Developing, Validating, and Incorporating a Rubric for Assessing the Construct of Integration

into an EAP Program

Pakize Uludag

A Thesis

In the Department

of

Education

Presented in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy (Education) at

Concordia University

Montreal, Quebec, Canada

June 2021

© Pakize Uludag, 2021

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Pakize Uludag

Entitled: Developing, Validating, and Incorporating a Rubric for Assessing the Construct of Integration into an EAP Program and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Education)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____Chair
Dr. Sandra Martin-Chang

_____External Examiner
Dr. Ute Knoch

_____External to Program
Dr. Calvin Kalman

_____Thesis Supervisor
Dr. Kim McDonough

_____Examiner
Dr. Heike Neumann

_____Examiner
Dr. Teresa Hernandez Gonzalez

Approved by   _____
Dr. Steven Shaw, Graduate Program Director

7/28/2021   _____
Dr. Pascale Sicotte, Dean
Faculty of Arts and Science

ABSTRACT

Developing, Validating, and Incorporating a Rubric for Assessing the Construct of Integration into an EAP Program

Pakize Uludag, Ph.D.

Concordia University, 2021

Due to their authenticity as an academic writing task, integrated writing tests are widely used for evaluating English for Academic Purposes (EAP) students' writing ability. However, apart from validation studies focusing on a common set of standardized test rubrics, little research has explored the construct of integration or how to assess it effectively in second language (L2) writing classrooms. Thus, this dissertation designed and validated an analytic rubric for assessing the construct of integration in L2 writing while also examining students' conceptualizations of integrated writing assessment in an EAP writing course.

Study 1 investigated which sub-constructs EAP instructors orient to when assessing L2 writers' integrated essays. Triangulation of data sources from instructor ratings, stimulated recall interviews, and textual analysis of the integrated essays informed the development of an analytic rubric with four categories (i.e., content, organization, source use and language use) for assessing classroom-based integrated writing tasks.

Study 2 focused on the validation of the rubric by employing a mixed-methods design, which involved a many-facet Rasch measurement analysis, semi-structured interviews with EAP instructors, and linguistic analysis of student essays for fluency, syntactic and lexical complexity, cohesion, and lexical diversity measures. Results from the Rasch model and textual analysis suggest the rubric is of good quality in terms of assessing one single construct and differentiating

the students' task performance across four levels. The instructor comments during the follow-up interviews contributed to the reformulation of descriptors.

Study 3 adopted a case study methodology and investigated L2 learners' conceptualizations of integrated writing assessment and their use of the analytic rubric for self-assessment in an EAP writing course. Data sources included integrated writing samples that were evaluated by the students and their instructor, a writing self-efficacy questionnaire, individual retrospective interviews, and course materials. Qualitative analysis revealed themes related to three aspects of classroom-based integrated writing assessment: task requirements, task conditions, and instructor feedback. There was an overlap between students' self-assessment and instructor evaluation of their integrated essays, suggesting that students could use the rubric effectively.

I save the last acknowledgement for my adoring husband Omer Baykan, the wind beneath my wings. His love, care, and confidence in me has made the journey I took more meaningful than ever.

Contribution of Authors

The original ideas of this dissertation were drafted based on an informal discussion with my thesis supervisor, Dr. Kim McDonough, and committee member Dr. Heike Neumann. As the first author, I took the lead on Study 1 and Study 2, which are coauthored with Dr. Kim McDonough. The writing samples used in Study 1 and Study 2 came from a corpus compiled by Kim McDonough, Heike Neumann, and Sarah Leu, but the rater data were collected especially for these studies. I was the primary researcher in each study and was responsible for collecting and analyzing the data and conducting the statistical analyses. I revised the manuscripts through many rounds of Dr McDonough's insightful feedback. Study 3 of this thesis is single authored.

Table of Contents

List of Tables and Figures

Glossary

Analytic rubric: Two-dimensional rubrics with levels of achievement as columns and assessment criteria as rows, which allow raters to assign points or weights to particular criteria, and then evaluate a writer's performance in each area.

Cohesive devices: Linguistic forms which are also called linking words, linkers, connectors, discourse markers or transitional words that facilitate sentence level cohesion in written texts (e.g., however, in addition to, in fact).

Discourse synthesis: A writer's process of making connections across ideas presented in a single source, or information from multiple sources when composing source-based writing.

Genre: Groupings of texts based on external criteria relating to the purpose, topic and social context of communication within discourse communities that share a set of discourses (e.g., research articles, creative writing, technical writing).

Holistic rubric: One-dimensional rubrics with predefined achievement levels that allow raters to assess a writer's overall performance rather than assigning points or weights to particular criteria.

Integrated writing: Written texts that require introducing, restructuring, and responding to source information to connect source ideas with writers' personal opinions about the topic.

Independent writing: Written texts in which arguments are built exclusively on a writer's prior knowledge and/or experience without using external sources.

Intertextuality: A writer's mental process of creating their own meaning by drawing upon information and the organizational structure of source materials.

Rhetorical acts: The rhetorical function (e.g., elaboration, drawing conclusions) served by a writer's comments or responses to source text information that facilitates integration and helps create authorial voice.

Source-based writing: An umbrella terms for all academic texts that require incorporating information from textual, auditory and/or visual sources (e.g., integrated writing, summary writing, response essays, research papers, literature reviews, and synthesis papers).

Source integration: The ability to incorporate source information effectively and appropriately into source-based writing.

Text-type: Groupings of texts based on internal criteria relating to certain linguistic forms rather than a specific context (e.g., cause-and-effect, argumentative).

Textual meta-function: A writer's process of organizing clauses, sentences, and paragraphs to create a thematic structure, information structure, and cohesion in any types of written text

Chapter 1: Introduction

Writing is a means of communication between writers and their potential audiences, and it is a domain for language use for students at all levels of formal schooling. Students in higher education demonstrate their knowledge and understanding of course subjects, learn language and content, and develop academic literacy skills though writing. Historically, the role of writing has evolved from being conceptualized as a single elementary skill to a multilayered and purposeful way of communication in educational settings (Hyland, 2015; Stahl et al., 2004). Traditional approaches to writing mainly adopted a structural orientation where writing is viewed as a cognitive activity (Applebee, 2000; Lavelle & Bushrow, 2007). The expansion of academic writing research into disciplinary genres has emphasized the interconnected nature of reading and writing skills for obtaining academic literacy (Leki & Carson, 1997; Leki, 2017; Polio & Shi, 2012).

Scholars have discussed *intertextuality* (i.e., composing a text based on the information and organizational structure present in other texts) as a key literacy skill to defining characteristic of university writing studies (Haswell, 2000; Hu, 2001; McAlpine & Amundsen, 2011). *Intertextuality* practices involve critical assessment and appropriate use of sources in academic texts. For university writers to create proper *intertextual* links, they need to understand information presented in written sources, and use textual borrowing strategies to incorporate source information in their texts (Polio & Shi, 2012; Shaw & Pecorari, 2013). First-year writing programs, for example, are designed to teach students how to incorporate a variety of source materials (e.g., scholarly articles, books) into their writing as they transition into upper-level courses (Howard et al., 2010). Undergraduate students are assigned tasks that require them to

synthesize information from sources in both general education and discipline-specific courses (Wette; 2010, 2017; Schuemann, 2008). In addition, most graduate students are expected to compose research-based argumentation using outside sources in their thesis and dissertation writing (Pecorari, 2003; Kwan, 2008).

In responding to the academic needs of international students studying at English-medium universities, English for Academic Purposes (EAP), as a branch of the broader field of English for Specific purposes (ESP), emerged in the 1970s as a pathway program (Hyland & Hamp-Lyons, 2002). Since the number of international students in such English-speaking countries as Canada and Australia has increased, universities have introduced different instructional models in EAP programs to facilitate study and research in English (Flowerdew & Peacock, 2001). For example, in Canada, English-medium universities implement different models of EAP programs to meet students' target needs as they embark on their degree courses (Tweedie & Kim, 2015). While some universities require students to complete intensive English programs before they begin their degree courses, others offer bridging programs that students can enroll in while taking one or two disciplinary courses simultaneously. In addition, some EAP programs run concurrent with undergraduate degree programs, which is the case in the instructional setting of this dissertation study. More specifically, the three studies in this dissertation were situated within an English-medium university in Montreal, Canada, where two academic writing courses are offered to students. While the first course focuses on paragraph-level writing skills with a strong emphasis on vocabulary and grammatical development, the second course targets the analytical skills needed for integrated writing. L2 students who do not meet the English language proficiency test requirements for admission register for the EAP courses, which are taken concurrently with their respective undergraduate degree programs.

Overall, although the objectives of EAP programs may be wide ranging, students have been shown to benefit from EAP instruction by obtaining higher levels of academic achievement and a greater likelihood of graduation (Dooey, 2010).

To prepare students for writing in their disciplines, pedagogical materials in EAP courses have evolved into more complex forms, such as authentic written texts, to teach students how to read and evaluate academic texts and produce text-responsible writing (Hyland & Shaw, 2016; Johns, 1991). Reflecting the strong emphasis on intertextuality practices and academic writing tasks in university courses, EAP programs began increasingly to incorporate *integrated writing* tasks in their curriculum and assessment (Plakans, 2009, 2010; Plakans & Gebril, 2012). Integrated writing tasks typically require that students incorporate written, auditory and/or visual input in their texts, which mirrors the writing requirements for most academic disciplines (Cumming et al., 2006). To write an essay using information from sources, writers must understand the arguments presented in the sources and then integrate content from those sources into their own piece of writing (Plakans, 2008, 2009). Integrating source information effectively helps writers convey their messages, ideas, or research claims to a specific audience in a smooth and successful manner.

Despite the pivotal role that source integration plays in establishing arguments and presenting research claims, little is known about the construct of integration or how to assess it effectively in L2 writing classrooms. In earlier L2 writing and assessment research, source integration has been often perceived as being synonymous with discourse synthesis based on Spivey's (1990) constructivist writing model, which explicates the role of three processes in textual transformation and meaning construction in synthesis writing: *organizing, selecting*, and *connecting.* Studies exploring L2 writers' source integration have focused on their discourse and

paragraph-level organization of textual information (Cumming, et al., 2006; Plakans, 2008; Wette, 2010). In addition, L2 writers' composing processes have been examined to identify how they synthesize discourse in integrated writing tasks (Ascensión, 2005; Plakans. 2009). However, researchers to date have not clarified the construct of source integration in localized assessment contexts, where students generate personal ideas about the topic and are expected to use source texts mainly to validate their propositions rather than writing based on source materials only. Furthermore, without a conceptual framework, which outlines the skills and competencies associated with source integration, it is not clear which rhetorical and linguistic resources L2 writers devote to decontextualizing source-text information in integrated writing tasks. In order to situate source integration in its broader context, the following section introduces a process model for integrated writing tasks. Importantly, this model mirrors the integrated writing tasks that are being used in this dissertation study context, where students are expected to connect source ideas with their personal opinions about the topic.

Developing a Process Model for Integrated Writing

To address the challenges associated with construct definitions, a process model for integrated writing is illustrated below (Figure 1), which outlines three operations involved in the composing process: (1) task environment, which stimulates the generation of personal ideas and pre-writing planning, (2) skills and competencies elicited for source integration, and (3) revision of the final product. The model is a modification of Flower and Hayes's Composing Processes Model (Flower & Hayes, 1981; Hayes, 1996), which consisted of three main cognitive processes/strategies: planning, translating, and reviewing. Along with outlining skills and competencies required for source integration, the model also accounts for the distinct cognitive processes that writers go through while completing classroom-based integrated tasks, where they

are expected to link source information to their personal ideas about the topic. For example, as part of task representation, writers develop attitudes towards the reported evidence which are influenced by the rhetorical purposes they would like to achieve in their texts, and the linguistic and functional requirements of different text-types (Hyland, 2005). In addition, they adopt positions on source content based on their purpose in writing, and the expectations of the audience (Harwood, 2009; Hyland, 2012). At the phase of idea development, writers activate their cognitive processes and develop strategies by engaging with the task environment and activating long-term memory. They develop authorial voice that they will establish in their writing and create a mental organization of the selected information into writing. Before integrating source information, writers also establish specific goals depending on what effect they want their texts to have on the readers (i.e., audience/ persona) (Flower & Hayes, 1981; Hayes, 1996). Writers' cognitive strategies are eventually translated into metacognitive strategies and help them monitor the writing process, produce information, and solve potential problems (Hacker et al., 1998; Roberts & Erdos, 1993). Because the composing process is complex, multi-layered, and recursive, writers move continually back and forth between these processes (e.g., idea development, reading of sources, source integration) while thinking their way toward a finished product.

Figure 1: A Process Model for Integrated Writing

Integrated Writing	Task Environment	**Task Representation** • Text-types • Topic • Writing prompt • Audience	**Idea Development** • **Generating personal ideas** - Activating prior knowledge - Developing stance • **Organizing** • **Goal setting** - Authorial voice (stance) - Writing plans
	Skills and Competencies	**Reading Sources** • Mental representation of source content • Evaluating & selecting relevant information • Accurate comprehension	**Source Integration** • **Introducing source information** - Determining source use purposes - Referencing to the source text (citation patterns & style) - Using linguistic devices (cohesive devices & reporting structures) • **Restructuring source information** - Performing linguistic revisions (paraphrasing & summarizing) - Accurate representation of source ideas • **Responding to source information** - Performing rhetorical acts using discourse markers - Following up with personal ideas
	Post-Writing		**Reviewing** • Revision & Proofreading

Monitoring

Despite being limited in number, studies have explored the impact of the task environment, such as pre-writing planning, and idea development, on L2 students' integrated writing task performance (e.g., Payant et al., 2019; Uludag et al., in print). In addition, among other skills outlined in the process model, reading source texts, and evaluating and selecting relevant information from different source materials have been shown to affect the quality of integrated texts (e.g., Esmaeili, 2002; Gebril & Plakans, 2013; Sawaki et al., 2013). On the other hand, in the absence of a comprehensive model or theory of source integration, evaluation of the three sub-constructs underlying source integration in EAP writing have not gained much research attention. Instead of adopting a holistic approach to source integration, researchers have focused on single aspects of it, such as linguistic revision and accurate representation of source ideas (e.g., Hyland, 2005; Gebril & Plakans, 2013; Uludag et al., 2019). Thus, this dissertation is primarily concerned with the assessment of three sub-constructs, namely introducing,

restructuring, and responding to source ideas, underlying the construct of source integration in EAP writing. The next section provides a description of the three sub-constructs.

