Assessing Writing 26 (2015) 5–19

Contents lists available at ScienceDirect

Assessing Writing

Keeping up with the times: Revising and refreshing a rating scale

Jayanti Banerjee[a,*], Xun Yan[b], Mark Chapman[c], Heather Elliott[d]

[a] Worden Consulting LLC., 115 Worden Avenue, Ann Arbor, MI 48013, USA
[b] University of Illinois—Urbana Champaign, 4080 Foreign Languages Building, 707 S. Matthews Avenue, MC-168, Urbana, IL 61801, USA
[c] The University of Wisconsin—Madison, WIDA Consortium, 181 Education Building, 1000 Bascom Mall, Madison, WI 53706, USA
[d] CaMLA, Argus 1 Building, 535 West William St., Suite 310, Ann Arbor, MI 48103-4978, USA

ARTICLE INFO

Article history:
Received 2 January 2015
Received in revised form 30 June 2015
Accepted 16 July 2015
Available online 6 August 2015

Keywords:
Rating scale design
Scale validation
Corpora
Discriminant function analysis

ABSTRACT

In performance-based writing assessment, regular monitoring and modification of the rating scale is essential to ensure reliable test scores and valid score inferences. However, the development and modification of rating scales (particularly writing scales) is rarely discussed in language assessment literature. The few studies documenting the scale development process have derived the rating scale from analyzing one or two data sources: expert intuition, rater discussion, and/or real performance.

This study reports on the review and revision of a rating scale for the writing section of a large-scale, advanced-level English language proficiency examination. Specifically, this study first identified from literature, the features of written text that tend to reliably distinguish between essays across levels of proficiency. Next, using corpus-based tools, 796 essays were analyzed for text features that predict writing proficiency levels. Lastly, rater discussions were analyzed to identify components of the existing scale that raters found helpful for assigning scores. Based on these findings, a new rating scale has been prepared. The results of this work demonstrate the benefits of triangulating information from writing research, rater discussions, and real performances in rating scale design.

1. Introduction

In the standardized assessment of writing, rating scale development is a ubiquitous activity. Regular monitoring and modification of the rating scale is also essential to ensure reliable test scores and valid score inferences. However, reports of scale development or revision are rare in language assessment literature. This presents a gap in our discussions of the challenges and opportunities presented during scale development and revision. Of the studies available, most describe the development of speaking assessment scales (e.g. Ducasse, 2009; Upshur & Turner, 1999; Fulcher, Davidson, & Kemp, 2011); there are relatively few studies that address the development of writing scales (Knoch, 2011; Lim, 2012; Sasaki & Hirose, 1999). That said, the scale development process for speaking and writing performances is largely similar and both are therefore relevant for the work presented here.

[*] Corresponding author. Tel.: +1 7346602767.
E-mail address: j.v.banerjee@gmail.com (J. Banerjee).

http://dx.doi.org/10.1016/j.asw.2015.07.001
1075-2935/© 2015 Elsevier Inc. All rights reserved.

Fulcher et al. (2011) describe the two most common approaches to constructing a rating scale: the measurement-driven, and performance data-driven approaches. The measurement-driven approach starts with the level descriptors. It focuses on the clarity of the descriptors and thus the usability of the rating scale. It also relies on the intuition of experts in language teaching and assessment (e.g., theorists, teachers, or raters) to develop the rating criteria (Hamp-Lyons, 1991). This approach is by far the most commonly used in scale development. However, views on the appropriateness of this approach are mixed. The criticisms of the approach include claims that the resulting scales can lack precision, specificity, and scalability (Fulcher et al., 2011). As the descriptors are often written in impressionistic, abstract, or relativistic language, the distinction between performances across score levels tends to be subjective or less consistent across raters (Knoch, 2009). Concerns have also been raised about the representativeness of the rating scales (Mickan, 2003; Upshur & Turner, 1995). Additionally, intuitively developed scales have been criticized for having descriptors that are inconsistent with theories of L2 development (Turner & Upshur, 2002). The involvement of expert raters in scale development tends to improve the usability of the scale compared with those derived directly from theory in a top-down fashion (Lowe, 1986). However, the intuitive nature of the measurement-driven approach requires no analysis of real performance prior to generating descriptors. This makes the resultant rating scales dependent upon post-hoc quantitative or qualitative analysis to ensure reliability of the descriptors and validity of the score inferences.

The performance data-driven approach, on the other hand, derives rating scales through analyzing real language performances. This approach starts with performances, and identifies traits or features that characterize and discriminate written texts or writers across proficiency levels. There are two sub-approaches within the performance data-driven approach (Council of Europe, 2001, p. 207): qualitative and quantitative methods. The qualitative method pre-tests the effectiveness of descriptors derived from the measurement-driven approach through detailed analysis of a small number of test performances. The quantitative method quantifies and cross-validates the qualitative evidence on a larger scale. The two methods are clearly complementary (Lim, 2012), and are thus recommended to be used in combination. Unlike the post-hoc reliability or validity analysis in the measurement-driven approach, the analyses in the performance data-driven approach are primarily exploratory in nature. That is, the analyses of performance data precede the development of the scale and are not aimed at confirming a pre-determined set of features. The advantage of this approach lies in the resulting scale's reflection of real performances. However, data-based analysis tends to be time consuming. Additionally, in a completely data-driven approach, especially when using corpus-based tools, the data tend to generate linguistic constructs that either bear complex mathematical formulae or become extremely difficult to operationalize by human raters (Fulcher, 2003). The level descriptors would need to be carefully written in order to ensure that the linguistic features are accessible to examiners. Additionally, rater training would need to be carefully structured so that divided and yet simultaneous attention to individual criteria is possible but not over-taxing for raters in real-time rating.

In addition to the aforementioned approaches, the literature on scale development has called for more theory-based practices in scale development (e.g., Fulcher, 1987; Knoch, 2011; McNamara, 2002). Lantolf and Frawley (1985) have argued that a lack of linkage between theories of L2 development and construct representation raises questions about the validity of the rating scale. Despite this there are no records of a scale development process using theory to inform its construction. This is perhaps, as argued by Knoch (2011) and Lantolf and Frawley themselves, due to the lack of a unified theory of L2 development or language proficiency. This makes it difficult to develop rating scales using a theory-based approach.

It appears, therefore, that the most defensible approach to rating scale development and revision would be to adopt an approach that combines our current understanding of the indicators of second language writing development (cf. Wolfe-Quintero, Inagaki, & Kim, 1998), expert intuition, and the empirical analysis of performance data. This is the approach that we have taken in the review and revision of the rating scale for the writing section of a large-scale advanced level English language proficiency examination. We have triangulated three data sources by: reviewing expert intuition and analysis to build a framework of the text features that are expected to predict writing proficiency; using corpus tools to analyze 796 real performances; and analyzing rater discussions during the scoring process.

2. Background to the study

The rating scale under review here is the assessment tool for the writing section of a large-scale English language proficiency examination designed for advanced-level learners, the Examination for the Certificate of Proficiency in English (ECPE). Developed by CaMLA (http://www.cambridgemichigan.org/), the exam comprises four sections, writing, listening, reading, and speaking. The results for each section are reported separately. The writing section is 30 min long and offers test takers a choice of two essay prompts. They choose one and are expected to write at least 300 words. Both prompts require test takers to give their opinion on a statement and to justify that opinion using supporting details or points. Test takers who pass the writing section are considered to be at C2 on the Common European Framework of Reference (CEFR, Council of Europe, 2001). They are able to communicate their ideas fully in clear, smoothly flowing language. They can structure their text logically to present an effective argument and can use grammatical structures and vocabulary flexibly in order to convey precise meaning. As such, the intended construct of the writing section includes breadth and depth of vocabulary knowledge, variety and accuracy of grammatical structures, ability to state and develop an argument, audience awareness, and text organization skills.

