differences between integrated and independent writing tasks in terms of writing processes (Plakans, 2007, 2008), scores (Brown, Hilgers, & Marsella, 1991; Delaney, 2008; Esmaeili, 2002; Gebril, 2009; Lewkowicz, 1994; Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012), and inter-rater reliability (Weigle, 2004). Fewer studies have focused on linguistic features in independent and integrated essays and how these features are predictive of task quality (i.e., human ratings of writing quality). Those that have assessed linguistic features have focused mainly on surface-level linguistic features such as word and text length (Cumming et al., 2005, 2006). Therefore, to clarify the construct coverage of integrated and independent writing tasks, validation studies that investigate a more comprehensive set of linguistic features are needed to establish empirically whether and how writing qualities differ between the tasks.

The purpose of this study is to assess the degree to which linguistic features in both integrated and independent essays are predictive of human judgments. Additionally, this study investigates differences between human judgments of integrated and independent essay quality based on these linguistic features. We focus not only on linguistic features at the surface level, but also deeper linguistic levels such as syntactic complexity, lexical sophistication, and text cohesion. Our linguistic features were calculated automatically using the computational linguistic tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012). Our primary research question is whether models of holistic judgments of writing quality for both independent and integrated essays based on linguistic features differ in terms of their underlying linguistic features. The empirical evidence yielded can help to clarify the construct coverage of the two tasks and thus to justify the concurrent use of both integrated and independent in testing situations (Chapelle, Enright, & Jamieson, 2008; Cumming et al., 2005, 2006).

1. Literature review

Because this study focuses on establishing the link between the linguistic features of TOEFL iBT integrated and independent essays and their human assigned scores, previous research on human rating is discussed first. This section then reviews empirical studies on relationship between linguistic features and essay scores with independent and/or integrated essays. Finally, a short account of research on how automated scores reported by *e-rater*® are comparable to human scores is provided.

1.1. Human ratings

Numerous studies have examined the decision-making processes of human raters as a means of investigating writing proficiency. Such an approach assumes that judgments of quality rendered by expert raters directly relates to the underlying writing skills of L2 learners. However, this approach is not without problems. For instance, think-aloud protocols have demonstrated that raters often use different approaches in arriving at their rating decisions (Brown, 1991; ReDemer, 1998; Vaugh, 1991). In addition, raters may vary in a number of different attributes that could affect their decision making processes, including their ESL experience (Brown, 1991; Cumming, Kantor, & Powers, 2002; Song & Caruso, 1996), linguistic and/or cultural backgrounds (Kobayashi & Rinnert, 1996; Shi, 2001), and rating experience (Huot, 1993). However, in terms of the effects of rater attributes and assessments of writing quality, mixed findings have been reported rendering strong conclusions virtually impossible. For instance, in some studies, significant differences between scores were reported for raters from different linguistic backgrounds (Hill, 1997), for raters with or without ESL experience (Song & Caruso, 1996), and for raters with different rating experience (Song & Caruso, 1996). In other studies, significant score differences were not found for raters with different linguistic backgrounds (Connor-Linton, 1995), for raters with or without ESL experience (Brown, 1991), and for raters with different rating experience (Shohamy, Gordon, & Kraemer, 1992). What is conclusive is that rater training improves intra-rater reliability (Weigle, 1994, 1999), and that, with training, human raters often can achieve high reliability with one another (Attali, under review). Further, regardless of rater attributes, all raters report that they attend to textual features such as fluency, complexity and accuracy of language, coherence, content, development, and organization when assessing writing quality (Cumming et al., 2002).

1.2. Independent and integrated essays

When exploring the link between observed scores and writing quality, most studies adopt a writing product oriented approach that focuses on how textual features in L2 writing vary as a function of essay scores. The vast majority of such research has traditionally focused on independent writing and relied mainly on surface level linguistic features. For instance, text length has often been found to be an important indicator of essay scores (Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Reid, 1990). Previous studies have also focused on textual features related to lexical sophistication and syntactic complexity. Such features include lexical diversity (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994; Yu, 2009), average length per word (Frase et al., 1999; Grant & Ginther, 2000; Reid, 1990; Reppen, 1994), subordination (Grant & Ginther, 2000), use of passive voice (Connor, 1990; Ferris, 1994; Grant & Ginther, 2000), and T-units (Song, 2007). These studies have found that higher rated independent essays often include longer words (Frase et al., 1999; Grant & Ginther, 2000; Reid, 1990;

Reppen, 1994), display greater lexical diversity (Crossley & McNamara, 2012; Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994), contain more complicated words such as nominalizations (Connor, 1990), and contain more passive structures (Ferris, 1994). However, some indices, such as T-unit indices, demonstrate no significant relationships with human ratings of writing quality (Song, 2007).

A number of other studies have looked at discourse features such as text cohesion (i.e., links between text elements) and their role in explaining human ratings of L2 writing quality. These studies have demonstrated that more proficient L2 writers tend to produce less cohesive text as measured by word overlap (Engber, 1995; Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). The effects of cohesion resulting from the use of explicit connectives have been less clear. For example, Jin (2001) and Connor (1990) showed that more proficient L2 writers used more connectives and thus produced more cohesive text, but Crossley and McNamara (2012) reported that higher rated L2 texts did not contain significantly more connectives.

Fewer studies have explored how linguistic features predict scores of integrated writing tasks (Cumming et al., 2005, 2006; Gebril & Plakans, 2009; Watanabe, 2001). Those studies that have examined how linguistic features can predict essay scores in integrated writing tasks have shown that higher-rated essays tended to contain more words (Cumming et al., 2005, 2006; Gebril & Plakans, 2009; Watanabe, 2001), a finding that has been repeatedly reported with independent writing tasks (e.g., Frase et al., 1999; Grant & Ginther, 2000; Reid, 1986). Similar to research examining independent writing tasks (Crossley & McNamara, 2012), research investigating integrated writing has also examined type-token ratio (TTR) and word length counts. This research indicates that more proficient L2 writers used a greater diversity of words as evidenced by TTR (Cumming et al., 2005, 2006) but that they did not necessarily use longer words (Cumming et al., 2005, 2006; Gebril & Plakans, 2009). One potential problem with these studies is the authors' reliance on simple TTR indices, which are strongly correlated with text length (McCarthy & Jarvis, 2010). As a result, the authors may not have been assessing lexical diversity, but rather simply the number of words in the essay.

Lastly, research into integrated writing tasks has examined the number of clauses per T-unit or sentence and the number of words per T-unit. This research has reported no statistical relationships between essay scores and the number of clauses per T-unit (Cumming et al., 2005, 2006) or per sentence (Gebril & Plakans, 2009). In reference to the number of words per T-unit, research findings have been inconclusive. Cumming et al. (2005, 2006) reported that higher rated integrated essays contained more words per T-unit while Gebril and Plakans (2009) reported no significant relationship between words per T-unit and integrated essay quality. Unlike studies of independent writing, explanations of integrated writing quality in reference to the effects of cohesion features seem to be missing.

