Research Note

Modelling Score Variation in Student Writing with a Big Data System: Benefits, Challenges, and Ways Forward

Lee McCallum, *University of Exeter*

Structured Abstract

- **Aim:** This theoretically-oriented research note has multiple aims. First, the note sets out approaches to studying the relationship between linguistic features and essay grades and how this has influenced approaches to score modelling in the domain of student university writing. Second, and more pressingly, the note makes the case for this work to continue by scholars questioning how they have gone about uncovering this language use and how they have used this knowledge to construct models of writing assessment and/or model what lies behind assessment scores. As part of this questioning, the note introduces the method of mixed-effects modelling as a robust alternative to traditional linear regression modelling techniques. This method allows us to determine how the use of language influences assessment scores, while taking account of individual writer and rater variables as well as contextual variables that include task and topic in this modelling process. The note encourages this research in the context of a first-year university writing program in a large U.S. university that has currently set up a "big data" text repository system to allow researchers the opportunity to carry out large-scale corpus examinations of language use. The note concludes by outlining some of the key challenges that scholars need to be aware of when undertaking such work.
- **Problem Formation:** This section explores how previous studies have attempted to examine the role of various assessment variables on the rating process. These studies focus on interconnected research strands: the role of the writer in completing the assessment task(s); and the role of the rater in

the assessment process. Among these strands, there has been a focus on the relationship between the linguistic features that students use to complete assessments and the grade score awarded and how this relationship may be mediated by other rater and writer characteristics. The review then narrows to analyze the methodological approaches in these studies and subsequently sets the scene to introduce and promote mixed-effects modelling as a viable method to model these assessment constructs and relationships.

- **Information Collection:** Building on the review of studies, this section begins by showcasing work that has used mixed-effects modelling in an attempt to minimize previous studies' methodological shortcomings. This section outlines how such exemplary work takes account of statistical dependency in corpus data sets and highlights the feasibility of using such mixed-effects modelling on the "big data" system at the University of South Florida (USF). Several theoretical and empirical points are made here in terms of considering practicality and the caveats involved in working with big data systems.
- **Conclusions:** Mixed-effects modelling appears to offer a reliable method that First-Year Composition (FYC) researchers can make use of in their study of numerous course and learner variables that influence multiple outcome variables in FYC programs. When we apply this method to the "big data" system at USF, it appears that the method can offer a robust and more accurate estimation of the relationship between student writers' language use, grades, and mediating course and learner variables. However, the method and the treatment of the data contained in such a system need to be considered cautiously, as the effects of sample size discrepancies across variable levels and the issue of missing data need further exploration.
- **Directions for Future Research:** Although the use of mixed-effects modelling is warranted from a theoretical and empirical evidence base, researchers need to take this work forward by also asking questions of the data structure and the variables contained within such data warehouse systems. Future research needs to examine how big data systems that are unbalanced can be used and how the presence of uneven data collection can influence the use of mixed-effects modelling.

Keywords: first-year writing, linguistic features, mixed-effects models, writing analytics

1.0 Aim

Many scholars have studied the scoring process and factors that influence grade/proficiency level variation in writing assessment as part of a wider drive to model different assessment constructs in first-year university/college writing and English for Academic Purposes (EAP) learning domains. Embedded in modelling the scoring process in these domains is a focus on understanding how the linguistic features that students use influence grade scores and at the same time, how these features are influenced by writer, rater, and contextual variables. This line of research inquiry has many important benefits for the study of university writing, especially the types of early university writing that feature in first-year composition programs in the U.S. These perceived benefits stem from the fact that little is known about how students use language in these programs (McDonald, 2007) and more importantly, even less is known about how this language use facilitates achieving task and program writing goals (Aull, 2017, 2019). In this sense, the promotion of language-oriented research would allow the first-year writing research community the opportunity to relate this language use to wider aspects of scoring and rater behavior and ultimately develop an understanding of writing assessment that balances socio-cultural and linguistic factors.

Considering these perceived benefits, this research note has several inter-related aims. First, the note aims to establish previous work that has focused on modelling the scoring process and the role of linguistic features and writer and rater variables. Second, the note aims to critically review the monofactorial methods behind these studies in order to explain their methodological shortcomings, and third, to introduce how the method of mixed-effects modelling can yield more conservative and inclusive estimations of the scoring process. The note addresses these aims by drawing on and reviewing a substantial body of EAP and first-year writing research that examines the variables that help "shape" a score and equally describes how students' language use varies across these score levels that represent proficient writing. Building on this emerging body of knowledge, the note explores the feasibility of adopting mixed-effects modelling within a large scale "big data" learner corpus that acts as an example data set that is typical of the teaching and assessment setup that we see in first-year writing programs in the U.S.

This research note concludes by offering several cautionary caveats that the international writing community needs to consider when choosing to implement mixed-effects modelling when using big data systems. It should be acknowledged that the aim of such a note is not to focus on the mathematical properties of such models (readers are referred to Snijders & Bosker,1999 and Hox,2002 for such descriptions), but to raise awareness of the effects that a hierarchical corpus structure potentially has on estimating and unpacking relationships between writers' language use and respective learner and course writing variables. A clear case is made that the fine-grained consideration of these variables will result in more robust measures of scoring, which can, in turn be used to inform automated scoring systems by incorporating these nuances into the systems and improve rater training and assessment decisions by raising awareness of features that characterize different levels of proficient writing.

2.0 Problem Formation

2.1 Models of Assessment, Language Description, and Our Understanding of the Scoring Process

Underpinning the current work on writing and its assessment is an awareness of how the assessment process has been outlined in several influential frameworks. These frameworks set out the key constructs and sub-constructs that are involved in assessment and how they are brought together. These frameworks have often emanated from U.K. contexts and gone on to influence the assessment of writing across international contexts, including those that assess early academic writing in Euro-centric EAP settings, as well as first-year university and college writing in the U.S. Before we examine such a framework, it is important to understand the shift in emphasis in early university/college writing teaching and assessment that contextualizes the influence and adoption of such a framework. In the U.S., the historical development of assessing writing began with direct essay tests of writing, then moved to treating writing as another endeavor that could be indirectly assessed as a kind of closed knowledge through multiple choice questions. More recently, the U.S. assessment community has returned to understanding writing assessment as a kind of socio-cultural endeavor that creates multiple types of knowledge and therefore must be assessed directly through extended written prose (Spolsky, 1994; Weir, Vidakovic, & Galaczi, 2013).