Constructs Underlying Source Integration

Source integration, which is the central focus of this dissertation, involves three major processes: introducing, restructuring, and responding to source ideas. When introducing source ideas, writers attribute information to sources to achieve a variety of rhetorical functions, such as creating content, validating propositions, and supplying authoritative information on topics about which they had no direct knowledge or experience (Gebril & Plakans, 2009; Mansourizadeh & Ahmad, 2011; Wette, 2017). In addition, they use the conventions of a particular citation style (e.g., APA, MLA) and a citation pattern, *integral or non-integral* (Swales, 1990) based on the type of information that they borrowed from the source materials. Integral citations emphasize the actions (analysis, methods) placing the cited authors within the sentence as subjects or agents, whereas the non-integral ones highlight research findings listing the cited authors in parenthesis (Coffin, 2006). Writers' reporting verb choices is also important to convey their stance and provide convincing arguments in their texts (Swales, 1990, 2014). Since the selection of reporting verbs differ by discipline (Hyland, 2019; Yilmaz & Erturk, 2017), understanding semantic and functional differences of reporting forms helps writers introduce source information effectively. Another possible linguistic means by which information from sources is introduced in the text is the use of cohesive devices which are also known as linking words, connectors, discourse markers or transitional words that facilitate sentence level cohesion in written texts (Celce-Murcia & Larsen-Freeman, 1999; Swales & Feak, 2004). Based on their pre-established goals, writers use cohesive devices to establish linkages between what the sources offer and what they have generated, which contributes to overall *textual meta-function* (i.e.,

7

organization of textual information to create a thematic structure, information structure, and cohesion) and enable logical flow across different source ideas (Liu & Braine, 2005).

Restructuring source language is about linguistic revision of source-text language and accurate representation of source content. Writers need to perform word-level and syntactic modifications on source information by either summarizing content from a single source or generalizing information from multiple sources (Polio & Shi, 2012; Thompson, 2001; Thompson & Tribble, 2001). In some cases, direct quotations are also used as an alternative to paraphrasing. When making linguistic modifications to the source content, writers need to draw upon their knowledge of grammar and vocabulary (Johns & Mayes, 1990) and avoid unacknowledged copying, patchwriting, and close paraphrasing, which might occur due to misconceptions of academic standards or lacking skills of appropriate borrowing (Keck, 2006; Pecorari, 2003; Shi, 2004). In addition to linguistic revision, restructuring source information also requires accurate representation of borrowed source content, which goes beyond understanding source information. If writers omit some key information or change the meaning of the cited information while making linguistic changes, it might lead to issues related to content generation and persuasiveness in integrated texts (Wette, 2017; Uludag et al., 2019).

Differing from restructuring, which involves the sub-constructs of linguistic revisions and accurate representation of source content, responding to source information elicits the skills of using rhetorical acts and following up with personal ideas. To respond to the source information and maintain textual coherence, writers need to establish logical connections between the ideas that they borrow from sources and their personal opinions. In addition to facilitating the introduction of source ideas, cohesive devices also help writers respond to source information and illustrate their perspective about the cited content. Writers' selection of specific cohesive

devices is informed by the rhetorical acts that they would like to perform, such as constructing a conclusion, explanation or prediction, or a synthesis or a generalization related to the information provided by source texts (Mateos et al., 2014). Writers might support the arguments borrowed from sources, or they could develop counter arguments to support their personal opinions and justify their research claims. Accomplishing such rhetorical skills help them express their voice and create authorial identity.

To summarize, the process model of integrated writing introduced above outlines the skills and competencies elicited by integrated writing tasks drawing on existing research, thereby offering a construct definition for source integration in EAP writing. One of the overarching goals of this dissertation is to determine which constructs underlying source integration (i.e., introducing, restructuring, and responding to source information) could be incorporated in a scoring rubric for assessing classroom-based integrated writing tasks. Earlier studies have utilized a variety of methodologies, such as analytic and holistic rating scales, textual analysis of students' essays, interviews, Likert scale items, and think-aloud protocols, to examine different aspects of source integration. Collecting evidence from multiple data sources is important to reflect on rater challenges and perceptions when designing and validating scoring rubrics, and better interpret the scores assigned to integrated writing tasks. Therefore, the first two studies in this dissertation adopted a mixed-methods approach to design and validate an evaluation criteria informed by the process model of source integration. The third study, which is different from the first two studies, adopted a case study methodology to investigate EAP students' conceptualizations of classroom-based assessment tasks. The next section provides a summary of each study.

Tying it All Together

Study 1 explored the sub-constructs EAP instructors orient to when evaluating L2 writers' integrated essays using an existing EAP program rubric. Triangulation of data sources from instructor ratings, stimulated recall interviews, and textual analysis of the integrated essays for aspects of source integration informed the development of an evidence-based analytic rubric with four categories (i.e., content, organization, source use and language use) for evaluating integrated writing tasks.

Study 2 aimed to validate the scoring rubric, which was developed based on the results from Study 1. The methodologies utilized for validation involved a many-facet Rasch measurement analysis, semi-structured interviews with EAP instructors, and linguistic analysis of student essays for fluency, syntactic and lexical complexity, cohesion, and lexical diversity measures. Results from the many-facet Rasch measurement analysis, and textual analysis suggest the rubric is of good quality in terms of assessing one single construct and differentiating the students' task performance across different levels. The instructor comments during the follow-up interviews contributed to the reformulation of descriptors.

Study 3 investigated L2 learners' perceptions of integrated writing assessment and their use of the analytic rubric for self-assessment in an EAP writing course. Data sources included integrated writing samples that were evaluated by the students and their instructor, a writing self-efficacy questionnaire, individual retrospective interviews, and course materials such as task instructions and syllabus. Qualitative analysis revealed themes related to three aspects of classroom-based integrated writing assessment: task requirements, task conditions, and instructor feedback. The themes were discussed in terms of students' test taking strategies and the use of available support systems in EAP contexts. In addition, findings indicated an overlap between

students' self-assessment and instructor evaluation of their integrated essays, suggesting that students could use the evaluation criteria effectively. The next chapters will document the rationale, methodology and results of the three studies.

Introduction to Study 1

The first study in this dissertation investigated which aspects of source integration outlined in the process model of integrated writing are most salient to EAP instructors while evaluating L2 students' integrated essays. Specifically, this study centralized a focus on the three processes by which writers introduce, restructure, and respond to source ideas and identified how EAP instructors define and evaluate source integration, as the intent with this manuscript was to develop a rubric for assessing classroom-based integrated writing tasks. One of the goals of this manuscript was to address the issue of rubric development through the use of a mixed-methods design. Thus, this study and Study 2 both rely on instructor ratings and perceptions as well as textual analysis of the essays to determine which skills should be represented in the evaluation criteria.

Chapter 2: Study 1

Exploring EAP instructors' perceptions of integrated writing performance: Defining source integration

For submission to TESOL Journal

By Pakize Uludag and Kim McDonough

Introduction

Reflecting the importance of writing using sources, integrated writing tasks have become the focus of greater attention in English for Academic Purposes (EAP) contexts (Haswell, 2000). Integrated writing assessments are widely used for college admissions and placement, while EAP students are often asked to write source-based essays (e.g., argumentative, cause-and-effect) as part of their transition to disciplinary courses (Hyland & Hamp-Lyons, 2002; Reid, 2001). Second language (L2) writing researchers have increasingly investigated the underlying skills required for integrated writing, such as reading (e.g., Belcher & Hirvela, 2001; Grabe & Zheng, 2013; Hirvela, 2004), and linguistic modification of source information (Cumming et al., 2006; Guo et al., 2013). However, the construct of integration itself has received less attention from researchers. As part of a larger study about the design and validation of an integrated writing rubric, the current study aims to identify how EAP instructors evaluate source integration in L2 writers' classroom-based integrated essays.

Defining Source Integration

Source integration involves three processes by which writers introduce, restructure, and respond to source ideas. When introducing source ideas, writers carry out rhetorical functions

such as creating content, validating propositions, and supplying authoritative information on topics about which they have no direct knowledge or experience (Gebril & Plakans, 2009; Mansourizadeh & Ahmad, 2011; Wette, 2017, 2018). In addition, they use the conventions of a particular citation style (e.g., APA, MLA) and a citation pattern, *integral or non-integral* (Swales, 1990) to refer to the source texts. Introducing source ideas also requires writers to use linguistic devices, such as reporting verbs, to convey their stance and provide convincing arguments (Hyland, 2016; Swales, 1990, 2014). Since reporting verbs usage differs by discipline (Hyland, 2019; Yilmaz & Erturk, 2017), understanding semantic and functional differences among reporting verbs helps writers introduce source information effectively. In addition, writers incorporate cohesive features, such as sentence connectors, or transitional words, for introducing source ideas to facilitate sentence level cohesion and establish links between source information and the writer's perspective (Liu & Braine, 2005; Swales & Feak, 2004).

Unlike introducing source information, which focusses on rhetorical purposes, restructuring source language is about linguistic revision and accurate representation of source ideas. Writers need to perform word-level and syntactic modifications on source-text language (Polio & Shi, 2012) either by summarizing content from a single source or by generalizing information from multiple sources (Thompson, 2001; Thompson & Tribble, 2001). In some cases, direct quotations are also used as an alternative to paraphrasing. When making linguistic modifications to source-text language, writers need to depend on their knowledge of grammar and vocabulary (Johns & Mayes, 1990) and avoid verbatim copying, and patchwriting, which might occur due to misconceptions of academic standards or lacking skills of appropriate borrowing (Keck, 2006; Pecorari, 2003; Shi, 2004). In addition to linguistic revision, restructuring source information also requires accurate representation of borrowed source

content, which goes beyond understanding source information. If writers omit some key information or change the meaning of the cited information while making linguistic changes, it might have a negative impact on the content and persuasiveness of integrated texts.

The third component of source integration is responding to source information, which requires that writers establish logical connections between source ideas and their personal opinions in ways that communicate authorial voice and maintain textual coherence. Using discourse markers help writers respond to source information and illustrate their perspective about the cited content. Writers' selection of specific discourse markers is informed by the rhetorical acts that they would like to perform. Within the constructivist approach, rhetorical acts are referred to as *meaning construction* as part of Mateos and colleagues' model (Mateos et al., 2014), which comprises the skills of a) connecting ideas or concepts in the text to examples from the writers' own experience or knowledge, and b) constructing a conclusion, explanation or prediction, or a synthesis or a generalization related to the information provided by cited information. Writers might support the arguments borrowed from sources, or they could develop counter arguments to develop their personal opinions and justify their personal claims. Accomplishing such rhetorical skills help them express their voice and create authorial identity and contribute to successful integration of source content.

Assessing Source Integration

Having defined the construct of source integration in terms of three sub-constructs (i.e., introducing, restructuring, and responding to source information), an important question is how EAP instructors orient to these sub-constructs when assessing students' integrated writing essays. Although researchers have recognized the importance of teacher input for test design and evaluation (Brindley, 2001; Kane, 2006; Shaw & Weir, 2007), few studies have examined rater

perceptions of integrated writing tasks while no research has examined how teachers assess source integration. For example, Cumming, Kantor, and Powers (2002) used think-aloud protocols to explore experienced raters' scoring of independent and integrated TOEFL writing tasks. They found that raters mostly concentrated on task completion, rhetorical organization, and content development as opposed to linguistic features. In terms of source integration, raters paid attention to the appropriate and creative use of source materials. Building on this body of research, Gebril and Plakans (2014) also examined rater processes using think-aloud protocols and interview data and reported that raters mostly employed judgement strategies related to source use (e.g., locating source ideas and citations in students' texts) while evaluating L2 integrated texts. Raters reported having difficulty distinguishing between source text language and the students' own words, scoring texts that include too many quotations, and evaluating linguistic revisions of source ideas. In sum, studies to date have shown that raters tend to focus on the sub-constructs of introducing and restructuring source information when assessing integrated writing texts. It is not clear whether EAP instructors also orient to these sub-constructs when evaluating their students' integrated writing essays.

Whereas few studies have examined rater perceptions of integrated writing texts, considerable research has carried out textual analyses to determine how students restructure source information. For example, studies have explored how L2 writers' linguistic revisions of source-text ideas by classifying their paraphrases along a continuum ranging from direct copying to substantial revision of source information (Campbell, 1990; Hyland, 2005; Keck, 2006; Plakans & Gebril, 2013; Shi, 2004). These studies reported that substantial modification of source language predicts the quality of written texts as assessed by textual analysis and using analytic rubrics. In addition, using automated textual analysis programs, researchers have

determined that lexical diversity and syntactic complexity (Cumming et al., 2006; Gebril & Plakans, 2016; Guo et al., 2013), as well as local and text cohesion (Crossley, Kyle, & McNamara, 2012; Yang & Sun, 2012) predict rater evaluation of integrated essays. In sum, linguistic analysis of integrated texts has demonstrated clear links between the restructuring of source language and judgements of text quality.

To explore how L2 writers incorporate ideas from sources, researchers have analyzed student writing in terms of the accurate representation and rhetorical purpose of source ideas. For example, the accuracy with which students incorporate source content into their text predicts scores on the Canadian Academic English Language test integrated writing task (Uludag et al., 2019). Considering the rhetorical purpose of source information, researchers have found that L2 writers most commonly used source ideas to introduce an idea or acknowledge the origin of information instead of engaging with it in a more complex manner (Neumann et al., 2019; Wette, 2018). Therefore, textual analysis of L2 writing has provided useful information related to rhetorical aspects of source integration not found in studies that have solely drawn on assessment criteria to explain the particular skills writers need to demonstrate.

Taken together, in the absence of a comprehensive model of integrated writing, researchers have focused on single aspects of source integration rather than taking a holistic approach to assessment. In addition, rater perception studies have provided limited insights into how specific constructs are evaluated in integrated writing assessment. Therefore, to address the challenges associated with construct definitions and contribute to score interpretations of classroom-based integrated writing tests, the current study follows a model for source integration that includes three main sub-constructs, as illustrated in Figure 2.

Figure 2: A Process Model for Source Integration

17

```
┌─────────────────────────────────────────────────────────┐
│          Introducing source information                 │
│       • Determining source use purposes                 │
│    • Referencing to the source text (citation patterns & style) │
│  • Using linguistic devices (cohesive features & reporting structures) │
└─────────────────────────────────────────────────────────┘
                          ▼
┌─────────────────────────────────────────────────────────┐
│          Restructuring source information               │
│  • Performing linguistic revisions (paraphrasing, summarising & direct │
│                      quotations)                        │
│      • Accurate representation of source ideas          │
└─────────────────────────────────────────────────────────┘
                          ▼
┌─────────────────────────────────────────────────────────┐
│          Responding to source information               │
│   • Performing rhetorical acts using discourse markers  │
│        • Following up with personal ideas               │
└─────────────────────────────────────────────────────────┘
```

To obtain further insight into EAP instructors' perceptions about the assessment of integrated writing, this study draws upon a variety of data sources in a mixed method approach to address the following research question: How do EAP instructors orient to introducing, restructuring, and responding to source ideas when assessing students' integrated essays?

