# Title: Teacher Assessment of Grammatical Ability in Second Language Academic Writing: A Case Study

**Author**: Heike Neumann

Faculty of Education, McGill University, 3700 McTavish Street, Montreal, Quebec, Canada. Current/Permanent affiliation: Department of Education, Concordia University, 1455 De Maisonneuve Blvd. West, LB 579, Montreal, Quebec, H3G 1M8, Canada. Tel: 1-514-848-2424 ext. 2443. Email: hneumann@education.concordia.ca

**Abstract**

 In the language assessment literature, grammatical ability is widely accepted as a key component of second language (L2) ability in general and L2 writing ability in particular. Indicators of grammatical ability have been investigated in L2 writing research, but the indicators L2 writing teachers attend to when determining grammatical ability levels of their students have not been studied. Furthermore, there is no research on what students know about their teachers’ assessment criteria and how that knowledge might affect their writing and learning process. This mixed methods triangulation study examines these questions in university L2 academic writing classes through a quantitative text-based analysis of academic essay exams, student questionnaires, and teacher and student interviews. The combined results of all data sources indicate that the teachers in this study focus primarily on accuracy when assessing grammatical ability. This leads to risk avoidance behaviour by students and may have a negative impact on their learning as students adapt their writing to meet above all their teachers’ expectations for grammatical accuracy.

**Keywords:** Second Language Writing, Academic Writing, Grammatical Ability, Assessment Criteria, Rating Scales

# Acknowledgements

This study was undertaken in partial fulfillment of the requirements of the author’s PhD in Second Language Education at McGill University. Earlier versions of this paper were presented at the Symposium on Second Language Writing in Phoenix, AZ in November 2009 and at the Language Testing Research Colloquium in Cambridge, United Kingdom in April 2010. The author would like to thank Alison Crump, Bonnie Barnett, Kerri Staples, and May Tan for their help in data transcription and analysis, and Carolyn E. Turner, Beverly A. Baker and the anonymous reviewers for their helpful comments on earlier versions of this paper. This study was supported by a doctoral bursary from the FQRSC (Fonds québécois de la recherche sur la société et la culture).

# Introduction

Cumming ([2004](#_ENREF_23)) has called for a “broadening” of the investigative scope of writing assessment because “we need to know more about the practices and principles of assessment in *ordinary contexts of teaching and learning* [italics added]” (p. 6), a call supported by McNamara ([2001](#_ENREF_58)) and Rea-Dickins ([2009](#_ENREF_71)), yet we still know very little about the classroom assessment of grammar in academic second language (L2) writing classrooms. This is despite the fact that grammatical knowledge is indispensable in accomplishing a task in an L2 ([Bachman & Palmer, 1996](#_ENREF_3); [Weigle, 2002](#_ENREF_87)) and an essential component of writing ability models ([Bachman, 1990](#_ENREF_2); [Bachman & Palmer, 1996](#_ENREF_3); [Grabe & Kaplan, 1996](#_ENREF_34); [Hayes, 1996](#_ENREF_39); [Hayes & Flower, 1980](#_ENREF_40); [Scardamalia & Bereiter, 1987](#_ENREF_72); [Weigle, 2002](#_ENREF_87)). Much of the L2 writing assessment research is focused on high-stakes testing situations because of the tests’ consequences for test takers and the tests’ importance for all stakeholders involved. On the other hand, classroom L2 writing research that is related to grammar often centers on pedagogical rather than assessment questions, such as how L2 writing teachers can provide effective feedback so that learning can take place (see [Ferris, 2002, 2003, 2007 for an excellent overview of this line of research](#_ENREF_30)) or how collaborative writing contributes to language and writing development ([e.g., Storch, 2002](#_ENREF_77); [Storch, 2005](#_ENREF_78); [Storch & Wigglesworth, 2007](#_ENREF_80); [Wigglesworth & Storch, 2009](#_ENREF_91), [2012](#_ENREF_92)). Little research, however, has been conducted on how L2 writing teachers assess grammar in writing classrooms. The current study examines this issue in the context of university L2 academic writing classes.

## Assessing Grammar in L2 Academic Writing

Students’ writing ability is usually assessed by means of a performance test. Although portfolio assessment has seen a surge in popularity since the 1990s as an alternative to impromptu essay exams and Byrnes and her colleagues have expanded the definition of what writing assessment tasks might look like ([Byrnes, 2002](#_ENREF_12); [Byrnes, Maxim, & Norris, 2010](#_ENREF_13)), a timed essay written in class, where students are presented with a topic and have a limited time frame to write their texts, is still the most commonly used assessment tool ([Barkaoui, 2010a](#_ENREF_6); [Crusan, 2010](#_ENREF_22); [Cumming, 2013](#_ENREF_24); [Hamp-Lyons, 2011](#_ENREF_38); [He & Shi, 2012](#_ENREF_41); [Weigle, 2002](#_ENREF_87)). Even writing assessment portfolios usually include at least one piece of timed writing ([Crusan, 2010](#_ENREF_22); [White, 1984](#_ENREF_89)). The continued use of the essay exam is due at least in part to two advantages of the timed essay: (a) One knows for certain who the writer is and that students submit their own writing ([Weigle, 2012](#_ENREF_88); [White, 1995](#_ENREF_90)), and (b) one obtains more accurate information about students’ language ability ([Weigle, 2012](#_ENREF_88)).

Student performance on an essay task has to be evaluated by qualified judges or teachers themselves ([Norris, Brown, Hudson, & Yoshioka, 1998](#_ENREF_62)), usually through the use of some kind of rating scale ([Crusan, 2010](#_ENREF_22); [Hamp-Lyons, 1991](#_ENREF_37); [Turner & Upshur, 2002](#_ENREF_84); [Weigle, 2002](#_ENREF_87), [2012](#_ENREF_88)). While it may be true that rating scales or rubrics are not used in all contexts, it is also certain that they *should* always be used to ensure good writing assessment. Rating scales appropriate for the assessment task improve reliability through clear articulation of assessment criteria for the benefit of both teachers and students ([Crusan, 2010](#_ENREF_22); [Weigle, 2002](#_ENREF_87), [2012](#_ENREF_88); [White, 1984](#_ENREF_89)).

 Rating scales generally belong to one of two categories: holistic scales, where one overall score is assigned for a piece of writing, or analytic scales, where subscores on different criteria make up the overall score. When L1 or L2 writing is rated by means of a holistic rating scale, grammatical accuracy is often correlated to the overall performance assessment ([Barkaoui, 2010b](#_ENREF_7); [Huang & Foote, 2010](#_ENREF_45); [Sweedler-Brown, 1993](#_ENREF_81)). When an analytical rating scale is used, grammar is often a separate criterion, but there is little research on what construct teachers base their assessment of grammar for classroom L2 writing assessments. In order to examine the teachers’ construct, it is necessary to look to the testing literature for construct definitions as a point of reference.

 Attempts to define this construct in the assessment literature within the context of L2 writing assessment emerge from models of writing ability based on cognitive science research for L1 writing ([Hayes, 1996](#_ENREF_39); [Hayes & Flower, 1980](#_ENREF_40); [Scardamalia & Bereiter, 1987](#_ENREF_72)) or the writing and testing literature for L2 writing ([Bachman, 1990](#_ENREF_2); [Bachman & Palmer, 1996](#_ENREF_3); [Grabe & Kaplan, 1996](#_ENREF_34); [Weigle, 2002](#_ENREF_87)). In both L1 and L2 writing models, knowledge pertaining to content, form, and process as well as theoretical language knowledge are essential to produce texts. An important component of this language knowledge, in particular in the language testing literature, is grammatical knowledge because it is seen as indispensable in accomplishing a task in a L2 ([Bachman, 1990](#_ENREF_2); [Bachman & Palmer, 1996](#_ENREF_3); [Weigle, 2002](#_ENREF_87)).

There are various models of grammatical or language ability (e.g., [Bachman, 1990](#_ENREF_2); [Bachman & Palmer, 1996](#_ENREF_3); [Canale & Swain, 1980](#_ENREF_14); [Purpura, 2004](#_ENREF_68); [Weigle, 2002](#_ENREF_87)), and while the format of these models differs, grammatical ability in all these models comprises the learner’s ability to use and apply theoretical grammatical knowledge accurately and meaningfully in language use situations. [[1]](#footnote-1) The grammar knowledge to be applied relates to graphology, lexicon, and a range of morphosyntactic forms and structures. The question is, however, whether teachers’ construct of grammatical ability resembles the theoretical construct definitions such models put forward.

## Indicators of Grammatical Ability in L2 Writing Research

The goal of the current study is to examine the construct of grammatical ability that is assessed in L2 academic writing courses. In order to examine this construct, an analysis of L2 texts is crucial because the texts, along with the rater who assigns the grade and the rating scale used by the rater, influence the score or grade for this performance ([McNamara, 1996](#_ENREF_57); [Upshur & Turner, 1999](#_ENREF_85)).

Both grammatical accuracy and grammatical complexity are key constructs in research studies dealing with the assessment of grammatical ability in L2 writing and analyzing L2 texts for indicators of grammatical ability. Many studies rely on operationalizations of grammatical accuracy in conjunction with grammatical complexity to obtain more comprehensive indicators of test takers’ language proficiency or level of linguistic performance ([e.g., Cumming et al., 2005](#_ENREF_25); [Ellis & Yuan, 2004](#_ENREF_27); [Housen & Kuiken, 2009b](#_ENREF_44); [Iwashita, Brown, McNamara, & O'Hagan, 2008](#_ENREF_50); [Kuiken & Vedder, 2008](#_ENREF_51); [Li, 2000](#_ENREF_53); [Storch, 2009](#_ENREF_79); [Wigglesworth & Storch, 2009](#_ENREF_91); [Wolfe-Quintero, Inagaki, & Kim, 1998](#_ENREF_93)). The expectation is that higher levels of accuracy and complexity are associated with increased proficiency or higher performance levels on a rating scale ([e.g., Educational Testing Service, 2004](#_ENREF_26); [International English Language Testing System, n.d.-a](#_ENREF_48), [n.d.-b](#_ENREF_49); [Polio, 1997](#_ENREF_66); [Wolfe-Quintero et al., 1998](#_ENREF_93)).