1.3. *Automatic scoring tools*

A variety of automatic scoring tools have been developed to assess essay quality such as such as Summary Street (Kintsch, Coccamise, Franzke, Johnson, & Dooley, 2007), *e-rater*® (Attali & Burstein, 2005), and the Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2000). For the purposes of this study, we will focus on *e-rater*®, which was developed by Educational Testing Services (ETS) and has been used to score TOEFL writing samples. *e-rater*®, like other automatic scoring tools, utilizes natural language processing (NLP) technology to measure dimensions/traits of writing quality to predict human scoring. The current version of *e-rater*® evaluates nine writing features and two content features (Ramineni et al., 2012). The nine writing features include four error features of grammar, usage, mechanics, and style, two organization and development features, two lexical complexity features (average word length and sophistication of word choice), and one feature indicative of good collocation density and preposition use. In addition, two prompt-specific features were developed to assess content with the first showing scores assigned to essays with similar vocabulary and the second illustrating similarity to essays receiving the highest scores. Each feature contains one or more underlying subfeatures. For instance, the mechanics feature includes subfeatures of spelling, capitalization, punctuation, fused, compound, and duplicated words while the style feature contains subfeatures of repetition of words, inappropriate words or phrases. The final *e-rater*® score is calculated using weighted average of feature scores. The weighting of feature scores is either determined by its construct relevance or empirically determined through regression models to predict human essay scores (Quilan, Higgins, & Wolff, 2009). In both cases, organization and development carry the highest weights among all the features included in the calculated e-rater® scores (Attali & Burstein, 2005; Attali, 2007).

e-rater® has demonstrated success in scoring independent and integrated essays written by L2 writers. For independent essays across prompts, the correlation between *e-rater*® scores and human scores for test and retest data (i.e., scores for two essays written by the same test-taker) has been reported as between *r* = .51–55 (Attali & Burstein, 2005), *r* = .56–65 (Attali, 2007), and *r* = .69–70 (Ramineni et al., 2012) depending on the weighting within the models (Attali, 2007) and whether the comparison is made to one or two human raters (Ramineni et al., 2012). For integrated essays across prompts, *e-rater*® scores report a correlation of *r* = .59–70 with human scores (Ramineni et al., 2012). Because of the strength of these correlations, Attali and Burstein (2005) have argued that *e-rater*® measures essentially the same scoring construct that human raters measure.

2. Methods

As discussed previously, there has been little research investigating the linguistic features common to integrated essays and how these features affect human judgments of writing quality. There has been even less work that directly compares independent and integrated writing samples in terms of their respective linguistic features. For those limited studies that have investigated integrated writing, many of the findings have often been inclusive and contradictory or have only focused on surface level features. Additionally, there may have been methodological problems with many of these studies with reference to the linguistic features analyzed (i.e., TTR) and the methodologies employed (i.e., lack of training and test sets: Crossley & McNamara, 2012; Witten, Frank, & Hall, 2011).

In the current study, we use Coh-Metrix (Graesser et al., 2004; McNamara and Graesser, 2012) to explore whether and how linguistic features related to lexical sophistication, syntactic complexity, and cohesion help to characterize L2 writing proficiency in the TOEFL iBT integrated and independent writing tasks, respectively. In addition, we compare differences between integrated and independent essays in order to better understand the task requirements and expectations for each. Our corpus comprises integrated and independent essays selected from the TOEFL iBT. Following Witten et al. (2011) and Crossley and McNamara (2012), each of the corpora was divided into training and test sets. With the training sets, correlations and linear regression were conducted to predict the assigned essay scores using the Coh-Metrix variables. The results from the linear regression models were then extended to the test set.

2.1. *Integrated writing task*

The integrated writing task contained two source texts with the listening and reading materials presenting on opposite effects of fish farming, respectively. The test-takers were required to summarize how the listening passage challenges the reading passage. The reading passage was presented first followed by the listening passage. The test-takers were allowed to take notes on both passages. The test-takers were then given 20 minutes to write an integrated essay. The reading passage was shown on the screen simultaneously when the test-takers were composing. The task recommended a response between 150 and 225 words.

2.2. *Independent writing task*

For the independent writing task, the test-takers were given 30 minutes to write an argumentative essay on the importance of cooperation in today's world as compared to cooperation in the past. The test-takers were expected to use specific reasons and examples to argue for the stance that they chose. The task recommended a minimum of 300 words.

Table 1
Participants by native languages.

Native language	Number	Percentage
Chinese	43	17.9
Spanish	29	12.1
Korean	21	8.8
Japanese	18	7.5
Arabic	14	5.8
German	13	5.4
French	10	4.2
Other languages[a]	92	38.3
Total	240	100

[a] Other languages include all the languages with fewer than 10 test-takers.

2.3. *Corpus collection*

Our selected corpus of integrated and independent essay samples was collected from one administration of TOEFL iBT. The essays were composed by the same group of 240 test takers who were stratified by quartiles for each task. The essays, the final scores, and the demographic information of the test takers were directly provided by ETS. The 240 test-takers included both ESL and English as a foreign language (EFL) learners. They were from a variety of home countries and from more than 10 different linguistic backgrounds (see Table 1 for number of participants sorted by their native languages).

The two types of essays were different in term of length as indicated by the task requirements. Table 2 presents descriptive information related to the text length for the two types of essays.

2.4. *Scoring rubric*

The integrated and independent scoring rubrics can be found at the website of http://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf. Both rubrics describe five levels of writing performance, scored 1–5. For the integrated task, the scoring rubric principally focuses on accurate and coherent presentation of the extracted information in the essays in addition to grammatical accuracy. An integrated essay with a score of 5 should successfully select important ideas from source text(s), present them coherently and accurately, and contain only minor and occasional language errors. An integrated essay with a score of 1, however, provides either little or no relevant content from the source text(s) or is difficult to understand due to the severity of language mistakes.

In the independent scoring rubric, linguistic sophistication at the lexical and syntactic levels is emphasized in addition to the logic and coherence of the arguments along with grammatical accuracy. An independent essay with a score of 5 should be a well-organized and developed response to the given topic, displaying linguistic sophistication and containing only minor language mistakes. An essay with a score of 1, on the other hand, has serious problems in organization, idea development, or language use.

2.5. *Human ratings*

Two expert raters trained by ETS scored each essay using the standardized holistic rubrics. The final holistic score of each essay were the average of the human rater scores if the two scores differed by less than two points. Otherwise, a third rater scored the essay, and the final score was the average

Table 2
Descriptive statistics for the length of the integrated and the independent essays.

Essay type	Mean	S.D.	Minimum	Maximum	Median
Integrated	197.12	50.834	54	388	192
Independent	312.37	77.457	85	592	315

Table 3
Number of test-takers at each score level and descriptive statistics of the scores.

Score	Integrated	Independent
5	35	25
4–4.5	57	66
3–3.5	56	100
2–2.5	50	45
1–1.5	42	4
M	3.148	3.471
S.D.	1.308	.910

of the two closest raters. While inter-rater reliability scores are not provided for the TOEFL iBT scores in the public use dataset, Attali (under review) reported that weighted Kappas for similarly double scored TOEFL writing samples was .70.

The average scores of the independent essays were higher than those of the integrated essays. Pearson correlations indicated that the two sets of scores were highly correlated at $r = .744$ ($p < .001$). Detailed information about the number of test-takers at each score level together with the descriptive statistics of the scores is presented in Table 3.