Method

Integrated writing essays

The EAP integrated essays ($N = 48$) were sampled from the Concordia Written English Academic Texts (CWEAT) corpus (McDonough et al., 2018), which consists of timed integrated writing exams written by students enrolled in an EAP class with a focus on integrated writing at Concordia University. As part of the corpus construction, the essays had been rated using a holistic rubric adapted from the TOEFL guidelines. The holistic ratings were used to sample argumentative essays written on the same topic using the same source texts reflecting a range of scores. The topic was about the role of governments in reducing economic inequality. The essays were written by students with a mean age of 22.36 years ($SD = 2.97$) who were studying in the faculties of Business (28), Arts and Science (10) and Engineering and Computer Science (10). The most frequent L1s were Chinese (21), Arabic (10), and French (9). Additional L1s included

Spanish (5), Romanian (2), and Persian (1). They wrote the essays during a three-hour final exam. Prior to the exam, students had read six readings relevant to the topic and taken notes on a template which encouraged them to paraphrase information from the readings, note key terms, and include any important quotations. At the examination, students selected one of two writing prompts and are expected to develop personal ideas about the topic and validate them using information from the sources to compose an argumentative essay. That is, students were expected to use source information authoritatively to support their propositions. They were allowed to use a paper-based English dictionary along with their notes while composing the essays.

EAP Instructors

Six EAP instructors (two men, four women) who taught EAP integrated writing at Canadian universities were recruited to rate the integrated essays. The instructors had MA degrees in applied linguistics or TESL and had taught in a variety of EAP programs (e.g., intensive programs, bridging programs), with two instructors from the EAP program at Concordia University. Except for one instructor from Turkey, they identified themselves as Canadian L1 English speakers. They ranged in age from 33 to 74 ($M = 43.6$, $SD = 15.1$) and had a mean of 14 years experience teaching English ($SD = 6.3$). Three instructors reported having experience as a professional rater or examiner for IELTS, DELNA and Cambridge exams, and three of the instructors had served administrative functions in their programs related to curriculum development and assessment.

Materials and Procedure

The instructors used an analytic rubric (Appendix A) to evaluate the argumentative essays. The rubric was created by the EAP program at Concordia University approximately 10

years ago and has been regularly updated based on instructor feedback. The rubric has three

dimensions (content and organization, grammar and vocabulary, and mechanics) with four score

levels. Aspects of integration are included in the descriptors for content and organization in

terms of selection and integration of source material, acknowledgement of sources, and accurate

interpretation of source information. At the lower score levels, the descriptors refer to verbatim

copying and overreliance on direct quotations. Integration is also assessed in mechanics where

the use of APA citing conventions is included. The rationale for using this rubric was to

contextualize the study within the EAP program where the sample essays came from.

The EAP instructors met the first researcher individually using an online meeting tool

(Zoom) for a 2-hour session that was recorded. After completing the consent and background

information forms (15 minutes), the instructors evaluated two sample essays and made notes

while reading (30 minutes). These essays were evaluated without using a rubric to elicit the

instructors' general perceptions about source integration before introducing them to the criteria.

The two sample essays represented high- and low-scoring essays that differed in terms of source

use (eight versus two citations, respectively) and length (727 and 442 words, respectively). After

evaluating the first essay, the instructors participated in a stimulated recall interview (Appendix

B) following the steps outlined in Gass and Mackey (2000). The instructors showed their

annotated essays to the researcher using the share screen function and explained their thought

processes while evaluating (20 minutes). The same process was repeated for the second essay.

After eliciting the instructors' general perceptions, the researcher then introduced the rubric and

provided training using two more high- and low-scoring essays from the corpus. After the online

training sessions, the instructors worked independently to evaluate all 48 essays using the rubric

and providing comments on each students' performance for each rubric dimension.

Analysis

Once the scores were obtained from the instructors, inter-rater reliability was calculated using two-way mixed intraclass correlation coefficients, which were .705 for content and organization, .710 for grammar and vocabulary, and .665 for mechanics. As reliability reached acceptable levels for content and organization, the mean ratings are reported in relation to the perception data. The stimulated recall sessions were transcribed and analyzed qualitatively. First, codes were assigned to data chunks related to source-text use, which resulted in the identification of 29 codes. After reviewing the codes, they were grouped into seven categories: verbatim copying from sources, close paraphrasing, unacknowledged source use, misrepresentation of source information, unsupported claims, overuse of source information, voice and tone, and misuse of citation conventions. The seven emergent categories were checked against the instructors' mean ratings and open-ended comments to gain more insight into their perceptions about integrated writing (Creswell & Plano Clark, 2017). After triangulating across data sources, three themes captured the instructors' perceptions about the assessment of integrated essays: (1) using source information to support personal claims, (2) incorporating source-text language in students' own words, and (3) representing source content accurately in their essays.

To establish a link between what instructors attended to while making scoring decisions and what students actually did in their essays, textual analyses were conducted for the sub-constructs of source integration. For this, each instance of source use, defined as specific events, ideas or information discussed in a source text, was identified in students' essays, and manually coded for different aspects of source integration. For the sub-construct of introducing source information, the purpose or rhetorical function served by the source use was coded into three categories used in prior research (Neumann et al., 2019): introduction of a new idea (to generate

21

new content), elaboration on a personal idea (to validate prepositions), and repetition of a previously mentioned idea. For the sub-construct of restructuring source information, the essays were coded for linguistic modification of source-text language using Gebril and Plakans' (2013) framework, which differentiates between indirect source use (paraphrasing and summaries) and direct source use (quotations and verbatim copying). Verbatim copying was operationalized as strings of four or more words copied from the source. Unacknowledged source use was included as an additional category as it emerged from the data. Restructuring was also considered in terms of how accurately the source information was communicated using a binary coding scheme (accurate representation versus misrepresentation) from previous research (Uludag et al., 2019). To confirm the reliability of textual analysis, a PhD student in Applied Linguistics coded 9 essays (20%) independently with interrater reliability of their coding decisions calculated using two-way interclass correlations for linguistic modification of source-text language, which was .81, and Cohen's Kappa for the categorical variables which were .76 for source use purposes, and .78 for source content representation.

Findings

The research question investigated how EAP instructors define and assess integrated writing. Although there were differences in the instructors' current teaching practices and previous experiences with the evaluation of integrated writing, the three themes that emerged from their data were (1) using source information to support personal claims, (2) incorporating source-text language in students' own words, and (3) representing source content accurately. The following sections describe each theme and provide excerpts from the essays to illustrate the instructors' perceptions about integration. Pseudonyms have been used when providing quotations from the instructors.

Using Source Information to Support Personal Claims

The instructors expected the essays to contain a balance of personal opinions and source use emphasizing that source information should be used to elaborate on personal ideas and validate propositions. Although the instructors overall viewed source use positively, they considered overreliance on source ideas as evidence of the inability to generate personal opinions. The textual analysis revealed that the number of source ideas in the essays ranged from 1 to 13 ($M_{per\ essay} = 5.7$, $SD = 2.9$), suggesting that some students depended on sources more than others. Essays that contained a high number of citations tended to receive lower content and organization scores ($r = -.49$). Excerpt (1) is from a lower scoring essay (1.5 out of 4) with ten citations where the source information functioned to generate ideas rather than elaborate on personal opinions.

(1) Overreliance on source-text ideas

In fact, the poorest 50% of the world's population has less than the money of the riches 85 people in the world (Fuentes, 2014). Economic inequality has many adverse impacts on many aspects of our lives (Fuentes, 2014). For example, it has been a threat to the social stability of countries, and the power of governmental authorities (Fuentes, 2014).

When rating this essay, the instructors commented on the lack of personal opinion and the overuse of source information:

I notice the overuse of citations one after another without connecting the ideas and including perspective [Dylan].

I didn't really find very many personal ideas in this essay which is a problem. This is an argumentative essay, but it reads more like expository [Cindy].

I cannot find any personal ideas in this essay. Source-texts were overused [Peter].

In addition to their negative reactions to source overuse, instructors were also critical of essays that lacked sufficient source information to support a personal opinion. Excerpt (2) is from a lower scoring essay (1.5 out of 4) where the student relied on personal ideas and used only one citation in the introduction. This excerpt is the opening sentence of the first body paragraph where the student is introducing the first reason in support of the thesis statement.

(2) Underuse of source-text ideas

> It is commonly argued that governments should not work to reduce economic inequalities as they find it is not their fault that wealth is unequally spread. Education, health, and safety are our government obligations. And a government who doesn't think primarily about the wellbeing of his people should be re-evaluated.

When commenting on this essay, the instructors remarked that the student should have provided an example to support the claims and failed to use information from sources as follows:

> *You are saying, it's commonly argued. So, if it's commonly argued, you should be able to cite an example* [Peter].
>
> *He's presenting this theory about education, and he doesn't have anything to back it up* [Amy]
>
> *Little evidence from sources* [Lily].

Although there is no existing research to identify the ideal number of source ideas that students should include in integrated-writing essays, Uludag et al., (2020) reported a positive relationship between CAEL integrated test band scores and the number of source ideas incorporated from two reading passages and one lecture into an opinion essay. However, direct comparisons with the current data are difficult because those students wrote shorter essays ($M_{per\ essay}$ = 255, SD = 52.0) and the mean number of source ideas was also lower ($M_{per\ essay}$ = 1.89, SD

= 1.49). Furthermore, Gebril and Plakans (2009) found that higher-level L2 writers used fewer source ideas, than lower-level writers on an integrated argumentative writing task, which included two reading passages. Thus, it appears that task type (e.g., argumentative vs. opinion essays), the number and nature of source materials and evaluation criteria (e.g., analytic vs. holistic) contribute to how raters determine the optimal number of source ideas.

In addition to orienting to the quantity of source information, reacting negatively to both over- and underuse, the instructors also considered the rhetorical purpose of source information in their evaluations. The textual analysis revealed that 68% of the source ideas in students' essays were used to elaborate on personal opinions, which the instructors viewed positively as illustrated previously. However, 24% of source information was used to introduce new content, which was associated with lower content and organization ratings ($r = -.38$). To illustrate, in excerpt (3), the student used the source information in place of a topic sentence. Rather than state an opinion and use the source information to support it, the student used the source information to introduce a new topic. As shown by their comments, the instructors viewed this negatively, with Kelly reporting *misuse of source information in para. 1* and Lily stating *missing topic sentence in para. 1.*

(3) Rhetorical purpose of introducing a new idea

Some people, especially those who are in high-income level, hold a view that the poor have to tackle the problems they created (Bolaria, & Wotherspoon, 2000). These erroneous opinions are kept even by people in developed countries today.

Although the rubric did not explicitly state reliance on source ideas or rhetorical purpose in the descriptors, it did include *exemplary selection and integration of source material*, which might have influenced the instructors' rating decisions. From the instructors' perspectives, elaboration

validates students' arguments, or propositions, thereby helping them establish a balance between personal opinions and source information. On the other hand, relying on the source information to introduce an idea or to repeat a previously mentioned idea might indicate a limited ability to synthesize source information.

The relatively low use of source information by these students to introduce new content contrasts with earlier studies which reported much higher use (Neumann et al., 2019; Wette, 2018). This discrepancy might be because the sample essays in this study come from a final examination which elicited an argumentative essay. Neumann et al. (2019) examined EAP students' source use functions in both cause-and-effect and argumentative essays, which were written at two different time intervals (i.e., midterm and final examinations). Wette's (2018) analysis, on the other hand, focused on post-EAP disciplinary texts written by L2 students with limited disciplinary knowledge. Thus, it is possible that these students relied on sources mostly to generate opinions, rather than supporting personal claims.

Incorporating Source-text Language in Students' Own Words

The second theme that emerged from the data was the instructors' belief that it was important for students to make linguistic revisions to source-text language. They paid attention to whether source ideas were paraphrased or summarized in students' own words or if the students had copied from the original text. The instructors were able to identify the copied strings either by going back to the original source to check or by inferring from the students' language use. For example, Amy remarked that the presence of little mistakes in a summary meant that *the student probably didn't copy from the source*. Although there were few instances of verbatim copying identified through the textual analysis ($M_{per\ essay} = 0.36$, $SD = 0.75$), four students copied

extensively on multiple occasions. Excerpt (4) illustrates an example of verbatim copying where the student kept the original sentence structure and copied several words from the source.

(4) Verbatim copying from sources

>*Original text:* Seven out of ten people live in countries where economic inequality has increased in the last 30 years (Fuentes-Nieva, 2014).

>*Student essay:* The recent survey indicates that <u>seven out of ten people live in countries</u> which <u>economic inequality has increased</u> during last 30 years (Fuentes-Nieva, 2014).

This essay, which included two additional instances of verbatim copying, received the lowest possible content and organization score (1 out of 4). In their comments about this essay, the instructors pointed to the occurrence of textual misappropriation as follows:

>*Some language of sources copied* [Cindy].

>*Verbatim copied from the source* [Dylan].

>*Some patchwriting in para. 2 and 3 [*Lily].

Overall, there was a negative relationship between the occurrence of verbatim copying and content and organization scores ($r = -.40$).

Although they regarded verbatim copying negatively, the instructors reacted positively to paraphrasing and summarizing. As compared to verbatim copying, both substantially modified paraphrases ($M_{per\ essay} = 2.42$, $SD = 1.89$) and summaries ($M_{per\ essay} = 2.43$, $SD = 2.39$) occurred more frequently in the essays. The rate of substantial revision found here is comparable to the amount reported by Uludag et al., (2019), who found that students used a mean of .67 ($SD = .42$) modified paraphrases on a CAEL integrated writing task. When evaluating the essays, instructors gave higher content and organization scores to texts with paraphrasing and summarizing, and the correlation between paraphrasing and ratings ($r = .43$) indicated a medium relationship.

Excerpt (5) illustrates paraphrasing in which the student changed both sentence structure (one efficient way…) and vocabulary (to mitigate financial inequality, available for no cost).

(5) Paraphrasing with substantial linguistic revisions

> *Original text:* "Free public health and education services are a strong weapon in the fight against economic inequality" (Seery, 2014).
>
> *Student essay:* Seery (2014) argues that one efficient way for governments to mitigate financial inequality is to invest in public services like education and health, available for no cost.

When reacting to this essay, the instructors provided positive comments such as *excellent integration of sources* [Peter] and *good treatment of source-texts* [Kelly].

Previous studies have reported a similar positive relationship between linguistic revisions of source-text language and L2 writers' integrated task performance (e.g., Plakans & Gebril, 2013; Shi, 2004). However, unlike raters in previous studies who reported difficulty assessing source-text language use (Gebril & Plakans, 2014), these instructors were able to evaluate how well the students had modified source language. Their ability to recognize paraphrasing is likely due to the fact that all students relied on the same six sources. In addition, these students did not have access to the sources when they were writing, which reduced the potential for verbatim copying. Nevertheless, it appears that a few students copied from the sources onto their note-taking template, which they were allowed to use during the exam.