Grammatical complexity is an important complementary construct to grammatical accuracy in order to operationalize grammatical ability comprehensively as both control (accuracy) and range (complexity) of the linguistic forms referred to in the language ability models mentioned above need to be assessed. In addition, the analysis of grammatical and/or syntactic complexity is important “because of the assumption that language development entails, among other processes, the growth of an L2 learner’s syntactic repertoire and her or his ability to use that repertoire appropriately in a variety of situations” ([Ortega, 2003, p. 492](#_ENREF_64)). This means (a) that learners have both basic and sophisticated structures at their disposal as their grammatical ability and proficiency increase and (b) that (it is assumed) L2 writers can then choose the structure that best fits the context and the purpose of the communicative situation ([Wolfe-Quintero et al., 1998](#_ENREF_93)). However, both Beers and Nagy ([2009](#_ENREF_8)) and Ortega ([2003](#_ENREF_64)) caution not to assume that linguistically more complex language leads to higher quality texts and argue complexity should be considered a necessary but not sufficient condition for good writing.

Another reason to assess both accuracy and complexity is the trade-off between accuracy and complexity for oral, task-based contexts in language performance ([Skehan, 1996](#_ENREF_74), [1998](#_ENREF_75), [2009](#_ENREF_76)). If the trade-off hypothesis also holds true for written tasks—and Li ([2000](#_ENREF_53)) presents partial evidence for this—accuracy measures on their own do not tell the whole story about grammatical ability in L2 texts. Accuracy has to be considered jointly with complexity. Furthermore, even when complexity is not explicitly referred to in rating scale descriptors, there is evidence that raters of L2 writing attend to grammatical complexity in conjunction with grammatical accuracy when assessing language use in L2 texts ([Lumley, 2005](#_ENREF_56)). A combination of accuracy and complexity measures is, therefore, the most suitable text-based method to examine the teachers’ construct of grammatical ability in L2 writing.

While grammatical accuracy is relatively easy to define as the absence of deviations from a particular linguistic norm ([Housen & Kuiken, 2009a](#_ENREF_43); [Wolfe-Quintero et al., 1998](#_ENREF_93)), the concept of linguistic or grammatical complexity is a little more controversial because it can be seen in terms of size, range, and variety of linguistic resources at a learner’s disposal ([Housen & Kuiken, 2009a](#_ENREF_43)). In terms of operationalizing grammatical complexity, Norris and Ortega ([2009](#_ENREF_63)) argue for a multidimensional definition of complexity that includes at the syntactic level measures of subordination and length, especially at the subclausal level, which points towards phrase-level elaboration. In other words, at least two measures of complexity are required in conjunction with accuracy measures for a comprehensive operationalization of grammatical ability.

## Voices of the Test Takers in L2 Writing Assessment Research

Leki ([2001](#_ENREF_52)) argued that there is a need for L2 writing research to uncover students’ experiences in L2 English writing classes, “to hear their voices talking about the problems and successes they encountered in their writing classes and their interpretation of why things went as they did” (p. 17). Similarly, there is a need to listen to students’ voices in regards to assessment in these L2 writing classes. Rea-Dickins ([1997](#_ENREF_70)) argued that the inclusion of learners’/ test-takers’ views of an assessment are particularly important to include because of the invaluable information they provide about how this group perceives a particular test task even though “their views are among the most difficult to make sense of and to use” (p. 306). Furthermore, the inclusion of the students’ voices can be expected to complement the teachers’ perspective in meaningful ways because these two different groups “will necessarily have different views on the need for or goals for assessment, the constructs to be measured in those assessments, and the benefit or harm to those being assessed” ([Baker, 2010, p. 10](#_ENREF_4)). Consequently, this study also focuses on understanding what students know about the assessment process of grammatical ability in L2 writing and how the teachers’ assessment criteria might affect them.

**Focus of the Study**

The present study contributes to research in L2 writing assessment in the following ways. First, there has been little research on the assessment of grammar in the context of the classroom as opposed to high-stakes assessment on proficiency tests. Although we can draw on theoretical definitions of the construct of grammar for the purposes of language assessment from the literature, there is a need to examine the construct that teachers assess when they assign a grade or score for grammar on their students’ writing. Second, students’ point of view also needs to be considered as students provide an important perspective on how they are assessed by their teachers in L2 writing classrooms and on what the effect of the teacher’s assessment criteria is on the students’ writing and learning.

# Method

## Research Questions and Design

The purpose of this study is to examine the following research questions:

1. What indicators of grammatical ability do teachers attend to in their students’ L2 academic writing when assigning grades for grammar on an analytical rating scale?
2. What knowledge do students have about the indicators their teachers look for when assigning a grade for grammar in writing?
3. What do students perceive to be the impact of their teachers’ assessment criteria on their way of writing and learning in the L2 writing classroom?

In a mixed methods triangulation research design ([Creswell & Plano Clark, 2007](#_ENREF_21)), qualitative (qual) and quantitative (quan) data sources were combined to answer these questions. Students’ essay exams were analyzed using quantitative text-based analysis measures to answer the first research question. In addition, student questionnaire and student- and teacher-interview data were analyzed qualitatively to answer the second and third research questions and to understand the context within which the exams took place. A case study approach was deemed most suitable for this study because it enabled the in-depth investigation of the assessment of grammar in L2 writing classrooms as a social phenomenon influenced by both teachers and students ([Yin, 2009](#_ENREF_94)). Figure 1 provides an overview of the research design using Morse’s ([2010](#_ENREF_45)) notation for mixed methods research to indicate the relationship between data sources and analyses.

[Insert Figure 1 about here]

## Research Context & Participants

The research questions were investigated in a case study of two academic English as a second language (ESL) writing classes focusing on academic writing and vocabulary development at the high-intermediate (Course 1) and advanced level[[2]](#footnote-2) (Course 2) at an English-medium university in Canada. This institution admits students with a TOEFL iBT score of at least 75. If students score between 75 and 89, they may be required to take ESL writing courses, depending on their performance on the institutional placement test. At the same time, these students are already permitted to commence their discipline-specific academic programs.

These L2 writing classes are general academic writing classes focusing on what Hyland ([2006](#_ENREF_47)) calls general English for Academic Purposes (EAP). Since the target students for these classes have only partially met the university’s proficiency requirement and therefore need further ESL training in an academic setting, students in these classes do not learn discipline-specific EAP or genres but develop their academic writing skills more generally, focusing on the “common core” ([Hyland, 2006, p. 11](#_ENREF_47)) of academic writing skills (i.e., the common underlying language forms and skills) in a course with students from different faculties and programs. The approach to teaching writing is what has been called the post-process approach (e.g., [Ferris & Hedgcock, 2005](#_ENREF_31); [Polio & Williams, 2009](#_ENREF_67)). That is to say, there is a focus on the cognitive processes involved in the production, revision, and editing of a piece of writing; at the same time, the product of the writing process is important. Essays are the primary focus in these classes, and students produce two drafts of all assignments written in class or at home. The focus of analysis for this study is the midterm essay exam, which is a task similar to class essay assignments and the final essay exam and serves both formative and summative purposes in these classes: formative by providing feedback to the students so that they can continue to improve for the remainder of the course and summative by measuring student achievement at the midpoint of the course and contributing 20% to the students’ course grade.

 In the ESL program where the study took place, teachers evaluate their own students’ essay exams. However, in order to be accountable to students and ensure similar standards of assessment across different sections or classes of the same course, the essays with the highest and lowest scores are also read by another teacher teaching the same course. The number of essays that are re-read for each class varies from course to course but typically ranges from 10 to 30%. For essay exams, all teachers have to use the ESL program evaluation grid[[3]](#footnote-3) (see Appendix A).

Thirty-three students in the two courses and their teachers participated in this study. All student participants were in the first two years of their degree, and half were in their first or second semester. All except one student were enrolled in an undergraduate program. Table 1 presents detailed background information on the student participants. As is evident, the student population in the two classes is comparable in terms of gender distribution, mean age, L1 backgrounds, and major programs. The two teachers were native speakers of English with degrees in teaching ESL and 10 and 33 years of experience teaching and assessing ESL.

[Insert Table 1 about here]

## Instruments

The student questionnaire (see Appendix B) was developed specifically for this study. The first part of the questionnaire was designed to obtain demographic data to describe the student population in this study. Most questions in the second part of the questionnaire were open-ended and served to answer the research questions as well as identify themes for the student interviews.

For the interviews, specific interview protocols were developed. For the student interviews, the questions were based on part 2 of the student questionnaire and were designed as follow-up and in-depth questions (see Appendix C). The teacher interview questions were formulated based on research and theory questions (see Appendix D). A detailed discussion and analysis of particular essays with the teacher participants was included as an interview intervention to clarify the teachers’ assessment criteria for grammatical ability. For this purpose, four essays that represented different levels of performance on the grammar grade were selected for each teacher to stimulate the discussion.

## Procedures

The study took place during one 13-week semester. At the end of week 7, the students wrote their midterm essay exam. In both classes, students could choose one of three topics that elicited the same rhetorical pattern (see Appendix E for the list of topics) and had 2.5 hours to write the essay. Before teachers marked the exams, the exams were photocopied for the text-based analysis (Phase 1 of the study, see Figure 1). After the teachers had assigned grades, the completed evaluation grid (see Appendix A) for each exam was photocopied. The teachers then returned the exams and evaluation grids to the students. In the following class, the student questionnaire was administered (Phase 2 of the study, see Figure 1). This ensured that students had had time to review their midterm exam and reflect on their grammar mark. Towards the end of the semester, the eight students who had agreed to be interviewed and have their interviews recorded and the two teachers were interviewed (Phase 3 of the study, see Figure 1). All interviews were recorded and then transcribed. Both teachers and students were compensated financially for their time to conduct the interviews.