In recognizing the need for a more socially-informed assessment framework, Weir et al.'s (2013) socio-cultural framework of writing assessment illustrates the move away from the narrow focus of indirect tests and recognizes the social and cognitive aspects of the writing process. The framework comprises three key aspects which are inter-related: scoring, cognitive, and contextual types of validity with the latter divided into two strands: setting and linguistic demands. The framework is set out in Figure 1.

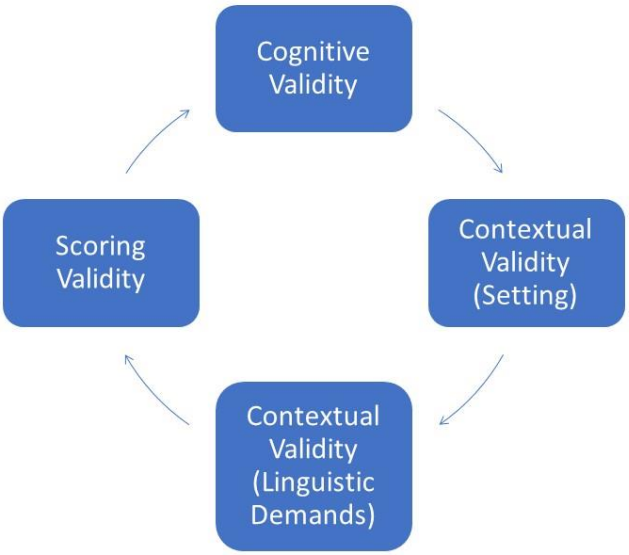

Figure 1. Socio-cultural model of writing (Weir et al., 2013).

These components include several sub-components which are detailed below:

- **Cognitive validity:** macro planning, organization, micro planning, translation, monitoring, and revising
- **Contextual validity (setting):** response format, purpose, knowledge of criteria, text length, time constraints, and writer-reader relationship
- **Contextual validity (linguistic demands):** task input and output, lexical resources, structural resources, discourse mode, functional resources, and content knowledge
- **Scoring validity:** criteria/rating scale, rater characteristics, rating process, rating conditions, rater training, post examination adjustment, and grading

Figure 1 allows us to observe the key components and considerations of assessing writing. It is important to draw attention to the appreciation of both context and assessment- or program-related concerns with scoring. Many scholars have sought to establish how these aspects of validity come together in the assessment process to "shape" the scores awarded to students. A key work in this area emanates from the Cambridge Assessment / Cambridge University Press-led *English Profile* project (Hawkey & Barker, 2004), with their main aim of initially identifying how learner language develops across Common European Framework of Reference (CEFR) proficiency levels and then later extending this in many CEFR grading contexts to show how language use shapes the scores awarded by raters.

In this respect, several scholars have modelled this notion of writing score or quality by incorporating aspects of Weir et al.'s (2013) framework as variables in their studies. In this work, there is an awareness that the understanding of the connection between the elements of scoring (e.g., scoring criteria, rater characteristics, and judgements) and cognitive and contextual factors (e.g., lexical/functional resources, task, and topic) is particularly underdeveloped. Under this view, the scoring criteria are said to be "underspecified" in the sense that how task achievement is fulfilled through specific language features is unclear,

with the decision-making processes of raters and their perceptions of this language use largely unknown. While this has an influence on providing adequate language description in grading rubrics, this underspecification also influences efforts to design automated scoring systems because it means distinguishing between features across proficiency levels is also unaccounted for and affects the success of the system in predicting essay scores (e.g., Chen, Fife, Bejar, & Rupp, 2016). Another consideration in this work has been that certain language features will be more prevalent at different proficiency levels and across different writing tasks, therefore introducing variables that straddle aspects of cognitive and contextual validity. This latter consideration led Hawkey and Barker (2004) and Hawkins and Filipovic (2012) in the *English Profile* to descriptively map out the linguistic features that characterize CEFR levels in terms of their frequency and functionality and how they differ across tasks.

In operationalizing language development and proficiency band scale components more generally, scholars have examined learner language under the domains of complexity, accuracy, and fluency (CAF). Broadly, complexity is concerned with the diversity and degree of sophistication that language items have, with diverse, sophisticated writing thought to center on being diverse in word types and less frequently used in that the more words from appropriate and/or specialized language modes or genres used, the more favorably the writing may be scored (Verspoor, Lowie, Chan, & Vahtrick, 2017). Accuracy relates to the extent the writing is error-free or the extent it conforms to native speaker norms (Evans, Hartshorn, Cox, & de Jel, 2014), and fluency broadly relates to how many words a student writes (Wolfe-Quintero, Inagaki, & Kim, 1998).

This profiling work has since led to a number of other Euro-centric CEFR, EAP, and more limited first-year writing-related studies being conducted to describe language use and model score variation in preparatory courses for university study, early university writing, and early postgraduate writing (e.g., Casal & Lee, 2019; Garner, Crossley, & Kyle, 2018a, 2018b; Granger & Bestgen, 2014; Krzeminska-Adamek, 2016; Lemmouh, 2008; Lu, 2011; Meara & Bell, 2001; Parkinson & Musgrave, 2014; Perin & Lauterbach, 2018; Staples & Reppen, 2016; and Taguchi, Crawford, & Wetzel, 2013). In attempting to map out these features across different assessment scales and determine their relationship to writing quality scores, some studies have taken a largely descriptive approach (e.g., Taguchi et al., 2013), while others have focused on following up on this descriptive "map" by looking at how these features can influence automated scoring systems via regression modelling. These quantitative studies have measured these relationships by asking empirical research questions such as:

- Are there identifiable linguistics features that are indicative of more proficient writing? (Taguchi et al., 2013).
- To what extent do high-, mid-, and low-rated research papers written by second language undergraduate students in a first-year writing course differ in their global, clausal, and phrasal syntactic complexity? (Casal & Lee, 2019).
- What is the relationship between lexical sophistication and independent writing task proficiency scores? (Kyle & Crossley, 2016).

Other more recent studies have also started to ask empirical questions about these feature-grade relationships when writer (e.g., writers' language backgrounds) and other contextual variables are taken into account (e.g., genre and topic):

- How does EFL students' writing differ in the measures of syntactic, lexical, and morphological complexity across different topics? (Yoon, 2017).
- How does task complexity, task type, and/or instructional focus impact the complexity and accuracy of the language use in global as well as specific features or structures? (Alexopoulou, Michel, Murakami, & Meurers, 2017).

