Language Learning & Technology
ISSN 1094-3501

October 2018, Volume 22, Issue 3
pp. 143–158

ARTICLE

Exploring relationships between automated and human evaluations of L2 texts

Joshua Matthews, *University of New England*

Ingrid Wijeyewardene, *University of New England*

Abstract

Despite the current potential to use computers to automatically generate a large range of text-based indices, many issues remain unresolved about how to apply these data in established language teaching and assessment contexts. One way to resolve these issues is to explore the degree to which automatically generated indices, which are reflective of key measures of text quality, align with parallel measures derived from locally relevant, human evaluations of texts. This study describes the automated evaluation of 104 English as a second language texts through use of the computational tool Coh-Metrix, which was used to generate indices reflecting text cohesion, lexical characteristics, and syntactic complexity. The same texts were then independently evaluated by two experienced human assessors through use of an analytic scoring rubric. The interrelationships between the computer and human generated evaluations of the texts are presented in this paper with a particular focus on the automatically generated indices that were most strongly linked to the human generated measures. A synthesis of these findings is then used to discuss the role that such automated evaluation may have in the teaching and assessment of second language writing.

Keywords: *Writing, Assessment/Testing, Language Teaching Methodology, Research Methods*

Language(s) Learned in This Study: *English*

APA Citation: Matthews, J., & Wijeyewardene, I. (2018). Exploring relationships between automated and human evaluations of L2 texts. *Language Learning & Technology*, *22*(3), 143–158. https://doi.org/10125/44661

Introduction

There is strong appeal in the idea that computers could be used to save time and reduce the costs associated with the labor-intensive process of assessing written work. Indeed, within contexts that require the evaluation of writing on a large scale, the effective employment of computational tools for this purpose may represent a crucial element of future best practice. The automated evaluation of text, although not uncontroversial (Ericsson & Haswell, 2006), is an area of ongoing interest in the research community (Deane, 2013; Weigle, 2013). Further, the current availability of computational tools that can be used to quantify multiple measures of text has made investigations involving the automatic evaluation of written samples more feasible than was previously the case.

Despite the capacity of computational tools to quantify a large range of indices representative of written texts (Graesser, McNamara, & Kulikowich, 2011; Graesser, McNamara, Louwerse, & Cai, 2004), there is a dearth of empirical data that can be used to inform the application of these indices in specific language teaching and assessment contexts. In short, having access to these readily generated measures of texts only goes a short way toward knowing how to apply these data in authentic teaching and learning contexts. An important starting point from which to begin addressing this gap in knowledge is to explore the strength of the interrelations between automatically generated measures of texts, and the measures ascribed to the same texts by human assessors (Crossley & McNamara, 2012). Such data can inform an understanding of the relationship between the relatively context-independent, objective evaluations of text produced by computers, and the relatively context-specific, and necessarily subjective estimations of quality ascribed to

texts by humans. Establishing a clearer understanding of this relationship is an essential starting point from which to assess the role that automated evaluation of text may play in established language teaching and assessment practice.

This study investigates the use of a particular automated evaluation system, Coh-Metrix, which analyzes texts on multiple levels of language through the application of various theoretical frameworks (Graesser et al., 2004; Graesser et al., 2011). As this system has a well-articulated theoretical and empirical research base (McNamara, Graesser, McCarthy, & Cai, 2014; McNamara, Louwerse, McCarthy, & Graesser, 2010), it is a strong candidate for research aimed at exploring the potential role of automated text evaluation in local teaching and assessment contexts. The specific evaluative task chosen to be at the center of the current study was the evaluation of short second language (L2) texts by academic language professionals working in the Australian higher education sector. Within this broad context, English as a second language learners are an important student cohort, and the largest group (29% of the total international student cohort in Australia in June, 2017) come from China (Department of Education and Training, 2017). Entry for these students into undergraduate and postgraduate degree programs is predominantly through standardized English proficiency tests, the most widely accepted of which is the International English Language Testing System (IELTS; O'Loughlin, 2015, p. 182). Therefore, exploring innovative approaches to evaluate samples of L2 writing produced by those with Chinese as a first language is an area of strong significance in the current study's local context. This is especially the case for L2 writing produced by those representative of this cohort in response to tasks that closely align with those typical of high stakes tests such as the IELTS.

The current study explores the evaluation of 104 standardized L2 texts via two means: automatic evaluation using Coh-Metrix and traditional human evaluation using an analytic scoring rubric. Automated and human evaluations both generated measures that were grouped under the broad categories of text cohesion, lexical characteristics, and syntactic complexity. Of central interest was to determine which automatically generated indices were most strongly associated with the values of quality ascribed to the texts by humans. The overarching objective of identifying these key automated indices was to address the potential role automated evaluation of text might play in the assessment and teaching of L2 writing in the study context.

Analytic Rubrics and the Evaluation of Key Measures of L2 Writing Performance

A common approach to assigning measures of quality to text involves the use of analytic scoring rubrics. In contrast to holistic approaches, which involve ascribing a single score that describes the overall quality of the text, the use of an analytic rubric involves the independent assignment of multiple scores, each of which align with discrete categories of writing quality (East, 2009). Analytic rubrics are composed of a number of categories of writing quality, otherwise known as performance criteria; each of these is assumed to represent an important writing performance construct that contributes to the overall quality of the text. Performance criteria are in turn divided into a number of performance levels, each of which possess performance band descriptors. These descriptions enable an assessor to evaluate the text by ascribing a score to each of the rubric's performance criteria. Analytic scoring rubrics can therefore provide information about discrete components of writing performance. This level of analysis is important for L2 teaching and learning, as L2 writers typically possess uneven levels of proficiency across these components (Weigle, 2002).