## Data Analysis

### Phase 1: Text-based analysis of exams (data to answer research question 1).

For the text-based analysis, first all production units (clauses, sentences, and T-units) were hand-tagged on all exams and then tallied by the researcher. Sentences were defined by the writer’s punctuation as phrases that are intercalated between two termination marks (i.e., periods, questions marks, or exclamation marks) and start with a capital letter ([Halliday & Matthiessen, 2004](#_ENREF_36); [Homburg, 1984](#_ENREF_42); [Hunt, 1965](#_ENREF_46); [Tapia, 1993](#_ENREF_82)). T-units were defined as one independent clause with all its dependent clauses ([Hunt, 1965](#_ENREF_46)). Following and accepting the writer’s punctuation when coding sentences and T-units meant that this sometimes led to errors in syntax (incomplete sentences or independent clauses not separated/joined appropriately). Quirk, Greenbaum, Leech, and Svartvik’s ([1985](#_ENREF_69)) definition of sentence limits as “wherever grammatical relations (such as those of subordination and coordination) cannot be established between clauses” (p. 48) was used to determine when the writer’s erroneous punctuation led to syntactic errors. For clauses, Quirk et al.’s ([1985](#_ENREF_69)) definition was adopted with two revisions: (a) all clauses consist of at least one finite or non-finite verb (i.e., verbless clauses were excluded) and (b) all verbs had to be accompanied by at least one other constituent. If verbs did not have a dependent constituent, they were considered part of another clause.

To assess the L2 writers’ ability to write accurately, morphological and syntactic errors were hand-tagged and tallied in order to calculate the errors per 100 words ratio. This measure is less commonly used than the errors per T-unit ratio ([Wolfe-Quintero et al., 1998](#_ENREF_93)), but it provides a more precise measure of the level of accuracy in L2 texts. Following Bardovi-Harlig and Bofman’s ([1989](#_ENREF_5)) definition, morphological errors were defined as “errors in nominal morphology (plural, number agreement, and possessive), errors in verbal morphology (tense, subject-verb agreement, and passive formation), errors in determiners and articles, and errors in prepositions” (p. 21). Syntactic errors “consisted of word order errors, errors resulting from the absence [or addition] of major and minor constituents, and errors in combining sentences” ([Bardovi-Harlig & Bofman, p. 21](#_ENREF_5)). Spelling and punctuation were ignored unless they led to morphologic errors (i.e., wrong verb or plural form) or syntactic errors (i.e., incomplete sentences).

Following Norris and Ortega’s ([2009](#_ENREF_63)) suggestion for a multidimensional definition of complexity, two complexity measures that gauge the level of subordination and phrasal elaboration were select: the clauses per T-unit ratio and the words per clause ratio. The clauses per T-unit ratio was chosen because it is a measure of subordination rather than coordination and therefore is more appropriate for higher-level L2 writers ([Norris & Ortega, 2009](#_ENREF_63)). Furthermore, it is the most commonly used subordination measure and has been found in previous research to increase as the writers’ proficiency level increases ([Ortega, 2003](#_ENREF_64)), albeit not always with statistical significance ([Wolfe-Quintero et al., 1998](#_ENREF_93)). The words per clause ratio measure was chosen because it assesses to what extent L2 writers use subclausal structures such as adjectives, adverbs, and prepositional phrases to increase clause length ([Norris & Ortega, 2009](#_ENREF_63)). The combination of these two measures is useful because two different patterns of complexification are hypothesized ([Halliday & Martin, 1993](#_ENREF_35)) and observed with higher-level learners: L2 writers first expand their writing through subordination (a pattern observed in many L2 writing studies with intermediate-level writers; see, for example, [Ortega, 2003](#_ENREF_64); [Wolfe-Quintero et al., 1998](#_ENREF_93)) before relying more on phrasal elaboration ([a pattern observed by Biber, 2006 with advanced L1 academic writers](#_ENREF_9)). Considering the level of the students in the current study—high-intermediate to advanced L2 and not advanced L1 writers—one would expect to see more expansion through subordination rather than at the phrase level.

Production units and morphological and syntactic errors were also hand-tagged blindly by a second coder with experience in teaching and rating L2 writing. After a training session, the second coder worked independently to code and tally clauses, sentences, T-units, and errors for 50% of the exams in each course. Interrater reliability was analyzed by calculating the absolute agreement for intraclass correlations with a two-way mixed model using SPSS. The interclass coefficients for single measures ranged from 0.96 for clauses to 0.88 for errors. Following Polio ([1997](#_ENREF_66)), discrepancies were resolved by averaging the tallies from both coders.

To examine the relationships between the accuracy and complexity measures and the teacher assigned grammar grade, Pearson product-moment correlation coefficients between the teacher-assigned grammar grade and the accuracy and complexity measures were calculated for each course. The grammar grade used for this analysis is the subscore assigned by the teacher on the grammar criterion on the program evaluation grid (see Appendix A). It therefore contributes to the overall score on the essay but is otherwise disarticulated from that overall grade.

### Phase 2 and 3: Questionnaire and interviews with students and teachers (QUAL) (data to answer research questions 1, 2, and 3).

I used open and axial coding techniques associated with developing grounded theory to identify themes and categories in the questionnaire and interview data ([Corbin & Strauss, 2008](#_ENREF_17)). Open coding requires the analysis of the data without drawing on preconceived codes or categories. Rather, the codes and concepts emerge from careful reading of, reflecting on, and thinking about the data and continue to develop as the process of analysis continues. The use and appearance of each coding category in the data is continuously cross-referenced with other instances to clarify and define each code and concept more clearly using the raw data ([Corbin & Strauss, 2008](#_ENREF_17)) and therefore keeps the researcher close to the raw data and not his/her preconceived notions ([Charmaz, 2003](#_ENREF_15)). In addition, I kept memos to record the reasoning and analysis behind the emerging concepts and categories ([Charmaz, 2003](#_ENREF_15), [2006](#_ENREF_16); [Corbin & Strauss, 2008](#_ENREF_17)). Furthermore, these memos helped me to clarify the concepts and categories I noticed in the data as the analysis of the transcripts progressed. At the same time, I used axial coding; this means that I related the emerging concepts to each other to understand the relationships among them ([Corbin & Strauss, 2008](#_ENREF_17)). Using this process, information on the teachers’ assessment criteria and process as well as problems, concerns, or issues expressed by both teachers and students emerged. In subsequent analyses and readings of these interviews, the concepts and categories were refined through more detailed cross-referencing between different participants’ accounts and the clarification of concepts and themes.

Two methods were used to validate these analyses: the triangulation technique and peer debriefing ([Creswell, 2009](#_ENREF_19); [Creswell & Miller, 2000](#_ENREF_20); [Lincoln & Guba, 1985](#_ENREF_55); [Teddlie & Tashakkori, 2009](#_ENREF_83)). For the student interviews, information from different participants was used to triangulate findings. The analysis of the student and teacher interviews was also conducted by a second researcher with a PhD in Language Assessment. She was familiar with the research project and reviewed the data, the emerging codes and themes, and the presentation of findings to assess the validity of the analysis.

# Results

## Phase 1: Text-based Analysis (QUAN)

Table 2 presents the descriptive statistics for the teacher-assigned grammar grade and the results for the accuracy and complexity measures for each course. Whereas the results in the two courses are similar for the two complexity measures, the figures for the accuracy measure and the grammar grade differ between the two courses. The mean and standard deviation for the grammar grade is higher in Course 2, but for the errors per 100 words ratio these values are lower for this course. In both courses, the variables have similar ranges with a certain degree of variance in relation to the grammar grade with a somewhat wider range for the errors per 100 words ratio in Course 1 and the words per clause ratio in Course 2. One outlier in Course 2 was removed from the data set.

[Insert Table 2 about here]

Correlation coefficients were computed between the teacher-assigned grammar grade and the accuracy and complexity measures. The results of the correlational analysis for each course presented in Table 3 show that only the strong negative correlation of -.848 between the grammar grade and the errors per 100 words ratio in Course 2 is statistically significant. The other correlations are moderate to weak and not statistically significant. Interestingly, there appears to be a moderate negative correlation between clause per T-unit ratio and the teacher-assigned grammar grade, albeit not statistically significant, whereas we would expect to see a positive correlation between these two variables.

[Insert Table 3 about here]

## Phase 2: Student Questionnaires and Interviews (QUAL)

 Thirty-three students completed the questionnaire, and eight students were also interviewed. The presentation of the findings focuses on the two most important themes that emerged from the analysis: the role of accuracy and complexity in the assessment of grammar and the impact of the teachers’ assessment criteria on students’ writing and learning. A more detailed discussion of additional themes can be found in Neumann ([2011](#_ENREF_61)). The findings are organized by theme rather than by data source, and, where possible, findings from both data sources have been included for each theme. For certain themes, however, only data from the interviews were available, in which case this is explicitly mentioned below.

### Grammatical accuracy, sentence variety, and complexity.

Grammatical accuracy emerged as an important theme in the questionnaire and interview data. About 85% of the student participants in Course 1 refer directly or indirectly to accuracy as the teacher’s assessment criterion for grammar. Some of these students simply list items from the evaluation grid that teachers can tick or circle as part of their evaluation, such as “sentence problems” or “verb structures” (see Appendix B). Since teachers use these lists during the assessment process to identify strong and weak elements, one can deduce that students believe their teachers assess the level of accuracy on the aspects of morphology and/or syntax that students mentioned on the questionnaire. Almost 75% of the student participants in Course 2 mention accuracy or grammar mistakes as important indicators of grammatical ability for the teacher.