Representing Source Content Accurately

Turning to the final theme that emerged from the data, the instructors expected students to avoid misrepresentation when incorporating source information into their essays. On average, 72 % of the source ideas were integrated accurately, which is comparable to the rates reported in

previous studies, which ranged from 70% to 80% (Uludag et al., 2019; Wette, 2018). Although they were less frequent (28%), instances of inaccurate source representation attracted the instructors' attention. Table 1 presents examples of accurate representation and misrepresentation. The student who misrepresented the content falsely attributed the reduction of unequal income distribution to increased jobs, which is not mentioned in the source, and equated virtual income with the improved job situation.

Table 1: Comparison of Original Text, Accurate Representation, and Misrepresentation

Original text	Essay with accurate representation	Essay with misrepresentation
Free public health and education services are a strong weapon in the fight against economic inequality. They mitigate the impact of skewed income distribution and redistribute by putting 'virtual income' into the pockets of the poorest women and men (Seery, 2014).	This will happen by providing the poor with a virtual income, which is provided as services, so they can benefit equally as rich people. Furthermore, the poor will not be spending all their incomes on health and education and will be willing to invest their money and enhance their revenues (Seery, 2014).	As Seery states in his paper we can reduce the impact of inequal income distribution by increasing job situation for men and women, which might be seen as virtual income for the poor (2014).

The instructors gave a low content and organization score (1.5 out of 4) to the essay with misrepresentation and commented that the student showed *a lack of understanding of the content*

being referenced [Lily] and that the source information was *misrepresented* [Amy], *misstated* [Kelly], or *distorted* [Peter].

The instructors may have oriented to accurate source ideas because the rubric included an explicit descriptor: *information from sources are accurately interpreted and acknowledged.* In their comments, they attributed source distortion or misrepresentation to poor reading comprehension skills. Prior studies have suggested that content inaccuracy occurs due to students' misinterpretation of the author's point (Neumann et al., 2019), which is related to reading comprehension. However, other researchers have suggested that it is caused by difficulties created when paraphrasing source information (Wette, 2017). Although this data does not provide direct evidence for the causes of source misrepresentation, it clearly shows that content accuracy is important to EAP instructors when assessing integrated writing.

Discussion and Conclusions

To summarize the main findings that emerged from the data, these EAP instructors oriented to two main sub-constructs in the model of source integration: introducing and restructuring source information. In terms of introducing source ideas, the instructors attended to students' source use purposes and awarded higher content and organization scores to the essays in which source ideas were introduced to elaborate on personal opinions. On the other hand, they appeared to place less emphasis on the components of referencing to the source texts and using linguistic devices. Although some of the instructors identified few minor errors in students' use of citations (e.g., the order of authors' names, missing date), and cohesive devices (e.g., wrong transition words), they tended to give precedence to rhetorical functions of source ideas in their scoring decisions. In the case of restructuring source information, both linguistic modification

and accurate representation of source ideas were considered as important aspects of source integration, as evidenced by instructor comments and correlation values.

However, these instructors did not report attending to aspects of responding to source information (i.e., rhetorical acts and following up with personal ideas). Analysis of students' essays for the rhetorical acts have shown that 42 % of the source ideas were followed up with an opinion that is not entirely related to the cited information. Excerpt (6) illustrates an example for this pattern, where the student avoids making an interpretive comment about the source information.

(6) Disengagement with source information

> Seery (2004) indicates that over 1.5 million lives were disappeared each year because of the income inequality. Another problem, lots of the poor cannot afford the basic necessary goods in their daily lives. Then those poor people working for several part-time jobs which are low-paying jobs.

In addition, there was a negative correlation between content and organization scores and following up with a diverging opinion ($r =. -28$). This finding indicates that responding to source information using rhetorical acts, such as elaborating on source ideas and drawing conclusions, might be considered as a characteristic and underlying construct of integrated writing tasks.

Although researchers have underscored a need for building a consistent theoretical framework to better interpret scores from integrated writing tasks (e.g., Knoch & Sitajalabhorn, 2013; Yu, 2013), a comprehensive model for assessing source integration has not been offered. The model introduced in this paper draws upon existing research and theories in L2 integrated writing. Our findings provide preliminary evidence for the interplay of the sub-constructs

discussed in the integrated writing model. Thus, we hope that it will benefit writing researchers with an interest in examining aspects of classroom-based integrated writing tasks.

Regarding the implications for teaching and evaluation of integrated writing in EAP contexts, in the light of our findings, we underscore the importance of refining rubric descriptors to better reflect instructor perspectives of source integration. Reaching a consensus as to which skills should be represented in the assessment criteria will not only help maximize the usefulness of test scores, but also reconcile the difference between assessment and classroom teaching practices. Integrated writing, as different from synthesis writing, response writing or literature review assignments, requires generation of personal ideas and supporting them with information from sources. This requirement should be emphasized pedagogically, raising students' awareness to sub-constructs elicited in integrated writing tasks. For this, the instructors could make use of modelling as an instructional strategy, incorporate actual student samples in their lesson plans, and provide students explicit feedback targeting their source integration patterns.

Although findings of this study provide preliminary evidence that EAP instructors orient to introducing and restructuring source information, there are some limitations that may limit its generalizability. The results are specific to the context where the sample essays and the rubric come from and may not be generalizable for all EAP models (e.g., intensive programs, bridging programs). In addition, the integrated essays used in this study were written on a final examination responding to an argumentative prompt, which required incorporation of personal claims. Also, following the exam protocol, the students were provided with the source materials before the examination so they had additional time to select and make notes of the source ideas they could possibly cite in their essays. Therefore, additional studies are needed to determine how varying writing conditions (e.g., (un)timed, with(out) access to sources) and different essay

types targeted in EAP programs (e.g., cause-and-effect) influence instructor perceptions of source integration. Finally, the EAP instructors were recruited among those who have experience in teaching integrated writing in a Canadian context. So, it is important to replicate these results in different contexts and with less experienced teachers to advance our knowledge and understanding of how integrated writing is contextualized and evaluated in ESL/EFL programs with different curricular and assessment objectives.

Connecting Study 1 to Study 2

Study 1 illustrated EAP instructors' perceptions about the assessment of integrated writing tasks using student samples and evaluation criteria from an EAP program. Based on the triangulation of data from instructor ratings, stimulated recall interviews, and textual analysis of students' essays, an analytic rubric for assessing integrated writing tasks was developed. Since the rubric was primarily aimed for summative evaluation of classroom-based integrated writing tasks, a decision was made to use an analytic scale with multiple criteria rather than a holistic one to encourage objective scoring. While rating holistically, raters are expected to decide on one general score, and their thought processes are not guided by a prescribed list of features, as opposed to analytic rating (Knoch, 2009). In other words, holistic scales encourage raters to use interpretation strategies to formulate a score (Barkaoui, 2010). Therefore, using analytic scales with clear and concise descriptors of writing components and refining the descriptors with potential users of the rubrics have been shown to minimize individual differences.

The constructs and performance levels in the new rubric were created based on the feedback from the EAP instructors who participated in Study 1. The qualitative data from the instructors indicated that the descriptors for source use (i.e., relevance and accuracy of selected source information, linguistic revision of source ideas and appropriate citation of the source information) need to be targeted as a separate category as different from the current EAP program rubric. Since students' content development is contingent upon their use of source information effectively in this particular EAP context, the descriptors for content were also separated from organization to allow consistent application of the analytic rubric. As for the performance levels, instructor notes highlighted few instances where students either showed no engagement with the topic or chose not to refer to source information in their essays. Thus, the

34

new analytic rubric defined a zero score for each category and included as a scoring option drawing on the instructors' qualitative feedback in Study 1.

The initial version of the rubric categories and performance descriptors were pilot tested by six EAP instructors, who participated in Study 1, and then refined based on their feedback (see Appendix D for the initial and final versions of the rubric). The rubric descriptors were reworked until a group consensus was reached among assessment professionals in the EAP program where the essays came from. The reformulation process through member checking with the EAP instructors and debriefing of the rubric with assessment professionals included the following methods: combining or removing duplicate descriptors, improving descriptor clarity by replacing technical language with commonly used terms, and revising descriptors to reflect features of an argumentative essay. The pilot-tested and revised rubric included four categories (i.e., content, organization, source use, and language use) scored from zero to four. Content, organization, and language use categories included generic and level-specific performance descriptors elicited by academic writing tasks. Rating criteria for source integration were incorporated in the descriptors for content and source use whereas organization and language use targeted general writing conventions. The purpose of Study 2 was to validate the rubric categories and level-specific performance descriptors using samples of student work from a final examination in an EAP program. Methodologies utilized included a Many-Facets Rasch Model, interview data from EAP instructors, and textual analysis of students' essays from three performance levels (low, average, high).

Chapter 3: Study 2:

Validating a Rubric for Assessing Integrated Writing in an EAP Context

For submission to Assessing Writing Journal

By Pakize Uludag and Kim McDonough

Introduction

There has been a resurgence of interest in the integrated assessment of language skills in recent years, particularly integrating reading into writing tasks (Grabe & Zhang, 2013). One of the advantages of integrated assessment tasks is that they have the potential to increase the validity and authenticity of the assessment of academic writing ability (Cumming, 2013; Yu, 2013). Because test-takers must display both receptive (reading) and productive (writing) skills, integrated writing tasks simulate higher education academic literacy tasks (Cumming et al., 2002; Weigle, 2004). However, there is still much debate in the field of assessment as to which distinct sub-skills are involved in integrated writing tasks.

In an attempt to further clarify the construct coverage of integrated writing tasks, researchers have determined that both reading and writing ability contribute to the scores, with writing ability being the stronger predictor (Trites & McGroarty, 2005; Watanabe, 2001). Furthermore, task characteristics, such as length and complexity of source materials, task type and the writing prompts have been shown to play a significant role in student performance (Ascención-Delaney, 2008; Ruiz-Funes, 1999). Although this line of research has shed light on the underlying factors that impact the assessment of integrated tasks, the validity of the rating criteria has not received much focus in L2 integrated writing studies. In fact, aside from few

studies focusing on standardized test rubrics, the design and evaluation of context-specific rubrics for assessing integrated tasks has not been a trend in the field of language assessment (Chan et al., 2015; Janssen et al., 2015). To ensure validity of integrated writing tasks, which are commonly used in EAP programs, it is important for the rubric descriptors to be validated using samples of actual student writing (Bruce & Hamp-Lyons, 2015; Inoue, 2009; Ohkubo, 2009). In addition, raters' perceptions about the assessment criteria serve the validation process and facilitate standardization of rubrics (Knoch et al., 2007). Therefore, as part of a broader project about the design and validation of an integrated writing rubric, this study aims to validate an evaluation rubric for integrated argumentative essays by considering EAP instructors' ratings and perceptions along with insights from a textual analysis of essays.

Literature Review

Integrated assessment, in its current mainstream use, requires application of at least two micro skills, such as reading and writing, to answer test questions (Lewkowicz, 1997). Integration of reading and/or listening with writing has provided an alternative to independent writing tasks, which do not require source use. It has been argued that for a writing task to be considered integrated, the written product needs to incorporate information from source materials, and source language needs to be linguistically modified when it is incorporated in the written product (Cumming, 2013). Although using integrated tasks positively contributes to the authenticity of writing tests, L2 writing researchers have noted assessment challenges with the use of such tasks, such as scoring validity and standardization of rater procedures (Knoch & Sitajalabhorn, 2013; Plakans & Gebril, 2013).

When students complete an integrated writing task, several factors determine which sub-skills are in use and whether those skills are measurable (Grabe, 2008). For example, studies

exploring the cognitive processes involved in integrated writing have found that L2 writers depend on source materials for idea development, language support, and textual organization, indicating that both reading and writing abilities are related to integrated writing performance (Leki & Carson, 1997; Watanabe, 2001). In addition, integrated writing tasks have been shown to elicit higher order sub-processes, namely organizing, selecting, and connecting abilities, suggesting that discourse synthesis is an important skill when writing from sources (Grabe & Zhang, 2013; Plakans & Gebril, 2017a). Therefore, it is important for evaluation rubrics to represent these underlying constructs as separate, measurable skills to better interpret scores from these tasks (Knoch & Sitajalabhorn, 2013).

An important challenge associated with the assessment of integrated writing tasks is local adoption of language proficiency scales originally created by applied linguists and language testing professionals (Chan et al., 2015; Janssen et al., 2015). Although these rubrics can isolate the key features of integrated writing ability, they may not serve the needs of the local users (Hudson, 2005; North, 2000). For example, writing scales which are usually anchored to specific language tests, such as TOEFL iBT, provide an accurate and valid description of integrated writing ability and distinguish different levels of proficiency. However, adopting such criteria without accounting for writing proficiency and curriculum objectives in a specific context will impact practicality and authenticity of the assessment. Therefore, during the last few decades, researchers have argued for taking an evidence-based approach for the design and validation of evaluation criteria that allow for valid interpretation and consistent application (McNamara, 1996; North, 2000). According to this approach, obtaining actual writing samples and illustrating how each scoring domain is related to underlying constructs is essential for gaining an understanding of students' ability levels (Crusan & Matsuda, 2018; Shin & Ewert, 2015).

One of the key elements in an evidence-based approach is the elicitation and interpretation of rater perceptions about evaluation criteria. Studies drawing on qualitative methods, such as interviews and think-aloud protocols, have shown that rater perceptions play an important role in the standardization of scoring rubrics (Knoch et al., 2007). For example, Cumming et al. (2001) found that while raters attended to rhetoric and content when scoring integrated writing tasks, they focused more on language use when scoring independent tasks. Gebril and Plakans (2014) also reported that raters oriented to content when assessing integrated writing by locating source information and evaluating the quality of source text information in students' essays. They also reported that raters faced challenges when rating essays in adjacent scoring bands and essays with a high incidence of quotations. Qualitative feedback on scoring rubrics is useful especially in educational contexts where the assessment criteria are mostly likely to be used by writing instructors. One example of this approach is the development and validation of the Common European Framework of Reference (CEFR) based on learning objectives and extensive qualitative feedback from the practicing teachers (Council of Europe 2001; North, 2000). Thus, the consideration of instructors' challenges and priorities when specifying achievement standards and validating evaluation rubrics helps bridge the gap between assessment and classroom teaching practices.