2.6. *Variable selection*

Text length, lexical sophistication, syntactic complexity, and cohesion features were chosen as construct relevant measures of essay quality to be included in the analysis. Indices related to these features were used to predict the human ratings for the independent and integrated essays. These features were chosen because they have been empirically proven to be correlated with L2 essay scores (e.g., Engber, 1995; Ferris, 1994; Grant & Ginther, 2000), correspond to L2 writing aspects human raters report attending to (Cumming et al., 2002), and are also the features that are meaningfully related to the TOEFL integrated and independent scoring rubrics (see http://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf). Different features under each category were included in the analysis because the development of language proficiency cannot be captured solely by one or two linguistic features and, instead, proficiency should be measured multi-dimensionally (Norris & Ortega, 2009; Tavakoli & Skehan, 2005). The preselected Coh-Metrix indices are presented in Table 4.

In reference to the validity of these indices in terms of L2 writing, these Coh-Metrix variables have previously provided valid assessments of beginning, intermediate, and advanced L2 writing proficiency (Crossley & McNamara, 2012; Crossley, Salsbury, McNamara, 2012) and general L2 language proficiency (Crossley & McNamara, 2011; Crossley, Salsbury, & McNamara, 2009, 2010a, 2010b; Crossley, Salsbury, McNamara, & Jarvis, 2011). Crossley et al. (2012) demonstrated that the lexical and grammatical indices reported by Coh-Metrix could successfully predict low, intermediate, and high proficiency L2 writers, while Crossley and McNamara (2012) demonstrated that the lexical indices in Coh-Metrix could distinguish between L2 writers of various proficiency levels. Automated indices similar to those found in Coh-Metrix have also been used in numerous past research studies that have examined L2 writing quality (e.g., Grant & Ginther, 2000; Jarvis, Grant, Bikowski, & Ferris, 2003) and, more specifically, writing quality as found in the TOEFL (Attali, 2007; Attali & Burstein, 2006; Burstein & Chodorow, 1999; Ramineni et al., 2012). For instance, Grant and Ginther (2000) successfully used an automated tagger that measured lexical, grammatical, and syntactic features to examine proficiency differences among L2 writers and Jarvis et al. (2003) used an automated tagger to examine the lexical, grammatical, and syntactic properties of high quality L2 texts. The *e-rater*® system, which was developed to provide automated assessments of TEOFL writing quality using linguistic measures related to grammar, organization, development, and vocabulary, has also shown strong correlations with human judgments of essay quality for low, intermediate, and high proficiency L2 writers (Attali, 2007; Attali & Burstein, 2006; Burstein & Chodorow, 1999; Ramineni et al., 2012).

In general, such studies demonstrate that computational tools provide reliable and valid estimates of linguistic features related to L2 writing quality. Such tools, though, are not without fault and many

Table 4
Summary of Coh-Metrix indices pre-selected for the regression analysis.

Categories	Coh-Metrix measures	Number of indices
Basic text information		
	Text length	4
Lexical sophistication		
	Average syllables per word	1
	Word hypernymy value	3
	Word polysemy value	1
	Lexical diversity	4
	Word frequency	6
	Word information (word concreteness, familiarity, imageability, & meaningfulness)	8
	Nominalizations	1
Syntactic complexity		
	Number of words before the main verb	1
	Number of higher-level constituents per word	1
	Number of modifiers per noun phrase	1
	Number of embedded clauses	1
	Syntactic similarity	3
	POS tags (lexical categories and phrases)	12
Cohesion		
	Causality	4
	Connectives	3
	Logical operators	1
	Lexical overlap	8
	Semantic similarity (LSA and LSA/given and new)	3
	Tense and/or aspect repetition	4

researchers have reported problems with computer analyses (e.g., Ferris, 1993; Frase et al., 1999; Granger, 2002). In particular, computational tools may be less reliable for use with L2 populations (as compared to L1 populations) in light of common non-typical lexical, grammatical, and syntactic errors. However, as computational tools develop and become more powerful, they provide an accessible and theoretically sound approach for evaluating texts from a quantifiable standpoint (Crossley et al., 2010a). This is especially true for TOEFL test takers whose focus is not specifically on rhetorical concerns, but rather on fluency, control, and sophistication, which can be used to demonstrate both language and writing ability (Deane, 2013; Weigle, 2013). Below we briefly discuss the indices we selected from Coh-Metrix. We refer readers to Crossley and McNamara (2009), Graesser et al. (2004), and McNamara and Graesser (2012) for greater detail.

2.6.1. *Basic text information indices*

Coh-Metrix reports basic textual information for the number and length of words, sentences, and paragraphs per text and the number of sentences per paragraph. Number counts relate to fluency while length counts relate to sophistication (i.e., longer words are usually less frequent and longer sentences are usually more syntactically complex).

2.6.2. *Lexical sophistication*

Coh-Metrix evaluates lexical sophistication of a given text by calculating syllables per word, lexical hypernymy and polysemy values, lexical diversity, word frequency, psycholinguistic word properties, and nominalizations.

Coh-Metrix calculates hypernymy and polysemy values for all words in a given text that have entries in WordNet (Fellbaum, 1998). The hypernymy values are calculated by counting the number of levels that is above a word in a conceptual taxonomic hierarchy. Words with more hypernymy levels tend to be more precise in signaling the intended meaning and less ambiguous than those with fewer levels. Polysemy values are calculated by counting the number of senses for each word, which indicates the lexical ambiguity of a given text.

Coh-Metrix estimates lexical diversity using two indices: MTLD (McCarthy & Jarvis, 2010) and D (Malvern & Richards, 1997; Jarvis, 2002). These indices are different from traditional measures of lexical diversity such as TTR because they avoid the problematic correlation with text length (Crossley & McNamara, 2012). A high lexical diversity score means that the given text contains a wider range of words and thus demonstrates greater lexical sophistication.

Coh-Metrix word frequency counts are based on CELEX (Baayen, Piepenbrock, & van Rijn, 1993), which consists of word frequencies taken from the early version of the COBUILD corpus of 17.9 million words. A higher word frequency score indicates that the input text contains more frequent words and is, thus, less lexically sophisticated.

Coh-Metrix reports word properties for concreteness, familiarity, imageability, and meaningfulness using human ratings provided by the Medical Research Council Psycholinguistic Database (MRC; Wilson, 1988). Concrete words are more tangible than abstract words while familiar words are more recognizable and frequent. Imageability indicates whether a word can easily evoke a mental image and meaningfulness relates to the number of associations a word has with other words (Toglia & Battig, 1978).

Lastly, Coh-Metrix computes the number of nominalizations in a text, which refer to abstract generic nouns that are derived from another part of speech via the addition of derivational morphemes (e.g., *-ment*, *-tion*, *-lity*, *-ness*; Biber, 1988). The higher the normalization score is, the more sophisticated the words of the given text are.