A small number of student participants in both courses also mention broader issues in relation to grammar assessment, as seen in the following statements: “That she will find stuff that will prove our understanding of her lessons.” (Pierre[[4]](#footnote-4), Course 1); “she tries to help us improving. After that, she’ll know which grammar points she needs to talk about in class” (Emilie, Course 2); or “I think the teacher look to improve our skills in English and help us learn from those mistakes.” (Gabriela, Course 2). These statements show the students’ awareness of the relationship between the teachers’ instructional goals and their assessment procedures and criteria.

Accuracy also emerged as a prevalent theme in student responses to the question of whether their teacher’s expectations affect the way they write their essays. Ten students in both courses indicated on the student questionnaire and/or in the interview that they try to avoid making grammatical mistakes because they know this is what the teacher looks for. The following statement from Xing’s (Course 2) interview is one example: “But I think the most important thing is grammar. If your grammar is perfect, maybe you can get an A- or B+”.

According to two of the eight student participants who were interviewed, sentence variety and complexity is another important assessment criterion for teachers. The teachers want students “to make [their] sentence[s] look nice” (Yi, Course 2) through the use of conjunctions and transition phrases. For this purpose, teachers provide students with support in order to enable them to use conjunctions and transition words. The goal of producing precise and varied sentences that capture the complexity of the writer’s ideas is not easily achieved, however, because one of the main concerns for students is to avoid errors. This means that these students have to negotiate and choose between “safe” linguistic options over which they have full control and “risky” options that would allow them to express more complex ideas and illustrate relationships between these ideas. Emilie explained this dilemma during the interview:

In fact what really happens if I really, if this idea is very important and I really want it to be in my essay because sometimes I have complex idea and I don’t really know how to write them? I just forget them and I say something easier unfortunately because of the grammar. But sometimes when I know that the rest of the essay is quite well written and there is no mistakes in it, I say ok I can make one sentence with a lot of mistakes in it I don’t mind. ... But it is like once or twice for an essay.

On essay exams, however, Emilie would not take this risk. Instead, she“will put the idea in simple sentences … [and] split [the idea] in various sentences”. Like Emilie, four other students report that they deliberately opt for simple sentence constructions and avoid compound, complex, and compound-complex sentences, all in an effort to circumvent making grammatical mistakes. These students resort to familiar sentence structures in order to avoid making mistakes with unfamiliar or new constructions, as the following quotes illustrate: “So in the exam I think the classmate is like me and we always write the things we are very familiar with to diminish the mistakes” (Xing) and “I try to use structure I am comfortable with to avoid mistakes” (Gabriela). Eight students also referred to this idea in their questionnaire responses. At the same time, these students hope that when they do select the more challenging and complex option that they will be rewarded for this by the teacher, as suggested by Emilie: “Because I hope the teacher ... will appreciate the fact that I am making an effort to write a complex sentence and maybe she will.” Emilie is uncertain, however, whether the teacher will appreciate her linguistic efforts and her risk-taking behaviour.

### Assessment and learning.

During the student interviews, it emerged that the grade assigned as a result of the assessment process has a negative influence on the willingness of some students in this study to take risks, and, ultimately, this may impact their learning. How and whether the grade affects the students depends on the relative importance of the grade in relation to the course mark. As illustrated in Emilie’s quote in the preceding section, the grammar grade sometimes keeps her from trying out new language, especially if the grade is important. She feels that she needs space to experiment with language without the fear of losing points for grammar mistakes. This is echoed by Xing’s comment: “Maybe just it is obviously a spelling mistake but it is always wrong. So the grades always discourage me”. However, if students feel they have the space to take risks without “being punished” for grammar mistakes and without a numeric assessment that counts towards the course grade, they feel that they have a better opportunity to learn as Xing points out:

[It is different] when we write … the sentence practice[[5]](#footnote-5) because not contain any marks in my final grade. So I can write what I want to write and maybe I am not so sure about this sentence but I can write it in the paper and let teacher to correct it. Yeah, so I think it is the good way for me to … improve my grammar.

The exchange with the teacher and her feedback are obviously very important for the students’ learning, but marking or grades may also have a negative effect. It seems that the students need a space without marks to really improve and expand their language repertoire because under exam conditions and in other assessment situations they may stick to the familiar and not take risks. According to the students of Course 1, both first and final drafts where graded, and there was little ungraded writing practice, so these students would have not had this opportunity. In Course 2, however, students completed ungraded sentence writing assignments with the target vocabulary for the course, and first drafts of essay assignments were ungraded.

### Assessment of grammar: A summary of the students’ perspective.

Students in this study agree with the teachers’ focus on grammatical accuracy as an assessment criterion for grammar. At the same time, this focus has unintended consequences. For a substantial number of these students, there is a trade-off between producing grammatically accurate language and the ability to use more complex sentence structures. Because these students know that grammatical errors lead to lower grammar grades, they avoid challenging language structures and stick to those structures that they know well. This avoidance behaviour affects their learning negatively because students have the impression that risk-taking behaviour is not rewarded, or at least they are not certain whether it is.

## Phase 3: Teacher Interviews (QUAL)

### Assessment criteria for grammar.

Grammatical accuracy also emerged as an important theme in the teacher interviews. Miriam, the teacher of Course 1, uses accuracy as the main criterion to assess her students’ grammatical ability in writing and has developed her own method of assigning scores on the evaluation grid:

 Typically what I do when I am going through an essay I actually tick errors that they are making … I’ll tally it all up and I look at it that way. … So if they have twenty tallies on [the evaluation grid] then I am just going to subtract from forty and that is how I will get their mark.[[6]](#footnote-6)

Her rationale for using this system is two-fold. First, it makes the assessment of grammar very clear and transparent to the students; they know exactly why they obtained a certain grade. Second, this system visually identifies and clearly communicates areas of weakness to the students: If students have more than five errors in one category, Miriam assigns extra work. As was evident in the student interviews, some students may opt to avoid making errors by choosing safer and simpler sentence constructions over which they have good control. According to Miriam, however, these students would not gain any advantage by doing that in her course because sentence variety and evidence of complex constructions are also criteria for her; she refers to the band descriptors that accompany the evaluation grid to ensure that students receive the appropriate grade that is reflective of the level of accuracy as well as complexity in their writing. During the discussion of the marking criteria for selected essays in her class, Miriam pointed particularly to the error count as a criterion for determining the grammar grade on the evaluation grid.

Elaine, the teacher of Course 2, looks for what she calls “tolerably acceptable language.” She gives low or failing marks “if the back of my head goes *screech*.” The biggest criterion for this intuitive reaction to her students’ writing is comprehensibility. In the discussion of the four essay exams, it became apparent that “acceptable language” also had to be accurate because Elaine pointed out different kinds of language errors as she was reviewing the essays during the interview. Like Miriam, Elaine marks errors on the evaluation grid, but she does not use a mathematical formula to determine the grammar grade. Rather, she looks not only at the level of accuracy but also for variety of expression and generally “elegant language” as reflected in vocabulary use, sentence structure, and the evident thought process. That is, Elaine considers the three constructs present in the evaluation grid descriptors: comprehensibility of the text, effectiveness of sentence construction, and accuracy (see Appendix A). For the low grades, comprehensibility and clarity of the ideas are always an issue. During the discussion of the marking criteria for particular essays in her class, Elaine makes reference to a number of criteria. While accuracy and the overall count of grammatical errors do have an impact, she also refers to concepts such as “workable language” or “variety of expression” when explaining why she assigned a particular grammar grade for essays. Therefore, in addition to accuracy, she assesses the ease with which she can understand the ideas expressed in the text (“workable language”) as well as the repertoire of linguistic expression that is evident in the student’s essay (“variety of expression”). The focus of grammatical accuracy as an instructional goal is also reflected in what Elaine refers to as her weekly sentence writing practice, where students have to submit individual sentences and focus on writing them as grammatically accurately as possible.

### Assessment of grammar: A summary of the teachers’ perspective.

Both teachers attend to accuracy in the students’ writing although this criterion plays a different role for each teacher in the assessment process. For Miriam, it seems that practical concerns have led her to adopt her method of tallying errors to determine the grammar grade. Elaine, on the other hand, relies on her intuitive judgement when she looks for “acceptable language.” This criterion means that the message should be comprehensible and that communication should not be impeded by linguistic errors. Consequently, Elaine also focuses on linguistic accuracy. The teachers’ focus on accuracy is motivated by their desire to encourage their students to improve their command of English grammar and their control of linguistic structures. That is, the teachers hope that the focus on accuracy will lead to student learning: Students will notice their mistakes and learn to avoid them in the future. The unintended consequence, however, is that students may avoid certain structures altogether in an effort to reduce grammatical errors. These avoidance strategies may ultimately have a negative impact on student learning, as the findings from the student interviews indicate.

# Discussion & Implications

## Answers to Research Questions (RQ)

### RQ 1: What indicators of grammatical ability do teachers attend to in their students’ L2 academic writing when assigning grades for grammar on an analytical rating scale?

The results from this study point to accuracy as an important indicator of grammatical ability for teachers in this study. Although the correlation between the grammar grade and the accuracy measure for Course 1 was not statistically significant, the teacher Miriam repeatedly mentioned accuracy as an assessment criterion during the interview, and she bases her assessment of grammatical ability on the overall error count for each essay. In contrast, the accuracy measure shows a strong negative correlation with the grammar grade in Course 2 despite the teacher’s comment that she was not focused on an overall error count. The lack of statistical significance of the correlation between the accuracy measure and the grammar grade for Course 1 may be at least partly due to the smaller variance of grammar scores as evident in the smaller standard deviation of 5.53 compared to that of 10.41 for Course 2 ([see Table 2; cf. Ortega, 2003](#_ENREF_64)). Although sentence variety and complexity are important criteria for higher grammar grades for Elaine according to the interview data, the negative correlation (albeit not statistically significant) between the clauses per T-unit ratio and the grammar grade points towards the possibility that the opposite may be the case: Complexity may in fact lead to lower grammar grades. There is only a weak positive trend observed in that course for the words per clause ratio. It could be argued that the descriptive statistics for the two courses potentially indicate different stages of development in terms of complexification of the L2 writers’ texts with the writers in Course 2 somewhat more advanced as apparent in the slightly lower mean for the clauses per T-unit ratio and slightly higher mean for the words per clause ratio in Course 2. This difference, however, is not evident in any differences in how the teachers describe their process of assigning the grammar grade or in different statistically significant correlations between these measures and the grammar grade for the two courses. Overall, the data point to sentence-level indicators of accuracy as the primary assessment criterion.