As a unified theory of language proficiency remains elusive (Knoch, 2011), it is unsurprising that the performance criteria used in analytical rubrics can vary significantly (Weigle, 2002). Despite differences, there are recurring themes evident in the analytic categories used in a range of published rubrics. Categories such as text cohesion, lexical characteristics, and syntactic complexity have been among those included in a number of rubrics devised for the assessment of L2 writing performance (East, 2009; Ruegg, Fritz, & Holland, 2011; Weir, 1990). Indeed, previous research provides evidence to support the link between overall quality of L2 written texts and measures of text cohesion, lexical characteristics, and syntactic complexity (Engber, 1995; Grant & Ginther, 2000; Yang & Sun, 2012). However, as briefly overviewed in

the following section, the links between automatically generated measures of text (e.g., cohesion, lexical characteristics, and syntactic complexity) and human judgements of L2 texts are less straightforward.

Attributes of L2 Text Quality

Measures of Text Cohesion

Research involving the construct of text cohesion and the role it plays in establishing coherent discourse representation in the reader's mind has been an area of enduring research interest (Halliday & Hasan, 1976). Cohesion can be defined as "the presence or absence of linguistic cues in the text that allow the reader to make connections between the ideas in the text" (Crossley, Kyle & McNamara, 2016a, p. 2). The role of cohesion in L2 writing quality, although not as broadly researched as cohesion in first language writing, has also been an area of substantive research. Previous studies have indicated that increased incidence of cohesive devices in L2 texts positively contributes to the perceived quality of those texts. For example, Yang and Sun (2012) reported a significant positive correlation between the correct use of cohesive devices and measures of students' L2 writing quality, regardless of student proficiency. Similarly, Liu and Braine (2005) demonstrated that the number of cohesive devices used in L2 student writing was significantly correlated with overall L2 written composition scores.

However, studies that have specifically investigated the relationship between automatically generated indices of text cohesion and the perceived quality of L2 written texts have presented contradictory results. For example, using texts of upper level L2 English for academic purposes students, Crossley et al. (2016a) demonstrated that automatically generated indices of cohesion and human judgements of text quality were generally positively correlated. Further, automatically generated indices of cohesion could predict 42% of the variance in human judgements. In contrast, Crossley and McNamara (2012) investigated a corpus of L2 texts written by graduating high school students from Hong Kong and concluded that highly proficient L2 writers produce texts with fewer cohesive devices. The results of these two studies suggest that the relationship between automatically generated indices of text cohesion and attributes of text quality are variable, and that these variabilities may relate to context specific factors such as proficiency level and study location.

Measures of Lexical Characteristics of Text

Lexical characteristics of L2 writing generally correlate well with the overall assessment of writing quality assigned to texts (Nation, 2001). The number and range of words known by a learner and the ability to use those words is a robust measure of meaningful engagement with the target language and thus overall proficiency level (Ellis, 2002). For example, Engber (1995) demonstrated a moderate to strong link between a measure of lexical richness in text and the corresponding quality of those texts as determined by a number of human assessors ($r = .57$, $p < .01$). A strong correlational link between L2 vocabulary knowledge and the scores achieved on a standardized English language test in L2 writing performance has also been demonstrated (Stæhr, 2008). Research involving automated modes of text evaluation also points to the importance of the lexical characteristics of L2 texts and the overall quality of L2 writing. For example, Crossley, Salsbury, and McNamara (2011) determined that automatically generated indices of L2 written texts including word imagability, word frequency, lexical diversity, and word familiarity could be used to reliably predict the proficiency level of L2 students based on their written texts. A key finding of this study was that as the proficiency level of the students increased, so too did the lexical diversity of the written texts. Crossley and McNamara (2012), also employing automatic means to generate lexical indices of L2 texts, demonstrated that texts which were judged to be of higher proficiency level contained lower-frequency words and possessed greater levels of lexical diversity.

Measures of Syntactic Complexity

Syntactic complexity can be defined as the variety and level of sophistication of the syntactic forms that are evident within a learner's language output (Ortega, 2003). As Ortega (2003) has emphasized, syntactic complexity is an important construct as L2 development generally entails the expansion of a language

learner's repertoire of syntactic forms, and the learner's capacity to use those forms in a variety of contexts. For example, previous research suggests that higher-rated L2 texts contain more subordination and more instances of passive voice than do lower-rated L2 texts (Ferris, 1994; Grant & Ginther, 2000). In more recent research, Lu (2011) demonstrated that indices of syntactic complexity, including complex nominals per clause and mean clause length, were useful discriminators of written texts belonging to adjacent categories of language proficiency level.

Other research findings that involve investigating the link between automatically derived measures of syntactic complexity and human judgements of L2 written quality are less straightforward. For example, Crossley and McNamara (2012) determined that automatically generated indices of syntactic complexity did not contribute to the predictive capacity of regression models seeking to explain variance within human evaluations of L2 text. Indeed, Crossley and McNamara reported that syntactic complexity was the only measure, among several investigated, which did not yield a significant correlation with measures of essay scores. Similarly, Crossley and McNamara (2014) concluded that most automated indices that signaled development in the complexity of L2 writing did not predict human judgements of writing quality.