2.6.3. Syntactic complexity

Coh-Metrix measures syntactic complexity using five indices related to the number of words before the main verb, number of higher-level constituents per word, number of modifiers per noun phrase, syntactic similarity, and number of embedded clauses. Coh-Metrix also reports syntactic categories for words (i.e., part of speech tags). The more words there are before the main verb, the more complex the sentence tends to be structurally. Sentences with difficult syntactic composition tend to have a higher ratio of high-level constituents per word than sentences with less complicated structure. The number of modifiers per noun phrases indicates how compressed the sentence structure is and signals the density of the information (Biber & Gray, 2010). The syntactic similarity index reported by Coh-Metrix compares the syntactic tree structures of sentences. A higher syntactic similarity score means a higher degree of similarity in syntactic structure of two adjacent sentences or among all sentences within a paragraph or a text and, thus, less syntactic variation (Crossley & McNamara, 2012). Coh-Metrix also reports on the number of embedded clauses of a given text as another measure of syntactic complexity. The higher the number is, the more complex the syntactic structure of the given text is as compared to one mainly containing simple sentences without embedding. Lastly, Coh-Metrix also generates frequency data for the major syntactic categories found in the Pen Treebank Tag Set. Phrasal level complexity, as measured by syntactic categories, is a strong index of language proficiency at the advanced level (Biber, 2006; Halliday & Martin, 1996; Norris & Ortega, 2009; Ortega, 2003).

2.6.4. Cohesion

Textual cohesion consists of linguistic devices that play a role in building links between ideas in a given text. These devices are, therefore, important in text processing and comprehension (Graesser, McNamara, & Louwerse, 2003; Halliday & Hasan, 1976). Coh-Metrix reports cohesion by examining causality, connectives, logical operators, lexical overlap, semantic similarity, and tense and/or aspect repetition in text.

Causality in Coh-Metrix is measured by calculating the number of causal verbs, causal particles (such as *as a result*, *because*), and causal connectives, which reflect the extent to which sentences are linked in a text. Connectives are mainly used to create links between ideas and clauses (Halliday & Hasan, 1976). Connective indices reported by Coh-Metrix include different types of cohesion such as causal connectives (e.g., *because*, *so*, *consequently*) and logical connectives (e.g., *or, actually, if*). Coh-Metrix also reports incidence counts for logical operators (*or*, *and*, *not*, and logical connectives), which differ from logical connectives in their relation to logical reasoning (as compared to connecting text segments). Coh-Metrix also reports four forms of lexical overlap between sentences: noun overlap, argument overlap (nouns, stems, and pronouns), stem overlap (nouns and stems, but not pronouns), and content word overlap. To assess semantic similarity between text segments (i.e., sentences and paragraphs), Coh-Metrix utilizes LSA. Coh-Metrix also estimates the proportion of new information each sentence provides by using LSA. Lastly, Coh-Metrix computes tense repetition, aspect repetition, and the combination of aspect and tense repetition in order to measure a text's temporal cohesion.

2.7. Statistical analysis

We first divided the two corpora (independent and integrated essays) into training and test sets following a 67/33 split (Witten et al., 2011). Thus, for both sets of 240 essays, we had a training set of 160 essays and a test set of 80 essays. For each set, we first conducted Pearson correlations to assess relationships between the selected variables and the human scores using the training set only. Those variables that demonstrated significant correlations with the human scores were retained as predictors in a subsequent regression analysis. Prior to inclusion, all significant variables were checked for multicollinearity to assure that the variables were not measuring similar constructs. Our cut-off for multicollinearity was $r \geq .70$. If two or more indices were highly correlated with each other, we selected the index with the highest correlation to the human raters for inclusion in the regression and removed the other, redundant variables. We next conducted a stepwise regression analysis using the training set only.

The selected indices were regressed against the holistic scores for the 160 essays with the essay scores as the dependent variable and the Coh-Metrix indices as the predictor variables. The derived regression model was then applied to the essays in the test sets to predict the scores.

3. Results

3.1. Integrated essays

3.1.1. Correlations training set

Correlations were conducted between the Coh-Metrix indices and the human scores for the 160 integrated essays in the training set. Nineteen Coh-Metrix indices demonstrated significant correlations with the human scores and did not demonstrate multicollinearity. Table 5 presents the 19 selected indices along with their *r* and *p* values in order of the strength of the correlation.

Table 5
Selected Coh-Metrix indices for regression analysis of the integrated essays: training set.

Coh-Metrix indices	Category	*r* value	*p* value
Number of words per text	Basic text information	.513	<.001
Word familiarity (content words)	Lexical sophistication	−.440	<.001
Past participle verbs	Syntactic complexity	.437	<.001
Word frequency (content words)	Lexical sophistication	−.436	<.001
Verbs in base form	Syntactic complexity	−.403	<.001
Nominalizations	Lexical sophistication	.357	<.001
Hypernymy values (nouns)	Lexical sophistication	.351	<.001
Verbs in non-3rd person singular present form	Syntactic complexity	−.344	<.001
Personal pronouns	Syntactic complexity	−.315	<.001
Semantic similarity (LSA sentence to sentence)	Cohesion	.296	<.001
Number of modifiers per noun phrase	Syntactic complexity	.264	<.050
Word concreteness (content words)	Lexical sophistication	.225	<.050
Number of sentences per text	Basic text information	.218	<.050
Noun overlap	Cohesion	.217	<.050
Verbs in 3rd person singular present form	Syntactic complexity	.194	<.050
Gerund or present participle verbs	Syntactic complexity	.186	<.050
Tense repetition	Cohesion	.174	<.050
Prepositional phrases	Syntactic complexity	.168	<.050
Verbs in past tense	Syntactic complexity	−.165	<.050

Table 6
Regression analysis findings to predict the integrated essay scores: training set.

Entry	Coh-Metrix index added	*r*	r^2	*B*	B	S.E.
Entry 1	Number of words per text	.513	.264	.009	.378	.001
Entry 2	Past participle verbs	.647	.419	.021	.258	.005
Entry 3	Word familiarity (content words)	.710	.504	−.055	−.206	.018
Entry 4	Verbs in 3rd person singular present form	.738	.545	.009	.133	.004
Entry 5	Semantic similarity (LSA sentence to sentence)	.747	.559	2.015	.146	.757
Entry 6	Verbs in base form	.756	.572	−.011	−.136	.005
Entry 7	Word frequency (content words)	.764	.584	−1.348	−.142	.651

Notes: *B* = unstandardized β; B = standardized; S.E. = standard error. Estimated constant term is 34.580.

3.1.2. Regression analysis training set

A stepwise regression analysis using the 19 indices as the independent variables to predict the human scores yielded a significant model, $F(1, 152) = 30.446$, $p < .050$, $r = .764$, $r^2 = .584$. Seven Coh-Metrix indices were included as significant predictors of the essay scores. The seven indices were: *number of words per text, past participle verbs, word familiarity (content words), verbs in 3rd person singular present form, semantic similarity (LSA sentence to sentence), verbs in base form*, and *word frequency (content words)*.

The model demonstrated that the seven indices together explained 58.4% of the variance in the evaluation of the 160 integrated essays in the training set (see Table 6 for additional information). *T*-test information for the seven indices together with the amount of variance explained are presented in Table 7.