Appendix A presents the current rating scale. It is a 5-point scale with three scoring criteria, each of which has five levels of performance. The rating scale is applied analytically; two examiners independently give performances a score of 1–5 in each of the scoring criteria. The examiner's scores are summed to arrive at the final score for each performance. The scores are then transformed onto a standardized 1000-point scale. Test takers receive their score on the 1000-point scale as well as a band score. There are five bands: three passing bands (A–C) and two failing bands (D & E). Rater agreement is monitored throughout the scoring process. If examiners give non-adjacent scores on any criteria, the performance is re-evaluated by a third examiner.

All CaMLA tests undergo regular review and revision. In 2014, we undertook to examine the reliability and usability of the current ECPE rating scale. Our aims were to ensure that the rating scale properly reflects the underlying construct of the ECPE writing section and can be effectively applied. We used the following research questions to guide our investigation:

RQ1: Which text features have been identified by the literature to most effectively predict writing proficiency?
RQ2: Which text features distinguish essays at each proficiency level?
RQ3: How do raters use and interpret the rating scale? What are the sources of agreement and disagreement between raters?
RQ4: Do raters encounter difficulties in applying the rating scale? If yes, what are the difficulties?
RQ5: Based on the findings from research questions 1–4, what revisions (if any) might be needed for the rating scale?

3. Methodology for the scale review

3.1. Data triangulation

We triangulated information from three sources: theoretical models of writing proficiency, corpus analysis of writing performance, and thematic analysis of rater discussion. Fig. 1 illustrates how the data were triangulated in the revision process. First, we reviewed the existing literature on writing proficiency in order to identify components of writing proficiency and representative linguistic and discourse features that predict writing proficiency.

For the corpus analysis of writing performances, we used Coh–Metrix to operationalize the linguistic and discourse features identified from the literature. Since Coh–Metrix provides a wealth of measures that operationalize these text features, we synthesized the results of existing Coh–Metrix studies to select the five indices that mostly reliably quantify the text features identified from the literature. In addition, computer programing scripts in a free statistical software package called R (R scripts) were used to operationalize prompt dependence as an additional construct measured specifically by the writing tasks. Then, a discriminant function analysis was performed to examine the extent to which the six quantified text features could predict essay band scores.

In the thematic analysis of rater discussions, rater discussions were qualitatively analyzed to investigate the usability of the current rating scale. Finally, patterns observed in rater discussions were triangulated with the results of the corpus analysis to identify the strengths and weaknesses of the current rating scale. Consistently discriminating features were retained as the strengths of the rating scale. Non-discriminating features, unreliable descriptors, and raters' difficulties with the rating scale were discussed for possible explanations and revision recommendations.

3.2. Corpus-based analysis of writing performance

The corpus for this study comprised 796 essays from two recent test administrations and covered responses to a range of essay prompts, selected to represent all possible score levels (Table 1). By analyzing essays written in response to different prompts we were able to investigate more general writing practices as well as prompt-specific language. Each essay had been assessed independently by two trained raters using the existing rating scale. The final score awarded was the result of exact agreement between the raters. To ensure that the findings of the study could be generalizable to the typical test-taking population, the sample was representative of typical gender and age distributions.

Table 1
The essay corpus.

	Score level	Number of essays
Pass	A	146
	B	199
	C	214
Fail	D	137
	E	100
Total		796

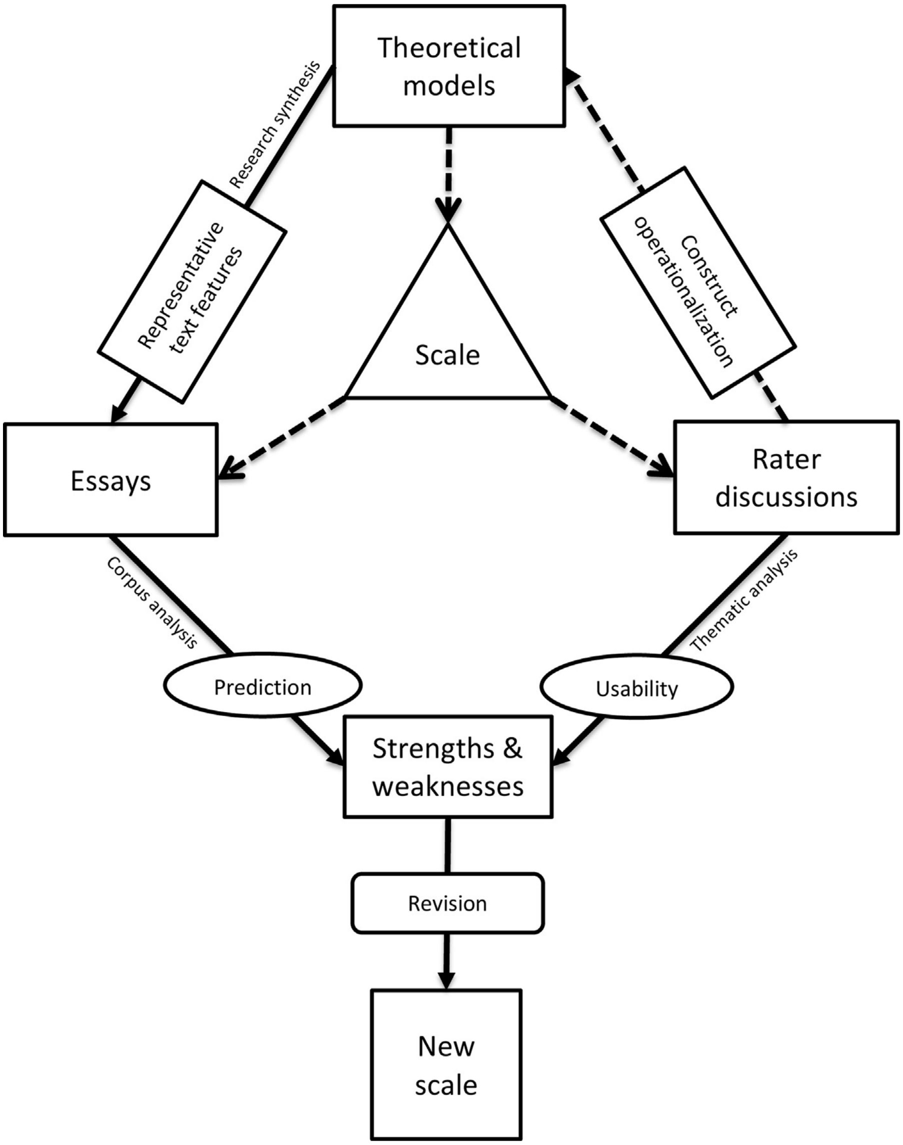

Fig. 1. Triangulation of data sources during the scale revision process.

Representative text features were identified from the literature and then the essays were analyzed using Coh–Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) and R scripts. Coh–Metrix produces 106 features, many of which measure similar constructs. Therefore, to avoid redundancy (as well as colinearity) in the features, we reviewed published Coh–Metrix studies and selected the Coh–Metrix index for each feature that significantly predicted writing proficiency across studies.