### RQ 2: What knowledge do students have about the indicators their teachers look for when assigning a grade for grammar in writing?

The students in this study are aware of their teachers’ focus on accuracy. They can enumerate particular areas of weaknesses that their teachers had previously identified in their writing using the evaluation grid.

### RQ 3: What do students perceive to be the impact of their teachers’ assessment criteria on their way of writing and learning in the L2 writing classroom?

The students in this study apparently attempt to understand their teachers’ expectations and then tailor their writing to meet those expectations. Rather than focusing on their learning and taking risks to experiment with new language and expand their linguistic repertoire, students may avoid particular structures in order to satisfy the teachers’ expectation for accuracy.

## Accuracy, Complexity, and the Teacher-Assigned Grammar Grade

Both the text analysis and the interview data indicate that grammatical accuracy is the main indicator for the teacher-assigned grammar grade. The errors per 100 words ratio is negatively correlated to the teacher-assigned grammar grade in Course 2. In other words, as the number of errors decreases and the level of accuracy increases, the grammar grade goes up. This finding was to be expected because rating scales for L2 writing tasks in high-stakes tests such as the TOEFL ([Educational Testing Service, 2004](#_ENREF_26)) or the IELTS ([International English Language Testing System, n.d.-a](#_ENREF_48), [n.d.-b](#_ENREF_49)) also include increasing demands for accuracy with every band increase on the rating scales.[[7]](#footnote-7) Just like the raters on those tests, the teachers in this study look for increased levels of accuracy in order to award higher grammar grades on the evaluation grid. Because of the emphasis on grammatical accuracy in class, in feedback, and on the program evaluation grids, students are aware of this evaluation criterion for grammar.

In contrast, there is no statistically significant correlation between the two complexity measures and the grammar grade. Interestingly, for one of the measures we can notice a trend towards a negative correlation between the clauses per T-unit ratio and the grammar grade in Course 2. This is surprising because increased complexity is usually considered a sign of increased proficiency and more advanced language use and control ([Norris & Ortega, 2009](#_ENREF_63); [Ortega, 2003](#_ENREF_64); [Skehan, 2009](#_ENREF_76); [Wolfe-Quintero et al., 1998](#_ENREF_93)). Consequently, when L2 writers produce more complex sentences, the teacher-assigned grammar grade should increase and not decrease, as appears to be the case in Course 2. This finding would also be in conflict with the descriptors of the program evaluation grid, where the top band describes the language as varied and complex (see Appendix B). Language use at the low bands, on the other hand, is characterized by frequent problems in or avoidance of complex structures (see Appendix B). This expectation for increased complexity at higher bands of a rating scale is also reflected in rating scales for high-stakes exams ([Educational Testing Service, 2004](#_ENREF_26); [International English Language Testing System, n.d.-a](#_ENREF_48), [n.d.-b](#_ENREF_49)). It could also be argued that this trend may also be explained by increased phrasal elaboration and increased complexity that is measured by the words per clause ratio. However, that correlation was not statistically significant either, and with a value of .276 this trend is noticeably weaker than the trend observed for the clauses per T-unit ratio. Ultimately, only the errors per 100 words ratio shows a strong statistically significant correlation with the teacher-assigned grammar grade.

One possible explanation is that increased levels of complexity may, in fact, be also associated with a higher number of errors. A statistically significant, moderate correlation between the clauses per T-unit ratio and the errors per 100 words ratio for Course 2 (r = .57, *p* < .05) provides support for this hypothesis. It appears that an increased error count, potentially as a result of more clauses per T-unit, may cause teachers to award a lower grammar grade despite higher levels of syntactic complexity. Consequently, the assessment criterion for accuracy seems to supersede any appreciation of complexity in L2 texts during the assessment process of grammatical ability in L2 writing. The correlations between the error per 100 words ratio and the words per clause ratio are not statistically significant in either course.

The inclusion and the importance of the accuracy criterion for both teachers in this study also led to a conflict between instructional goals and the effect on the students. The accuracy criterion is included on the evaluation grid because higher proficiency in an L2 is associated with an increased control over or mastery of the structures of this language. Consequently, L2 writers should make fewer mistakes to reach higher levels on rating scales. The L2 teachers in this study focus on grammatical accuracy for this reason; they also wish to encourage their students to learn by pointing out particular areas of weakness (as Miriam, the teacher of Course 1, does) or through focus on sentence-level accuracy (as Elaine does in the weekly sentence writing practice). The goal for both teachers is to help students improve and learn to avoid making particular grammatical errors. This is clearly a positive instructional goal. This focus, however, also has a potentially negative result. First, students may actually take a rather pragmatic approach to meeting this criterion by avoiding risk-taking behaviour and employing safe language structures, thereby making fewer grammatical errors. Second, students may avoid expressing complex ideas that would require the use of more sophisticated or challenging linguistic structures so that they do not have to take a chance and risk grammatical errors, especially on important assignments or exams. As a result, teachers do not see the potential of what their students might be able to do (i.e., their full range of linguistic potential) but indirectly encourage the students to retreat to a linguistic safe zone.

Ultimately, the focus on accuracy to the exclusion of complexity as an assessment criterion for grammatical ability in this study contravenes the goal in L2 pedagogy to encourage risk-taking. In L2 acquisition research, risk-taking has been found to be associated with L2 learning success ([Brown, 2000](#_ENREF_10); [Gass & Selinker, 2008](#_ENREF_33); [Lightbown & Spada, 2013](#_ENREF_54); [Oxford, 1990](#_ENREF_65); [Skehan, 1989](#_ENREF_73)). For this reason, language teaching methodology books ([e.g., Brown, 2007](#_ENREF_11)) and language curricula ([e.g., Ministère d'Éducation, Loisir et Sport, n.d.](#_ENREF_60)) recommend that teachers encourage student risk-taking behaviour to facilitate and promote language learning in the classroom.

The data show that the teachers in this study had difficulty balancing the assessment criteria for grammatical accuracy and complexity. This finding is similar to what Fritz and Ruegg ([2013](#_ENREF_32)) found in their study of the rating of vocabulary in writing. Although raters had been trained to assess accuracy as well as range, scores were only related to lexical accuracy but not range. Achieving a balance between grammatical accuracy and complexity is just as important in L2 writing assessment as Li ([2000](#_ENREF_53)) had argued for L2 writing instruction. For this reason, some developers structure rating scales so that L2 test takers are not penalized for taking risks but are, in fact, encouraged to show the full range of their linguistic abilities. This approach has been adopted for the rating scales of the *Deutsche Sprachdiplom* (German Language Diploma, GLD)[[8]](#footnote-8) with separate criteria for linguistic accuracy and linguistic range or repertoire. For the criterion *Strukturen* (structures), raters judge the level of complexity and the linguistic repertoire in texts, whereas they assess the level of morphosyntactic accuracy under the criterion *Korrektheit* (accuracy) ([Zentralstelle für das Auslandsschulwesen, 2012](#_ENREF_95)). Furthermore, the GLD handbook specifically discusses the relationship between these two criteria and how raters should address them jointly. This directive is based on the principle that learners’ risk-taking should be rewarded and not discouraged so that a negative washback on instruction can be avoided ([Zentralstelle für das Auslandsschulwesen, 2012](#_ENREF_95)).

# Implications: The Assessment of Grammatical Ability in Practice.

The assessment of grammatical ability in academic writing is an important component of the overall assessment of students’ ability to write in an L2. The definition of grammar and grammatical ability as a construct in the literature is relatively complex, but the findings of this study indicate that the construct assessed in the classroom by academic L2 writing teachers is reduced to accuracy on the sentence level. Although Elaine, the teacher for Course 2, and certain students in this study refer to complexity and variety as an assessment criterion, the text-based analysis does not provide confirmation for this. In fact, if anything, there appears to be an opposite trend in the data for Course 2. There appears to be evidence then that the construct assessed by the teachers does not resemble the model put forward in the testing literature. For this reason, the assessment of grammatical ability is based on what Messick ([1989](#_ENREF_59)) calls an underrepresentation of the theoretical construct. This finding should be cause for concern because of the mismatch between the labelling of the construct (grammar) and its operationalization (accuracy only). It appears that for the teachers in this study grammar equals accuracy whereas the language testing literature defines the construct more broadly and comprehensively.

For rating scale developers, the findings of this study mean that, despite all best intentions, the rating scale may not be used as planned and a different construct is actually assessed. The structure of a rating scale may lead to a simplified construct operationalization of grammatical ability in practice despite a complex operationalization of the construct in the rating scale development process when grammatical accuracy, grammatical complexity, and effectiveness and variety of syntactic constructions are included under one criterion. If grammatical accuracy, however, is assessed separately from other aspects of grammatical ability, as Fritz and Ruegg ([2013](#_ENREF_32)) suggest for vocabulary based on the findings of their study, rating scale designers could ensure positive washback through writing assessment to promote learning more broadly rather than generate a focus on error avoidance strategies by the students. Maybe this emphasis would help both L2 writing teachers and their students to move away from “an exaggerated emphasis on grammatical accuracy” ([Iwashita et al., 2008, p. 47](#_ENREF_50)). This would potentially allow the students to deal with language more creatively and to have the room to experiment with the language so that their linguistic abilities could grow and develop. It would, therefore, be worth considering the adoption of this approach to rating scale design more widely, especially for classroom assessment.