The Current Study

In summary, the previous brief overview provides two general points that rationalize aspects of the current study. Firstly, previous research has generally suggested that analytic measures of cohesion, lexical characteristics, and syntactic complexity are important constructs that have links with L2 writing performance. This general finding, coupled with the recurring presence of these categories in published analytic scoring rubrics, provides a rationale for using these categories in the current research. Secondly, a range of alternative results has emerged from previous research that has investigated the relationship between the global quality of texts and corresponding automatically generated measures of cohesion, lexical characteristics, and syntactic complexity. These sometimes conflicting results suggest that the nature of the relationship between automatically generated text-based indices and human based measures of L2 text quality are likely to depend on contextual factors such as L2 proficiency, task type, and study location. These findings motivate research that investigates the relationships between key automatically derived indices and the quality of L2 text as determined by locally relevant assessment practices. Investigating these relationships is likely to provide information about how computer-based evaluation of texts may inform existing approaches to teaching and assessment.

Research Questions

The following research questions are addressed:

1. What is the relationship between automatically generated indices and human assessors' scores?
2. What is the difference in automatically generated indices between lower- and higher-quality texts?
3. Do automatically generated indices predict human assessors' scores?

Method

Texts

The texts used in this study were drawn from a corpus of L2 text samples. The texts were all written in English by tertiary level students with Chinese as a first language. The texts chosen for analysis were each written in response to the same task question, which required writers to provide a short argumentative essay presenting their views on youth employment (see Appendix). The texts used were written under timed conditions (40 minutes). In terms of the local context of this study, the specific writing task was similar to the IELTS Task 2 Writing, in which students are provided with a topic and are required to write a persuasive text, giving and justifying an opinion on a topic, supporting the answer with examples from their own experiences. In order to standardize the word length of the texts that were evaluated as part of the study,

only texts that were between 250 and 350 words in length were selected (*M* = 290 words, *SD* = 24.91).

Automated Evaluation of Texts

The automated evaluation of texts was carried out with the freely available, online tool, Coh-Metrix (Version 3.0). Coh-Metrix facilitated the analysis of text against a range of tools and information sources widely used in computational linguistics including "lexicons, part-of-speech classifiers, syntactic parsers, templates, corpora, latent semantic analysis, and other components" (Graesser et al., 2004, p. 193). The originators of Coh-Metrix describe this computational tool as a "linguistic workbench that researchers … can use to obtain information about their texts on numerous levels of language" (McNamara et al., 2014, p. 1). The resultant output from the Coh-Metrix platform used in this study provided 108 indices that fell within 11 broad categories of information. To refine the scope of this study, of these 11, six were selected to be used as shown below.

- *Referential cohesion* (representative of text cohesion)
- *Connectives* (representative of text cohesion)
- *Lexical diversity* (representative of lexical characteristics of texts)
- *Word information* (representative of lexical characteristics of texts)
- *Syntactic complexity* (representative of syntactic complexity)
- *Syntactic pattern density* (representative of syntactic complexity)

The selection of these six categories was made in order to ensure that the broad categories central to the research objectives of the present study (viz., cohesion, lexical characteristics, and syntactic complexity) were adequately represented in the automated outputs. Further, this selection was made in an effort to align the categories of automated output with the performance criteria of the analytic scoring rubric used in the study. A brief overview of the indices generated within each of the six categories is provided below.

Indices Generated by Coh-Metrix for Each Measure

Referential Cohesion

A range of indices, both local and global, that reflects the degree to which content words overlap within a written text, is generated. Local indices show the degree of content word overlap between adjacent sentences, whereas global indices show the degree of overlap between all sentences in the text. Indices are generated for a number of different forms of content words. For example, the indices may relate to exact noun overlap, argument overlap (includes overlap between nouns and pronouns), and stem overlap (includes overlap between nouns and lemmas).

Connectives

Indices related to five classes of connectives are generated: causal (e.g., *because*, *so*), logical (e.g., *or*, *and*), contrastive (e.g., *although*, *whereas*), temporal (e.g., *first*, *until*), and additive (e.g., *moreover*, *and*). Further, indices are provided that distinguish between the presence of positive (e.g., *also*, *moreover*) and negative connectives (e.g., *but*, *however*; see McNamara et al., 2014, pp. 67–68).

Lexical Diversity

Three indices of the lexical diversity of a text are generated. The first is type–token ratio, which is the total number of unique words (types) divided by the total number of words of a text (tokens). Indices are supplied in relation to type–token ratios for content words and all words respectively. Two additional indices of lexical diversity are generated: measure of textual lexical diversity and vocd (McCarthy & Jarvis, 2007, 2010). These last two indices take into account the influence of the number of words of the text on type–token ratio.

Word Information

A range of word information is generated including the incidence score of parts of speech categories (nouns, verbs, adjectives, and adverbs), and the incidence of personal pronouns (e.g., first-, second-, and third-

person pronouns). Other indices generated within this measure include word frequency, word familiarity, imagability, meaningfulness, and polysemy.

Syntactic Complexity

Indices are generated which indicate the number of words before the main verb and the mean number of modifiers present in noun phrases. Indices based on the concept of minimal edit distance are also generated (McCarthy, Guess, & McNamara, 2009). These evaluate the degree of similarity between sentences in relation to their semantic (i.e., words present in the text) and syntactic features (i.e., the position of the words present in the sentence; see McNamara et al., 2014).

Syntactic Pattern Density

Indices are generated which measure the relative density of phrase types (noun, verb, adverbial, and prepositional), passive voice forms, negation expressions, gerunds, and infinitives.