3.1.3. Regression analysis test set

We used the model reported for the training set to predict the human scores in the test set. To determine the predictive power of the seven variables retained in the regression model, we computed an estimated score for each integrated essay in the independent test set using the B weights and the constant from the training set regression analysis. This computation gave us a score estimate for the essays in the test set. A Pearson's correlation was then conducted between the estimated score and the actual score assigned on each of the integrated essays in the test set. This correlation together with its r^2 was then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

The regression model, when applied to the test set, reported $r = .730$, $r^2 = .533$. The results from the test set model demonstrated that the combination of the seven predictors accounted for 53.3% of the variance in the assigned scores of the 80 integrated essays in the test set, providing increased confidence for the generalizability of our model.

3.2. Independent essays

3.2.1. Correlations training set

Correlations were conducted between the Coh-Metrix indices and the human scores for the 160 essays in the training set. Twenty-one Coh-Metrix indices demonstrated significant correlations with

Table 7
t value, *p* values, and variance explained of the seven significant indices for the integrated essay scores: training set.

Coh-Metrix indices	*t* value	*p* value	r^2
Number of words per text	6.964	<.001	.264
Past participle verbs	4.176	<.001	.156
Word familiarity (content words)	−3.080	<.050	.085
Verbs in 3rd person singular present form	2.193	<.050	.041
Semantic similarity (LSA sentence to sentence)	2.662	<.050	.014
Verbs in base form	−2.081	<.050	.013
Word frequency (content words)	−2.071	<.050	.012

Table 8
Selected Coh-Metrix indices for regression analysis of the independent essays: training set.

Coh-Metrix indices	Categories	*r* value	*p* value
Number of words per text	Basic text information	.691	<.001
Nominalizations	Lexical sophistication	.521	<.001
Noun hypernymy values	Lexical sophistication	.475	<.001
Past participle verbs	Syntactic complexity	.464	<.001
Verbs in non-3rd person singular present form	Syntactic complexity	−.441	<.001
Word familiarity (all words)	Lexical sophistication	−.419	<.001
Lexical diversity *D*	Lexical sophistication	.415	<.001
Word meaningfulness (all words)	Lexical sophistication	−.365	<.001
Embedded clauses	Syntactic complexity	−.339	<.001
Number of modifiers per noun phrase	Syntactic complexity	.337	<.001
Average syllables per word	Lexical sophistication	.309	<.001
Aspect repetition	Cohesion	−.308	<.001
Personal pronouns	Syntactic complexity	−.297	<.001
Word frequency (content words)	Lexical sophistication	−.295	<.001
Content word overlap	Cohesion	−.289	<.001
Verbs in base form	Syntactic complexity	−.281	<.001
Conditionals connectives	Cohesion	−.245	<.050
Number of paragraphs per text	Basic text information	.209	<.050
Word polysemy values	Lexical sophistication	−.170	<.050
Word concreteness (content words)	Lexical sophistication	.167	<.050
Word imageability (all words)	Lexical sophistication	−.156	<.050

Table 9
Regression analysis findings to predict the scores of independent essays: training set.

Entry	Coh-Metrix index added	*r*	r^2	*B*	B	S.E.
Entry 1	Number of words per text	.691	.478	.007	.577	.001
Entry 2	Average syllables per word	.753	.568	1.511	.179	.448
Entry 3	Noun hypernymy values	.785	.616	.359	.199	.094
Entry 4	Past participle verbs	.800	.641	.016	.165	.005
Entry 5	Conditional connectives	.807	.650	−.020	−.104	.010

Notes: *B* = unstandardized β; B = standardized; S.E. = standard error. Estimated constant term is −3.097.

the human scores and did not demonstrate multicollinearity. Table 8 presents the 21 selected indices along with their *r* and *p* values in order of the strength of the correlation.

3.2.2. *Regression analysis training set*

A stepwise regression analysis using the 21 selected indices to predict the variance in the essay scores was conducted for the training set of 160 independent essays. The regression yielded a significant model, $F(1, 154) = 57.325$, $p < .001$, $r = .807$, $r^2 = .650$. Five Coh-Metrix indices were significant predictors in the regression: *number of words per text, average syllables per word, noun hypernymy, past participle verbs*, and *conditional connectives*.

The model demonstrated that the combination of the five variables accounted for 65.0% of the variance in the evaluation of the training set of 160 independent essays (for additional information see Table 9). Table 10 presents *t*-test information of the five indices that were retained in the regression model and the variance explained by each index.

Table 10
t value, *p* values, and variance explained of the five significant indices for the independent essay scores: training set.

Coh-Metrix indices	*t* value	*p* value	r^2
Number of words per text	11.194	<.001	.478
Average syllables per word	3.376	<.050	.090
Noun hypernymy values	3.837	<.001	.048
Past participle verbs	2.992	<.050	.025
Conditional connectives	−2.071	<.050	.010

3.2.3. Regression test set

The regression model, when extended to the test set of the independent essays, yielded $r = .758$, $r^2 = .574$, demonstrating that the combination of the five significant predictors identified in the training set regression model accounted for 57.4% of the variance in the human scores assigned to the 80 independent essays in the test set, providing increased confidence for the generalizability of our model.

4. Discussion

The regression analyses provide evidence that linguistic features can significantly predict holistic evaluations of writing quality for the TOEFL iBT integrated and the independent essays. The analyses also demonstrated that the models established on the training sets can be extended to the independent data sets (the test sets), achieving similar predictive accuracy and demonstrating the strength of the models to predict human scores in independent data sets. Thus, the results of the study lend reliable support to the notion that linguistic features can significantly predict essay scores for both of the writing tasks. Furthermore, the findings help to demonstrate that the two tasks elicit different scoring patterns on the part of the expert raters, illustrating that the construct coverage of the two tasks does not fully overlap. The results provide empirical evidence for the validity arguments for the concurrent use of both integrated and independent in testing situations (Chapelle et al., 2008). We discuss the predictive ability of our models for both integrated and independent essays below followed by a discussion of how the results relate to the *e-rater*® predictors as well as the differences in scoring between the two tasks as reflected in the human ratings.

4.1. Integrated essays

Similar to previous studies on integrated writing (Gebril & Plakans, 2009; Watanabe, 2001), this study demonstrated that textual length has a large effect on the essay scores assigned (defined as Pearson's correlations ≥.50, Cohen, 1988). Longer essays were scored higher. In fact, as shown in Table 7, text length was the strongest predictor of essays quality among the seven indices that were retained in the regression model, accounting for 26.4% of the variance of the human scores for the integrated essays. Although text length alone cannot signify good writing quality, many of the features of highly scored essays (e.g., substantial supporting details and idea development) are difficult to embed in a shorter essay (Chodorow & Brustein, 2004) and that fluency of text production skills is an important element in perceived writing quality (Chenoweth & Hayes, 2001; Ransdell & Levy, 1999). Furthermore, in studies on power of *e-rater*® features in predicting human scores, text length has also been found to be highly correlated with features of organization and development such as the number and length of discourse units (Attali, 2007; Attali & Powers, 2008).

The next strongest predictor for integrated essays was past participle verbs with higher rated integrated essays containing a greater incidence. Past participle verbs are normally used to construct passive voice or to indicate present or past aspect. A closer examination of the essays revealed that the past participle verbs in integrated writing often occurred in the construction of passive voice. Examples from the integrated essays include:

> The professor supports that species of fish that are **used** to feed the farm-raised fish are usually not **eaten** by people.
> Humans are "**exposed** to harmful or unnatural long-term effects" when consuming farm-raised fish, which are **fed** with growth-inducing chemicals.