For teacher educators, these results have implications in terms of what students should learn in their courses. It seems that a concerted effort may be required by teachers to resist the temptation to reduce grammar in writing to accuracy only. Rating scales that separate the constructs of accuracy and complexity may help teachers to do that ([e.g., the GLD rating scales, Zentralstelle für das Auslandsschulwesen, 2012](#_ENREF_95)), but this would have to be explored in future research. Finally, teachers may not have considered how their marking criterion with its focus on accuracy affects the strategies students employ to satisfy their teacher’s expectations. Teachers, consequently, may not see what students can do, but only what students think teachers want them to produce. This raises the question to what extent the evaluation grid used in this ESL program is useful in assessing grammatical ability. Does this scale hinder or assist teachers in the comprehensive assessment of grammatical ability in academic writing? It seems the former may be the case.

Finally, L2 writing teachers probably need to find a way that allows them to promote grammatical accuracy and increased complexity at the same time. If that is not possible, it may be worth exploring if the assessment focus in terms of grammar can shift depending on the writing task or assignment, which would, of course, have to be clearly communicated to the students. Should program and/or university policy not permit teachers to do this, ungraded work could be used to encourage students to take risks and try out new and/or more complex linguistic structures with which they are less familiar. This may be realized, for example, by not grading first drafts, as Elaine did in this study. Either of these practices would provide a space for students in which they could safely experiment with language and push their limits without fearing the reprisal of grades and the loss of points for every grammatical mistake that they make.

# Strengths and Limitations of the Study

This study has two strengths that may be considered limitations by some. First, the findings are based on a case study of 2 teachers and their 33 students in one particular educational context. Although this context is similar to other educational settings in terms of instructional goals, purpose of the program, and its student clientele, the findings may be a product of the interaction between the research participants in this study and therefore not valid beyond the current research context. Yin ([2009](#_ENREF_94)) argues, however, the findings of any empirical study (and not just case studies) have to be confirmed by further research. Second, the study is based on a mixed methods research design. Such a design “offsets the weaknesses of both quantitative and qualitative research” ([Creswell & Plano Clark, 2007, p. 9](#_ENREF_21)) but also requires the merging of different data sources and the consolidation of potentially contradictory findings ([Creswell & Plano Clark, 2007](#_ENREF_21); [Teddlie & Tashakkori, 2009](#_ENREF_83)). Depending on one’s worldview, this can be a strength or a limitation. The results of the current study, however, provide substantive evidence of the usefulness of the design because the overall finding for grammatical accuracy as an important assessment criterion for grammatical ability is confirmed by different data sources across the two courses.

Another limitation of this study is the context within which it was conducted. This study took place in university academic writing classes with a focus on developing students’ ability to write academic essays, and the students in this study are fairly proficient users of ESL. Although the essay exams analyzed in this study are not the only writing task that students in these classes complete, the focus of these courses is on essay writing tasks. The findings of this study might be different in writing courses that include a wider range of writing tasks (e.g., research papers) or have student participants from other age groups and/or with higher and/or lower proficiency levels.

# Conclusion

The current study of classroom assessment of grammar in L2 academic writing has two important findings. First, this case study discovered a gap between the theoretical definition of grammatical ability in the language testing literature and the construct that is assessed in the L2 writing classrooms in this study, despite the teachers’ attempts to assess grammatical ability more broadly. Second, the findings reveal that even classroom assessment creates a washback effect on how and what students may learn. Alderson and Wall ([1993](#_ENREF_1)) formulated possible washback hypotheses, one of which states that “a test will influence **what** learners learn; and a test will influence **how** learners learn” (p. 120). This is not new to teachers who consciously use tests and other forms of assessment in an attempt to positively influence student behaviour ([Wall & Alderson, 1993](#_ENREF_86)). This study revealed, however, that the relationship between the teachers’ pedagogical goals, assessment criteria, and learning objectives for students and the resulting student behaviour is not necessarily straightforward. The teachers in this study want their students to improve their grammatical ability, so the teachers encourage them to increase the level of grammatical accuracy in their writing. As the findings from the interviews with students in this study indicate, however, students may try to meet these objectives in ways not intended by the teacher: a focus on error avoidance instead of risk-taking, the latter of which is associated with L2 learning.

# Acknowledgements

This study was undertaken in partial fulfillment of the requirements of the author’s PhD in Second Language Education at McGill University. Earlier versions of this paper were presented at the Symposium on Second Language Writing in Phoenix, AZ in November 2009 and at the Language Testing Research Colloquium in Cambridge, United Kingdom in April 2010. The author would like to thank Alison Crump, Bonnie Barnett, Kerri Staples, and May Tan for their help in data transcription and analysis, and Carolyn E. Turner, Beverly A. Baker and the anonymous reviewers for their helpful comments on earlier versions of this paper. This study was supported by a doctoral bursary from the FQRSC (*Fonds québécois de la recherche sur la société et la culture*).

# References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129. doi: 10.1093/applin/14.2.115

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, B. A. (2010). *In the service of the stakeholder: A critical, mixed methods program of research in high stakes language assessment.* (Unpublished doctoral dissertation). McGill University, Montreal, Quebec, Canada, Montreal. Retrieved from <http://digitool.Library.McGill.CA:8881/R/?func=dbin-jump-full&object_id=96783>

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*(1), 17-34. doi: 10.1017/S0272263100007816

Barkaoui, K. (2010a). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515-535. doi: 10.1177/0265532210368717

Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54 - 74. doi: 10.1080/15434300903464418

Beers, S., & Nagy, W. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*(2), 185-200. doi: 10.1007/s11145-007-9107-5

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

Brown, H. D. (2000). *Principles of language learning and teaching* (4th ed.). White Plains, NY: Longman.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy* (3rd ed.). White Plains, NY: Pearson.

Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate foreign language curriculum. *Language Testing, 19*(4), 419-437. doi: 10.1191/0265532202lt238oa

Byrnes, H., Maxim, H. H., & Norris, J. M. (2010). Realizing advanced FL writing development in collegiate education: Curricular design, pedagogy, assessment. *Modern Language Journal, 94*(Monograph Issue). doi: 10.1111/j.1540-4781.2010.01147.x

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47. doi: 10.1093/applin/I.1.1

Charmaz, K. (2003). Qualitative interviewing and grounded theory analysis. In J. A. Holstein & J. F. Gubrium (Eds.), *Inside interviewing: New lenses, new concerns* (pp. 311-330). Thousand Oaks, CA: SAGE.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: SAGE.

Corbin, J. M., & Strauss, A. L. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (3rd ed.). Thousand Oaks, CA: SAGE.

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>

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice, 39*(3), 124-130. doi: 10.1207/s15430421tip3903\_2

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: SAGE.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: The University of Michigan Press.

Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly, 1*(1), 5-18. doi: 10.1207/s15434311laq0101\_2

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and Perils. *Language Assessment Quarterly, 10*(1), 1-8. doi: 10.1080/15434303.2011.622016

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43. doi: 10.1016/j.asw.2005.02.001

Educational Testing Service. (2004). Scoring Guides (Rubrics) for Writing Responses. Retrieved April 8, 2007, from [www.ets.org](http://www.ets.org) under "iBT TOEFL Scoring Guides"

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), 59-84. doi: 10.1017/S0272263104026130

Ferris, D. R. (2002). *Treatment of error in second language student writing*. Ann Arbor: University of Michigan Press.

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Ferris, D. R. (2007). Preparing teachers to respond to student writing. *Journal of Second Language Writing, 16*(3), 165-193. doi: 10.1016/j.jslw.2007.07.003

Ferris, D. R., & Hedgcock, J. (2005). *Teaching ESL composition: Purpose, process, and practice*. Mahwah, N.J.; London: Lawrence Erlbaum.

Fritz, E., & Ruegg, R. (2013). Rater sensitivity to lexical accuracy, sophistication and range when assessing writing. *Assessing Writing, 18*(2), 173-181. doi: <http://dx.doi.org/10.1016/j.asw.2013.02.001>

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course*. New York: Routledge/Taylor & Francis.

Grabe, W., & Kaplan, R. B. (1996). *Theory & practice of writing: An applied linguistic perspective*. New York: Addison Wesley Longman.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Pittsburgh: University of Pittsburgh Press.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). London: Arnold.

Hamp-Lyons, L. (1991). What is a "writing test"? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex.

Hamp-Lyons, L. (2011). Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing, 16*(1), 3-5. doi: <http://dx.doi.org/10.1016/j.asw.2010.12.001>

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Randsdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Erlbaum.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464. doi: 10.1177/0265532212436659

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87-107.

Housen, A., & Kuiken, F. (2009a). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473. doi: 10.1093/applin/amp048

Housen, A., & Kuiken, F. (2009b). Complexity, accuracy, and fluency in second language acquisition [Special issue]. *Applied Linguistics, 30*(4), 461 - 626

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors’ holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3), 219-233. doi: 10.1080/15434300903540894

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. Urbana, IL: The National Council of Teachers of English.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. London: Routledge.

International English Language Testing System. (n.d.-a). IELTS Task 1 writing band descriptors. Retrieved June 10, 2009, from <http://www.ielts.org/researchers/score_processing_and_reporting.aspx>

International English Language Testing System. (n.d.-b). IELTS Task 2 writing band descriptors. Retrieved June 10, 2009, from <http://www.ielts.org/researchers/score_processing_and_reporting.aspx>

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49. doi: 10.1093/applin/amm017

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*(1), 48-60. doi: 10.1016/j.jslw.2007.08.003

Leki, I. (2001). Hearing voices: L2 students' experiences in L2 writing courses. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 17-28). Mahwah, NJ: Lawrence Erlbaum.

Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System, 28*(2), 229-245. doi: 10.1016/S0346-251X(00)00009-9

Lightbown, P. M., & Spada, N. (2013). *How languages are learned* (4th ed.). Oxford: Oxford University Press.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: SAGE.