Discriminant Function Analysis (DFA) was used to analyze the predictive power of all the text features because (1) the independent variables, i.e., text features, were quantitative in nature, and (2) the dependent variable, i.e., essay band scores, is a categorical variable measured on an ordinal scale (Warner, 2013). DFA accounts for all the independent variables and reduces them to a few discriminant functions (similar to factors in factor analysis) to predict the dependent variable. Depending on the number of levels (k) in the dependent variable (e.g., the number of band scores) and the number of independent variables (q) (e.g., the number of text features), the number of discriminant functions created is the smaller number between q and $k-1$. However, only statistically significant discriminant functions are retained. DFA produces both discriminant scores (likened to a weighted score of latent writing proficiency) on an interval scale and predicted membership (i.e., band essay scores) as a categorical variable.

Cross validation of the DFA results is recommended to establish repeated sampling reliability, especially for exploratory studies. To achieve this, some studies (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011) split one sample into two subsamples: a training set for predictor selection and a testing set for cross validation. However, repeated sampling reliability for variables selected in this way can be dismissed as the sampling procedure is not actually repeated; and it can be argued that the two sets actually come from one sample. Therefore, in this study, all the results were cross-validated through the leave-one-out method and output into a classification table, which allows for examination of prediction accuracy through classification errors. The DFA was performed to assess how well final essay band scores could be predicted by the selected text features.

3.3. Qualitative analysis of rater discussions

The rater discussion data comprised nearly 20 h (1188 min) of recordings of five expert raters talking about their ratings of individual essays. These recordings had been made during the course of preparing rater training materials and checking for rater consistency over time. The recordings were transcribed, coded, and analyzed in an inductive approach using NVivo, version 10 (QSR International, 2012). The inductive approach requires the researcher to begin the analysis without any preconceived hypotheses and thus allows themes and patterns to emerge from the raw data (Miles & Huberman, 1994). This approach was desirable since the study aimed to assess how raters used the rating criteria and to establish whether they experienced any difficulties with the rating scale. Specifically, rater discussions were closely read several times to generate coding categories (see Appendix B, for details of the categories and sub-categories of the coding scheme). Next, we revisited the data, and dissected and labeled different pieces of data using the emergent codes. Then, we analyzed the data in light of how raters use individual categories, what difficulties raters encountered with the rating scale, and what features they used to distinguish essays across levels.

4. Results of the scale review

4.1. Representative text features of writing proficiency

We explored RQ1 (which text features have been identified by the literature to most effectively predict writing proficiency?) in three steps. First, we reviewed models and frameworks of writing proficiency to identify relevant text features. The Grabe and Kaplan (1996) taxonomy of language knowledge is probably the most widely recognized framework of writing proficiency. Building on communicative competence models (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972), Grabe and Kaplan postulate that writing requires knowledge from three domains, (i.e., linguistic, discourse, and sociolinguistic) in order to generate and monitor the quality of the text. Weigle (2002: 36) argues for a hierarchy among the three knowledge domains; linguistic knowledge is foundational while discourse and sociolinguistic knowledge are higher-order constructs. Linguistic knowledge entails a range of features representing writing fluency, (lexical and syntactic) complexity and accuracy (e.g., Wolfe-Quintero et al., 1998). Limited linguistic knowledge can disrupt a writer's composing process. Rather than focusing on the content, organization, or audience, less proficient writers are pre-occupied with labored searches for and control of syntactic structures and lexical items. However, Crossley and McNamara (2011) argue that a quick access to discourse knowledge (e.g., cohesion and coherence) is equally important for success in L2 writing. This ability tends to distinguish L2 writers from L1 writers as well as across different proficiency levels for L1 and L2 writers. Therefore, for the corpus analysis, we proceeded with text features that represent the linguistic and discourse knowledge domains.

To operationalize these features as measures we looked to Coh–Metrix, an automated text evaluation tool. Coh–Metrix produces 106 measures. Many of these tend to measure similar constructs so caution should be exercised in using the measures in order to avoid redundancy (issue of colinearity). Moreover, it appears that the majority of the Coh–Metrix studies were conducted by authors involved in the development of the tool. It was necessary, therefore, to synthesize the research to address questions typically raised regarding the redundancy and statistical stability of Coh–Metrix indices. Appendix C summarizes the studies that have used Coh–Metrix to investigate the text features that predict overall writing proficiency. Appendix D summarizes the Coh–Metrix studies that have investigated the indicators of lexical proficiency. Using this research synthesis we selected indices that: (1) are consistently significant in predicting writing proficiency across studies; (2) represent a textual feature in the categories of linguistic and discourse knowledge; (3) are the only index selected to represent a particular textual feature. We identified five consistently significant indices that represent distinct linguistic and discourse features of written text. These indices relate to four aspects of the writing construct assessed by the exam:

- *Fluency*—In two of the three Coh–Metrix studies that examined timed writing, text length was the most effective predictor of writing fluency. It is important to note that fluency (as measured by the number of words produced) is associated primarily with timed writing assignments. It is perhaps a less relevant aspect of the writing construct in untimed writing. However, it is possible to argue that time constraints exist even in untimed writing conditions. For example, in higher education settings, undergraduate ESL students tend to have numerous course assignments; in that case, the amount of time they can spend on writing tends to be limited. Therefore, the ability to produce written texts within time constraints is an important construct subsumed under overall writing proficiency.
- *Lexical sophistication*—Lexical diversity and lexical frequency emerged from the Coh–Metrix synthesis as two dimensions of lexical sophistication. For this study, vocd-D (a mathematical variation of type-token ratio that accounts for the impact of text length) was selected to operationalize lexical diversity. Lexical frequency was operationalized by CELEX log word frequency, i.e. a log transformation of the averaged frequency of all words in each essay based on the word frequencies in

Table 2
Summary of predicting variables used in the DFA.

Predictor	Explanation	Range	Computing tool
Length	Total number of words	0–∞	Coh–Metrix
Lexical diversity	VOCD D statistics	0–∞	
Lexical frequency	CELEX log word frequency	0–1000,000	
Cohesion	Proportion of content word overlap across all sentences	0–1	
Syntactic complexity	Number of modifiers per noun	0–∞	
Prompt dependence	Proportion of prompt-related formulaic sequences used by total number of words	0–1	R

the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), a corpus of 17.9 million words assembled along the same criteria as the British National Corpus.

- *Cohesion*—The most effective index of cohesion was content word overlap, which is a global measure of referential cohesion (McNamara et al., 2014). The index is interpreted as the proportion of explicit content words repeated across all sentences. Low-level writers tend to rely on explicit cohesive devices to create relationships between ideas, whereas high-level writers can use a variety of lexical devices to link different ideas to form a coherent argument (cf. Banerjee, Franceschina, & Smith, 2007; Kennedy & Thorp, 2002).
- *Syntactic complexity*—There were two consistently significant indices for syntactic complexity: left embeddedness (i.e., mean number of words before the main verb, an index for subordination and coordination) and noun phrase complexity (i.e., mean number of modifiers per noun). However, as Biber, Gray, and Poonpon (2011) argue, subordination and coordination is more functional in speaking for the purpose of achieving fluency, whereas "complex noun phrase constituents (rather than clause constituents) and complex phrases (rather than clauses)" are more related to the abstraction of ideas, a distinct characteristic of written discourse (p. 5). Therefore, in this study, the mean number of modifiers per noun was selected to represent syntactic complexity.