Human Evaluation of Text

Human Assessors and the Analytic Rubric

Two assessors, each with over 30 years of experience in the assessment and teaching of L2 writing, were involved in the evaluation of the texts. The analytic rubric used was a public version of writing band descriptors produced by the British Council (n.d.-b). The rubric contained four performance criteria: coherence and cohesion (CC), lexical resources (LR), grammatical range and accuracy (GRA), and task achievement (TA). The analytic category of CC provided descriptors that related to the degree to which the information in the text was organized logically, how ideas were linked (e.g., through transition words and phrases and through reference chains), and whether paragraphs were well formed. This analytic category was aligned with the automated indices of referential cohesion and connectives. The analytic category of LR provided descriptors that considered the accuracy and sophistication of word use. This included the success with which an appropriate range of vocabulary, including less frequent words, was used in the text. This category was aligned with the automated indices of lexical diversity and word information. GRA descriptors were concerned with the complexity and range of use of a variety of grammatical forms, including complex and simple sentence structures and subordinate clauses. This category was aligned with the automated indices of syntactic complexity and syntactic pattern. Finally, TA descriptors considered the extent to which the student's response answered the set question and presented a clear and consistent position and how the ideas were supported and developed. This holistic category was not explicitly aligned with a particular set of automated indices, but was assumed to potentially have an indirect relationship with a number of automated indices. Each performance criterion contained qualitative descriptions of each performance level: 0, in which no attempt was made; 1, a non-user; 2, an intermittent user; 3, an extremely limited user; 4, a limited user; 5, a modest user; 6, a competent user; 7, a good user; 8, a very good user; and 9, an expert user. This rubric was selected for use in the current study as the performance criteria of the rubric were quite well aligned with the measures of text cohesion, lexical characteristics, and syntactic complexity. Furthermore, the rubric is freely available in the public domain, thus making the procedures in this study repeatable in future work.

Evaluation of Texts With the Rubric

There are often significant differences between the decision-making processes of each individual assessor engaged in the evaluation of texts (Baker, 2012; Cumming, Kantor, & Powers, 2002). For this reason, it is important that assessors be provided adequate training and support leading up to evaluation processes, including those that involve the use of analytic rubrics (Lumley, 2002). To cater to this requirement, assessors were trained in a number of phases. All training activities involved the use of sample texts that were written by L2 learners from the same population as the main study and that were also written in response to the same task prompt. These texts were well suited for training as they ensured training activities were adequately representative of the main evaluation tasks.

Training occurred in two phases. First, 20 sample texts were each individually assessed. This provided assessors with a substantive opportunity to engage in evaluation events strongly representative of those of the main marking sessions. In the second phase, each assessor attended a group session facilitated by the first author, which involved assessors discussing any disagreements in their evaluations of the texts across the four performance criteria. These discussions provided assessors with opportunities to refine and calibrate their use of the rubric through specific references to samples of evaluated text. Such discussion was important in helping assessors establish a common understanding of how the rubric was to be interpreted (Trace, Meier, & Janssen, 2016). In addition to the initial marking of the 20 sample texts, exemplar sample texts which typified various performance levels as described by the rubric were provided to assessors. These were offered as an additional point of reference during the main rating sessions.

Once training was complete, assessors individually rated the 104 texts, resulting in four discrete scores for each text: TA, CC, LR, and GRA. These scores, and their cumulative totals for each assessor, were collated. Correlation between the total scores from each assessor was strong (r = .76, p < .001), indicating an acceptable level of agreement (Lumley, 2002, p. 253).

Scores for each of the four analytic performance criteria were summed and divided by 2 to establish mean scores for TA, CC, LR, and GRA. These mean scores were summed to provide a global measure of text quality (GMTQ). Thus, there were five criterion scores that were procured as a result of human evaluation, with GMTQ held as the main criterion variable in subsequent analyses.

Results

Mean values for the cumulative and analytic scores assigned by the assessors are provided in Table 1. Mean analytic scores for GMTQ for the cohort were distributed relatively normally between a lower performance range of band 4 and an upper performance range of band 6.5 (See Supplementary File 1). These results position the cohort's writing proficiency at approximately the B1 to B2 level of the Common European Framework of Reference (CEFR; British Council, n.d.-a). Assessors were also asked to consider the CEFR descriptors for reports and essays (Council of Europe, n.d., p. 62), and they determined that the descriptors of levels B1 and B2 aligned well with the writing samples produced by the cohort.

Table 1. *Mean Scores for GMTQ, TA, CC, LRA, and GRA*

Score	*M*	*SD*
GMTQ	21.94	2.18
TA	5.22	0.68
CC	5.42	0.68
LR	5.66	0.61
GRA	5.64	0.68

Word count of the texts, which was identified as a potential confound, was shown to have no significant correlation with either analytic scores or key automated indices (see Supplementary File 2).

Research Question 1. What Is the Relationship Between Automatically Generated Indices and Assessors' Scores?

The first step of the analysis involved exploring the strength of correlation between the mean GMTQs generated through human evaluation and the automated indices generated by Coh-Metrix. Correlational analysis showed that of the 57 indices explored, 13 of these reached a level of statistical significance (see Table 2). As can be noted, the magnitudes of the resultant Pearson correlation coefficients were all small to medium (Cohen, 1992). None of the six connectives indices investigated as part of the cohesion category reached statistical significance. A complete list of all indices with corresponding information can be found

in Supplementary File 3.