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

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

McNamara, T. F. (2001). Rethinking alternative assessment. *Language Testing, 18*(4), 329-332. doi: 10.1177/026553220101800401

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Ministère d'Éducation, Loisir et Sport. (n.d.). Programme de formation de l'école québécoise: Enseignment secondaire, deuxième cycle (English as a Second Language). [Quebec education program: Secondary school education, cycle two] Retrieved from <http://www1.mels.gouv.qc.ca/sections/programmeFormation/secondaire2/>

Neumann, H. (2011). *What's in a grade? A mixed methods investigation of teacher assessment of grammatical ability in L2 academic writing.* (Unpublished doctoral dissertation). McGill University, Montreal. Retrieved from <http://digitool.Library.McGill.CA:8881/R/?func=dbin-jump-full&object_id=103454>

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. doi: 10.1093/applin/amp044

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518. doi: 10.1093/applin/24.4.492

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143. doi: 10.1111/0023-8333.31997003

Polio, C. G., & Williams, J. (2009). Teaching and Testing Writing. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 486-517): Wiley-Blackwell.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London & New York: Longman.

Rea-Dickins, P. (1997). So, why do we need relationships with stakeholders in language testing? A view from the UK. *Language Testing, 14*(3), 304-314. doi: 10.1177/026553229701400307

Rea-Dickins, P. (2009). Classroom-based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7: Language testing and assessment, pp. 257-571). New York: Springer.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, Volume 2: Reading, writing, and language learning* (pp. 142-175). Cambridge: Cambridge University Press.

Skehan, P. (1989). *Individual differences in second-language learning*. London: Arnold.

Skehan, P. (1996). A Framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62. doi: 10.1093/applin/17.1.38

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510-532. doi: 10.1093/applin/amp047

Storch, N. (2002). Relationships formed in dyadic interaction and opportunity for learning. *International Journal of Educational Research, 37*(3–4), 305-322. doi: 10.1016/s0883-0355(03)00007-7

Storch, N. (2005). Collaborative writing: Product, process, and students’ reflections. *Journal of Second Language Writing, 14*(3), 153-173. doi: 10.1016/j.jslw.2005.05.002

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103-118. doi: 10.1016/j.jslw.2009.02.003

Storch, N., & Wigglesworth, G. (2007). Writing tasks: The effects of collaboration. In M. d. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 157-177). Clevedon, England: Multilingual Matters.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing, 2*(1), 3-17. doi: [http://dx.doi.org/10.1016/1060-3743(93)90003-L](http://dx.doi.org/10.1016/1060-3743%2893%2990003-L)

Tapia, E. (1993). *Cognitive demands as a factor in interlanguage syntax: A study in topics and texts.* (Unpublished doctoral dissertation). Indiana University, Bloomington. Retrieved from <http://proquest.umi.com/pqdweb?did=746315871&sid=3&Fmt=2&clientId=10843&RQT=309&VName=PQD>

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: SAGE.

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. doi: 10.2307/3588360

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), 82-111. doi: 10.1177/026553229901600105

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41-69. doi: 10.1177/026553229301000103

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

Weigle, S. C. (2012). Assessing writing. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 218-224). Cambridge: Cambridge University Press.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35*(4), 400-409. doi: 10.2307/357792

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication, 46*(1), 30-45. doi: 10.2307/358868

Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing, 26*(3), 445-466. doi: 10.1177/0265532209104670

Wigglesworth, G., & Storch, N. (2012). What role for collaboration in writing and writing feedback. *Journal of Second Language Writing, 21*(4), 364-374. doi: <http://dx.doi.org/10.1016/j.jslw.2012.09.005>

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

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Thousand Oaks, CA: SAGE.

Zentralstelle für das Auslandsschulwesen. (2012). Deutsches Sprachdiplom der Kultusministerkonferenz: Handreichungen für die schrifltliche Kommunikation (Niveaustufe B2/C1). [German language diploma of the education ministers' conference: Documents for the written communication (Level B2/C1)] Retrieved from <http://www.bva.bund.de/DE/Organisation/Abteilungen/Abteilung_ZfA/Auslandsschularbeit/DSD/Handbuch/download_DSD_II_SK_Hand_2012_T2.pdf?__blob=publicationFile&v=2>

Table 1

*Gender, Age, Degree Program, and L1 Background of Student Participants*

|  |  |  |
| --- | --- | --- |
|  | **Students in Course 1** | **Students in Course 2** |
| **Gender** | 10 female, 9 male | 7 female, 7 male |
| **Age** | 18-40 (M = 24.5, SD = 6.3) | 18-39 (M = 23.1, SD = 6.0) |
| **L1 Background**  | 4 Chinese, 3 Arabic, 3 French, 2 Farsi, & 7 others | 5 Chinese, 3 Arabic, 2 French, & 4 others |
| **Degree Program** | 10 Finance, Business, & Accounting3 Engineering2 Social Science2 Science1 Languages & Literature1 Art History & Fine Arts | 6 Finance, Business, & Accounting2 Engineering2 Social Science2 Science1 Language & Literature1 Art History & Fine Arts |

Table 2

*Means and Standard Deviations for the Grammar Grade & Accuracy and Complexity Measures by Course*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Course 1** |  | **Course 2** |
|  | ***M*** | ***SD*** | **Range** |  | ***M*** | ***SD*** | **Range** |
| **Errors per 100 words** | 6.84 | 3.55 | 1.65-12.52 |  | 4.6 | 2.2 | 1.75-8.72 |
| **Clauses per T-unit** | 2.36 | 0.35 | 1.84-3.00 |  | 2.13 | 0.30 | 1.47-2.28 |
| **Words per Clause** | 6.48 | 0.67 | 5.06-7.63 |  | 7.73 | 1.22 | 5.97-9.54 |
| **Grammar Grade** | 71.32 | 5.53 | 63-85 |  | 76.0 | 10.41 | 60-86 |

Table 3

*Correlations between the Grammar Grade and the Accuracy and Complexity Measures by Course*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Course 1** |  | **Course 2** |
| **Errors per 100 words** | -.433 |  | -.868\*\* |
| **Clauses per T-unit** | .266 |  | -.532 |
| **Words per Clause** | -.015 |  | .299 |

\* *p* < .05, \*\* *p* < .01

# Appendix A

## Evaluation Grid for Course 1

(The evaluation grid for Course 2 has the same format, but the weighting of the criteria differs.)



GUIDE FOR COMPLETING THE EVALUATION GRID (Course 1)

Content: Ideas & Information (25%)

1. Thesis statement (explicit, identifiable; appropriate to essay type or topic; predictive)
2. Topic development (depth and quality/originality of information)
3. Support (relevant, sufficient, detailed; general vs. specific support, fact vs. opinion)
4. Information level/value

Excellent (A+, A, A-) Very clear and appropriate thesis, defined and supported with sound generalizations and substantial, specific, and relevant details; distinctive, original content for maximum impact; excellent information level; strong introduction and conclusion.

(Very) Good (B+, B, B-) Clear and appropriate thesis; selects; suitable and appropriate content with sufficient details; informative; occasional minor problems with focus, depth, and/or unity; good introduction and conclusion.

Satisfactory (C+, C, C-) Thesis may be unclear (e.g. too broad/narrow); acceptable topic development; some support points may be vague, insufficient, obvious, unconvincing; satisfactory introduction and conclusion.

Weak (D+, D, D-) Thesis not apparent or weak; poor topic development; lacking in substance; many support points are insufficient, irrelevant and/or repetitive; low information level; weak conclusion.

Fail (F)lacks main idea; unacceptable topic development; too vague, insufficient, unconvincing, or off-topic; not enough to evaluate.

**Organization & Text Structure (20%)**

1. Presence and logical sequencing of introduction, body paragraphs, and conclusion
2. Use of relevant patterns of organization (related to topic or essay type)
3. Coherent and unified relationship of ideas (NB: grammatical accuracy related to cohesive devices is considered under Grammar & Language Use)

Excellent (A+, A, A-) - exceptionally clear plan connected to thesis; well organized, effective and logical sequencing; smooth flow of ideas; excellent use of transition techniques; clarity of message enhanced by organization.

(Very) Good (B+, B, B-) - appropriate pattern of organization relevant to topic or essay type; generally smooth flow of ideas and appropriate use of transition techniques; overall organization good; most transitions used appropriately but would benefit from more frequent and varied use of transitions; sequencing generally logical.

Satisfactory C+, C, C-) - shows understanding of pattern of development; somewhat choppy; relationships between ideas not *always* clear; overall organization satisfactory, but some elements may be loosely connected or lacking in transitions; most points logically sequenced but some problems in organization still exist.

Weak (D+, D, D-) - problems with pattern of organization; disjointed; ideas do not flow well and relationships between ideas are oftennot clear; ideas difficult to follow because they are *often* not logically sequenced and/or are unrelated

Fail (F) - does not show understanding of pattern of organization; no clear organization: confusing, vague, or seemingly unrelated ideas; pattern of organization not pertinent to topic/essay type; ideas not developed in separate paragraphs; not enough text to evaluate

Grammar & Language Use (40%)

1. Sentence structure (coordination and subordination; variety)
2. Sentence problems (fragments, comma splices, run-ons)
3. Verb structures (agreement, tense, form)
4. Phrase structure
5. Articles, pronouns, prepositions

Excellent (A+, A, A-) – sentences skillfully constructed, effectively varied with simple and complex forms; harmonious agreement of content and sentence design; hardly any errors in basic sentence or grammatical forms

(Very) Good (B+, B, B-) – sentences accurately and coherently constructed with some variety; good use of complex constructions; only a few errors in grammatical forms; meaning not affected by errors.

Satisfactory (C+, C, C-) - effective but simpler constructions and/or problems with complex constructions; meaning generally clear; several errors in grammatical forms.

Weak (D, D+, D-) - some problems in simple constructions and/or frequent problems in complex constructions, or avoidance of complex structures; clarity weakened by awkward grammatical structures; many problems in grammatical forms.

Fail (F) - many problems in sentence structures (both simple and complex) and/or absence of complex structures; frequent sentence structure errors which confuse and distract the reader; frequent errors in grammatical forms; not enough text to evaluate.