Importantly, accuracy features (i.e., grammatical accuracy and appropriate use of lexical items) were not included in the corpus analysis as they are not easy to automate.

Finally, because the current rating scale includes prompt dependence, this feature was added to the analysis and operationalized as the proportion of prompt-related formulaic sequences used. Although the use of generic formulaic sequences can be seen as a marker of writing fluency (Ellis, 1996), overreliance on sequences that are either already provided in or are closely related to the prompt can indicate a narrow range of vocabulary and, more importantly, an inability to develop the topic beyond the prompt (e.g., Staples, Egbert, Biber, & McClair, 2013).

To analyze prompt dependence, a bottom-up approach was used to create a master list of formulaic sequences. That is, the list was extracted from the essay corpus and then used to analyze individual essays. This approach was chosen because (1) the formulaic sequences used were thus representative of the writers and corpus examined; and (2) the prompt-related formulaic sequences allowed us to operationalize prompt dependence. The master list was compiled as follows: an R programing script was written to retrieve all the three- to five word sequences from the essay corpus and rank the sequences by frequency. Sequences occurring more than 30 times were retained and manually screened. Duplicates were avoided by only including the longer sequences (e.g., both "on the other" and "the other hand" were duplicates of "on the other hand"). Next, the sequences were classified into two subcategories: prompt-related and non-prompt related. Prompt-related sequences were defined as sequences that either appeared verbatim in the prompt or were closely related to the prompt in meaning. Only prompt-related sequences were retained in the final list, which consisted of 47 formulaic sequences (see Appendix E for the full list). To operationalize prompt dependence, a second R script was written to record the frequency of each prompt-related sequence in each essay. Prompt dependence was computed as the proportion of formulaic words in each essay using the following formula:

$$\text{Proportion of formulaic sequences} = \frac{\sum \left(N_{\text{frequency of each sequence}} \times N_{\text{number of words in each sequence}}\right)}{\text{Total number of words in each essay}}$$

Table 2 presents the final list of text features for the analysis.

It is important to note here that the text features identified from the literature review related primarily to those demonstrating linguistic and discourse knowledge. Also important are features related to sociolinguistic knowledge such as management of the writer-reader relationship and authorial voice (Zhao, 2013). These are addressed in Section 5.

4.2. Text features predicting the essay scores

To answer RQ2 we quantified the text features identified in Table 2 and then performed a DFA on the results. Prior to the DFA, we screened the data for the statistical assumptions of normality, homoscedasticity, multicollinearity, and independence. Appendix F shows that the scores for prompt dependence were not normally distributed at any score level. However, DFA is robust against violations of the normality assumption as long as the sample size is large (Warner, 2013). Therefore, the variable of prompt dependence was retained in the analysis. A pooled within-groups correlation matrix

Table 3
Classification results of essays at individual score levels by DFA.

			Predicted group membership					
		Score level	A	B	C	D	E	Total
Original[a]	Count	A	78	55	12	1	0	146
		B	40	88	60	9	2	199
		C	11	51	133	11	8	214
		D	1	23	71	16	26	137
		E	3	4	26	10	57	100
	%	A	53.4	37.7	8.2	.7	.0	100
		B	20.1	44.2	30.2	4.5	1.0	100
		C	5.1	23.8	62.1	5.1	3.7	100
		D	.7	16.8	51.8	11.7	19	100
		E	3.0	4.0	26.0	10.0	57.0	100
Cross-validated[b]	Count	A	77	56	12	1	0	146
		B	41	85	61	10	2	199
		C	11	51	130	13	9	214
		D	1	23	72	15	26	137
		E	3	5	27	12	53	100
	%	A	52.7	38.4	8.2	.7	.0	100
		B	20.6	42.7	30.7	5.0	1.0	100
		C	5.1	23.8	60.7	6.1	4.2	100
		D	.7	16.8	52.6	10.9	19.0	100
		E	3.0	5.0	27.0	12.0	53.0	100

[a] 46.7% of original grouped cases correctly classified.
[b] 45.2% of cross-validated grouped cases correctly classified.

among the six features (see Appendix G) shows that no correlation exceeded .7 (none were higher than .59 in absolute value), suggesting the absence of multicollinearity (Tabachnick & Fidell, 1996).

Homogeneity of variance/covariance matrices was tested with Box's *M* statistics. Ideally, we would expect a non-significant result for Box's *M* test to indicate homoscedasticity. Using the $\alpha = .01$ significance level, the Box *M* tests was significant for the prediction of band scores ($M = 458.75$, $F(84, 757{,}701) = 5.37$, $p < .01$). However, since the sample size was sufficiently large, violations to the homogeneity of variance/covariance assumption tend to have a small impact on the validity of the results (Warner, 2013).

When comparing essays across the five band scores, four discriminant functions were created. The test for the combined discriminant function was statistically significant: $\chi^2(24) = 591.16$, $p < .001$. Wilks's Λ was .47, suggesting that 53% of the variance in DFA scores was explained by the between-group differences on the six text features. The prediction accuracy of the combined discriminant function (shown at the bottom of Table 3) reached 46.7% in the original and 45.2% with cross-validation.

When individual discriminant functions were examined, Discriminant Functions 3 and 4 were weakly correlated with predicted group membership. The chi-square test for the combined predictive value of Discriminant Functions 3 and 4 was not statistically significant: $\chi^2(8) = 9.75$, $p = .28$ (see Table 4). Although the addition of Function 2 to Discriminant Functions 3 and 4 made the chi-square test statistically significant: $\chi^2(15) = 42.03$, $p = .00$, Function 2 contributed little to the overall prediction. Therefore, this function was removed from the model and only coefficients for Function 1 were interpreted (see Table 5).

The canonical structure coefficients (third column of Table 5) for length (.74), lexical diversity (.39), lexical frequency (−.37), and cohesion (−.32) were above .3 (absolute value), suggesting that these features make important contributions to the meaning of the discriminant function. Therefore, the discriminant function can be interpreted as a combination of each writer's writing fluency, vocabulary knowledge, and the ability to vary the use of content words to express similar meanings. Note, however, that the canonical structure coefficient for lexical frequency and cohesion were negative, indicating a negative correlation between lexical frequency and the discriminant function and between cohesion and the discriminant function. The negative coefficient for cohesion is probably because writers of higher proficiency rely less on explicit cohesive ties (e.g., repetitions) to achieve textual coherence when compared with lower scoring essays. Rather, they are more capable of using

Table 4
Significance test of discriminant function(s).

Function(s)	Wilks' Λ	χ^2	df	p
1 through 4	.47	591.16	24	.00
2 through 4	.95	42.03	15	.00
3 through 4	.99	9.75	8	.28
4	1.00	.52	3	.92

Table 5
Summary of information about individual predictors for DFA.

Predictor	Standardized canonical coefficient	Canonical structure coefficient	Λ	η^2	F
Length	.89	.74[a]	.64	.36	$F(4, 791) = 111.42$[***]
Lexical frequency	−.51	−.37[a]	.87	.13	$F(4, 791) = 29.32$[***]
Syntactic complexity	.25	.20	.95	.05	$F(4, 791) = 10.62$[***]
Cohesion	−.21	−.32[a]	.90	.10	$F(4, 791) = 22.79$[***]
Lexical diversity	.10	.39[a]	.87	.13	$F(4, 791) = 30.55$[***]
Prompt dependence	.03	−.20	.95	.05	$F(4, 791) = 9.67$[***]

[***] $p < .001$.
[a] Canonical structure coefficient for Discriminant Function 1 is above .3.