Since passive voice is one of the markers for formal academic writing style (Hinkel, 2002), this finding suggests that the higher rated integrated essays include more linguistic devices that are characteristic of general academic writing. Although not included in the final regression model, the significant positive correlation between nominalizations and the integrated essay scores (see Table 5) also confirms that higher rated integrated essays bore more resemblance to formal academic writing than the lower rated ones.

Our third strongest predictor was word familiarity. Word familiarity, as a measure of lexical sophistication, explained 8.5% of the variance in the human scores. Another index of lexical sophistication, word frequency, which was the lowest significant predictor, explained 1% of the variance. In both cases, test-takers who used less frequent and less familiar words received higher scores on their integrated essays. While these results are in line with previous empirical research on writing quality (e.g., Nation, 1988), it is interesting to note that the scoring rubric for integrated essays contains no component related to word sophistication in the responses. The findings, therefore, illustrate a phenomenon discussed by Lumley (2005): even expert raters likely attend to many features beyond what is included in the scoring rubric.

Our next strongest predictor was the use of 3rd person singular verbs with higher quality essays characterized by significantly more verbs in 3rd person singular present form. The frequent use of 3rd person singular form is likely related to citing sources (referring to the article or the author) and staying on topic (in this case, staying on the topic of fish farming which would require the use of the singular 3rd person form rather than focusing on farmers or consumers in general which would require the use of 3rd person plural form). Thus, this finding likely indicates that higher rated essays contained more occurrences of correctly marked verbs for citing the source and staying on the topic while, at the same time, conveying expected information in a detached manner (i.e., without using first or second person pronouns).

An index of semantic similarity calculated through LSA values was also a significant predictor of integrated essay quality. The semantic similarity index indicated that conceptual similarity between adjacent sentences is a significant predictor of essay quality with higher rated integrated essays having a higher conceptual similarity than essays that were judged to be of a poorer quality. Such a finding indicates that maintaining semantic cohesion throughout an integrated essay is an important component of writing quality.

Our last predictor of integrated essay quality was verbs in the base form. The analysis demonstrated that essays including a greater incidence of verbs in base form received lower scores. The low rated integrated essays that scored high on this index were pulled out from the corpus of essays for further examination. A qualitative analysis of low rated essays that contained a high proportion of verbs in base form showed that the majority of verbs in base form were actually grammatical errors wherein the test-takers failed to correctly indicate the subject of the sentence or did not provide the correct suffixes for the verbs. The following example is provided to illustrate this finding:

> . . .because the fishes from the fish farm aren’t **produce** to release into the wild, but rather for commercial purposes.

This analysis suggests that, in terms of verb forms, the integrated essays that contained more grammatical errors were rated lower, indicating that grammatical accuracy plays a role in the evaluation of the integrated writing, which is consistent with findings from Cumming et al. (2005, 2006) and Gebril and Plakans (2009).

In general, this analysis demonstrates that linguistic features do vary with the scores in the integrated writing. The regression analysis showed that in the integrated task, writing quality was partially determined by the number of words written, whether the expected content was presented (as evidenced by verbs in 3rd person singular present form), the level of lexical sophistication, whether the information was presented cohesively (as evidenced by semantic similarity), and grammatical expectation (in terms of verb forms).

4.2. Independent essays

Similar to previous studies on independent writing (e.g., Ferris, 1994; Frase et al., 1999; Reid, 1990) as well as the analysis of the integrated essays, the analysis of the independent essays also indicated that text length is the most significant predictor of the essay scores, accounting for 47.8% of the variance. As shown in Table 10, text length was actually the strongest predictor among the five indices that were retained in the regression model. Similar to the integrated essays, longer independent essays were also rated higher.

Average syllables per word and noun hypernymy values were the next two strongest predictors of human ratings of the independent essays. In accordance with previous studies examining lexical properties in relation to writing quality (Crossley & McNamara, 2012; Frase et al., 1999; Yu, 2009), these findings suggest that lexical sophistication is integral to predicting human judgments of essay quality. These findings also provide support for the scoring rubric of the independent writing. As described in the rubric, the test-takers who were classified as more proficient writers used more words that display a high level of sophistication (as evidenced by being less specific and frequent) as compared to those who received lower scores.

Similar to the findings made in the integrated essays, past participle verbs were also found to be a significant predictor of the essay scores and were positively correlated with the independent essay scores. The test-takers who were judged to be more proficient produced significantly more cases of past participle verbs in comparison to those who were judged to be less proficient. As mentioned earlier, the more frequent use of past participle verbs in the test-takers' essays was correlated with the use of passive voice, a feature of general academic writing (Hinkel, 2002). This finding indicates that more proficient writers employed more passive voice structures in their essays than the writers who were judged to be less proficient.

Our last predictor of independent essay quality was conditional connectives, a cohesion index. The negative correlation between the conditional connectives and the essay scores (see Table 10) demonstrated that essays containing fewer cases of conditional connectives obtained higher scores. This finding indicates that the test-takers who were rated to be more proficient actually produced essays with fewer cohesive devices. Additional support for this notion can be found in the correlation analysis (see Table 8). For instance, the essays composed by the more proficient test-takers not only included fewer conditional connectives but also had lower scores for two other cohesive devices: aspect repetition and content word overlap. This particular finding counters previous L2 studies (e.g., Connor, 1990; Jin, 2001) that found that high proficiency writers produce more cohesive devices in their writing when compared to those with lower proficiency (Connor, 1990; Jin, 2001). Similar findings have been reported in L1 and L2 writing studies that have utilized Coh-Metrix indices (Crossley & McNamara, 2010, 2011, 2012). One possible explanation is *a reverse cohesion effect* in which more proficient writers assume that their readers are high knowledge readers that benefit more from less cohesion texts and, as a result, produce fewer cohesive devices (McNamara, Kintsch, Songer, & Kintsch, 1996).

4.3. Relation to e-rater® predictors

With *e-rater®*, discourse features (development and organization) accounted for the biggest proportion of score variance (Attali, 2007). As previously mentioned, since discourse features are highly correlated with text length, it is not difficult to understand why in the current study, text length was found to be the strongest predictor of the essay scores. The second similarity is that lexical sophistication was identified as an aspect of writing that contributed significantly to the independent essay scores. However, unlike *e-rater®*, our analysis did not demonstrate that syntactic complexity features were significant predictors of essay scores for either the integrated or independent essay scoring. A potential reason for this finding could be that syntactic features cannot strongly differentiate writing proficiency at the advanced level (Norris & Ortega, 2009). Since most of TOEFL test takers are applicants for undergraduate and graduate programs, they have likely studied English for many years making many advanced English learners. In much the same way as the syntactic complexity indices, phrasal level complexity (number of modifiers per noun phrase) was not a significant predictor of essay scores although phrasal level complexity is a feature of advanced academic writing skill (Biber & Gray, 2010). Thus, while TOEFL test takers might be advanced English learners, they are unlikely to be advanced academic writers. Lastly, cohesion was found to be a significant predictor of the essay scores in the current study, but it was not included in the feature set chosen by *e-rater®*.