Despite the importance of the standardization of rubrics for teaching and learning activities, rubric validation research combining both quantitative (e.g., scores) and qualitative analyses (e.g., rater feedback) is notably sparse in educational settings. One of the few exemplars of detailed account of contextualized integrated writing rubrics has been contributed by Chan et al. (2015). Initially outlining the theoretical construct definition of "reading-into-writing ability" for developing integrated tasks, they designed and pilot-tested analytic rubrics for use with

Trinity College London reading-into-writing tasks. Using a mixed-methods design, the researchers applied a multi-facet Rasch model for the analysis of rater reliability. As part of ongoing validation, they conducted automated textual analysis and elicited rater perceptions through questionnaires and group interviews to refine the rubrics based on test specifications. The outcome of their process yielded rating rubrics at four proficiency levels (i.e., CEFR A2 to C1). Ewert and Shin's (2015) study also documented the process of developing an empirically derived binary choice, boundary defined (EBB) scoring tool for placement into university writing courses. They involved the instructors in the program to develop the EBB scale which focused on "reading-to-write tasks" and provided details of teachers' challenges in the development and validation process, which included articulating the writing prompts and internalizing curriculum objectives. The results supported the use of an EBB rating scale in the local context of the specific placement task. Describing the overall processes of rubric creation and validation for different purposes, these two studies have offered important perspectives into the contextualization of evaluation criteria when assessing integrated writing tasks.

Situated within this line of research, the current study aims to validate an integrated writing rubric for assessing classroom-based integrated essays in an EAP program. As the first step in this process, an initial exploratory study (Study 1) examined EAP instructors' perceptions about the assessment of integrated writing tasks using student samples from an EAP program. Using a theoretically motivated process model for source integration (see Figure 2 for the model), while triangulating data sources from instructor ratings, stimulated recall interviews, and textual analysis of students' essays, they developed an analytic rubric for assessing integrated writing. The initial version of the rubric categories and performance descriptors was pilot tested by six EAP instructors and then refined based on their feedback. The credibility of the rubric was

established by reworking the descriptors until a group consensus was reached among assessment professionals in the EAP program where the essays came from. The reformulation process included the following methods: combining or removing duplicate descriptors, improving descriptor clarity by replacing technical language with commonly used terms, and revising descriptors to reflect features of an argumentative essay.

The pilot-tested and revised rubric included four categories (i.e., content, organization, source use, and language use) scored from zero to four. Content, organization, and language use categories included generic and level-specific performance descriptors elicited by academic writing tasks. Rating criteria for source integration were incorporated in the descriptors for content and source use whereas organization and language use targeted general writing conventions. Within the source use category, the descriptors referred to the following aspects: relevance and accuracy of selected source information, linguistic revision of source ideas (i.e., paraphrasing/summarizing with substantial modification, verbatim copying from sources, unacknowledged source use, overreliance on direct quotations), and appropriate citation of the source information. One of the descriptors in the content referred to supporting personal claims with information from sources.

As the next step in rubric evaluation, the current study aims to validate the rubric categories and level-specific performance descriptors using samples of student work from a final examination in an EAP program. Methodologies utilised included a Many-Facets Rasch Model (MFRM), interview data from EAP instructors, and textual analysis of students' essays from three performance levels (low, average, high). Over the past two decades, the MFRM has been used to evaluate the quality of rating criteria, particularly in writing and speaking assessment contexts (e.g., Janssen et al., 2015; Youn, 2015). Previous research has provided evidence of

rater severity in their evaluation of L2 writing performance (Eckes, 2008; Kondo-Brown, 2002). In case of integrated writing assessment, since raters attend to both source materials and student writing (Ewert & Shin, 2015) and focus on different aspects of student writing (Gebril & Plakans, 2014), accounting for rater severity is important for score reliability. Therefore, the MFRM was used as a statistical means to engage multiple perspectives from EAP instructors.

Although data from the MFRM can be used to construct measures of the integrated writing ability, qualitative approaches, such as rater interviews, has been proven to help explain potential sources of variance in raters' application of the rubric criteria in terms of determining its usefulness (Barkaoui, 2010; Han & Huang, 2017). Finally, additional empirical support for the rubric's ability to differentiate among performance levels can be obtained through textual analysis of students' essays. Researchers have shown that linguistic features such as textual length, type-token ratio, instances of verbs and academic words predict human judgments of L2 integrated writing quality (Gebril & Plakans, 2016; Guo et al., 2013). The current study addresses the following research question: How effectively does an integrated writing rubric differentiate among EAP students' integrated writing performance?

Method

Integrated Writing Essays

The integrated essays ($N = 48$) were sampled from the Concordia Written English Academic Texts (CWEAT) corpus (McDonough et al., 2018), which consists of source-based argumentative and cause-and-effect essays written by EAP students enrolled in an integrated writing course at an English-medium Canadian university that had been rated using a holistic rubric during corpus construction. The sampled argumentative essays had been written in

42

response to the same prompt (i.e., the role of governments in reducing economic inequality) and source texts and received a range of holistic scores. These essays were written by students with a mean age of 24.4 years ($SD = 3.81$) who were studying in the faculties of Business (26), Arts and Science (13) and Engineering and Computer Science (9). The most frequent L1s were Chinese (18), Arabic (9), French (8) and Spanish (7). Additional L1s included Russian (2), Turkish (1), Ukrainian (1), Persian (1), and Bengali (1).

The students wrote the argumentative essays as a final exam in their EAP course, which was the second in a two-course sequence. The instructional objectives were to improve the students' general academic language skills, including reading and writing strategies, and targeted high-level academic tasks, such as critical reading, synthesizing, and integrated writing assignments. Following the assessment procedures designed by the EAP program, students were assigned six readings relevant to the topic prior to the exam. They were allowed to complete one note sheet per source to take notes on the main idea and key supporting details using a note-taking template. Student were allowed to record the source citation, key terms or new lexical items, and important quotations on this template. At the examination, students were provided with two writing prompts, and they chose which one to write about using information from the sources. Although they did not have access to the source materials, they were allowed to use their notes and a paper-based English dictionary during the exam.

EAP Instructors

The EAP instructors (7 women, 3 men) have taught integrated writing courses in a variety of EAP models across four different English-medium universities in Canada. Three instructors were familiar with the rubric from their involvement in the initial stage of the project (Study 1) and two instructors were from the EAP program where the sample essays came from.

All instructors had advanced degrees in fields such as TESL, Applied Linguistics, and Communication Studies. They identified themselves as Canadian (6); Iranian (2), Korean (1), and American (1). Except two instructors who reported Persian as their L1s, all instructors were L1 English speakers. They ranged in age from 30 to 53 (M = 43.8, SD= 7.0) and had a mean of 17.5 years experience teaching English (SD = 5.3). Other than teaching, they had also participated in administrative, curriculum development, and assessment tasks in their EAP programs. Five instructors also reported having worked as a professional examiner or rater for IELTS and CELPIP exams.

Rating Procedure

The first researcher scheduled individual data collection sessions with the instructors using an online meeting tool (Zoom). After signing the consent form and completing a background questionnaire, the instructors participated in a rubric familiarization session which focused on achieving a common understanding of the construct being measured and the rubric descriptors for the sub-constructs at each score level. The instructors were initially introduced to rubric categories, the task instructions, and the source texts. To minimize the influence of biases, they were presented with detailed rating guidelines and three essays with preset scores to help them internalize the rubric. The instructors then worked independently to evaluate two sample essays, with high and low holistic scores, making notes on the rubric descriptors, identifying relevant parts in source texts, and assigning a score for each rubric category. After evaluating the first essay, the instructors showed their annotated rubrics and the essay to the researcher via the share screen function and explained their rating process and scoring decisions. They repeated the same process after scoring the second essay. Disagreements regarding the instructors' approach to specific rubric descriptors were resolved.

Following the online training, the instructors had two weeks to score 48 essays independently using the rubric and highlighting the descriptors which corresponded to their evaluation of a construct. After finalizing the rating procedure, they participated in a 30-minute semi-structured interview (Appendix C), which elicited their perspective of and experiences with using the rubric (Mendoza & Knoch, 2018). The instructors were encouraged to reflect on their personal EAP teaching and rating experiences to support their opinions of the rubric during the interviews.

Analysis

To examine the functioning of rubric categories and rater effects (i.e., severity), a Many-Facets Rasch Model (MFRM) was implemented using FACETS software (version 3.71.4), which is an extension of the basic Rasch model (Linacre, 2014). The model specifications included three facets, which can potentially influence the reliability of the integrated writing rubric: raters (EAP instructors), test-takers (integrated writing essays) and rubric categories (content, organization, source use and language use). The model was tested for the assumptions of unidimensionality (i.e., only one underlying ability is measured at a time) and local independence of responses by examining the mean-square statistics for each facet (Linacre, 2010). Infit and outfit mean-square values greater than 0.6 and less than 1.5 were considered acceptable and to fit the model (Lunz et al., 1990). The appropriate use of the rubric categories was determined by checking the rating category statistics which provides estimates for average measure values, threshold statistics, and the outfit mean square of each level (Bond & Fox, 2015; Linacre, 2004).

To provide additional evidence for the rubric's ability to differentiate among performance levels, the argumentative essays were analyzed for five linguistic features (see Table 2), which

were associated with higher quality of integrated writing in previous research when they were used more extensively (e.g., Gebril & Plakans, 2016; Guo et al., 2013).

Table 2: Linguistic Features

Category	Measures	Description
Fluency	Response length	Total number of words per essay
Lexical diversity	Type-token ratio	The number of unique words divided by the number of tokens of these words
Syntactic complexity	Instances of verbs	The proportion of verbs, including verbs in base form, past participle verbs and 3rd person singular verbs to the total number of words per essay
Cohesion	Semantic similarity (LSA sentence to sentence)	Conceptual similarity between adjacent sentences
Lexical complexity	Academic words (AWL)	The proportion of words in the Academic Word list (Coxhead, 2000) to the total number of words per essay

To analyze the linguistic features, two online automated software tools were used. For response length, type-token ratio, instances of verbs, and semantic similarity, typed versions of the students' hand-written essays were submitted to Coh-Metrix (http://cohmetrix.com). The

measure of AWL was obtained using Compleat Lexical Tutor (https://www.lextutor.ca/cgi-bin/range/texts/index.pl). Textual length, which showed variation across the essays, was controlled for using the proportion scores in the analysis.

Finally, the EAP instructors' interviews were transcribed, and the transcriptions were read recursively to detect initial codes during a preliminary exploration stage. Classification of the interview segments in the light of initial codes and using open and axial coding (Corbin & Strauss, 2015) led to the identification of EAP instructors' perceptions about a) the clarity of the rubric descriptors, b) potential overlap between the categories, and c) differentiating among performance levels.

Results

The MFRM Analysis

The Wright Map (Figure 3) shows a summary of the three facets (test-taker, raters, and rubric categories) on the same logit scale. As seen in the first column, the measurement ruler spans from −1 to 4 logits, indicating that the EAP instructors separated the essays across five logits. The second column provides an estimate of the spread of test-takers (i.e., each star represents one test-taker) along a continuum, which resembles a normal distribution. In the third column, it can be observed that the instructors 1, 3, 9 and 8 were more severe, with logit positions above zero, whereas 10, 4 and 2 were more lenient. The instructors 5, 6, and 7 were positioned closer the origin, showing intermediate severity. In the fourth column, we can observe that source use was the most difficult category and organization the easiest. The last column displays the performance levels ranging from 0 to 4.

Figure 3: Wright Map

```
+----------------------------------------------------------------------+
|Measr|+test-takers|-raters |-categories                          |Scale|
|-----+------------+--------+--------------------------------------+-----|
|  4  +            +        +                                      + (4) |
|     |            |        |                                      |     |
|     |            |        |                                      | --- |
|     |     *      |        |                                      |     |
|     |     *      |        |                                      |     |
|     |     *      |        |                                      |     |
|  3  +            +        +                                      +     |
|     |     *      |        |                                      |     |
|     |     *      |        |                                      |     |
|     |    **      |        |                                      |     |
|     |            |        |                                      |     |
|     |    **      |        |                                      |  3  |
|  2  +            +        +                                      +     |
|     |   ****     |        |                                      |     |
|     |   ****     |        |                                      |     |
|     |     *      |        |                                      |     |
|     |   ***      |        |                                      |     |
|     |  *****     |        |                                      |     |
|     |    **      |        |                                      |     |
|     |     *      |        |                                      |     |
|     |   ****     |        |                                      |     |
|  1  +  *******   +        +                                      +     |
|     |     *      |        |                                      | --- |
|     |     *      |   1    |                                      |     |
|     |     *      |        |                                      |     |
|     |    **      |   3  9 | Source                               |     |
|     |            |   8    |                                      |     |
|     |     *      |   5    |                                      |     |
| *   0  * *       |  * 6   |                                      *  *  |
|     |            |   7    | Content        Language              |     |
|     |            |        | Organization                         |     |
|     |            |  10    |                                      |     |
|     |            |   4    |                                      |  2  |
|     |            |        |                                      |     |
|     |            |   2    |                                      |     |
|     |     *      |        |                                      |     |
| -1  +            +        +                                      + (0) |
|-----+------------+--------+--------------------------------------+-----|
|Measr| * = 1      |-raters |-categories                          |Scale|
+----------------------------------------------------------------------+
```
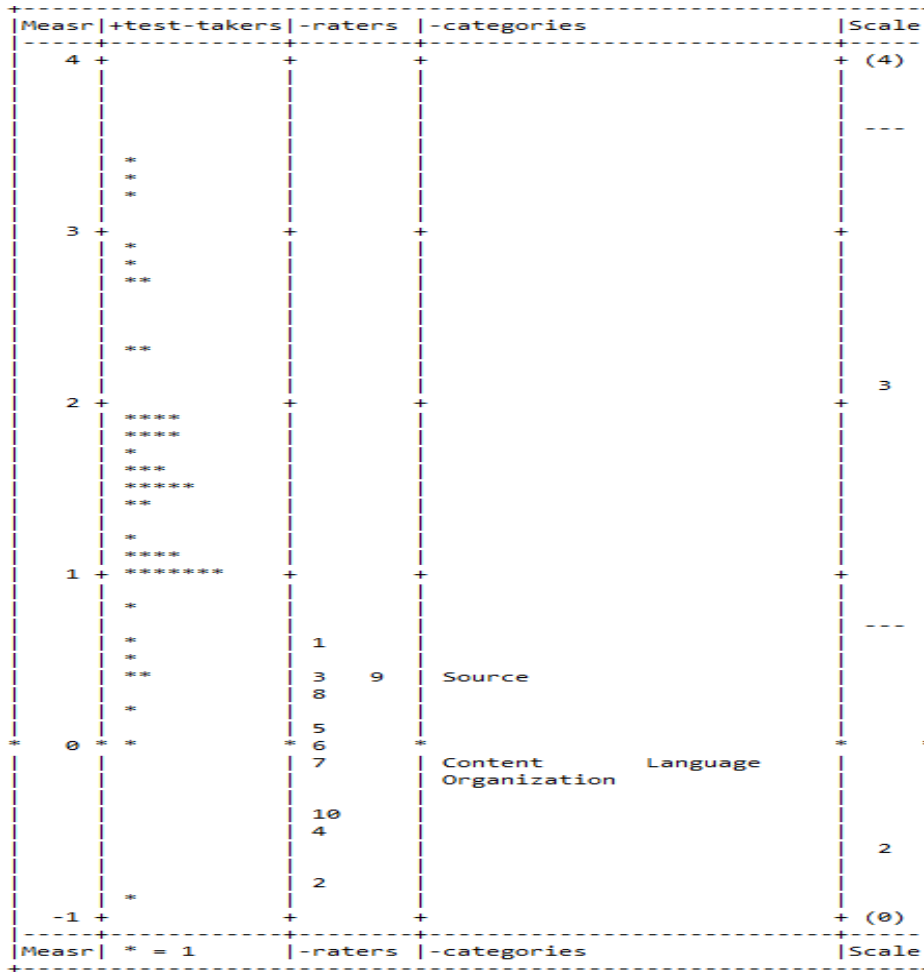
The mean-square values for test-takers ranged between 0.51 and 1.5, with all test-takers showing adequate outfit and infit levels (Lunz et al., 1990). The test-takers were sorted into over four distinct performance levels (Strata 4.23) with a reliability of .93, indicating that the integrated writing task reliably separated the 48 students' varying abilities.