Through these research questions, we see nuanced models of scoring: a focus on the perceived simplistic relationship between linguistic features and writing quality; and a focus that incorporates writer and contextual variables into this relationship. The nuanced differences between these strands are now illuminated via a series of study reviews. A summary of the aims, research questions, and methodological approaches taken in these studies is presented in Table 1.

In their study of a corpus of TOEFL integrated and independent task essays, Kyle and Crossley (2016) use their automated tool TAALES (Tool for the Automatic Analysis of Lexical Sophistication) to determine how its measures of lexical sophistication, across word and multi-word n-grams, are related to proficiency scores in both respective tasks. Using stepwise multiple regression, they enter the strongest lexical sophistication correlates of proficiency scores into a multiple regression model for each task type. They note how measures of range (which tap into the number of reference texts that a word appears in to measure how widely the word is used) and bigrams (two-word combinations that often have salient specific meanings) are important measures that are able to explain 34.5% of the variation in scores for independent task essays. However, these lexical sophistication measures were unable to explain much variation in the integrated task essays.

In an analysis of a Chinese sample (1,198 essays) of the ICNALE (International Corpus Network of Asian Learners of English) corpus, Yoon (2017) used univariate MANOVAs to show that the effect of task type (two argumentative tasks) had the largest effect on writers' language use and writing quality. Univariate MANOVAs also found that topic (two topics were used) had a statistically significant main effect on the production of numerous linguistic features across proficiency levels.

Using the big data system corpus EFCAMDAT (EF-Cambridge Open Language Database), Alexopoulou et al. (2017) studied the influence of task design features on the production of written linguistic features and how this varies across CEFR levels. EFCAMDAT consists of over a million texts, with each student contributing multiple texts. In their analysis, they included three tasks: narrative (e.g., "write a movie plot"), descriptive (e.g., "write a letter of complaint"), and professional (e.g., "write a job advertisement"), and found that task type was a statistically significant moderating variable in examining the relationship between linguistic complexity measures and proficiency levels.

In a rare exploration of community college writing, Perin and Lauterbach (2018) studied the ability of several Coh-metrix measures of lexical complexity and cohesion to explain proficiency scores across a sample of 211 college essays. Across two tasks (a persuasive essay and a summary), they found that a limited number of measures were able to predict

both tasks. Lexical diversity (type-token ratio) and argument overlap were able to explain 20% of the variation in proficiency scores in the essay task, and content word overlap, lexical diversity (VOCD), and word familiarity were among the measures able to explain 22% of the variation in proficiency scores in the summary task.

In their study of L2 first-year writing, Casal and Lee (2019) examine the relationship between syntactic complexity and writing quality in a corpus of 280 research papers. They operationalized writing quality by dividing the papers into high, medium, and low proficiency levels and used a one-way MANOVA test to determine that there was little variation in clausal subordination and coordination across the proficiency levels. However, there were statistically significant differences in complex nominal densities, mean clause length, and mean t-unit length production, with fewer of these features found in lower-rated research papers.

Bringing these studies together, we can see that this research agenda of modelling scoring through examination of the role of linguistic features and grades appears to produce substantive information across various assessment contexts. By extension, we can also question the strength of this body of evidence when differences in feature use and grades are measured over aggregated means that do not model for group *and* individual variation at writer level, or indeed group and individual variation at rater level, with both individual levels known to be sources of score variation and learner language development and acquisition (Alexopoulou et al., 2017; Barkaoui, 2008; Murakami, 2016). As the research note will demonstrate below, the multifactorial method of mixed-effects modelling allows us to determine how the use of language influences assessment scores, while taking account of individual writer and rater variables as well as contextual variables that include task and topic in this modelling process.

Table 1

Feature-Grade Studies

Study	Study aim	Research questions	Selected linguistic features/ key language proficiency components	Methodology
Kyle and Crossley (2016)	To explore the relationship between newly developed indices of lexical sophistication and holistic scores of writing proficiency in both independent and source-based tasks	1. What is the relationship between lexical sophistication and independent writing task proficiency scores? 2. What is the relationship between lexical sophistication and source-based task proficiency scores?	**Construct**: Lexical Sophistication **Key Measures**: Word frequency, n-gram measures, academic word list measures	Stepwise multiple regression to determine to what extent lexical sophistication measures can explain grade score variations
Yoon (2017)	To explore the validity of syntactic, lexical, and morphological complexity measures in capturing topic and proficiency differences in L2 writing	1. How does EFL students' writing differ in the measures of syntactic, lexical, and morphological complexity across different topics? 2. How does EFL students' writing differ in the measures of syntactic, lexical, and morphological complexity across proficiency levels? 3. How do the measures of syntactic, lexical, and morphological complexity tap distinct constructs?	**Constructs**: Syntactic, lexical and morphological complexity **Key Measures**: Measures reported in the Syntactic Complexity Analyser (Lu, 2011) to measure syntactic complexity. Measures reported in Coh-metrix to measure lexical sophistication (word length and word frequency) and lexical diversity (VOCD). Measures reported in the Morpho complexity tool that target the variation in the forms of morphological components and the diversity of inflectional morphemes attached to a base form.	Two-way repeated-measures MANOVA and a series of univariate analyses to investigate topic and proficiency variable interactions

Study	Study aim	Research questions	Selected linguistic features/ key language proficiency components	Methodology
Alexopoulou et al. (2017)	To provide a conceptual and methodological example of how to connect the analysis of task effects in learner corpora with insights from task-based learning. The aim is also to relate this work along proficiency scales.	1. How do task design features and instructional focus affect the written language used by L2 learners when they try to meet the non-linguistic goal of a task?	**Construct**: Linguistic Complexity **Key Measures**: Accuracy (error counts) **Syntactic Complexity**: Average sentence length, mean clause length, the number of subordinate clauses per t-unit. **Lexical Diversity**: Measure of Textual Lexical Diversity (MTLD)	Descriptive comparison of feature counts across proficiency levels and then across task types
Perin & Lauterbach (2018)	To determine whether or not variables from an automated scoring system are predictive of human scores on writing quality rubrics	No specific research questions reported.	**Cohesion and Lexical Complexity**: Various sub-constructs measured, including lexical diversity: type-token ratio (TTR) and word overlap and word familiarity as sub-constructs of lexical cohesion and lexical sophistication.	ANOVAs, stepwise regression