A comparison between the models reported in this analysis and the models reported in *e-rater®* literature demonstrate similar levels of success in explaining the variance found in human ratings. *e-rater®* has reported correlations ranging from $r = .51–.70$ for independent essays across prompts. Our model for independent essays scoring reported $r = .730$, but this was a prompt specific model.

Table 11
Significant predictors for integrated and independent essay scores.

Coh-Metrix indices	Category	Integrated	Independent
Number of words per text	Basic text information	Yes	Yes
Past participle verbs	Syntactic complexity	Yes	Yes
Word familiarity (content words)	Lexical sophistication	Yes	No
Verbs in 3rd person singular present form	Syntactic complexity	Yes	No
Semantic similarity (LSA sentence to sentence)	Cohesion	Yes	No
Verbs in base form	Syntactic complexity	Yes	No
Word frequency (content words)	Lexical sophistication	Yes	No
Average syllables per word	Lexical sophistication	No	Yes
Noun hypernymy values	Lexical sophistication	No	Yes
Conditional connectives	Cohesion	No	Yes

Likewise, *e-rater*® reports correlations of $r = .59–.70$ with human ratings of integrated essays across prompts. Our model reported a correlation of $r = .758$, but this was also prompt specific.

4.4. Comparison between integrated and independent models

A direct comparison of the significant predictors for the integrated and the independent essays is presented in Table 11. Several similarities are evident when comparing the predictive indices of the integrated essays with those of the independent essays. For both tasks, higher rated test-takers tended to write longer essays, to use more past participle verbs, and use less frequent words (i.e., to use less frequent and familiar words in the case of integrated essays and use words with fewer syllables in the case of independent essays). These linguistic features are thus likely indicators of general writing ability as perceived by expert raters.

The integrated and independent predictors also exhibited differences. For example, the integrated essays that were rated higher demonstrated greater semantic similarity (LSA sentence to sentence similarity), but semantic similarity was not a significant predictor of independent essay quality. Thus, conceptual similarity between sentences is an important element of human judgments of integrated, but not independent essay writing. In addition, a greater number of 3rd person singular present forms and fewer verbs in the base form were significant predictors of the integrated essay scores but not of the independent essay scores, indicating that integrated writing benefits from information conveyed in a detached manner. This difference between the two sets of essays might have to do with the notion that the two types of writing are of different genres. The independent writing task calls for a lengthy argumentative essay relying on writers' personal opinion and life experiences. The integrated writing task, on the other hand, provides source materials and is a compare/contrast essay that is more academically oriented. Due to the differences in genre and the presence of source texts, test takers, in composing integrated essays, may tend to model the structures of the source texts (Plakans & Gebril, 2012) leading to a more detached style. Furthermore, the features of verb forms also suggest that grammatical mistakes (i.e., not correctly marked verb forms) are predictive elements of integrated writing but not for the independent essays. This may indicate that raters assess grammatical mistakes more severely when source materials are provided because content is readily available, unlike in independent essays where writers have to struggle with content and language simultaneously.

In the case of independent essays, fewer conditional connectives were significant predictors of essay scores. Thus, in judgments of independent writing quality, cohesion appears to negatively affect quality. Support for this notion can be found in the correlational evidence (see Table 8) in which other cohesion devices (e.g., aspect repetition and content word overlap) demonstrated a negative correlation with essay quality. This is unlike judgments of integrated essay quality in which cohesive devices were positively correlated with essay quality (e.g., semantic similarity, noun overlap, and tense repetition). Thus, cohesion seems to be an important property of human ratings of integrated essays, but not independent essays. This finding is likely the result of text integration, which requires the writer to maintain links between similar ideas taken from different sources. Alternatively, integrated essays may be more expository in nature and thus be more cohesive (Crossley, 2013). Lastly, an index of noun hypernymy was a significant predictor of independent essay quality, but not integrated essay quality. The index was positively correlated with human judgments of essay quality indicating that independent essays with less specific words were scored higher (i.e., abstract words). Such a finding may relate to the more abstract and generalized prompts used in independent writing samples.

The similarities between the tasks indicate that text length, use of past participle verbs (mainly used to construct passive voices), and word frequency are integral to distinguishing the score levels, regardless of the task type. Although not specified in the scoring rubrics, these features seem to be attended to by raters when deriving judgments of writing quality and thus demonstrate overlapping construct coverage for the two sets of scores. The differences in cohesive devices, verb forms, and lexical features demonstrate that evaluations of the two tasks pinpoint different elements of writing proficiency. In grading the integrated essays, as specified in the scoring rubric, greater importance was attached to whether the expected information was presented by the test-takers in a cohesive and detached way. In contrast, independent writing samples benefit from less cohesive writing that is less specific. These differences indicate that the two writing tasks also tap into different elements of writing quality and, therefore, the construct coverage of the two writing tasks do not completely overlap. These findings help to substantiate the notion that the concurrent use of the integrated and the independent tasks in the TOEFL iBT writing section diversify the measurements of writing ability, thus helping justify the inclusion of both tasks.

5. Conclusion

This study provides evidence that linguistic features can predict human ratings of the essays for the TOEFL iBT independent and integrated writing tasks. The findings indicate that the integrated and the independent writing tasks share construct coverage and, at the same time, tap into different elements of writing, thus justifying the combined use of the two tasks in a single test. The findings, from a product perspective, also help to validate the scoring rubrics by verifying that some of predictors are meaningfully related to the writing aspects specified in the rubrics. Finally, the findings also complement many of the features used in *e-rater*®.

Unlike other L2 studies that have investigated linguistic features in relation to human ratings of essay scores (e.g., Cumming et al., 2005, 2006; Engber, 1995; Gebril & Plakans, 2009; Grant & Ginther, 2000), this study utilized the advanced computational tool, Coh-Metrix, which not only looked into surface level linguistic features, but also deeper level linguistic features. Additionally, the large quantity of TOEFL iBT essays from real administrations allowed us to not only use more rigorous statistical methodology (training and test sets) to control for issues like over-fitting, but also added to the task validity of the findings. Given that the models established on the training sets can also be extended to the test sets (the independent data set) with similar predictive accuracy, we have more confidence that the findings are generalizable, at least for the test-taker population investigated. Follow up studies should investigate a greater number of prompts with the understanding that test-takers might demonstrate a varied use or different types of writing performance when responding to different prompts/source texts (Yang, 2009). Additionally, future studies should give more attention to grammatical accuracy of the writing products. As revealed in previous studies on integrated writing, grammatical accuracy often exerts an important influence on the score assigned (Cumming et al., 2005, 2006). Due to the limitations of Coh-Metrix, grammatical accuracy was not directly examined. Given that grammatical errors tend to be one of the characteristics of L2 writing (Frase et al., 1999), such information may be vital to a better understanding of human judgments of writing proficiency.

Acknowledgments

This research was supported in part by the English Testing Services and the International Research Foundation for English Language Education (TIRF). Any opinions, findings and conclusions or recommendations made in this paper are those of the authors and do not necessarily reflect the views of the ETS and the TIRF. This research was also supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

Attali, Y. (2007). *Construct validity of e-rater® in scoring TOEFL essays (ETS Research Report No. RR-07-21)*. Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater® v.2.0 (ETS Research Report No. 04-45)*. Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4* (3), 3–30.