The raters displayed an acceptable range of severity, as shown in Table 3, in which the raters were ranked according to their severity. The range between the most severe (Rater 1) and the most lenient rater (Rater 2) is from -0.79 to 0.63 (about 1.42 logit spread), which conforms to the expectations of the Rasch model. In addition, despite variation in raters' severity, all infit and outfit values were within the acceptable range of 0.6 and 1.5, ensuring internal consistency and predictable behavior of the raters.

Table 3: Rater Measurement Report

Raters	Severity Logit	Model S.E	Infit MnSq	Outfit MnSq
1	0.63	0.12	0.72	0.71
3	0.39	0.12	1.28	1.27
9	0.39	0.12	0.97	0.98
8	0.29	0.12	0.90	0.87
5	0.11	0.12	1.23	1.23
6	0.00	0.12	0.71	0.70
7	-0.12	0.12	0.91	0.91
10	-0.41	0.12	0.89	0.89
4	-0.48	0.12	1.04	1.03
2	-0.79	0.12	1.32	1.31
Mean	0.00	0.12	0.10	0.99
SD	0.43	0.00	0.21	0.21

Notes: Reliability = 0.93; Separation = 3.49; Fixed chi-square = 123.3 (df = 9; p <.0001)

Table 4 presents the measurement report for the rubric categories. Among the four categories, source use was the most difficult with 0.40 logits while organization was the easiest with –0.20 logits. The difficulty measures demonstrated a 0.60 logit spread between source use and organization. Differing levels of difficulty across four categories suggest that the raters were able to differentiate them consistently. As for the quality of the mean-square statistics, infit and outfit values ranged between 0.88 and 1.08, indicating that the criteria functioned as intended and measured one unidimensional construct (Bond & Fox, 2015).

Table 4: Rubric Category Measurement Report

Categories	Difficulty Logit	Model S.E	Infit MnSq	Outfit MnSq
Source Use	0.40	0.80	1.02	1.00
Content	-0.12	0.80	0.88	0.89
Language Use	-0.10	0.80	1.08	1.06
Organization	-0.20	0.80	1.00	1.00
Mean	0.24	0.80	1.00	0.99
SD	0.44	0.00	0.70	0.60

Notes: Reliability = 0.91; Separation = 3.01; Fixed chi-square = 123.3 (*df* = 3; *p* <.0001)

The rubric category statistics are summarized in Table 5. The raters rarely used 0 in their ratings, which is not surprising given that the essays were written as part of a final examination. The most frequently assigned score was 3 (55%), followed by 2 (31%) and 4 (13%). As shown in the third column, the average measures advance monotonically, with the scores increasing as the score level goes up. This indicates that more proficient test-takers were awarded higher scores while less-proficient test-takers were given lower scores (Bond & Fox, 2007). In the fourth column, the threshold measures (i.e., the lowest possible test-taker ability measure at a score level) also show a steady increase across the score levels. The outfit values were 1.0, which means that the categories are functioning appropriately, and the scores were assigned in a consistent manner (Barkaoui, 2013; Linacre, 2004).

Table 5: Rubric Category Statistics

Score	Category Total (%)	Quality Control Avge Meas	Outfit MnSq	Threshold Measures

0	6 (0%)	-1.43	0.4	None
1	30 (2%)	0.25	1.0	-2.48
2	593 (31%)	0.96	1.0	-1.79
3	1047 (55%)	1.64	1.0	0.70
4	244 (13%)	2.40	1.0	3.53

The Instructors' Perceptions of the Rubric

To expand on the statistical results from the Rasch analysis and identify the instructors' challenges with distinguishing between the rubric elements, a follow-up interview was performed with the EAP instructors ($N = 10$). The interview protocol contained 11 questions concerning the instructors' experiences with the rubric criteria, descriptors, and score levels. The themes emerged from the data were related to the clarity of the rubric descriptors, potential overlap between the categories, and differentiating among performance levels. The instructors generally commented favorably on the rubric in terms of its usefulness in integrated writing teaching and assessment. They remarked that the descriptors were informative and provided them sufficient guidance in their scoring decisions:

> *I think the rubric is very good. I think it is comprehensive and gives a holistic picture of the student's writing. It is easy for "language" teachers to fixate on language and grammar issues, which are important, however, they are not the only component in good academic writing* [Rater 5].

> *It is a well-defined rubric that captures the required elements of an academic essay. It was easy working with the rubric* [Rater 4].

I think the categories were well chosen and separated with well-chosen descriptors. I had

no problem categorizing students into different levels [Rater 3].

Three instructors who were familiar with the existing EAP rubric from their involvement in the

rubric development process were asked to compare their experience with using the current rubric

and the existing EAP rubric. Their comments were as follows:

This rubric covered all the segments that were required to successfully write the   essay

and I was able to highlight different components of different categories to achieve

an appropriate grade. It is definitely much better than the rubric used in the first study

[Rater 10].

The major difference is the source use category and the separation between content and

organization. These differences made the marking easier [Rater 2]

The rubric used in the first study was too simplified and did not include enough variables

nor components in order to grade appropriately. The current rubric is better and/or

closer to an accurate grading system [Rater 6].

When asked specifically about the clarity of the rubric descriptors, three instructors

expressed concerns about this descriptor under content: "Student's stance is clear in the essay".

One of the instructors who was teaching at the EAP program where the essays come from made

this comment:

Stance is not a term I use with students for this essay. While some students showed a

stance, the degree to which they analyzed the arguments was not assessed based on the

wording in this point [Rater 2].

In response to the above-mentioned feedback, this descriptor was reformulated as "Student's

position is clear in the essay" to make it more accessible to the instructors in the program.

In addition, two instructors expressed disagreement with the classification of "APA citations are accurate" under the category of source use saying *Punctuation related to APA citation seem to overlap with punctuation* [Rater 9], and *Accuracy of APA citations might be assessed as part of mechanics* [Rater 1]. After careful consideration, we decided to remove APA citation from the rubric and reworded the descriptor as "Source information is cited properly" and noted that this descriptor should be clarified as part of rater training. Finally, all instructors agreed that the rubric is suitable to use in the classroom to help EAP students improve their integrated writing skills. Two instructors suggested that *The rubric should be deconstructed with examples for classroom use* [Rater 3] and *Certain elements could be quantified for students* [Rater 7]. Overall, the EAP instructors' experiences with the rubric supported the assumption that it can be used as a valid instrument upon careful rater training.

Textual Analysis

Having shown that the rubric categories functioned well, and the instructors viewed the rubric favorably, additional insight into the reliability of the assessment was obtained through textual analysis. To reflect on whether the rubric could differentiate between high and low performance groups, the ratings from 10 instructors were averaged for each essay and the mean total score (11.16 out of 16) was used to classify students into three groups. Students who scored more than +1 standard deviation above the mean were classified as high scorers ($n =14$) while students who scored less than -1 standard deviation below the mean were classified as low scorers ($n =13$). The average level included those who scored closer to the mean ($n = 21$). As shown in Table 6, there was as steady increase in lexical diversity (length and type-token ratio) and syntactic complexity (verbs) values as the performance level increased. Despite some

inconsistency, cohesion (semantic similarity) and lexical complexity (AWL) values also

increased in upper levels.

Table 6: Comparison of Textual Features at each Performance Level

Measures	Performance Levels	Mean	SD
Length (number of words per text)	High	663	81.4
	Average	589	87.0
	Low	571	96.8
Type-token ratio	High	0.60	0.06
	Average	0.57	0.18
	Low	0.53	0.15
Instances of verbs	High	124.0	13.6
	Average	121.2	16.6
	Low	112.5	27.7
Semantic similarity (LSA sentence to sentence)	High	0.23	0.09
	Average	0.24	0.04
	Low	0.21	0.06
AWL	High	4.3	1.3
	Average	4.3	2.2
	Low	4.0	2.6

In sum, the textual analysis comparisons provide additional evidence that the rubric successfully

discriminates between student performance levels.

Discussion and Conclusions

The purpose of this study was to validate an integrated writing rubric using actual student samples from a classroom-based final examination by incorporating EAP instructors' feedback into the validation process. Results from the MFRM and the textual analysis demonstrated the effectiveness of the rubric for assessing EAP integrated writing tasks. In addition, the instructor comments during the follow-up interviews shed light on the statistical results and constituted evidence of the usefulness of the rubric.

One goal was to provide an example of a mixed-method approach to standardization of evaluation rubrics in educational settings. EAP instructors played an active role in the process of rubric development and validation, which helped us contextualize the criteria to reflect curriculum objectives and assessment practices in an EAP context (Crusan & Matsuda, 2018). First, employing a MFRM allowed us to engage EAP instructors' multiple perspectives on the unique characteristics of each individual essay (Hamp-Lyons 2007; Lumley, 2005; McNamara, 1996). In contrast to classical test theory, in which rater disagreement is considered undesirable, the MFRM determines the quality of a scoring rubric by using rater disagreements to estimate rater severity and adjusting test-taker measures accordingly after calculating measurement error (Bond & Fox 2015; Eckes, 2011). Previous L2 writing studies have found that raters' linguistic backgrounds and previous rating experiences may influence their scoring decisions (e.g., Kondo-Brown, 2002; Lumley & McNamara, 1995). Although training sessions help raters to develop a shared understanding of the construct (Turley & Gallagher, 2008; Wilson, 2007), students' individual performances are seen as too complex to be perfectly matched to a set of descriptors. These instructors represented the diversity of the EAP programs in Canadian context and their L1 backgrounds, education levels and professional experience varied. Thus, conducting a MFRM

analysis, we could adjust for differences in instructor severity when they were constructing subjective judgements of students' writing performance.

Second, eliciting the instructors' perceptions as part of the validation process contributed useful information for reformulation of two rubric descriptors. The feedback from the instructors were generally positive although additional training and benchmark performances are needed before using the rubric in an EAP setting. As such, adjustments to rubric categories may be needed if it is to be used for the evaluation of different essay types (e.g., descriptive, cause-and-effect) with distinct linguistic, structural and discourse features. Third, textual analysis of lexical, syntactic, and cohesive features of the essays also demonstrated the rubric's ability to differentiate across performance levels. Confirming the findings of previous research (Gebril & Plakans, 2016; Guo et al., 2013) that reported a relationship between linguistic features and L2 integrated test scores, our results have shown that higher-scoring students wrote longer texts with a higher rate of type-token ratio. In addition, as the performance level went up, semantic similarity, instances of verb use, and AWL use showed an overall increase. Whereas these findings suggest the selected features are meaningfully related to performance levels identified by the rubric, they should be interpreted with caution given the small sample size. Nevertheless, the textual analysis provided supplemental evidence for the scoring validity. Including textual analysis as part of an evidence-based approach to rubric design and validation may be helpful, especially in EAP contexts where the focus is on academic reading, vocabulary and writing skills (Hyland & Hamp-Lyons, 2002).

In addition to being relatively small in scale, this study has other limitations which should be considered when interpreting the findings. One of the constraints was that the rubric was tested on an argumentative essay only with samples from a final examination. Thus, the rubric

validity should be established on different essay types, such as cause-and-effect, and on a full test taker population before adopting it for more widespread use in an EAP program. Notably, the rubric validation approach outlined in this study was part of an ongoing research project, not part of an EAP program's evaluation efforts. Therefore, the effectiveness of this approach needs to be confirmed by the stakeholders and practitioners in localized contexts drawing on multiple trials and extensive rater training with practice materials. Finally, one important future direction is to perform a more substantial textual analysis, which includes fine grained indices of the same construct (e.g., lexical diversity, cohesion) in terms of differentiating the distinctive features at different proficiency levels.

In conclusion, this study has provided an example of the process by which an integrated writing rubric can be validated contextually and situationally for use in an EAP program. It would be interesting to examine how the procedures reported in this study might inform rubric development in university-level EAP programs following different models (e.g., intensive programs, bridging programs). With the growing interest in the assessment of L2 integrated writing, further research into the development and validation of evidence-based scoring rubrics for assessing classroom-based tasks is needed.

Connection of Studies 1 and 2 to Study 3

The first two studies in this dissertation adopted an empirically-based approach to the development and validation of an integrated writing rubric. Both studies focused on EAP instructors' scoring decisions and perceptions of the assessment of classroom-based integrated writing essays. The results helped confirm the validity of the analytic rubric in terms of capturing the underlying constructs within integrated writing. Study 3 takes this further by exploring EAP students' conceptualizations of integrated writing assessment in an EAP writing course and their use of the analytic rubric for self-assessment. Adopting an inductive case study approach, this exploratory research aims to contribute to the validity of the rubric from students' perspectives while reflecting on their challenges and concerns with the task representation, task conditions, and teacher feedback. In this sense, this study represents pedagogical extension of the score validity issues targeted in the first two studies in this dissertation.

Chapter 4: Study 3

Investigating EAP students' perceptions of integrated writing assessment

For submission to Language Education & Assessment Journal

By Pakize Uludag

Introduction

Integrated writing tasks have become a standard component of large-scale English proficiency tests, such as the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT), and the Canadian Academic English Language (CAEL) test to ensure the validity and authenticity of writing assessment. Reflecting the strong emphasis on academic writing in university courses, English for Academic Purposes (EAP) programs have also incorporated integrated writing in their curriculum and assessment (Plakans, 2009, 2010; Plakans & Gebril, 2012) to help students transition to disciplinary writing. The use of integrated writing tasks for EAP pedagogy and standardized writing tests has not only improved the measures of writing ability, but also contributed to students' engagement with academic tasks that are fundamental for disciplinary contexts in higher education (Cumming et al., 2006; Lewkowicz, 1997; Weigle, 2004).