Study	Study aim	Research questions	Selected linguistic features/ key language proficiency components	Methodology
Casal & Lee (2019)	To explore the relationship between syntactic complexity and writing quality in source-based research papers produced by ESL undergraduate writers in a first-year writing course	**1.** To what extent do high-, mid-, and low-rated research papers written by second language undergraduate students in a first-year writing course differ in their global, clausal, and phrasal syntactic complexity? **2.** To what extent do high-, mid-, and low-rated research papers written by second language undergraduate students in a first-year writing course differ in the normalized frequency of five specific types of noun-modifier based phrasal complexity measures?	**Syntactic Complexity:** Various global, clausal, and phrasal measures, including mean length of t-unit, sentence coordination ratio, t-unit complexity ratio, mean clause length, and complex nominal density	One-way MANOVA and post-hoc Tukey test to test for differences between high, medium, and low proficiency levels

In this respect, although the research trajectories that have shaped score modelling have mostly relied on monofactorial inferential analyses (e.g., ANOVAs, discriminant analysis, and multiple regression), many of these studies have made concluding remarks that acknowledge the caveats of using such analyses and promote a move towards mixed-effects models. Kyle and Crossley (2016) note that they did not measure the moderating influence of prompt in their analysis of TOEFL texts, while Yoon (2017) and Alexopoulou et al. (2017) both acknowledge the limitations of monofactorial methods. Yoon (2017) notes a need to explore the interaction between writing topic and L1/cultural backgrounds "systematically" but does not provide specific guidance on how this may be built into analyses. Similarly, Alexopoulou et al. note that the use of multifactorial methods with large-scale "big data" learner corpora is essential since they can "tease apart the impact of distinct factors and yield rich inventories of features modelling development as well as task" (2017, p. 203). However, these studies do not provide specific guidance on how these multifactorial methods may be used and how they explicitly differ from monofactorial methods.

In order to move this suggestion forward, this research note now moves on to present the key theoretical underpinnings of the multifactorial method of mixed-effects modelling and how it addresses the perceived limitations of monofactorial methods. Then, the note explains how these theoretical advantages have been used only sporadically to model score variability, and considers the feasibility of such use in a "big data" system whose structure and content typifies the learning and assessment setup in many U.S. first-year university writing programs.

3.0 Information Collection

3.1 Understanding Mixed-Effects Modelling and What it Offers Score Modelling

Multiple linear regression, along with discriminant factor analysis and logistic regression, is among the most commonly used methods of analyses that have been used to study feature-grade relationships (Barkaoui, 2008, 2010). Multiple linear regression considers the strongest correlations between a dependent variable (e.g., holistic/analytical grades) and a series of independent variables (e.g., linguistic features including features of subordination and coordination, collocations, and different word types) in a regression model that can best predict or explain the amount of grade variation that is accounted for by the linguistic features (Jeon, 2015; Tabachnick & Fidell, 2014). This regression model is represented in equation (1):

$$Y = a + b_1(x_1) + b_2(x_2) + \ldots b_k(x_k) \qquad (1)$$

Where Y= predicted value of the dependent variable, a = the regression constant (intercept), b= partial regression coefficients, and k = the number of independent variables in the equation. We can see that the equation is extended depending on the number of independent variables or predictors.

This equation allows the value of Y (the grade) to be estimated from the values of a and b. Simply, the independent variables ($b_1(x_1)$) (e.g., linguistic features) are able to predict the grade scores of essays for the data set being tested and can also be generalized to other data sets in the population (Jeon, 2015; Tabachnick & Fidell, 2014). Inside this equation, the data points that make up the regression line are assumed to be individual cases where each case is independent from all others.

In modelling this relationship, there is an assumption that the variation that is estimated in the equation is based on independent observations, and therefore, the independent variables (e.g., b_1 (x_1)) consist of independent data points that all share the same variance estimations. However, in essay-based regression modelling, the data points are unlikely to be completely independent from each other. This assumption exists because, as Barkaoui (2010) points out, corpus-based essay research is often derived from a corpus that has a hierarchical structure where observations at lower levels are nested within observations at higher levels. These levels relate to the wider educational context where the writing takes place. In educational contexts, examples of this nesting occur with students in classes, classes in schools, and schools in states or districts.

Applying this hierarchical structure to essay scores, the hierarchical nature of the corpus means ratings that are assigned by the same rater are nested within this rater, with each individual rater effectively having their own cluster or nest of data, and so that cluster shares the same variance profile. The ratings within a nest or cluster are likely to be more related to each other than those ratings belonging to a different cluster (Barkaoui, 2010; Hox, 2002; Raudenbush & Bryk, 2002). In the case of most first-year writing programs in the U.S., operational constraints mean that no single rater grades all the essays, and therefore, the data points will not be independent because the ratings have been assigned by multiple different raters who grade across projects, classes, and student populations. Osbourne clarifies the idea of a hierarchy by stating that "individuals that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population because they share the same environment" (2010, p. 60). When we refer back to the studies in Table 1, we see how these monofactorial designs do not account for this individual variation, and instead, this variation is concealed by only modelling across group aggregated means. This distinction between group and individual variation manifests itself in the mixed-effects modelling literature as "fixed" and "random" effects, with fixed effects remaining constant across the corpus sample, while the random effects are variables that introduce experimental variation into the sample via the corpus' hierarchical structure. This clustering or nesting can be demonstrated in Figure 2. Figure 2 highlights how essay ratings operate at level 1 and raters operate at level 2.

Figure 2. Two-level nesting.

This data independence violation has important implications for reporting the strength of correlations and the regression model's ability to explain/predict variation, with a greater chance that the resulting model is more susceptible to Type 1 measurement errors. These Type 1 errors raise the possibility that significance claims/variation explanations are inflated because they are only based on group aggregations (Heck & Thomas, 2000; Hox, 2002; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). In other words, when we use multiple linear regression on dependent data points, we run the risk of obtaining statistically significant correlations between variables that are false positives due to our failure to take into account random variation from moderating or intervening variables that arise from the corpus hierarchy.

The next section of the note provides a seminal example of how previous score modelling has been carried out by taking account not only of fixed essay factors but also random rater factors that come together to influence score variability.

3.2 An Illustration of Mixed Modelling in Writing Assessment Research

In making the rationale for the use of mixed-effects modelling clear, there are few studies that we can directly draw on. There have been very few studies that have used this modelling in first and second language acquisition (e.g., Meunier & Littre, 2013; Murakami, 2016) and first and second language assessment studies (e.g., Barkaoui, 2010), with the majority of guiding work emanating from general education studies (e.g., Goldstein & Tomas, 1996; Leckie & Baird, 2011; Yang, Goldstein, Browne, & Woodhouse, 2002) that have an interest in assessment/achievement scores as well as how variables such as schools, subjects, and classes influence these scores at group and individual levels of analyses.