Attali, Y., & Powers, D. (2008). *A developmental writing scale (ETS Research Report No. RR-08-19)*. Princeton, NJ: Educational Testing Service.

Attali, Y. (2013). *e-rater® performance for TOEFL iBT Independent essays*. (under review)

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (Eds.). (1993). *The CELEX lexical database (CD-ROM)*. Philadelphia, PA: Linguistic Data Consortium.

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

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

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*, 2–20.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*, 587–603.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communications*, *8*, 533–556.

Burstein, J., & Chodorow, M. (1999). Automated Essay Scoring for nonnative English speakers. In: *Proceedings of the ACL99 workshop on computer-mediated language assessment and evaluation of natural language processing*.

Camp, R. (1993). Changing the model for the direct writing assessment. In: M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45–78). Cresskill, NJ: Hampton Press Inc.

Carson, J. (2001). A task analysis of reading and writing in academic contexts. In: D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading–writing connections* (pp. 48–83). Ann Arbor, MI: The University of Michigan Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In: C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 145–186). NY: Routledge.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, *18* (1), 90–98.

Chodorow, M., & Brustein, J. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL essays (TOEFL Research Report No. 73)*. Princeton, NJ: Educational Testing Service.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Connor, U. (1990). Linguistic/rhetorical measures of international persuasive student writing. *Research in the Teaching of English*, *24*, 67–87.

Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, *14*, 99–115.

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in second language writing. *Journal of Second Language Writing*, *17* (2), 119–135.

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

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

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of semantic relations in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics*, *7*, 55–74.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, *21* (2/3), 170–191.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing*, *28* (4), 561–580.

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

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices? *Language Testing*, *29* (2), 240–260.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, *46* (2), 256–271.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while scoring ESL/EFL compositions: A descriptive model. *Modern Language Journal*, *86*, 67–96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper (TOEFL Monograph Series, Report No. 18)*. Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in writing-only and reading-to-write prototype tasks for next generation TOEFL. *Assessing Writing*, *10*, 5–43.

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

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7–24.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*, 140–150.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4* (2), 139–155.

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *The Canadian Modern Language Review*, *58* (4), 599–622.

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

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

Ferris, D. R. (1993). The design of an automatic analysis program for L2 text research: Necessity and feasibility. *Journal of Second Language Writing*, *2* (2), 119–129.

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*, 414–420.

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written english (TOEFL Research Report No. 64)*. Princeton, NJ: ETS.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Journal of Language Testing*, *26*, 507–531.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *7*, 47–84.

Granger, S. (2002). A bird's-eye view of learner corpus research. In: S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Philadelphia, PA: John Benjamins.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In: A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford.

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

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123–145.

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

Halliday, M. A. K., & Martin, J. R. (1996). *Writing science: Literacy and discursive power*. London: Falmer Press.

Hamp-Lyons, L. (1991). Introduction. In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 1–4). Norwood, NJ: Ablex.

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

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In: A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LRTC 96* (pp. 275–290). Jyvaskyla, Finland: University of Jyvaskyla.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 71–86). Norwood, NJ: Ablex.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S., & Schedl, M. (2008). Prototyping a new test. In: C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 187–225). NY: Routledge.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In: M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206–236). Creskill, NJ: Hampton.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*, 57–84.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*, 377–403. http://dx.doi.org/10.1016/j.jslw.2003.09.001

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-taker's choice: An investigation of the effect of topic on language test performance. *Language Testing*, *16* (4), 426–456.

Jin, W. (2001). *A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels (ERIC Document Reproduction Service No. ED 452 726)*.

Kintsch, E., Coccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary street: Computer-guided summary writing. In: T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 263–278). Mahwah, NJ: Lawrence Erlbaum Associates.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical patterns and readers' background. *Language Learning*, *46*, 397–437.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*, *15*, 27–31.

Leki, I., & Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly*, *31*, 36–69.

Lewkowicz, J. (1994). Writing from sources: Does source material help or hinder students' performance? In: *Paper presented at the Annual International Language in Education conference, Hong Kong (RIC Document Reproduction Service No. ED386050)*.

Lewkowicz, J. (1997). *Investigating authenticity in language testing*. University of Lancaster. (Unpublished doctoral dissertation).

Lumley, T. (2005). *Assessing second language writing: The raters' perspective*. Frankfurt, Germany: Peter Lang.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In: A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.

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

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

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

Nation, P. (1988). *Word lists*. Victoria: University of Wellington Press.

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

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Plakans, L. (2007). *Second language writing and reading-to-write assessment tasks: A process study*. The University of Iowa. (Unpublished doctoral dissertation).

Plakans, L. (2008). Comparing composing process in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111–129.

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

Quilan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater® scoring engine (ETS Research Report No. 09-01)*. Princeton, NJ: Educational Testing Service.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater® scoring engine for the TOEFL independent and integrated prompts (ETS Research Report No. 12-06)*. Princeton, NJ: Educational Testing Service.

Ransdell, S., & Levy, C. M. (1999). Writing, reading, and speaking memory spans and the importance of resource flexibility. In: M. Torrance & G. C. Jefferey (Eds.), *The cognitive demands of writing: Processing capacity and working memory in text production* (pp. 99–113). Amsterdam: Amsterdam University Press.

ReDemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*, 7–29.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In: C. Stansfield (Ed.), *Technology and language testing* (pp. 167–188). Alexandria, VA: TESOL.

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

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective*. Northern Arizona University. (Unpublished doctoral dissertation).

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303–325.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*, 27–33.

Song, M. Y. (2007). *A correlational study of the holistic measure with the index measure of accuracy and complexity in international English-as-a-second-language (ESL) student writings*. University of Mississippi. (Unpublished doctoral dissertation).

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, *5* (2), 163–182.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In: R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). Amsterdam: John Benjamins.

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

Vaugh, C. (1991). Holistic assessment: What goes on in the rater's mind? In: Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Norwood, NJ: Ablex.

Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions*. University of Hawaii. (Unpublished doctoral dissertation).

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *Modern Language Journal*, *84*, 171–184.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11* (2), 197–223.

Weigle, S. C. (1999). Investigating rater/prompt interations in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.

Weigle, S. C. (2002). *Assessing writing*. New York, NY: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*, 27–55.

Weigle, S. C. (2013). English Language Learners and Automated Scoring of Essays: Critical Considerations. *Assessing Writing*, *18* (1), 85–99.

White, E. (1994). *Teaching and assessing writing* (2nd ed.). San Francisco, CA: Jossey-Bass Publishers.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instrumentation and Computers*, *20* (1), 6–10.

Witten, I. A., Frank, E., & Hall, M. A. (2011). *Data mining*. San Francisco, CA: Elsevier.

Yang, H. C. (2009). *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches*. The University of Texas at Austin. (Unpublished doctoral dissertation).

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31* (2), 236–259.

Liang Guo is a lecturer at Shanghai University of Finance and Economics. Her research focuses on writing instruction and assessment, discourse analysis and corpus linguistics.

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

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