During the last few decades, second language (L2) writing researchers have explored the link between observed scores and integrated writing quality to address the validity issues associated with the use of integrated writing tasks. Few studies have focused on rater perceptions to examine the scoring validity and standardization of rater procedures (Cumming et al., 2002; Gebril & Plakans, 2014). In addition, conducting textual analysis, researchers have demonstrated that both reading and writing abilities are pertinent to integrated writing test performance (Leki

& Carson, 1997; Watanabe, 2001). Although these studies have shed light on the factors that impact the assessment of integrated tasks, it remains unexplored how students conceptualize classroom-based integrated writing assessment and whether they can use integrated writing assessment criteria. Therefore, to establish clearer assessment standards and improve writing instruction in an EAP context, this case study, as part of a larger study about the design and validation of an integrated writing rubric, explored students' insights into classroom-based integrated writing assessment and their use of an analytic rubric for self-assessment.

Literature Review

One of the concerns in the field of language testing is the development of tasks that represent the characteristics of the "target language use (TLU) domain" (Bachman & Palmer, 1996), which refers to the "situation or context in which the test taker will be using the language outside of the test itself" (p. 18). Researchers working within the assessment for learning perspective have increasingly recognized the importance of incorporating stakeholder perspectives into the development of tasks that reflect real-life domains (e.g., Rosenfeld et al. 2001; Malone & Montee, 2014). Studies focusing on the use of formative and ongoing assessments have reported that incorporating student feedback into revision of assessment tasks and test rubrics positively impacts teaching. In addition, articulation of students' thoughts and experiences with assessment have been shown to enhance student motivation and promote self-regulated learning (Benson, 2007; Butler, 2016; Rea-Dickins, 2006).

Although studies from the perspectives of EAP students have been limited, researchers have explored how teachers and assessment professionals understand assessment constructs as compared to student understanding. For example, Sato and Ikeda (2015) explored Japanese and Korean college students' interpretations of the skills targeted in a high-stakes college entrance

60

examination that used a multiple-choice format for assessing receptive and productive language skills. Students did not have a clear understanding of test items that measured the ability to read between the lines and writing ability. Similar findings have been reported for the use of alternative assessment methods in different EFL contexts. For example, Vlanti (2012) revealed discrepancies between the interpretations of the Greek junior high-school students and their instructors in terms of the use of self- and peer-assessment for grading in language classrooms. Thus, researchers have emphasized student involvement in assessment decisions to obtain positive washback on language teaching and learning.

More recently, studies have incorporated student perceptions into language assessment literacy (LAL) discussions (Butler at al., 2021). This strand of research has demonstrated that students are major stakeholders (Erickson & Gustafsson, 2005) who can reflect on and influence the assessment processes (Butler, 2019). For example, Butler et al. (2021) investigated fourth- and sixth-grade Chinese students' knowledge and understanding of assessment purposes, skills, and principles depending on existing LAL models. Administering a mock English test and eliciting student perceptions through semi-structured interviews, they found that students had substantial prior knowledge and experience with assessment. In addition, despite being young, they were capable of articulating their wants and needs (e.g., communicative-based assessment) and identifying construct-irrelevant factors, such as anxiety, which might affect their test performance.

Situated within this line of research, a number of researchers have elicited student perceptions to confirm the validity of large-scale English language tests. For example, Winke et al. (2018) interviewed both L1 and L2 English speaking children (ages 7 – 9) to investigate the cognitive validity of the Young Learners Tests of English administered by Cambridge Michigan

Language Assessments. They found that incorrect responses by L1 children resulted from age-related cognitive constraints and lack of assessment literacy, which were both considered to be construct-irrelevant variance. Similarly, Cheng and DeLuca (2011) incorporated student perspectives into validation of four different high-stakes language tests used for entrance and certification at an English-medium university in Asia. The students' testing experiences concerned both construct representation (e.g., testing consequences) and construct-irrelevant variance (e.g., test format and administration). Both studies have confirmed that insights from test-takers contribute to test validity and help promote a more balanced and comprehensive understanding of LAL.

Focusing more specifically on writing tasks, researchers have compared the TOELF iBT writing section and target language use (TLU) tasks (i.e., academic writing tasks in university courses) drawing on interview data from teachers and students. For example, Malone and Montee (2014) used stimulated recall interviews to elicit university students' perceptions of TOEFL iBT items, including the integrated writing task. They found that majority of the students were able to identify the similarities and differences between TOEFL integrated writing task and the type of tasks they previously encountered in university courses. In a follow-up study, which focused specifically on TOEFL iBT writing tasks, Llosa and Malone (2017) obtained student and instructor perceptions, through which they revealed a convergence between the abilities assessed in TOEFL writing tasks and underlying TLU tasks, such as summary and synthesis writing. Taken together, assessment validation studies have confirmed the importance of obtaining student perspectives to enhance validity arguments and understand the outcomes of large-scale English language tests.

Despite the growing interest in classroom-based assessment, student perceptions of integrated writing test characteristics and scoring criteria have rarely been explored in previous research. Studies of L2 integrated writing have largely examined the individual and contextual factors that might play a role in students' writing processes. For example, Plakans (2009) used think-aloud protocols and interviews to examine L2 students' integrated writing processes on an ESL placement test. The results indicated that students' composing processes were influenced by their prior experiences and background knowledge of writing along with their topic familiarity. In a qualitative case study, Zhu (2005) investigated a graduate Chinese student's experiences with integrated writing tasks and found that the student's task representation was shaped by task purposes. In addition, the opinions expressed in the source texts facilitated the student's idea development and aided the direction of the responses. The function of source texts was also investigated by Plakans and Gebril (2013) using a mixed method design, which included interviews and questionnaires. The study found that students across different proficiency levels used source texts for similar purposes such as developing personal opinions about the topic and for language support. Although not focused on assessment constructs and uses, L2 students' perceptions and practices were also elicited in numerous case studies to explore source use variables, such as patchwriting and verbatim copying (e.g., Harwood & Petri, 2012; Hirvela & Du, 2013; Li & Casanave, 2004). Although these studies have shed light on students' textual borrowing and citation behavior while working on integrated writing tasks, they do not shed light on how students conceptualize integrated writing assessment or how they benefit from assessment tools, such as analytic rubrics, in classroom settings.

Another important issue that merits attention from researchers concerns L2 students' attitudes and reactions to the use of integrated writing rubrics in EAP classrooms. Within the

63

cycle of test development and validation, rubrics are reporting mechanisms that show students how their work is assessed and what skills they need to achieve specific success levels (Ferris & Hedgcock, 2014; Hamp-Lyons 2014; Weigle, 2002). Describing language learning objectives and accomplishments in qualitative terms, evaluation criteria not only serve the purpose of transparency in assessment (Crusan, 2010, 2015; Hudson, 2005) but also help students become autonomous and responsible for their own learning by means of self-monitoring (Jonsson, 2014; Panadero & Jonsson, 2013). Researchers exploring the instructional role of evaluation criteria in writing classes have suggested rubrics may constitute corrective feedback (Hyland, 2003) and help students improve their revision strategies in multi-drafted writing (Jonsson & Svingby, 2007). On the other hand, concerns have been raised about whether instructor-oriented rubrics can facilitate students' interpretation of test results and encourage self-monitoring (Alderson & Hamp-Lyons, 1996; Cheng, 2005). In fact, most discussions on integrated writing rubrics have concerned the reliability of the assessment criteria (Knoch et al., 2007), with little attention to whether rubrics are reflective of curriculum objectives or help students identify their strengths and weaknesses with source integration. Therefore, studies are needed to explore how students engage with integrated writing rubrics while assessing their own writing. Establishing such evidence for the usefulness and practicality of evaluation criteria could promote positive washback and facilitate student participation in integrated writing assessment.

In summary, student perspectives have contributed to test validation and score interpretation in educational assessment contexts, such as large-scale achievement or proficiency tests. On the other hand, although research on students have yielded an increased recognition of the key role that students play as stakeholders, classroom-based assessment constructs and procedures have been rarely explored from students' perspectives. In particular, the research is

scant on students' beliefs, experiences, and expectations related to classroom-based integrated writing assessment and evaluation criteria. Therefore, this exploratory case study focuses on students' experiences with and perceptions of integrated writing tasks in an EAP writing course and their use of an analytic rubric for self-assessment.

To gain a deeper understanding of the students' perceptions of integrated writing assessment, this study adopted an inductive case study approach. A case study design, commonly used in educational research, involves an in-depth analysis of the cases in a real-life context and draw meaningful conclusions by taking an interpretive orientation (Simons, 2009; Yin, 2014). Placing an emphasis on the process (e.g., task design, teacher feedback) as well as the outcomes, the current study examined perceptions of individual students about integrated writing assessment by engaging multiple data sources in an organized and systematic way. The researcher, who had no supervisory or administrative role in the EAP program, acted as a participant-observer in a naturalistic EAP setting (Patton, 2014). The research questions are as follows:

1. What are students' perceptions of integrated writing assessment in an EAP course?

2. How well can EAP students apply an integrated writing rubric for self-assessment?

Methodology

Study Context

The study was situated in an EAP program at an English-medium university in Canada, where two six-credit academic writing courses are offered to students. While the first course (Course 1) focuses on paragraph-level writing skills with a strong emphasis on vocabulary and grammatical development, the second course (Course 2) targets the analytical skills needed for integrated writing. L2 speakers who do not meet the English language proficiency test

requirements for admission register for the EAP courses concurrent with their respective undergraduate degree programs (see Table 7). Some students are exempted from Course 1 based on their performance on an in-house placement test. The current study took place in Course 2, which meets twice a week for 2.75 hours per meeting over a 13-week semester. The course objectives are to introduce students to source-based writing tasks, comprised of summary writing, cause-and-effect, and argumentative essays, which require them to synthesize academic content, and develop and support their views on academic subjects.

Table 7: English Proficiency Admissions Requirements

Test	Admission without EAP courses	Admission with EAP courses
TOEFL iBT	90 or higher with no component score under 20	75-89 with combined speaking and writing score of 34 or higher
IELTS	6.5 with no component scores under 6.5	Overall score of 6.5 with no component scores under 6.0

Participants

The study was announced to the students registered in the same section of the EAP writing course (Course 2) during the first week of the Winter 2021 semester. The course, originally designed for face-to-face instruction, was delivered in a synchronous mode using an online meeting tool (Zoom) due to the pandemic. Out of 24 students registered in the course, three students, whose names were replaced by pseudonyms (Liu, Qian, and Moza) agreed to participate in this case study. After reading and signing the consent forms, they completed a background questionnaire, which elicited specific information about their English learning background (Table 8).

Table 8: Participant Background

Name*	Liu	Qian	Moza
L1	Mandarin	Mandarin	Arabic
L2	English	English	French
Gender	Female	Male	Female
Age (years)	45	34	18
Length of residence in Canada	13 years	1.5 years	8 months
Length of English study in home country	12 years	10 years	13 years
Length of English study in Canada	2 years	1 year	8 months

Instructional Design

Pedagogical materials utilized in Course 2 were compiled in the course-pack by the EAP program coordinators and consisted of a) theme-based academic texts, b) authentic articles from newspapers and news magazines, and c) vocabulary and grammar topics which led to unit-final integrated writing tasks. The theme-based academic texts accompanied by vocabulary exercises were selected based on their content. Authentic articles were intended to offer different perspectives or type of information that students could use when composing integrated essays on relevant topics. As for the writing skills, students received explicit instruction on how to paraphrase and summarize source information, in addition to learning citation skills using the APA publication manual (American Psychological Association, 2020).

Throughout the 13-week semester, students carried out several individual and group assignments, such as graded vocabulary and grammar quizzes, and summary and paraphrasing

exercises. In addition, they completed three ungraded online quizzes, which were planned as formative assessment to address students' challenges with source integration. Informed by the results from a previous research project about the design and validation of an integrated writing rubric (see Study 1 and Study 2), three aspects of source integration were targeted in these quizzes: (1) linguistic revision of source-text language, (2) accurate representation of source content, and (3) the use of rhetorical acts to respond to source-text ideas. Regarding the structure, students were initially introduced to source use features based on exemplars from a locally developed EAP student corpus (McDonough et al., 2018). Following that, they were asked to evaluate the use of target features in sentences derived from the student corpus using selection items. They received immediate automated feedback on their selections. In the end, they were asked to complete short writing exercises to practice source integration features, and they received individualized feedback on their responses.

As for the EAP program's regularly scheduled battery of summative tests (see Figure 4) the students wrote a midterm exam, which elicited summary writing, and three classroom-based integrated writing tests, which targeted two different genres (i.e., cause-and-effect and argumentative), and required them to integrate information from sources and acknowledge the use of these sources through in-text citations. Following the assessment protocol in the EAP program, two weeks prior to each exam, students were assigned a reading list with six to seven sources from the course-pack. They discussed these readings with the course instructor and prepared notes using a note-taking template. Students could refer to the note-taking sheet in the course-pack which is filled in with notes on a source reading from the same course-pack. This provided them with an example of how to transfer information from sources to the note-taking sheet. On the exam day, the students were given two integrated writing prompts and chose which

one to write about. They were allowed to use one note-taking sheet per source as well as an English-only dictionary.

Figure 4: Timeline for the Summative Assessment Tasks
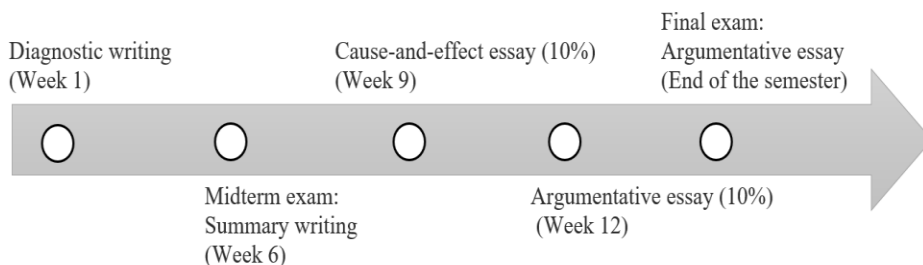
Data Sources

To identify the students' perceptions of classroom-based integrated writing assessment, the researcher, as a participant-observer, was granted access to the online course page which included pedagogical materials, such as course syllabi, rubric, and task instructions. After writing the cause-and-effect integrated writing test (week 9), the students who volunteered for this study were asked to complete a writing self-efficacy questionnaire adapted from Abdel-Latif (2015), which contained 18 items. There were 8 items in the first section to assess students' judgement about their integrated writing skills using a 5-point Likert scale (1 = strongly agree; 5 = strongly disagree). The second section with 10 items asked students to rate how confident they are when performing various tasks when writing from sources (e.g., sentence structure, organization of ideas) using a 5-point Likert scale (1 = very unconfident; 5 = very confident). Scores could range from 18 to 90, with high scores indicating high self-efficacy. Along with the self-efficacy questionnaire, the students answered some open-ended questions about the EAP writing course to obtain an understanding of their conceptualizations of the formative and summative assessment tasks prior to meeting with the researcher for an interview.