In order to clarify the benefits of mixed-effects modelling to language assessment and the modelling of the scoring process, Barkaoui's (2010) study provides an example of work that differs from the studies in Table 1. Barkaoui recognizes that the multidimensional nature of scoring does not only include essay factors that emanate from the writer or the context (e.g., task, topic, exam conditions) but is also influenced by rater variables since it is the rater's judgement of these linguistic features and other writer and context factors that guides and governs the scoring process.

Barkaoui (2010) combines essay and rater factors to explain score variation in a sample of ESL essays that had been analytically scored (see Table 2 for a study overview). As noted in Section 3.1, in corpus studies that use essays, there are a minimum of two levels in the corpus hierarchy: ratings at level 1 (the basic unit of analysis), and these are nested within the higher level 2 raters. This hierarchy is what underpins Barkaoui's (2010) modelling work. Under level 1, he considers multiple essay factors that act as fixed effects, with linguistic features consisting of the five analytical grading categories: communicative quality, organization, argumentation, linguistic accuracy, and linguistic appropriacy, alongside essay length and content. Under level 2, he introduces the random effect of rater experience as a predictor in the model because it is suspected of introducing experimental variation into the modelling process. In the study, 31 novice and 29 experienced raters are sampled; novice raters are those who just finished teacher training, and experienced raters are graduate students with at least five years of teaching experience. When modelled in this way, the introduction of rater characteristics allows the researcher to account for between-rater and within-rater variation in scoring, therefore tapping into systematic group differences via between raters but also tapping into individual variation via within raters. This research design significantly differs from previous ANOVA and monofactorial multiple linear regression methods in which the analysis rests on aggregated means across group differences only.

In this case, regression equation (1) changes from $Y = a + b_1(x_1) + b_2(x_2) + \ldots b_k(x_k)$ to an extended equation (2), which generally translates to reading simply as $y=X\beta+Ub+\epsilon$, with the "Xβ" being the traditional fixed effects marked as "$b_1(x_1)$" in equation (1); however, the extension comes from the "Ub" component, which introduces the random effects modelling part of the equation plus the related errors of measurement, which is represented by the "ϵ" component. Through this analysis, Barkaoui (2010) shows how scoring takes on different trajectories across novice and experienced raters at group and individual levels. Experienced raters awarded lower scores overall and paid more attention to linguistic accuracy, whereas novice raters showed more individual score variation and paid more attention to argumentation.

Like the earlier monofactorial studies, we can see how the assessment of language becomes a much more fine-grained, robust effort that pays attention to the fact that the relationship between linguistic features and grades is multifactorial. In Barkaoui (2010), there is a shift in emphasis methodologically through mixed-effects modelling, but there is also a conceptual shift to recognize that variability in the essay scoring process is not one-sided with the writer wholly responsible for this variation through producing certain linguistic features, but instead is also influenced by rater characteristics.

Table 2

Overview of Barkaoui (2010)'s Approach to Explaining Score Variation with Mixed Effects

Study	**Study aim**	**Research questions**	**Selected linguistic features/ language focus**
Barkaoui (2010)	To examine the contribution of rater and essay factors to variability in ESL essay holistic scores and to illustrate the use and potential contributions of multilevel modelling to research on essay score variability	1. What is the proportion of between-rater relative to within-rater variance in ESL essay holistic scores? 2. Which essay features (and to what extent) account for variance in the holistic scores? 3. Do the relationships between essay features and holistic scores vary significantly across raters? 4. What is the main effect of rater experience on the holistic scores? 5. What are the effects of rater experience on the relationships between essay features and holistic scores?	Essay features are five features from the analytical grading rubric (communicative quality, organization, argumentation, linguistic accuracy, and linguistic appropriacy).

Barkaoui's (2010) approach also has important implications for the modelling of rubric descriptions as well as automated scoring systems. In the former research area, this type of design becomes important because it allows assessment administrators to better align novice and experienced raters' practice and allows rubric creators to realign focus on particular categories of writing proficiency (e.g., linguistic accuracy and argumentation). Barkaoui's (2010) study also highlights that raters are not in tune with regards to how global aspects of writing proficiency (e.g., argumentation and task completion) and local aspects (e.g., linguistic accuracy) are weighted or given importance in the scoring process. In the latter area of research, this type of work also informs automated scoring because it potentially allows rater variability to be recognized in the system and allows the system to react to rater variation in a more realistic manner that mimics actual scoring practices.

3.3 Modelling Scoring in First-Year Writing Contexts: A Recent Attempt to Use Big Data Systems

Thus far in this theoretical research note, we have developed an understanding of the feature-grade relationship and how mixed-effects modelling may be a robust conservative alternative method that allows us to account for individual- and group-level variation in such a relationship. Within this work, several preparatory university courses or university-level contexts have been examined. In the case of U.S.-based university contexts, there are few studies that have fully focused on this writing context in its entirety. For example, although we examined the hierarchical corpus structure within IELTS/TOEFL university entrance/preparation studies, at a more local level, there is a pressing need to understand how writing programs need to embrace this work more empirically because (a) it allows first-year writing stakeholders to develop descriptive profiles of learner language across proficiency levels/program achievement levels, and (b) it allows these profiles to be embedded into score modelling work when their rater, writer, and contextual variables are also incorporated into the model. As a collective whole, first-year writing programs represent a clear domain where mixed-effects modelling is both possible and much needed. Given the size and learning/assessment structure of these programs, there are several potential moderating variables that need to be factored into language feature work. In augmenting (a) and (b), we take the first-year writing program at USF as an example to show the types of modelling work possible and the inter-connected learner and course variables that need to be embedded into such modelling.

The first-year program at USF is a typical representative of the writing programs that operate across U.S. universities in terms of program aims, student population, grading system, and teaching and assessment philosophies. The program at USF operates as a sequence of modules, with the first module providing an introduction to university academic writing and the second module building on the first by varying task difficulty and variety. The first module focuses on setting out scholarly arguments, whereas the second module takes this further by asking students to critically analyze these arguments and attempt to suggest compromise between stakeholder arguments. The program has in excess of 4,500 students per term, with native and non-native speakers of English enrolled in the program. Writing is assessed holistically on a 15-point scale, with students receiving grades A–F,

where A indicates a maximum score of 15. Analytical grades operate on an 8-point scale that evaluates style, organization, formatting, and adherence to academic conventions (Durrant, Moxley, & McCallum, 2019).