**Vocabulary (Terminology) (10%)**

1. Word forms
2. Word choice (precision)
3. Register
4. Idiomatic usage
5. Range

Excellent (A+, A, A-) high level of sophistication; impressive range; effective use of vocabulary to express ideas; only a few minor errors with word choice/form/idiom.

(Very) Good (B+, B, B-) – (very) good range and variety in the use of vocabulary; effective word/idiom choice and usage; appropriate register; several minor errors related to word choice/form/idiom.

Satisfactory (C+, C, C-) – adequate range in the use of vocabulary; occasional errors of word choice/form/idiom or usage, meaning generally clear (some minor ambiguity).

Weak pass (D+, D, D-) - limited range; frequent errors of word choice/form/idiom and usage; meaning sometimes unclear or ambiguous as a result of errors.

Fail (F) - very limited range; words recycled, reused, or too general; frequent errors of word choice/form/idiom and usage may obscure the meaning; problems with basic vocabulary; not enough text to evaluate.

**Mechanics (5%)**

1. Punctuation
2. Spelling
3. Capitalization
4. Presentation (NB: punctuation involving fragments, comma splices and run-ons are considered under
5. Grammar & Language Use)

Excellent, (A+, A, A-) – very few errors either in punctuation, spelling, or capitalization; correct indentation; neat presentation.

(Very) Good (B+, B, B-) - only a few minor errors in punctuation, spelling, and capitalization; clarity of message never affected by errors; correct indentation; legible handwriting.

Satisfactory (C+, C, C-) - occasional errors in punctuation, spelling or capitalization, problems with indentation; meaning still clear despite errors; handwriting hard to read but basically legible.

Weak (D+, D, D-) - many errors in punctuation, spelling, capitalization; meaning sometimes unclear as result of mechanical errors; absence of indentation; nearly illegible handwriting affecting text comprehension

Fail (F) - dominated by errors in punctuation, spelling, indentation and capitalization; illegible handwriting.

# Appendix B:

## Student Questionnaire

**Information about you**

1. What is your age? (Please state) \_\_\_\_\_\_\_\_\_\_\_\_
2. What is your gender? (Please check) \_\_\_\_\_ female \_\_\_\_\_ male
3. Which language did you FIRST learn to speak? (Please state)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Which other languages do you speak? (Please state)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How long have you lived in Canada? (Please check)

\_\_\_\_\_ 1 year or less

\_\_\_\_\_ 2-4 years

\_\_\_\_\_ 5 years or longer

\_\_\_\_\_ I was born in Canada

1. What is your current student status at Concordia University? (Please check)

\_\_\_\_\_ Undergraduate student

\_\_\_\_\_ MA student

\_\_\_\_\_ PhD student

1. What is your major field of study? (Please state)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. What is your current semester in your degree program at Concordia University?

\_\_\_\_\_ 1st semester

\_\_\_\_\_ 2nd semester

\_\_\_\_\_ 3rd semester

\_\_\_\_\_ 4th semester

\_\_\_\_\_ 5th semester or higher

1. Is this your first ESL course at Concordia University? \_\_\_\_\_ yes \_\_\_\_\_ no

If you answered NO, which courses have you already taken? (Please state)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. How did you learn English? (Please check ALL that apply)

\_\_\_\_\_ English classes in my country. Please state country: \_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_ English classes in Canada

\_\_\_\_\_ With friends in Canada

\_\_\_\_\_ At work in Canada

\_\_\_\_\_ Other, please state: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. What is the total number of years you have been studying English since you started school? (Please state) \_\_\_\_\_\_\_\_\_\_\_\_
2. How often do you use English outside of classes? (Please check only ONE)

\_\_\_\_\_ Never

\_\_\_\_\_ Almost never

\_\_\_\_\_ A few times a week

\_\_\_\_\_ At least once every day

\_\_\_\_\_ All the time

1. For what purposes do you use English? (Please check ALL that apply)

\_\_\_\_\_ At university

\_\_\_\_\_ At work

\_\_\_\_\_ With my friends

\_\_\_\_\_ At home

\_\_\_\_\_ Other, please state: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Questions on the Assessment of Grammar**

When your teacher evaluates your essays, he/she uses an evaluation grid. One of the marks on the grid is for *language use & grammar*. The following questions all refer to this grammar mark on the evaluation grid.

1. What is grammar to you?
2. What do you think your teacher looks for in your essays when he/she assesses grammar?
3. Do you understand why you receive a particular grade for grammar on your essays?

\_\_\_\_\_ Yes \_\_\_\_\_ No \_\_\_\_\_ Don’t know

Comments:

1. What do you focus on when you’re writing an essay exam in English?
2. Do you think about the grammar mark when you write an essay exam in English?

\_\_\_\_\_ All the time

\_\_\_\_\_ Sometimes

\_\_\_\_\_ Not really

\_\_\_\_\_ Don’t know

Comments:

1. Do teacher’s expectations for grammar affect the way you write? Explain why or why not.

# Appendix C

## Protocol for Student Interviews

Welcome and chat with the student.

Thank you very much for taking the time today to come here for this interview. It is very important for me to hear what you have to say about the assessment of grammar. I really appreciated your responses on the questionnaire. I would like your help in understanding some of them in more depth.

1. Before you agreed to participate in my study, had you ever thought about how teachers assess students’ work?
2. Have you given some thought to what grammar represents or means to you?

1. What do you think your teacher looks for in your essays when she assesses grammar?
2. Have you thought about why you receive a particular grade for grammar on your essays?
3. What do you focus on when you’re writing an essay exam in English?
4. Do you think about the grammar mark when you write an essay exam in English?
5. Do your teacher’s expectations for grammar affect the way you write? Explain why or why not.
6. I have asked you a lot of questions and you’ve been very helpful. But do you think there’s anything we’ve missed out?
7. Do you have any other comments about what we’ve discussed?

# Appendix D

## Protocol for Teacher Interviews

Greet and chat with the teacher.

Thank you very much for taking the time today for this interview. I would like your help in understanding your expectations in terms of grammar performance…and your beliefs. It is very important for me to hear what you have to say about the assessment of grammar as it will allow me to understand the assessment process much better.

1. When you sit down to assess your students’ essays, what is the biggest problem or challenge you face?
2. What do you tell your students they should focus on when writing an essay?
3. When assigning the grammar mark, what do you look for in an essay?
4. What is the most important aspect for you when deciding which grammar mark to give?
5. Which criteria do you use to distinguish between the different grades levels A, B, C, etc.?
6. What are the characteristics of this particular essay that cause you to assign it Grade X for grammar?

For each of the four essays

1. If the student who wrote this essay would ask you about why he/she got this particular mark on grammar, what would you tell him/her?
2. When you discuss grammar marks for your students’ essays with your colleagues, what do you usually agree on?
3. When you discuss grammar marks for your students’ essay with your colleagues, what do you usually disagree on?
4. I have asked you a lot of questions and you’ve been very helpful. But do you think there’s anything we’ve missed out?
5. Do you have any other comments about what we’ve discussed?

# Appendix E

## Essay Topics on the Midterm Exam

Topics for Course 1

1. Write an essay with examples illustrating the drawbacks of living gin the digital age, i.e., a time in which electronic devices are prevalent.
2. There are different ways to escape the stress of modern life. Write an essay using specific examples to support his statement.
3. There are several important life skills that people need in order to be successful in the workplace. Write an essay using specific examples to support this statement.

Topics for Course 2

1. We have all experienced stress in our lives. It comes from our work, our family life, and our relationship with people. Discuss the positive effects of stress in our lives.
2. Fast-food restaurants are everywhere. They are in shopping malls, office buildings, city centers, and neighbourhoods. Discuss EITHER the causes OR the effects of the popularity of fast food restaurants.
3. Despite the social safety net and various assistance programs run by governments and charity organizations, we see many homeless people in streets, in parks, and at metro stations. Discuss the causes of homelessness.
1. Although there are alternative ways of defining and analyzing grammar, such as the systemic-functional linguistic framework ([Halliday & Matthiessen, 2004](#_ENREF_36)), I draw here specifically on definitions of grammar in the language assessment literature, within which the current study is situated. [↑](#footnote-ref-1)
2. These labels refer to institutional ESL program levels. [↑](#footnote-ref-2)
3. Although the term rating scale is more commonly used in assessment research, I will use the term evaluation grid when referring to the program evaluation grid for two reasons. First, it coincides with program terminology and facilitated communication with student and teacher participants. Second, because of its feedback function, this grid and how it is used seems different from large-scale assessment contexts where much L2 writing assessment research takes place. [↑](#footnote-ref-3)
4. All names are pseudonyms. [↑](#footnote-ref-4)
5. Students write individual sentences using target vocabulary for the teacher in class. These sentences are then corrected by the teacher for grammatical mistakes but are not graded. [↑](#footnote-ref-5)
6. This teacher’s approach may seem unusual. However, quantification of grammatical errors is usually included in writing rubrics (e.g., see the ESL program evaluation grid in Appendix B or the TOEFL writing rubrics). The sole focus on this error count may appear curious at first, but the text-based analysis indicated that in fact the teacher of Course 2 adopted a similar approach, without being aware of it or stating it explicitly. [↑](#footnote-ref-6)
7. It is worth pointing out that both the TOEFL and IELTS rating scale are holistic scales, so a number of criteria have to be met to reach a top band score. The evaluation grid used in the ESL program in this study, however, is an analytical scale; therefore, only indicators relevant to the assessment of grammar should determine the subscore on this criterion. [↑](#footnote-ref-7)
8. The GLD was developed and is administered by the German *Conference of Education Ministers*. This high-stakes diploma serves as a certificate of proficiency in German for students who attend German-medium high schools or learn German through other educational programs outside the German-speaking countries. The DSD is offered at the A2, A2/B1, and B2/C1 levels of the *Common European Framework of Reference for Languages* ([Council of Europe, 2001](#_ENREF_18)). [↑](#footnote-ref-8)