Table 2. *Significant Correlations Between Automatically Generated Indices and GMTQ*

Category	Coh-Metrix Measure[a]	Coh-Metrix Index	*r*	*M*	*SD*
Cohesion	Referential cohesion (10)	Noun overlap (local)	.30**	0.45	0.22
		Noun overlap (global)	.31**	0.40	0.21
		Argument overlap (local)	.29**	0.63	0.18
		Argument overlap (global)	.33**	0.55	0.19
		Stem overlap (local)	.30**	0.52	0.22
		Stem overlap (global)	.30**	0.48	0.22
		Content word overlap (local)	.24*	0.12	0.47
		Content word overlap (global)	.29**	0.10	0.04
		Content word overlap *SD* (global)	.27**	0.10	0.02
Lexical characteristics	Lexical diversity (4)	Type–token ratio (content words)	-.28**	0.72	0.06
	Word information (22)	Incidence of pronouns (second-person)	-.29**	3.01	6.83
Syntactic characteristics	Syntactic complexity (7)	Left embeddedness (Mean number of words before main verb of main clause)	.22*	4.61	1.35
	Syntactic pattern density (8)	Incidence of passive voice forms	.20*	11.50	6.30

*Note. *p < .05, **p < .01*

[a]*Number of indices explored are in parentheses*

Measures of referential cohesion yielded the greatest number of indices reaching a significant level of correlation with GMTQ (nine in total). These correlation coefficients were positive, indicating that increases in automatically derived measures of referential cohesion were associated with small to moderate increases in GMTQ. Evidently, indices reflecting noun, argument, and stem overlap (both local and global) were those that most strongly aligned with GMTQ (for descriptions, see Supplementary File 4). An investigation of the correlation among these indices of referential cohesion indicated a high level of correlation (r = [.80, .94]). Thus in order to avoid issues of collinearity in forthcoming analyses, the strongest relative correlate from the referential cohesion measures, argument overlap (global), was chosen as the representative metric for referential cohesion.

Two indices, one each from lexical diversity and word information, reached a level of statistical significance: type–token ratio of content words and the incidence of personal pronouns in the second person, respectively. There was a weak to moderate inverse relationship between each of these indices and GMTQ. In terms of syntactic complexity, two indices, left embeddedness and the incidence of passive voice forms respectively, were each weakly and positively associated with GMTQ. Thus, there was a small yet significant linear relationship between syntactic complexity and human assessors' impressions of global text quality.

In summary, indices with the strongest relative magnitude of significant correlation with GMTQ were identified as the following: argument overlap, type–token ratio, incidence of second-person pronouns, left embeddedness, and incidence of passive voice forms. These indices are henceforth collectively referred to as *key indices.* A correlation matrix between the key indices can be found in Supplementary File 5.

In order to address in more detail the nature of the relationship between the key indices and those measures ascribed to the texts by human assessors, correlational analysis was undertaken, which involved the mean scores for each of the performance criteria of the analytic rubric (see Table 3).

Table 3. *The Strength of Correlation Between Key Indices and Analytic Measures*

Key Index	GMTQ	TA	CC	LR	GRA
Argument overlap	0.33**	0.26**	0.36**	0.26**	0.20*
Type–token ratio	-0.28**	-0.34**	-0.36**	-0.13	-0.08
Pronouns in the second person	-0.29**	-0.24*	-0.38**	-0.21*	-0.13
Left embeddedness	0.22*	0.11	0.22*	0.18	0.23*
Passive voice forms	0.20*	0.14	0.14	0.27**	0.14

Note. $*p < .05$, $**p < .01$

From Table 3, it can be noted that argument overlap was the only variable that correlated significantly with all four performance criteria (TA, CC, LR, and GRA). This result seems to suggest the relative importance of referential cohesion in relation to human assessors' evaluations of the texts as the other key indices did not correlate significantly across all performance criteria in the same way. In contrast, type–token ratio possessed a significant negative correlation with only two of these criteria, TA and CC. It is of some interest that type–token ratio did not reach a level of statistical significance with LR despite this criterion describing use of uncommon lexical items and a wide range of vocabulary as important elements of a writer's LR. Pronouns in the second person possessed a negative and significant correlation with TA, CC, and LR.

Left embeddedness correlated significantly with two analytic measures (CC and GRA), and passive voice forms positively correlated with just one (LR). Neither left embeddedness nor passive voice forms possessed a significant correlation with TA. Apparently indices of syntactic complexity were not significantly aligned with assessors' perceptions of the degree to which the texts addressed the task and presented relevant and well developed ideas.

Research Question 2. What Is the Difference in Automatically Generated Indices Between Lower- and Higher-Quality Texts?

The next step of analysis sought to identify the key indices that were most useful in differentiating texts with relatively low and relatively high GMTQ. This first entailed the categorization of the texts into relatively low and relatively high groups. This was achieved by grouping all 104 texts into three groups based on 33rd percentile cut-off scores for GMTQ. This yielded three groups of which the relatively low GMTQ group ($n = 36$, $M = 19.58$, $SD = 0.25$) and the relatively high group ($n = 39$, $M = 24.03$, $SD = 0.16$) were of primary interest. From here, a multivariate analysis of variance between the low and high groups was undertaken. This analysis indicated that there was a significant difference between low and high groups when the key variables of argument overlap, type–token ratio, second-person pronoun, left embeddedness, and incidence of agentless passive were jointly considered (Hotelling's $T = .37$, $F(5, 69) = 5.12$, $p = .000$ partial $\eta^2 = .27$). A separate univariate ANOVA was then undertaken for each dependent variable. A Bonferroni corrected alpha level of .01 was used (viz., a standard α of 0.05 divided by 5 comparisons). These analyses indicated that there was a significant difference between low and high groups on three of the five comparisons: mean argument overlap, mean incidence of second-person pronouns, and mean type–token ratio for content words.