Table 6
Bonferroni adjusted post-hoc group comparisons.

Predictor	Significant Bonferroni-adjusted group comparisons
Length	A > B > CD > E
Lexical frequency	A < B < CDE
Lexical diversity	AB > CD > E
Cohesion	AB < CDE
Syntactic complexity	A > BCDE
Prompt dependence	AB < CDE

different words to express similar meanings. The negative coefficient for lexical frequency suggests that writers of higher proficiency tend to use less frequent (more rare) words.

The standardized canonical coefficients for Discriminant Function 1 in Table 5 rank length (.89) and lexical frequency (−.51) as the more discriminating predictors of band scores. The effect size (η^2) for individual predictor variables (fifth column of Table 5) indicate that length (.36), lexical diversity (.13), and lexical frequency (.13) contributed the most to the scores awarded to the essays. Cohesion, syntactic complexity, and prompt dependence had lower η^2, indicating that these variables did not effectively discriminate between essays at each score level. Bonferroni adjusted post-hoc group comparisons (see Table 6) confirmed results from the DFA in that only length, lexical diversity, and lexical frequency significantly distinguished essays across multiple score levels.

Table 6 also showed the following patterns:

- As and Bs are significantly different on text length, lexical frequency, and syntactic complexity;
- In terms of cohesion, both As and Bs demonstrated lower proportion of repetition of content words across all sentences;
- As and Bs also tend to rely less on prompt-related formulaic sequences, which suggest the ability of higher proficiency writers to develop an argument beyond the prompt;
- Bs and Cs are significantly different on text length, lexical diversity, lexical frequency, cohesion, and prompt dependence;
- Ds and Es are significantly different on text length and lexical diversity.
- However, there was no significant difference between Cs and Ds on any of the six text features.

A closer examination of the classification accuracy by the model in the DFA (Table 3) revealed that the prediction accuracy was high at the A, C, and E levels, slightly lower at the B level, but rather problematic at the D level. Within the D level, the majority of the essays (51.8% original, 52.6% cross-validated) were misclassified[1] as Cs. The absence of significant differences between Cs and Ds seems to suggest that the real differences of essay quality between Cs and Ds are small. This preliminary conclusion was examined by analyzing the rater discussions.

4.3. Qualitative analysis of rater discussions

Approaches to scale development that focus on the analysis of real performances (performance data driven approaches) are conducive to generating discriminating measures. However, they can be difficult for raters to operationalize. Therefore, it is important to include raters' voices in scale development; an analysis of rater discussions can reveal whether or to what extent examiners have difficulties with the scale. In this study, the rater discussions captured two groups of three or four raters reviewing and debating the scores awarded to approximately 100 essays. All the essays had been rated using the current scale (Appendix A). These discussions were analyzed thematically to examine whether raters applied similar constructs and used these constructs to distinguish writing performances similarly. The thematic analysis of rater discussions addresses RQs 3 and 4. Three major themes related to the usability of the scale emerged: (1) key features distinguishing

[1] It should be pointed out that misclassification means that the essay received one band score but the analysis of linguistic features placed it in another band score. The original (human scored) band score is presumed correct and the categorization based on linguistic features is presumed incorrect.

essays across levels, (2) categories that facilitate or inhibit rater judgments, and (3) difficulties raters encountered with the rating scale.

4.3.1. Key features distinguishing essays across levels

The rater discussions illuminated how raters viewed the different text features of essays at different band scores. Lexical features were most salient at the highest and lowest band scores. In the raters' opinion, As differ from Bs mostly in the effective use of vivid and specific words (often referred to as "lexical richness" by the raters). Vocabulary use in B-level essays tend to be either cautious but appropriate, or more adventurous but less accurate. Es tend to demonstrate limited range of vocabulary.

Raters applied the concepts of syntactic complexity and accuracy together. For instance, both As and Bs demonstrate syntactic complexity. However, A-level essays tend to display nearly perfect morphosyntactic control, while B-level essays may contain minor errors in complex structures. C-level essays may display some level of syntactic complexity, but with inconsistent morphological control. Indeed the analysis revealed that raters frequently noted accuracy features (i.e., grammatical accuracy) when distinguishing between levels. For instance, in the case of E-level essays, the accumulation of errors in the language makes the text mostly incomprehensible. Reconstruction of meaning often fails. Accuracy features had not been included in the corpus analysis as they are not easy to automate. However, the rater discussions showed that these features are important scoring criteria.

The rater discussions characterized the discourse features of the essays as argument development and organization. In the evaluation of essays at different band scores, these two criteria were viewed as connected but separate. For instance, essays at the A- and B-levels typically present clear, coherent, and well developed arguments. However, C-level essays, though well-organized, are generally not sufficiently developed. Further, D-level essays are meaningful mostly at the sentence level but the language is vague, repetitive or incoherent, requiring raters to reconstruct the meaning and the main argument.

4.3.2. Categories that facilitate or inhibit rater judgments

Interestingly, the discourse features of the essays (i.e. textual coherence, topic relevance, and argument development) were the greatest source of disagreement between the raters. The case of a D-level essay illustrates the difficulty of rater alignment. This essay featured an incoherent main message, poor morphological control, overreliance on memorized chunks, and inappropriate use of formulaic discourse markers. During the discussion, R2 became aligned with R1 on the poor morphological control and inappropriate use of formulaic discourse markers. However, R2's evaluation clearly diverged from R1's on the clarity of meaning and the coherence of argument. Even after an extended debate, the raters could not reach a consensus on whether the essay presented a clear and coherent message:

> **R1**: To me, the fact that you have to do that [rewriting the essay in her head]. . .makes the difference between a "C" and a "D". With a "C", they make their point. With a "D", they only make their point with me rewriting in my head.
>
> **R2**: I don't feel like I have to rewrite it. I think the meaning is clear. . .I think the weaknesses are with plurals and articles. . .but I'm not sure those problems are serious enough to condemn it to a "D".
>
> . . .
>
> **R1**: You've got to throw out all these prefabricated chunks which give the sense of fluency but don't add to meaning.
>
> **R2**: Well, those prefabricated chunks didn't have that effect on me. I mean, I read it and I didn't feel like I had to do much to put the meaning together.
>
> (Transcript, meeting 5)

Although the use of various discourse features is crucial to the construction of a clear and coherent argument, these features tend to overlap and interact with each other. For example, when discussing the rating of topic relevance, argumentation, and organization, the raters' discussion suggests the difficulty of operationalizing these discourse features.

> **R3**: I mean, [the essay] was a slight off-topic.
>
> **R1**: The way this rating scale has been written. . .I see that as "problems with connection," because, if you think of connection as not just being connector words, but as being connections between ideas. . .if something goes off-topic, it's not connected back to the main point of the prompt.
>
> **R2**: Yeah, it says, "organization well-controlled and appropriate to the material" for Rhetoric under "A", and "connection is smooth." So, perhaps it wouldn't achieve either of those.
>
> (Transcript, meeting 7)

Arguably, the effectiveness of discourse features varies according to a rater's world knowledge, expectation, and interpretation (McNamara, Crossley, & McCarthy, 2010; Tapiero, 2007). Therefore, rater disagreement is to be expected.