Thus far, little language-oriented research has emanated from USF with respect to the kinds of description and score modelling work that this research note has so far outlined. This disjuncture may be the result of a strong reliance on implementing process-writing models of pedagogy. In the case of USF, process writing is implemented throughout the program. Students practice writing various drafts of assessed work and are expected to give and receive multiple rounds of peer and instructor feedback online. Students also meet face-to-face with their peers and instructors to orally discuss revisions and draw up revision plans before finally submitting a final draft that is graded by an instructor. Although class time and feedback may focus on language features, a substantial focus on the process of writing, editing and final revisions takes precedence over explicit language feature instruction. While important, the emphasis on writing processes may not be attending to specific language features as inherent in effective revision.

However, scholars of first-year writing are starting to recognize this limitation. There is growing awareness that more research on student language use in first-year writing is needed. McDonald (2007) also recognizes this lack of focus on language and argues that this gap is a barrier to academic success which leaves students unable to see how language, writing style, and form combine to achieve specific rhetorical goals within their respective communities of practice. These scholars point out that although the Council of Writing Program Administrators (CWPA), National Council of Teachers of English (NCTE), and National Writing Project (NWP) have issued guidelines and frameworks to ensure students develop the necessary skills needed to enhance their chances of academic success at university, their underspecification hampers our efforts to accurately describe student language, how it is perceived by raters, and ultimately how it ties into achieving the rhetorical goals that first-year writing programs promote. The CWPA, NCTE, and NWP (2011) jointly-developed *Framework for Success in Postsecondary Writing* and the CWPA's (2014) Outcomes Statement are two especially pertinent examples of this underspecification. While there is an undertone of how language plays a role in students meeting program outcomes, it is perhaps clearest when we examine the development of *Rhetorical Knowledge* and '*Knowledge of Conventions*. In developing rhetorical knowledge, students are expected to:

- Develop facility in responding to a variety of situations and contexts calling for purposeful shifts in voice, tone, level of formality, design, medium and/or structure (CWPA, 2014)

In developing this knowledge, instructors are expected to guide students towards learning the expectations of readers, the main features of the genres, and the main purposes of composing in their fields.

Reference to language becomes more explicit when the Outcomes Statement sets out its *Knowledge of Conventions* guidance. The Outcomes Statement defines conventions as the formal rules and informal guidelines that define genres, and in doing so, shape readers' and writers' perceptions of correctness or appropriateness. Under this knowledge, the WPA Outcomes Statement (2014) expects students to:

- Develop knowledge of linguistic structures, including grammar, punctuation, and spelling, through practice in composing and revising
- Understand why genre conventions for structure, paragraphing, tone, and mechanics vary
- Gain experience negotiating variations in genre conventions
- Learn common formats and/or design features for different kinds of texts
- Explore the concepts of intellectual property (such as fair use and copyright) that motivate documentation conventions (CWPA, 2014)

In setting out these student goals, instructors are expected to help raise awareness of:

- The reasons behind conventions of usage, specialized vocabulary, format, and citation systems in their fields or disciplines
- Strategies for controlling conventions in their fields or disciplines
- Factors that influence the ways work is designed, documented, and disseminated in their fields
- Ways to make informed decisions about intellectual property issues connected to common genres and modalities in their fields (CWPA, 2014)

Although these guidelines are intended to be loose enough to be applied across individual institutions, there is a lack of clarity in how the statement "teachers can provide opportunities and guidance for students" is operationalized to help students develop rhetorical and convention knowledge. The current guidelines do not, for example, elaborate on what opportunities for practice involve, what form they might take, and what language features are best used to show or develop these aspects of rhetoric or convention knowledge. Equally, when it comes to assessment, there are few examples of how language use facilitates and achieves these areas of knowledge. Having access to this information is useful for assessment purposes because we can begin to understand how these language features and their associated rhetorical functions play a role in score variation and ideas about "good" writing at this level of study.

In recognition of this, efforts have started to come together at USF to create a big data system that allows researchers and first-year writing practitioners to investigate these gaps. Alexopoulou, Geertzen, Korhonen, & Meurers (2015) highlight that big data systems are characterized by two properties: they have significantly larger amounts of data than standard resources in the field and contain data that is generated through a real-life activity rather than being collected for research purposes. In the case of the USF data warehouse, these two properties are clearly present. The USF data warehouse is a text repository system that stores first, second, and final draft assignments that students complete in the program's two modules (ENC 1101 and ENC 1102). Students voluntarily allow their texts to be stored in the warehouse along with a plethora of peer and teacher essay feedback comments and student

metadata which is obtained through a voluntary demographic survey. The survey collects metadata such as students' age range, language background, year of study, and intended major. This structure means that for each student, the warehouse allows research questions to be answered on a group and individual basis because researchers have access to variables on a case by case (i.e., student by student) basis as shown in Figure 3.

class_code	student_user_id	project_id	project_name	draft_id	Draft	Grader_User_Id
23844	7506	801	Project 1	2114	Early	1462
23844	7506	837	Project 1	2184	Early	1462
23844	7514	837	Project 1	2184	Early	1462
11506	14691	801	Project 1	2114	Early	5570
10234	8999	801	Project 1	2114	Early	5570
10234	8117	801	Project 1	2114	Early	5570
10234	3456	801	Project 1	2114	Early	5570

Figure 3. A snapshot of the data warehouse text information.

The data warehouse has been used across multiple non-language-focused studies (e.g., Moxley, 2012, 2013) to holistically study instructor commentary (e.g., Dixon & Moxley, 2013) as well as differences in student and instructor grading approaches (e.g., Moxley & Eubanks, 2015). However, taking this big data resource into consideration, limited inferential studies have started to emerge that symbolize the first attempts at modelling and understanding the scoring process at USF with the relationship between linguistic features and grades. Durrant, Moxley and McCallum (2019) used the final draft texts from the USF system to analyze the relationship between features of lexical sophistication and holistic and style grades awarded by raters in a corpus of almost 7,000 texts. Using Principal Component Analysis on the TAALES tool's measures of lexical sophistication, they identified measures which were not correlated with each other and so were able to tap into different lexical sophistication constructs (from TAALES' almost 500 measures, they retained a total of fewer than 20 measures). Then, they assessed the relationship these measures have with style grades and overall holistic grades, with the former concerning a student's use of appropriate word choice and diction and the latter concerning the overall impression of organization, argument/evaluation, and the use of conventions and formatting. They found that the measures correlated more strongly with holistic scores, suggesting that raters may not perceive the "style" grade as merely aligning with these measures of lexical sophistication. Interestingly, and perhaps most illuminatingly, the study found a distinct difference in the strength of the correlations between these lexical measures and grades across modules, with the first module having stronger correlations between these measures and grades than the second module. This finding may indicate that considerations for style grades may change focus over the course of the whole first-year writing program.