- The mean argument overlap value for the relatively low group ($M = 0.48$, $SD = 0.19$) was significantly lower than the relatively high group ($M = 0.63$, $SD = 0.16$; $F(1, 73) = 14.18$, $p = .000$, partial $\eta^2 = .16$).
- The mean incidence of second-person pronouns for the relatively low group ($M = 5.89$, $SD = 10.26$) was significantly higher than the relatively high group ($M = 1.15$, $SD = 2.58$; $F(1, 73) = 7.82$, p

=.007, partial η^2 = .10).

- The mean type–token ratio value for the relatively low group (M = 0.74, SD = 0.06) was significantly higher than the relatively high group (M = 0.70, SD = 0.06; $F(1, 73) = 11.69$, p = .001, partial η^2 = .14).
- The mean incidence of passive voice forms for the relatively low group (M = 9.56, SD = 5.80) and the high group (M = 12.12, SD = 7.02) did not reach the level of significance ($F(1, 73) = 2.95$, p = .090, partial η^2 = .04).
- The mean left embeddedness for the relatively low group (M = 4.39, SD = 1.31) and the high group (M = 5.14, SD = 1.38) did not reach the level of significance ($F(1, 73) = 5.88$, p =.018, partial η^2 = .08)

From these results, three key indices appear to be of strongest value in differentiating relatively low and high performance as determined by human global evaluation of these texts: argument overlap, type–token ratio, and pronouns in the second person.

Research Question 3. Do Automatically Generated Indices Predict Human Assessors' Scores?

The last phase of analysis explored the predictive capacity of the automated indices in relation to GMTQ through use of a hierarchical multiple regression model. The predictor variables used in this analysis were those indices identified as most useful in differentiating low- and high-quality texts: argument overlap, incidence of second-person pronouns, and type–token ratio. The dependent variable for the analysis was GMTQ. Initial analysis was undertaken to ensure that there were no violations of the assumptions relating to linearity, multicollinearity, normality, or homoscedasticity. Variables were entered into the regression model in order of their strength of correlation with the GMTQ (see Table 2 and Supplementary File 6). Step one involved entry of argument overlap into the model, and explained 10.69% of the variance in GMTQ. Incidence of second-person pronouns was entered in the second step, and explained an additional 4.75% of the variance in GMTQ. Adding type–token ratio in step three added no additional significant predictive capacity to the model. Thus, multiple regression analysis indicated that two predictors collectively explained 15.45% of the variance in GMTQ in the second step of the model, (r^2 = .15, $F(2, 101) = 9.24$, p < .000). Of the two indices which contributed to the predictive capacity of the model, argument overlap had a beta value of greater magnitude (β = .27, p = .005), when compared to the value for incidence of second-person pronouns (β = -.23, p = .019).

Discussion

This research explored the relationship between automated text-based indices and measures of quality ascribed to text by human assessors. The underlying objective of this exploration was to draw insight on how automated evaluation may be applied in teaching and assessment activities in contexts such as that described in the current study. A number of the automated indices explored in the study were significantly correlated with global assessments of text quality; however, the magnitude of these linear relationships did not exceed moderate levels. Additionally, the capacity of these key indices to predict variance in global assessments of text quality was relatively constrained (r^2 = .16). Given these findings, at least within the context of this study, the prospect of using such automated evaluation for high stakes summative assessment decisions is clearly untenable. This finding is not unexpected in light of previous work that has made clear the difficulty of establishing linear relationships between the linguistic features of texts and assessments of L2 writing quality (Bulté & Housen, 2014; Crossley & McNamara, 2014; Jarvis, Grant, Bikowski, & Ferris, 2003). Moreover, computer evaluations of written texts cannot account for all linguistic, semantic, or discourse features that contribute to the quality of a text such as metaphor or disciplinary knowledge (Graesser et al., 2011).

Although there were clear limits on the capacity of automated indices to predict human assessments of L2 writing quality, there were also trends observed within the data that can inform teaching and assessment

practice. Largely, what is suggested here is the value of automated indices and their comparison to GMTQs for low stakes, formative assessment. Clearly, automated evaluation does not resolve all the difficulties associated with evaluating L2 texts, but it does present a potentially valuable, immediately available range of data that can inform teaching and assessment practice. For example, the most significant finding of the current study was the consistent and generally moderate relationship between measures of referential cohesion and human assessment of L2 text quality. Both local and global measures of referential cohesion significantly correlated with GMTQ (See Table 2). Furthermore, global argument overlap correlated significantly with each of the analytic categories ascribed to texts by human assessors (See Table 3). Global argument overlap was also the variable that was most predictive of GMTQ. These findings allude to the overall importance of referential cohesion in these texts in relation to their global quality, and this type of information holds the potential to usefully inform teaching practice. Such results may motivate teaching that heightens learners' awareness of the importance of cohesion in written discourse, which may also help learners to develop the skills necessary to input adequate levels of argument overlap in their writing.

Similar conclusions may be drawn in relation to other indices that had a significant correlational relationship with GMTQ. For example, both left embeddedness and incidence of passive voice forms were positively correlated with GMTQ. In terms of teaching practice, such findings, at a minimum, suggest that language learners could benefit from explicit instruction in these forms of syntactic complexity and practice in developing these linguistic features in their own written texts. In addition, L2 learners can be provided with models of high- and low-rated texts to identify these linguistic features and raise awareness of their functions in written English.