4.3.3. *Difficulties raters encountered with the rating scale*

Indeed, raters generally had more difficulty in classifying and operationalizing the discourse features implied by the rhetoric criterion on the rating scale. For instance, they struggled with the concept of logical reasoning. They debated whether it should be included in the operationalization of rhetoric or whether rating for logic introduced construct-irrelevant variance:

> **R4**: We're venturing into the territory of actual aptitude and cognitive capabilities, rather than linguistic capabilities. Like, if you're a very good speaker of a language, but you aren't particularly excellent at stringing your thoughts together in a cohesive written form, in the way that we require for an "A", we're testing both their proficiency [in] written English [and] their proficiency of logic.
>
> **R1**: This is one of the problems with language, that there comes a point where language literacy practices and cognitive development are hard to disentangle.
>
> **R2**: So at the C2 [level], there has to be a little bit of interplay between their cognitive ability and linguistic ability.
>
> (Transcript, meeting 8)

In support of R1's and R2's claims, the CEFR also specifies "effective logical structure" in the descriptors for overall writing proficiency of C2 (Council of Europe, 2001, p. 27). Since the exam and the rating scale target the C2 level, it is reasonable to argue that, at this level of language performance, the boundary between "pure" language skills and cognitive abilities may be less clear. However, if this definition of language proficiency (i.e. C2 on the CEFR) is to be fully operationalized, the relationship between logical reasoning and language abilities needs to be addressed in the rating scale.

Another area of difficulty emerged in discussions of topic relevance. Though not explicitly mentioned in the scale, topic relevance was frequently used by raters to evaluate the essays. Raters referred to it when marking down essays that were not on-topic. Since topic relevance is not explicitly included in theoretical models of writing proficiency or language proficiency, this was a matter of concern for R2:

> **R2**: [Topic relevance] tends not to be in the literature. If you look at the definition of writing proficiency. . .not many of them that I've looked at include relevance to the topic.
>
> (Transcript, meeting 7)

However, R1 linked topic relevance to audience awareness, which is often listed under the sociolinguistic knowledge required when writing (e.g., Grabe & Kaplan, 1996). According to R1, it is reasonable for a reader to expect the essay to stay relevant to the topic or prompt. Therefore, a writer with a high level of audience awareness will make efforts to meet the readers' expectations in their writing.

> **R1**: It's something that we need to look at the Common European Framework for. . .when you are writing at a very high level, part of getting your message across is telling people what they're expecting to hear about. . .which is the prompt. The prompt tells people what to expect. . .and then your answer should be related [to] what they expect.
>
> **R3**: Right. . .the "communication."
>
> **R2**: Yeah.
>
> (Transcript, meeting 7)

The CEFR descriptors for writing proficiency at C1 state that writing performance at this level should be able to underline "the relevant salient issues" (Council of Europe, 2001, p. 27) and use "relevant examples". Therefore, it is not unreasonable to include topic relevance as part of the construct in the rating scale, especially if raters frequently use this feature to evaluate the essays.

4.4. *Triangulation of quantitative and qualitative results*

The findings from the quantitative (DFA) analysis and the qualitative (rater discussion) analysis were triangulated to answer RQ5. Table 7 shows the alignments and discrepancies between the findings from the two analysis strands.

The quantitative and qualitative results agreed in three major respects. First, linguistic features were more reliable and discriminating than discourse features. This is perhaps because linguistic features tend to be more concrete and have a long-standing traditional place in both language assessment and language instruction. In contrast, certain discourse features are more difficult to operationalize and might not have been clearly defined in the existing scale. Second, lexical diversity and lexical frequency were discriminating features across writing proficiency levels. This corresponds with previous research, suggesting lexical knowledge is an important component of writing proficiency. Third, the C-D distinctions were less stable. The analysis of rater discussions suggests that the distinction between the two levels lies in the presence of a coherent main argument. That is, C-level essays require little effort from the raters to infer the main argument while Ds tend to require reconstruction of meaning across sentences. However, since raters often disagreed on the rating of discourse features such as coherence and argument development, it is unsurprising that raters operationalized the boundaries between Cs and Ds less reliably.

Table 7
Triangulation of quantitative and qualitative results.

	Quantitative		Qualitative
Alignments	Linguistic features were stronger predictors of essay band scores than discourse features	Linguistic vs. discourse features	Raters agreed more easily and frequently on linguistic features than on discourse features
	A-level essays displayed significantly more low-frequency words than B-level essays E-level essays demonstrated significantly lower lexical diversity than D-level essays	Lexical knowledge	A-level essays showed more vivid and specific words (which tend to be of lower-frequency than generic words) than B-level essays E-level essays showed limited linguistic (lexical and syntactic) resources compared with D-level essays
	There were no significant difference in the text features between Cs and Ds	C–D distinctions	Raters expressed difficulty in differentiating C- and D-level essays
Discrepancies	Syntactic complexity only significantly differentiated A-level essays from B-level essays	Syntactic complexity	Raters frequently used syntactic complexity criterion to distinguish A-, B, C-, and D-level essays
	C-level essays did not show significant difference in the use of prompt-related sequences from D- and E-level essays	Prompt dependence	Raters considered prompt dependence as a typical feature of D- and E-level essays, but not C-level essays

However, there were two key discrepancies between the quantitative and qualitative analyses. First, the triangulated results raised questions about the operationalization of syntactic complexity as a scoring criterion. Syntactic complexity (operationalized as number of modifiers per noun by Coh–Metrix) did not significantly distinguish essays at multiple (i.e. the B, C, and D) levels. This is probably because, in our analysis, syntactic complexity was reduced to a single measure. Our operationalization of syntactic complexity was probably insufficiently comprehensive and did not fully capture the range of complexity features in the essays, a point we shall return to at the end of this paper. In addition, accuracy features were not included in the Coh–Metrix analysis. However, when applying the syntax criterion, raters examined not only complexity but also accuracy. If syntactic accuracy were included it might help distinguish essays across levels. The second discrepancy related to the prompt dependence measure (operationalized as the proportion of prompt-related formulaic sequences). This failed to show a significant difference between the C-, D-, and E-levels. Rater discussions suggest that, although C-level essays may display the use of formulaic sequences, the formulaic language tends to be used appropriately. This is a reasonable explanation for why C was not significantly different from D and E on the prompt dependence measure. Even though C-level essays may include generic language use closely related to the prompt, writers at this level are more capable of using the prompt-related formulaic sequences effectively.

5. Revisions to the rating scale

Though many aspects of the current scale proved robust under scrutiny, the scale review yielded a number of recommendations for revision:

- Lexical knowledge was confirmed to be a strong predictor of writing proficiency. However, the current descriptors for vocabulary should be elaborated to better account for lexical diversity and lexical frequency. In addition, appropriateness and accuracy of vocabulary use should be included in the scale.
- Syntactic complexity failed to distinguish essays across multiple score levels. However, this is probably because the Coh–Metrix analysis reduced syntactic complexity to a single measure and accuracy features were not included in the DFA. Rater discussions show that raters frequently used a combination of syntactic complexity and accuracy to differentiate essays across multiple levels. For example, morphosyntactic control of complex sentences appeared to distinguish between B and C levels; and the amount of basic morphosyntatic errors distinguished between Ds and Es. These features should be stated more clearly in the descriptors for Grammar and Syntax.
- Raters appeared to adopt different operationalizations of certain discourse features, and use those features interchangeably or indistinguishably with other features. A possible solution to this problem would be to divide the rhetoric category into two smaller but distinct subcategories: argumentation and organization. Argumentation entails comprehensibility, clarity of argument and the development of the argument with supporting details. Organization refers to the connection both within and between paragraphs. It is concerned with the structure of individual paragraphs and the overall essay. This division of the rhetoric category may help create more reliable and distinct descriptors for each subcategory.
- Arguably, skills in the sociolinguistic knowledge domain are an integral part of writing proficiency. Although not included in the current rating scale, raters seemed to apply, to a varying extent, sociolinguistic features (e.g., audience awareness) in essay scoring. Higher proficiency writers tend to achieve better communicative effects through original content and reader engagement. Therefore, it may be effective to incorporate sociolinguistic features in the revised rating scale.