While this study shows how student language relates to awarded grades, the study does not examine the relationship beyond the level of group mean and still considers the data points underlying the linguistic features and the awarded grades as being independent from each other. A close examination of the USF data warehouse construction in Figure 3 reveals that feature-grade relationships may be influenced by the clustered nature of the raters who rate

multiple essays per class and operate across classes. For example, the rater 1462 marks several entries in a single class, while the rater 5570 crossed over between classes.

The penultimate section of this research note considers the practical difficulties that may arise when we have such a big data system that contains these types of nested or clustered and crossed interactions within the hierarchy. The section also considers the implications for matching up this complex structure with students' demographic information which, because it is obtained voluntarily, may be missing or incomplete.

3.4 Taking Advantage of Big Data Systems Through Mixed-Effects Modelling

In setting out the need for mixed-effects modelling to be used across first-year writing studies and for the research community to rely more on big data sets, there is a danger that we present a picture that is overly simple and "easy" to put into practice. When we refer back to Figure 4, we see a complex picture that includes multiple texts written by multiple individuals which are rated by "floating" raters who cut across these texts that belong to certain modules, task assignments, and writers. It is this very operational setup that makes the adoption of mixed-effects modelling both a desired method aligned to a specific curricular design as well as a challenge to carry out. As previous sections illuminated, mixed-effects modelling should be promoted because it can handle these complex data sets that represent the realities of providing large-scale writing programs in the U.S. However, this research note has also pointed out the scarcity of mixed-effects modelling in language assessment and pedagogy research thus far. This scarcity may lie in the fact that the method of mixed-effects modelling requires a degree of methodological literacy on the part of the researcher; but more tellingly, as Gries (2015) notes, there is a danger that the innovative nature of the method and its relatively limited application into language research means a number of methodologically thorny issues do not have "standards" to draw on.

In this respect, two areas that offer pause for thought in the literature are particularly relevant here: sample size and missing data. Sample size for each level in the hierarchy becomes important because it reduces the likelihood that estimations at the hierarchical levels (e.g., estimations about raters and estimations about ratings) can be generalized. Sample size advice remains unclear, with few researchers reaching consensus on this. A survey of the literature reveals advice that some researchers see sample size as most important for higher-level variables (e.g., Level 2;McCoach, 2010), while others suggest a minimum number of clusters that need to be available for mixed-effects analysis to take place more broadly (Maas & Hox, 2005). However, in language-related work, these issues have received less attention, with few doctoral-level theses and published works providing clear-cut field-specific advice (Barkaoui, 2008, 2010; Cunnings, 2012; Linck & Cunnings, 2015).

Another contested issue is that of how the modelling deals with missing data and how much missing data the method is able to handle without causing analysis problems. Some researchers (e.g., Linck & Cunnings, 2015) promote mixed-effects modelling for its ability to handle missing data, while others refer to this ability as only holding true when missing data is "random" (e.g., Collins, Schafer, & Kam, 2001) because missing data that is non-random can influence parameter estimates and inflate Type 1 error rates. The literature appears to remain fuzzy on this issue with the definition of "non-random" often left unspecified.

Both issues are relevant to the USF data warehouse with respect to learner variables that we may wish to incorporate in our estimations of score modelling. Many of these variables have been obtained via students answering the voluntary demographic survey, and although these variables (e.g., language background, gender, age, and intended major) can provide a rich variable set for our analyses, there remains a concern that missing data through non-completion of the survey may influence the possible sampling frame and thus the level of generalization possible in final work.

4.0 Conclusion

This research note began by tracing the continuous need for lines of research inquiry that tap into the relationships between language features and writing quality/proficiency grade scores. In undertaking such inquiries, the research note set out how an understanding of these relationships can yield valuable mapping information that allows us to map out profiles of language use across writing quality grade scores and begin to better understand how these features play a role in rating and score modelling across university writing as a whole and first-year writing in particular . This research was traced along CEFR and EAP contexts with these approaches then linked to U.S. first-year writing contexts. In making this connection, several observations and limitations with the current approaches to mapping were made:

- FYC contexts appear to have similar shortcomings to those CEFR and EAP contexts, whereby the connections between assessment criteria and language features remain under-developed.
- Previous studies had heavily relied on descriptions of language use and/or mono-factorial designs that looked simply at the linear relationship between features and grades without accounting for variations from the types of learner and contextual variables that also feature in rich socio-cultural frameworks of writing assessment (e.g., task type, rater characteristics).

To take account of these variations, a multifactorial method, namely mixed-effects modelling, was introduced. This kind of modelling can "even out" the stated relationship claims by accounting for shared variance in the model that can be attributed to the types of learner and context variables that Weir et al. (2013) attempt to give recognition to. This modelling takes on further importance because, unlike methods used in previous studies, it does not assume that the data points that make up the variables are independent of each other, but realizes they are dependent (because of the hierarchical nature of the corpus) and takes this into account in its measurements.

5.0 Directions for Future Research

While this research note illustrated the hierarchical levels at play in USF's big data system and their potential influence on our interest of linguistic feature-grade relationships, future work in this area should build on these illustrations by *empirically* investigating how these relationships vary when these hierarchies are considered. In taking such an approach, researchers should be able to provide more valid representations of these relationships and in turn make more concrete recommendations for the teaching and assessment of language in

FYC programs. Researchers are encouraged to develop lines of research inquiry that examine rater perceptions of language use and tap into score modelling by carrying out mixed-effects modelling in the USF data warehouse.

In beginning to develop a research agenda for this modelling work, several possible ideas can be taken forward. This research note pertinently identified several key linguistic areas that could be included in first-year contexts (e.g., syntax, lexis, and phraseology); however, as indicated in the WPA Outcomes Statement (2014), there is a need to link the quantitative features of interest back to functional usage that taps into the broader rhetorical goals of first-year programs. In this respect, an avenue of interest may lie in linguistic features that are considered features of metadiscourse that tap into evaluation language (e.g., hedging language) as well as language that plays a role in text organization (e.g., lexical bundles). Analyses of these linguistic features may provide specific results that help determine appropriate features that influence assessment grading and become teaching points in first-year programs.