The significant negative correlation between incidence of second-person pronouns and GMTQ also presents an example of how automatically generated indices may inform teaching and assessment practice. Evidently, the prevalence of second-person pronouns held a significant measurable capacity to predict GMTQ. However, it was of interest that reference to personal pronouns of any sort was not present in the descriptors of the analytic rubric. Although somewhat speculative, from discussion with the two assessors after completion of the present study, it seems that increased incidence of second-person pronouns may have been an indicator of texts that conveyed less adequate control of appropriate levels of formality (e.g., register). These findings present an example of how automated indices may cast light on less-obvious aspects of written texts that are associated with assessors' appreciations of text quality. Such findings may alert assessors and teacher trainers to linguistic factors which are seemingly linked to evaluation decisions, but which are based on rationales not explicitly referred to in the scoring rubric criteria. This type of information may result in adjustments to the way assessors are asked to interpret scoring rubrics, or may potentially inspire alterations to the scoring rubrics themselves. Such insight can also inform teaching practice. For example, the concept of register and the role that use of personal pronouns plays in conveying appropriate levels of formality can be put forward as targeted teaching points.

As with other research looking at the relationship between attributes of linguistic features and text quality, some aspects of the current research findings were problematic. For example, the negative correlation observed between content word type–token ratio and GMTQ seems to conflict with other studies that have concluded that automatically generated indices of lexical diversity were positively correlated with L2 text quality (e.g., Crossley & McNamara, 2012; Crossley et al., 2011). Such findings speak to the limitation of models that assume that key linguistic indices will have independent, linear relationships with GMTQs. Such models do not factor in the potential for interactivity between the key variables. For instance, it may be the case that use of a broader variety of lexical items in text is generally indicative of text quality (e.g., Engber, 1995). However, it seems that, in the instance of the specific task that was at the center of the current study, high levels of referential cohesion, such as global argument overlap, may have brought about lower levels of lexical diversity. An assessment of the direction and magnitude of correlation between global argument overlap and type–token ratio adds some support to this assertion (r = -.53, p < .001). This may represent an instance of what Jarvis et al. (2003) refer to as *complementarity*, where although there are "a number of linguistic features that contribute to the overall quality of a written text, high levels of some features may bring about low levels of other features" (p. 399).

In order to establish a more generalized picture of how to practically apply automated evaluation of text in authentic contexts, further investigations of automated and human evaluations of text in a broader array of contexts are warranted. Future studies in this field will benefit from reference to the substantial body of research on L2 learner corpora that already exists (e.g., Connor-Linton & Polio, 2014; Paquot, 2017). For example, future studies that analyze a common corpus from alternative methodological angles will enable groups of researchers to compare their results with less concern for the potential confound of population difference (Connor-Linton & Polio, 2014). Further, to ameliorate potential issues associated with comparisons across different study populations, future investigations need to ensure a thorough approach to the reporting of methods so that research protocols can be adequately replicated (Polio & Shea, 2014). Despite these important considerations, in our view, learner corpora, which comprise data of immediate relevance to local contexts, can be relatively easily compiled by teachers and researchers. Such corpora can include learners' texts written in response to a range of tasks, the automated indices of the texts, and measures of locally relevant human assessments of text quality. As there are various contextual factors that influence the patterns of linguistic features used in student texts (Jarvis et al., 2003; Weigle & Friginal, 2015), it makes sense that the practical application of automated evaluation in the language classroom needs to be based on data gathered at the local level. Over time, an analysis of such corpora may enable the context-specific relationships between automated indices and human evaluations of text quality to be more fully understood and more effectively applied.

The current study has investigated the relationship between automated indices and human evaluation of texts in a relatively constrained context. Although this refined scope enabled focused analysis and a feasible research design, it also represented a core limitation to the current study. The scope of similar future studies can be usefully broadened in a number of ways. A broader array of L2 text types, drawn from a larger and more diverse cohort over a longer period of time, may provide insight, which is likely to go beyond that presented in the current study. For example, the corpus used in the current study contained texts that had a relatively homogenous range of proficiency levels (e.g., approximately B1 to B2). Although speculative, this limited range of proficiency levels may have contributed to the weak to moderate relationships observed between the automated indices and human evaluations of texts. To test this assertion, replication studies involving corpora containing a broader range of proficiency levels (e.g., B1 to C2) would be required.

Further, the use of alternative computation tools, an alternative range of automatically generated indices, and alternative modes of human assessment are also likely to provide additional depth to future research. For example, it has been noted that automated measures of referential cohesion were those with the strongest and most consistent correlation with human evaluations of text. For this reason, use of alternative text analysis tools, such as the Tool for Automatic Analysis of Cohesion (Crossley, Kyle, & McNamara, 2016b), which offer a greater array of cohesion indices than Coh-metrix, may be a fruitful option for future research. Repeating the basic methodological approach discussed in the current study using a broader array of cohesion indices may enable a more fine-grained understanding of the link between human assessment of text quality and various computationally derived measures of text cohesion.

Although broadening the scope of future research is suggested, there is also cause to consider the potential value of replication studies in this field. As part of this study, we have attempted to employ methodological procedures and tools that are readily available to other teachers and researchers. It is hoped that this study provides an accessible point of reference for other research that seeks to explore the types of relationships discussed in the present paper in a variety of local contexts.