5.1. Revision process

The process of revising the rating scale occurred in three broad stages. First, the results of the scale review were discussed in detail. In addition to examining the recommendations generated by the scale review process, the project team also reviewed the descriptors of C1 and C2 writing laid out in the CEFR, taking particular note of repeated references to "logical structure" and "well-structured." Finally, we independently re-read papers on rating scale design (Grabe & Kaplan, 1996; Knoch, 2011; Turner & Upshur, 2002; Weigle, 2002) and on authorial voice (Zhao, 2013).

Subsequently, the process of drafting and revising the new rating scale was begun. Working separately we each developed scoring criteria that we felt best reflected the consolidation of these sources, which would then be used to inform the new rating scale. At the first scale revision meeting we found that the features we had separately identified were broadly similar. While terminology differed, the team was in agreement that the vocabulary and grammar/syntax criteria should be retained, and that the rhetoric criterion was too broad; the latter should be divided into the more specific criteria of content/topic development and organization/connection of ideas. Finally, some team members felt that a category discussed in the literature but not addressed by the current scale concerned the writer-reader relationship, and might be expressed as "audience awareness", "communicative effect", or "voice." While there was some skepticism that this type of sociolinguistic knowledge could be objectively evaluated or that clear score levels could be adequately described, we decided to include a stylistic measure in the initial draft to see if it could be operationalized. The group's decisions were incorporated into Draft 1 of a revised rating scale. More detailed descriptors were written for the vocabulary and grammar/syntax criteria and rhetoric was divided into two new criteria, "Content and Topic Development" and "Organization and Connection of Ideas." Finally, the criterion "Authorial Voice" was added, largely informed by the discussion of sociolinguistic awareness and the work of Zhao (2013).

During the review of Draft 1, the new descriptors were inspected for clarity and consistency of wording. Concern remained about overlap between the criteria "Content & Topic Development" and "Organization and Connection of Ideas". For instance, did the feature 'development of argument' (in "Content & Topic Development") assume the feature 'appropriate macro-informational structuring' (in Organization & Connection of Ideas")? However, the majority of the Draft 1 review meeting was spent discussing the category "Authorial Voice". We were concerned that the descriptors overlapped with other macro categories like topic development, or that "Authorial Voice" could be interpreted by raters to refer to which essays they personally found most interesting. Therefore, arriving at a shared conception of what was meant by "Authorial Voice" was an important discussion point.

Based on the Draft 1 discussion, a second draft of the scale was prepared. The names of each criterion and descriptors for each score point were refined and clarified. For instance, the criterion "Content and Topic Development" was revised to "Topic Development". This is because "Organization and Connection of Ideas" and "Authorial Voice" also deal with essay content. Another change was to the order of the criteria; the macro features dealing with the essay as a whole such as "Topic Development" were placed before micro criteria such as "Vocabulary". As a result of these edits, Draft 2 was a descriptively consistent, theoretically grounded rating scale for writing at the C2 level.

Once we were satisfied with the wording and flow of the descriptors, it was necessary to evaluate how readily the scale could be applied. The scale was field-tested on a sample of 30 essays with original scores representing both the actual test population and the full range of potential scores. Two readers evaluated the essays independently and then met to discuss their experiences applying the scale. The readers found that the two lowest-performance score points were under-described and not appropriately aligned with the abilities of the test population. Consequently, few essays were being awarded those scores. This resulted in substantial revision to the descriptors for score points 1 and 2, resulting in Draft 3 of the scale.

Draft 3 was then field-tested by a third reader, who applied it to the same 30 essays. This generated more areas for revision. References (in the "Authorial Voice" criterion) to whether the response is "of interest to the reader" were removed because they were found to be too subjective. Other edits resolved inconsistencies in descriptor wording. These edits resulted in the final draft, a revised rating scale which is theoretically sound, designed to capture performances at C2 on the CEFR, and is relevant to the exam's target population. The revised rating scale (Appendix H) consists of five criteria: topic development, organization and connection of ideas, grammar and syntax, vocabulary, and authorial voice. In order to create internally consistent descriptors at each score point, three distinguishing features were identified for each criterion. For instance, the distinguishing features for the grammar and syntax criterion are: complexity of syntax, accuracy of syntax, and clarity of meaning. Together, the criteria represent the underlying construct of the writing test. They can be used to examine a test taker response for linguistic, discourse, and sociolinguistic knowledge.

It is important to note, however, that the scale has not yet been comprehensively validated. In a subsequent phase, the scale will be applied to a larger sample of essays. Rater reliability will be calculated both at the trait level and at the overall score level. Additionally, many-faceted Rasch measurement (MFRM) analysis will be used to confirm that the newly defined criteria are distinct measurement facets.

6. Reflections

Despite the rigorous approach adopted, the purpose of the scale (to assess advanced-level writing proficiency in the context of a large-scale English language proficiency examination) and the limitations of automated text analysis leave us with four issues to consider. The first is the division of the scale into five levels. Since the exam targets advanced-level

(C2) language proficiency, it is questionable whether the rating scale can meaningfully distinguish five performance levels. Admittedly, the decision to describe five levels of performance was an artifact of the larger exam of which the writing test is a part. Each section on the exam describes five levels of performance: three passing levels and two failing levels. This could be viewed as a practical constraint that lacks theoretical support. Interestingly, however, both the quantitative and qualitative analyses showed that five levels could be described.

The challenges of level definition are greatest at the pass-fail boundary (between levels C and D). This is the second issue that must be considered. Test takers at levels C and D attempt complex sentence constructions and try to use less frequent vocabulary. However, D-level responses are less successful; they are grammatically and lexically less accurate and require the reader to expend effort to understand the intended meaning. Measures of error (grammatical and lexical accuracy) would therefore be useful. However, a review of the literature suggested that measures of accuracy tend to be less accurately automated and are more reliably human scored (Attali & Burstein, 2004). We therefore relied primarily on theory and the rater discussions for distinctions of grammatical and lexical accuracy. Arguably, this is not ideal and additional analyses in which the corpus is hand-coded for errors would be a profitable approach. Perhaps a grammar and lexis 'error measure' might be a useful extension of this work. But this would be a complex undertaking that would have to define the nature of the error, its severity (perhaps in relation to the intended meaning but also in relation to what test takers at this level might be expected to have mastered) and also its effect upon the text as a whole.