A second avenue for research lies in building in understandings of rater characteristics. In first-year contexts, we can take into consideration variables such as rater experience because grades are awarded by experienced professors along with adjunct and graduate teaching assistants (GTAs). However, a specific consideration may also be given to rater workload. Figure 3 helps highlight how grading load may be unequal across individual raters, and this should be a valuable factor that is built into future score modelling.

Author Biography

Lee McCallum is an EdD candidate at the University of Exeter and a Subject Teacher and In-Sessional Tutor at the University of Stirling in the U.K. She has extensive teaching experience in EAP from the Middle East, Europe, and China. Her research interests include language assessment and writing instruction with a focus on how corpus-based methods can enhance these areas. Her most recent work, forthcoming in 2020, is a co-authored book titled *Understanding Development and Proficiency in Writing: Quantitative Corpus Linguistics Approaches*, which will be published by Cambridge University Press.

References

Alexopoulou, T., Geertzen, I., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research. *International Journal of Learner Corpus Linguistics*, *1*(1), 96–129.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180–208.

Aull, L. L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. *Journal of Writing Analytics*, *1*, 1–47. Retrieved from https://wac.colostate.edu/docs/jwa/vol1/aull.pdf

Aull, L. L. (2019). Linguistic markers of stance and genre in upper-level student writing. *Written Communication*, *36*(2), 267–295.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Unpublished Ph.D. thesis). University of Toronto, Toronto, Canada.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing*, *27*(4), 515–535.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62.

Chen, J., Fife, J. H., Bejar, I. I., & Rupp, A. A. (2016). *Building e-rater scoring models using machine learning methods*. (ETS Research Report RR-16-04). Princeton, NJ: Educational Testing Service.

Collins, L. M., Schafer, J. L., & Kam, C. M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.

Council of Writing Program Administrators (CWPA). (2014). Outcomes statement for first-year composition (3.0). Retrieved from http://wpacouncil.org/positions/outcomes.html

Council of Writing Program Administrators (CWPA), National Council of Teachers of English (NCTE), & National Writing Project (NWP). (2011). *Framework for success in postsecondary writing*. Retrieved from http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382.

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*(4), 241–256.

Durrant, P., Moxley, J., & McCallum, L. (2019). Vocabulary sophistication in freshman composition assignments. *International Journal of Corpus Linguistics*, *24*(1), 31–64.

Evans, N. W., Hartshorn, K. J., Cox, T. L., de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing*, *24*, 33–50.

Garner, J., Crossley, S., & Kyle, K. (2018a). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*. doi: https://doi.org/10.1515/iral-2017-0089

Garner, J., Crossley, S., & Kyle, K. (2018b). N-grams and L2 writing proficiency. *System*, *80*, 1–37. doi: 10.1016/j.system.2018.12.001

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society*, *159*(1), 149–163.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics*, *52*(3), 229–252.

Gries, S. Th. (2015). *Statistics for linguistics with R: A practical introduction* (2nd edition). Berlin & New York: De Gruyter Mouton.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.

Hawkins, J. A., & Filipovic, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Heck, R. H., & Thomas, S. L. (2000). *An introduction to multilevel modelling techniques*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jeon, E. H. (2015). Multiple linear regression. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp.130–158). New York & London: Routledge.

Kreft, I., & de Leeuw, J. (1998). *Introduction to multilevel modelling*. London: Sage.

Krzeminska-Adamek, M. (2016). Lexis in writing: Investigating the relationship between lexical richness and the quality of advanced learners' texts. In M. Pawlak (Ed.), *Classroom-oriented research: Reconciling theory and practice* (pp. 195–197). Switzerland: Springer.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24.

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency and rater experience. *Journal of Educational Measurement, 48*(4), 399–418.

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies, 7*(3), 163–180.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed effects models in second language research. *Language Learning, 65*(1), 185–207.

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.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology, 1*(3), 86–92.

McCoach, B. (2010). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in social sciences* (pp.123–141). London: Routledge.

McDonald, S. P. (2007). The erasure of language. *College Composition and Communication, 58*, 585–625.

Meara, P., & Bell, H. (2001). P-Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5–19.

Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual "corpus cum experimental data" approach. *Modern Language Journal, 97*, 61–76.

Moxley, J. M. (2012). Aggregated assessment and objectivity 2.0. *Proceedings of the EACL 2012 Workshop on Computational Linguistics and Writing,* Avignon, France, 19–26.

Moxley, J. (2013). Big data, learning analytics and social assessment. *The Journal of Writing Assessment, 6*(1), 1–12.

Moxley, J. M., & Eubanks, D. (2015). On keeping score: Instructors vs. students; rubric ratings of 46,689 essays. *Writing Program Administration, 39*(2), 53–80.

Murakami, A. (2016). Modeling systematicity and individuality in non-linear second language development: The case of English grammatical morphemes. *Language Learning, 66*(4), 834–871.

Osbourne, J. W. (2010). Correlation and other measures of association. In G. R. Hancock & R. O. Mueller, (Eds.), *The reviewer's guide to quantitative methods in social sciences* (pp. 55–71). London: Routledge.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purpose*s, *14*, 48–59.

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education,* doi: 10.1007/s40593-016-0122-z

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modelling*. London: Sage.

Spolsky, B. (1994). Policy issues in testing and evaluation. *The Annals of the American Academy of Political and Social Sciences, 532*, 226–237.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32*, 17–35.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th edition). Harlow, UK: Pearson.

Taguchi, N., Crawford, W., & Wetzel, D. Z. (2013). What linguistic features are indicative of writing quality? A case of argumentative essays in a college composition program. *TESOL Quarterly, 47*(2), 420–430.

Verspoor, M., Lowie, M., Chan, H. P., & Vahtrick, L. (2017). Linguistic complexity in second language development: Variability and variation at advanced stages. *Recherches en didactique des langues et des cultures, 14*(1), 1–28.

Weir, C. J., Vidakovic, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English language examinations 1913-2012*. Cambridge: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii Press.

Yang, M., Goldstein, H., Browne, W., & Woodhouse, G. (2002). Multivariate multilevel analyses of examination results. *Journal of the Royal Society, 165*(1), 137–153.

Yoon, H-J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System, 66*, 130–141.