Conclusion

Currently there are freely available computational tools that can be used to generate a large range of indices reflective of the linguistic features of text. Despite the potential these forms of automated evaluations may hold, questions remain about how best to apply such data in established language teaching and assessment contexts. This study has explored the relationships between computationally generated indices and the corresponding human assessments of the global quality of L2 texts. The results indicated that indices such

as argument overlap, incidence of second-person pronouns, and type–token ratio were significant correlates with GMTQs, but these linear relationships were weak to moderate in magnitude. Despite these limitations, explorations of the relationship between automated indices and measures of global text quality can be immediately applied to good effect in the language classroom. The current study has put forward some examples of how these types of explorations can be carried out and how they may be used to inform teaching and assessment practices.

Automated evaluation of text does have a potentially important role to play in current and future language assessment practice; however, far more research is required to take advantage of this potential. Future research in the field should proceed with an appreciation that automated evaluations of texts are fundamentally *linguistic fingerprints*. These fingerprints possess a great deal of information, but this information is limited in that it is only reflective of the linguistic features evident within the texts themselves. It is clear that the quality of written text can be partly defined by its linguistic features; however, establishing a global evaluation of a text must necessarily go beyond the text itself. Such evaluation depends on nuanced, socially embedded decisions that judge the degree to which the linguistic elements of the text align with important contextual factors such as the task parameters, institutional requirements, and local assessor expectations. Such evaluation, which transcends the linguistic features of the text and which depends on reference to external contextual elements, currently exceeds the capacity of computer algorithms. On the other hand, this type of evaluation is something that humans can do relatively well. It is therefore important for future research to continue to bridge the gap between automated evaluations of text and human impressions of text quality. A body of research that explores these relationships within a range of locally relevant language assessment settings, over the longer term, will be an important foundation for future progress in this field.

Acknowledgements

This research project was supported by a research grant obtained from the Association for Academic Language and Learning (AALL) in 2016. The authors wish to sincerely thank the AALL for their support. The authors also wish to sincerely thank Lyndall Nairn for the contribution she has made to this project.

References

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, *9*(3), 225–248.

British Council. (n.d.-a). *Common European Framework equivalencies*. Retrieved from http://takeielts.britishcouncil.org/find-out-about-results/understand-your-ielts-scores/common-european-framework-equivalencies

British Council. (n.d.-b). *IELTS Task 2 Writing band descriptors*. Retrieved from https://takeielts.britishcouncil.org/sites/default/files/IELTS_task_2_Writing_band_descriptors.pdf

Bulté, B., & Housen, A. (2014). Conceptualiing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9.

Council of Europe. (n.d.). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Retrieved from https://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication and their relations to judgments of essay quality. *Journal of Research in Reading, 35*(2), 115–135.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing. *Journal of Second Language Writing, 32*, 1–16.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67–96.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24.

Department of Education and Training. (2017). *International student data monthly summary*. Retrieved from https://internationaleducation.gov.au/research/International-Student-Data/Documents/MONTHLY%20SUMMARIES/2017/Jun%202017%20MonthlyInfographic.pdf

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing, 14*(2), 88–115.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143–188.

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

Ericsson, P. F., & Haswell, R. H. (Eds.). (2006). *Machine scoring of student essays: Truth and consequence*. Logan, UT: Utah State University Press.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414–420.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, 36*(2), 193–202.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145.

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

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377–403.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96.

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*(4), 623–636.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

McCarthy, P. M., Guess, R. H., & McNamara, D. S. (2009). The components of paraphrase evaluations. *Behavior Research Methods*, *41*(3), 682–690.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*(4), 292–330.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

O'Loughlin, K. (2015). "But isn't IELTS the most trustworthy?": English language assessment for entry into higher education. In A. Ata & A. Kostogriz (Eds.), *International education and cultural-linguistic experiences of international students in Australia* (pp. 181–194). Samford Valley, Australia: Australian Academic Press.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*. https://doi.org/10.1177/0267658317694221

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, *26*, 10–27.

Ruegg, R., Fritz, E., & Holland, J. (2011). Rater sensitivity to qualities of lexis in writing. *TESOL Quarterly*, *45*(1) 63–80.

Sanders, J. (2014). *IELTS - The best writing correction*. Morrisville, NC: Lulu Press.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading, and writing. *Language Learning Journal*, *36*(2), 139–152.

Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubric category interpretations through score negotiation. *Assessing Writing*, *30*, 32–43.

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

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85–99.

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

Weir, C. J. (1990). *Communicative language testing*. New York, NY: Prentice Hall.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, *23*(1), 31–48.

Appendix. Writing Task

The written texts are responses to the following task (adapted from Sanders, 2014):

> Present a written argument to an educated reader with no specialist knowledge of the following topic.
>
> *In many countries children are engaged in some kind of paid work. Some people regard this as completely wrong, while others consider it as valuable work experience, important for learning and taking responsibility.*
>
> *What are your opinions on this?*
>
> You should use your own ideas, knowledge, and experience and support your arguments with examples and relevant evidence.
>
> You should write at least 250 words and no more than 350 words.

About the Authors

Dr. Joshua Matthews is a lecturer at the University of New England, Australia. His major research interests include computer assisted language learning, L2 vocabulary, L2 teaching, and language testing. His previous publications have appeared in journals including *Computer Assisted Language Learning*, *ReCALL*, *System*, and *Language Testing*.

E-mail: joshua.matthews@une.edu.au

Dr. Ingrid Wijeyewardene is a lecturer at the University of New England, Australia, where she teaches online courses in academic literacy. Her research interests include systemic functional linguistics and its application in discourse analysis, academic language and learning, L2 teaching, and technology-based language learning.

E-mail: iwijeyew@une.edu.au