The third issue to consider is a natural corollary of the challenges associated with defining five levels of performance for each assessment criterion. Not all the criteria are salient at every level on the scale. For instance, audience awareness becomes salient at the higher score levels because higher proficiency writers are more likely to have the sociolinguistic tools to engage with the reader. Additionally, morphosyntactic accuracy is typically more salient at lower levels of language proficiency while morphosyntactic appropriacy is more salient at higher levels of language proficiency. It could be argued, therefore, that the revised rating scale should have different criteria at different levels rather than degrees of the same criteria across the levels. Such scales, however, are not commonplace in standardized writing assessment and would present a number of challenges. For instance, examiners would need extensive training to ensure that they knew when to apply particular criteria, i.e. when is the essay so weak that you can ignore the audience awareness criteria? Additionally, the interpretation of final scores would become complex since different criteria had been used to assess essays at different levels of proficiency. However, for local or smaller-scale writing assessments, a modified (compromise) approach to this issue is possible. That is, identify benchmark performances at each level and provide well-structured rater training to attune raters to the typical benchmark performances across levels rather than conceptualizing and operationalizing a continuum of marked difference for each criterion along the rating scale.

A final issue to consider is the Coh–Metrix measure of syntactic complexity adopted for this analysis. As previously noted, the measure was not helpful in distinguishing essays at different levels. It is possible that syntactic complexity is present in all essays at the advanced-level and that it cannot help us to meaningfully distinguish essays of different quality within the broad category of 'advanced'. Additionally, there are so many facets within syntactic complexity, indeed even within the measure of noun phrase complexity that we adopted, that a single indicator might not be sufficiently sensitive to differentiate between levels. However, it is also possible that the measure provided by Coh–Metrix cannot accurately represent the range of syntactic features in writing. Even within the comprehensive investigations of syntactic complexity, there is a division in terms of appropriateness and representativeness between holistic measures (e.g., T-unit measures, Coh–Metrix indices) and more fine-grained analyses of individual syntactic features. As an aspect of linguistic knowledge, syntactic complexity would benefit from an approach that combines fine-grained and holistic analyses, looking at multiple indicators and using multi-dimensional analysis.

7. Conclusion

We began this paper with the comment that discussions of scale design and construction are rare in the literature. As a consequence, test developers have very few examples of the scale development process to guide their work. Our work provides a concrete example of one approach. Additionally, our approach is extremely unusual. Scale designers rarely combine theory, analysis of performance data, and rater input. This study demonstrates the usefulness of such a three-pronged approach to scale development and revision. The result is a revised rating scale that more clearly represents the test's intended construct as well as writer proficiency. We have showcased the opportunities provided by automated text analysis and listed the additional analyses that could be profitable when refining and validating scale descriptors. Our findings indicate that this approach provides rich, theory- and data-driven justifications for rating scale criteria.

In its current iteration, the rating scale is ready for validation using Rasch Analysis to confirm that the newly defined criteria are distinct measurement facets. Although this approach requires time and resources, the outcomes are directly relevant to the scale's construct definition and, by extension, to the construct definition of the test. Since the focus of each criterion has been refined, the scale better supports the rating process. The findings also provide preliminary evidence for the construct representativeness of the scale. Finally, our results have yielded diagnostic information that could be used in the development of score profiles. Therefore, though this was not our primary intention, the results are also relevant for score reporting.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.asw.2015.07.001.

References

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater V.2.0. In *Presented at the annual meeting of the international association for educational assessment* June, 2004, Philadelphia, PA,. Retrieved on July 16, 2014 from ⟨https://www.ets.org/Media/Products/e-rater/erater_IAEA.pdf⟩

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

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

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels. In S. Walsh (Ed.), *IELTS Research Report* (Vol. 7) (pp. 241–309). Canberra: British Council and IDP:IELTS Australia.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5–35.

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

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Retrieved from ⟨http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf⟩.

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, *20*(4), 271–285.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282–311.

Ducasse, A. M. (2009). Raters as scale makers for an L2 Spanish speaking test: Using paired test discourse to develop a rating scale for communicative interaction. In A. Brown, & K. Hill (Eds.), *Task and criteria in performance assessment: Proceedings of the 28th language testing research colloquium* (pp. 1–22). New York, NY: Peter Lang.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition*, *18*, 91–126.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287–291.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing*. New York, NY: Longman.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. NJ: Ablex.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth: Penguin.

Kennedy, C., & Thorp, D. (2002). A corpus investigation of linguistic responses to an IELTS academic writing task. In L. Taylor, & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 316–377). Cambridge: UCLES/Cambridge University Press.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304.

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*, 81–96.

Lim, G. S. (2012). Developing and validating a mark scheme for Writing. *Cambridge ESOL: Research Notes*, *49*, 6–9.

Lantolf, J. P., & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal*, *69*(4), 337–345.

Lowe, P. (1986). Proficiency: Panacea, framework, process? A reply to Kramsch, Schulz, and particularly Bachman and Savignon. *Modern Language Journal*, *70*(4), 391–397.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh–Metrix*. Cambridge, MA: Cambridge University Press.

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

McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221–242.

Mickan, P. (2003). *'What's your score?' An investigation into language descriptors for rating written performance*. Canberra: IELTS Australia.

Miles, M. B., & Huberman, A. M. (1994). Early steps in analysis. In M. B. Miles, & A. M. Huberman (Eds.), *Qualitative data analysis: An expanded sourcebook* (2nd ed., Vol. 7, pp. 50–89). Thousand Oaks, CA: Sage.

QSR International. (2012). *NVivo 10 [Computer software]*,. Available from ⟨http://www.qsrinternational.com⟩.

Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, *16*(4), 457–478.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and academic writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York, NY: Harper Collins.

Tapiero, I. (2007). *Situation models and levels of coherence: Toward a definition of comprehension*. New York, NY: Taylor & Francis Group, LLC.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49–70.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language testers. *English Language Teaching Journal*, *49*(1), 3–12.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 84–111.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques*. London, UK: SAGE Publications.

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

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawaii: Second Language Teaching and Curriculum Center.

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing*, *30*(2), 201–230.

Jayanti Banerjee Jayanti has 25 years of experience in teaching and assessing writing at secondary and tertiary levels of education. She has delivered workshops on the assessment of listening and speaking, research methods, and qualitative approaches to data analysis. Jayanti has published in the areas of language testing and English for academic purposes and presented papers at a number of international conferences. Her recent research activities include DIF investigations, reviews of writing and speaking rating scales, and an exploration into alternative approaches to testing listening.

Xun Yan is an assistant professor of linguistics at University of Illinois at Urbana-Champaign, specializing in language testing, second language acquisition, and world Englishes. His current research interests include the development and quality control of post-admission language assessments, assessment literacy for language teachers, formulaic language acquisition and lexical development by L2 speakers, L2 pronunciation and intelligibility, and test score use in educational settings.

Mark Chapman Mark directs test development for the WIDA test programs at The University of Wisconsin—Madison. He coordinates test development with The Center for Applied Linguistics (CAL) to ensure the quality of all test content. He also contributes to the development of formative language assessment tools for the pre-K Early Years program. Mark's research interests include writing assessment (prompt effect, scale development and revision, and approaches to rater training), score report design, and the development of enhanced item types.

Heather Elliott Heather is an Assessment Specialist at CaMLA. She writes and reviews test items, coordinates exam compilation, and reviews test item performance. Heather holds a BA in English and a TEFL Certificate from University of Wisconsin-Eau Claire and an MFA in creative writing from Minnesota State University, Mankato. She has taught English and writing in a variety of contexts, including primary school students in Dalian, China, university students in Minnesota, and adult English language learners at a language school in Michigan.