After administering the online questionnaire, the researcher scheduled individual data collection sessions with each student using an online meeting tool (Zoom). During this meeting, a semi-structured interview was conducted as an initial step, which lasted approximately 60 minutes. The introspective open-ended questions elicited students' perceptions about the integrated writing assessment tasks, teacher feedback and their use of available support systems in the EAP program. In addition, information was collected about students' cognitive activities, challenges, and strategies while writing from sources (e.g., generating ideas, selecting information from sources, incorporating citations, structuring the essay, and language-related reflections). To facilitate recall of their thought processes specific to source integration, the researcher referred to students' cause-and-effect essays, which they recently completed as part of the summative assessment. Before the interview, the essays had been analyzed qualitatively for the aspects of source use (i.e., accurate representation of source ideas, linguistic revision of source language, and source use purposes).

In the second part of the interview, students were asked to use an integrated writing rubric (Appendix D) to self-assess their cause-and-effect essay performance. The researcher introduced the rubric categories (i.e., content, organization, source use, and language use) and asked students to work independently to apply the rubric to their own writing, assigning a score for each category. Importantly, this rubric had been designed and validated as part of an ongoing research project in the EAP program (see Study 1 and Study 2), and it was different from the current EAP program rubric in terms of targeting source use as a separate category and including more detailed and task specific descriptors. After completing the self-assessment (around 20 minutes), the students showed their annotated rubrics and the essay to the researcher via the

share screen function and discussed their scoring decisions. All interviews were audio- and video-recorded and subsequently transcribed for analysis.

Students' cause-and-effect integrated essays were also evaluated by an EAP instructor from the same program using the same analytic rubric. The instructor had been teaching a different section of the same EAP course at the time of the study. In addition, she had contributed to the development and validation of the rubric. Thus, she was familiar with the test task instructions, prompts, source materials, and the rubric criteria. The scores from the instructor and the students were compared to identify similarities and differences in their perceptions.

Analysis

Following a case study methodology, data sources were analyzed taking an inductive approach to discover themes and patterns from unique cases (Creswell & Poth, 2016; Patton, 2014). An inductive approach emphasizes an initial examination of individual cases before combining those cases, and thus, the primary research focus was placed on the analysis of introspective interviews, and self-assessment from each student. For this purpose, self-efficacy questionnaire items were scored from 1 to 5 and then summed for each student. The interview data were analyzed qualitatively using open and axial coding (Corbin & Strauss, 2015). This process entailed reading the transcripts recursively and making notes to identify initial codes for establishing tentative and provisional categories. Using the initial codes as a guide, axial coding was carried out to review, refine, and group the initial codes into more meaningful categories. Finally, the key themes extracted from individual cases were combined and included in the final report. All analyses were conducted by the researcher and then verified by a trained research assistant to ensure the veracity of findings.

Findings

Individual profiles below provide details about students' background, self-efficacy ratings, and general perspectives about integrated writing tasks. The first set of findings discussed after the student profiles pertains to students' conceptualization of classroom-based integrated writing tasks, which were discussed under three themes: a) task requirements, b) test conditions, and c) feedback from the instructor. These themes are individually described below and discussed in relation to students' test preparation, test taking strategies and the use of available support systems, such as attending office hours. The second set of findings concern students' self-assessment of their own writing and engagement with the rubric categories. These findings were examined in relation to results from the textual analysis and instructor evaluation of the essays.

Student Profiles

Liu – with Low Self-efficacy

Before immigrating to Canada 13 years ago, Liu had completed a master's degree in business administration. She was in middle school when she first started learning English in China. At the time of the study, it was her second semester as an undergraduate student in accountancy. Based on her performance on the placement test, she was required to take both writing courses (Course 1, and 2) in the EAP program. When asked to compare Course 1 and Course 2, Liu said: "This semester I feel a little bit more difficult compared to last semester. I am very nervous before the tests this year because I really have no idea how I could prepare". Her responses to the open-ended questions before the interview suggested she felt more confident in her receptive skills (reading and listening) compared with writing and speaking abilities. On the other hand, she remarked that her reading comprehension skills do not help her attain excellence

in writing: "In my study, I am reading and reading and reading but it is not enough to write the main idea and construct essay". During our one-to-one meeting, Liu discussed her experience with integrated writing assessment focusing on her physical and cognitive reactions (e.g., feelings of panic, test anxiety) to assessment tasks. Her writing self-efficacy score was 56, which was the lowest among the three students in this study.

Qian – Questioning Test Purposes

Holding a bachelor's degree in business administration in China, Qian worked as a professional for 10 years and actively used English for work purposes in his home country. A year after moving to Canada, he was admitted to the university for a bachelor's degree in accountancy. During his first semester as an undergraduate student, he was required to register in the EAP writing course (Course 2) along with two other courses in his degree program. While describing his experience with the coursework, Qian said: "So far, to tell you the truth, I did not think I gained much from the course". His writing self-efficacy score was 71, which is much higher compared to Liu, although he identified writing in English as one of his "weaknesses". The crux of our discussion focused on his challenges with incorporating information from source materials along with personal opinions. Qian also questioned test purposes in terms of assessing background knowledge, which seemed to distract him from engaging with course content and attaining higher scores in exams.

Moza – a Motivated Learner

Right after finishing high school in Lebanon, Moza moved to Canada to study Human Resource Management as an undergraduate student. In her first semester, apart from the EAP writing course (Course 2), she was registered in prerequisite courses, macroeconomics and math, offered by the business department. She reported speaking French as an L2, adding that she has

73

focused on communicating in English since she moved to Canada, which helped her with the EAP coursework. Regarding her background in academic writing, she remarked that despite taking an academic writing course in Grade 12, she did not have to cite information from the sources. Even so, she acknowledged that the EAP writing course is not very challenging for her, which reflected on her self-efficacy score (82). Overall, Moza's experience with the assessment tasks in the EAP course seemed to be rather positive. She considered integrated writing as an important skill for university courses as she said: "I definitely think I need to do well in the exams to be successful in my other courses".

Keeping in mind the students' writing background, self-efficacy, and general perceptions of their writing course, the following section expands on their perceptions of integrated writing assessment by exploring their thoughts on the task requirements, and how the test conditions and instructor feedback impacts their task performance.

Student Perceptions of Integrated Writing Assessment

Task Requirements

An important concern related to task requirements raised by these students was moving from a summary task, elicited in the midterm exam, to writing an integrated cause-and-effect essay. Because summary writing does not require incorporating personal opinions, as different from integrated writing, the students felt disoriented since they were unsure how much of the information in their cause-and-effect essays should come from the source materials and whether they were allowed to discuss their personal opinions. For example, Liu reported preparing for the integrated writing test by reading the pre-assigned source materials carefully and then summarizing the main ideas in her own words so that she could cite them in the exam. She

described her experience with the classroom-based cause-and-effect integrated writing test as follows:

> *So, I wrote about population in the first paragraph and then summary from the readings. Then, in the second paragraph, I also gave another summary but according to the teacher's requirement, maybe I should give my opinion in every paragraph. I did not know that.*

Qian, on the other hand, was aware that he needed to develop personal arguments, but he was not certain if his opinions needed to be validated by source-text information. His comment below illustrates his confusion in terms of establishing a balance between personal ideas and information from sources:

> *In the exam, I was not sure when I needed to cite. When I write my own opinion, it seems that it is not that supportive. If I know the point of view is from a source, I cite it. But sometimes I just express the same or similar opinions without reading the articles. I just expressed my own opinions, with no citation and maybe it was wrong. Sometimes, I added paraphrases but not my idea. I do not know if you meet this kind of situation. So, what should we do? Am I wrong?*

The written task instructions for the cause-and-effect essay included that students must use "a minimum of three sources to support their ideas" and they should write "a clear thesis statement that connects logically with the three supporting paragraphs". After the midterm examination, which targeted summary writing, the students completed a pedagogical task, which required them to write a cause-and-effect essay using information from sources, and the course instructor provided them feedback on their responses using the current assessment criteria from the EAP program. They had also taken two ungraded quizzes, which targeted linguistic revision

75

of source-text language, and accurate representation of source content. On the other hand, it appears from these comments that their particular challenges were associated with retrieving relevant information from the task environment, rather than integrating source information (Plakans, 2009). This perhaps encouraged their efforts to address the prompt within their own capacity.

In case of Moza, despite reporting that she had tried to "extract numbers, examples and important stuff" from the source materials while preparing for the test, she acknowledged struggling with organizing the discourse, as shown in this comment:

It is kind of hard to be inspired by the texts, but at the same time have your own idea. Like it is hard to separate them. So, you have just a bit here and here, and then you mix them in your own writing. That was my problem while writing this essay with time pressure.

Developmental issues, such as Moza's, which concern the application of knowledge from summary writing to integrated writing tasks could be assisted by extended instruction, and targeted practice. On the other hand, Liu and Qian's misconceptions point towards a problem in respect to task representation – a process that entails understanding the task instructions, establishing goals, and devising strategies to achieve these goals (Cheng, 2009; Plakans & Liao, 2018). Previously, researchers have shown that task type can impact L2 students' task representation, and textual borrowing strategies, such as the number of words borrowed from the source materials (Spivey, 1991; Shi, 2004). In addition, understanding and following task instructions have been found to predict students' integrated writing test performance (Plakans & Gebril, 2013). Therefore, students' insights into task instructions are useful to address the factors affecting test validity.

Test Conditions

Test conditions were discussed in relation to background knowledge and psychological factors by these students. First, they described their concerns about the high-stakes nature of classroom-based integrated writing tests. Timing, more specifically, contributed to a perception of increased anxiety and fatigue within testing conditions. For example, Liu, who commented: "We have 3 hours and 1 shot! The time is clicking, stressful and I feel I cannot organize my paragraph". Similar comments were made by Moza, who believed that she could not perform well enough on the test due to time pressure and feelings of exhaustion:

> *Honestly, I do not think I did any paraphrasing to the last example from the reading because after three hours, I was so tired, and I did not put all my focus on it. If I can write a second draft, definitely I would do better on this essay.*

Qian's challenge, on the other hand, was tethered to not having an opportunity to use vocabulary and grammar creatively while focusing on other aspects of integrated writing, which made him question the test purposes:

> *I think grammar is my strong point, but I do not have a very large academic vocabulary. When writing this essay, I tried to use citations with good grammar and vocabulary, but I needed time for fixing other things like my topic sentence, so I cannot support my essay with enough ideas from readings. I do not know what exactly teacher expects from us.*

Moza and Qian also noted that having no background information about the prompt increased their stress and yielded to lower performance on the test. Moza remarked that if she had prior knowledge about the topic, she could have done a better job with paraphrasing:

> *In the exam, we had to write about the environmental impact of bottled water on the environment, and I had no knowledge about this subject, so it was hard for me to*

understand the readings and use my own words when paraphrasing. In the first weeks,

the topics were much better for me like nutrition and happiness. I was able to write better

as I am familiar with these subjects.

In case of Qian, he was challenged by the lack of choice in topics as he expected the prompts to be more directly relevant to the pre-assigned readings:

What we need to do is to the first is to read the materials from the course-pack before the

test. But the two topics teacher gave us to choose was different from what I read in the

materials. So, that is another reason I did not do well enough in this test. Is it for testing

my knowledge or ideas from other people's passages?

These perceived challenges suggest that timed exam conditions can be cognitively demanding for the students and impact their task performance. Integrated writing tasks have been shown to elicit more sophisticated linguistic and organizational features compared to independent writing (Cumming, 2013; Plakans & Gebril, 2017b). Additionally, these tasks require relying on source materials for validating content in students' texts, which indicates that reading ability is an important underlying construct in integrated writing assessment (Leki & Carson, 1997; Watanabe, 2001). Because these students needed to divide their limited attentional resources to multiple subskills (e.g., generating personal ideas, organizing essay structure, linguistic revision of source ideas), exam conditions possibly limited their ability to demonstrate their learning from classroom activities. This was particularly evident in Liu's comments about the time pressure and Qian's questions about test purposes. Similar findings have been reported in DeLuca and Cheng's (2011) research, in which students perceived psychological factors such as anxiety and fatigue as detrimental to their test performance. Because assessment conditions tend to play a role in students' test appeal, and task performance (Davies et al., 1999), it is

78

important for classroom instructors and test developers to minimize the impact of construct-irrelevant variance by obtaining evidence from test-takers.

Feedback from the Instructor

The students discussed the effect of feedback on their task performance in terms of understanding the course instructor's expectations and the usefulness of scoring criteria. In general, the students seemed to engage with teacher feedback; however, they held both positive and negative views of its usefulness. For example, Liu, among the three students had the most favorable opinion about teacher feedback:

> *I think to improvement for me is when I get the feedback from the professor on my writing. So, I know where I need to more to focus on all and where is my weakness. This is a great help to improve my writing.*

When asked specifically about what types of feedback she had received from the instructor on the pedagogical tasks, Liu said:

> *The professor tells me to use the words and construct sentences in proper way. I think that she knows what I mean, what I want to express, but she wanted me to correct my language for better express my ideas and it takes time. Also, the professor wanted me talk about only one idea in a paragraph. I know I need more ideas.*

As different from Liu, both Qian and Moza regarded feedback as discouraging and unclear, at times. For example, Qian commented:

> *I did not think I did so bad until I got it the feedback on the classroom cause-and-effect essay. But the result proved different from what I thought. The teacher said I cannot always grasp the key point. What I think is important in the reading passage, the teacher*

do not think so. I am not sure how I change my understanding of what is important and

what is not.

Moza's experience, on the other hand was pertinent to not being able to address teacher feedback

without making major changes in her essay. She expressed her concern saying: "When the

teacher points out the mistakes, then I feel that I should change like my whole text, not just the

mistakes." Elaborating on the kinds of feedback that she had received from the instructor, Moza

said:

Like every time I get feedback from the teacher, she said "wording" both in summary and

other writing assignments. I do not know what exactly "wording" is. I never asked her. I

asked her so many things, but I never thought to ask what "wording" is. Maybe I tend to

sophisticate my language and I add ideas and ideas, but at the end of the day, no, that is

not good. Maybe I should ask her before the next test.

These varying perceptions about feedback suggest that it was not entirely clear to the

students how they could address specific comments effectively from the instructor. Although the

instructor held weekly office hours and encouraged students to attend, all three students

remarked that they preferred asking their questions during regular class meetings. They also

reported consulting with their classmates about the feedback and looking up sample essays

online to improve their writing performance. Regarding the feedback they received on formative

tasks (i.e., ungraded quizzes targeting source use features), the students agreed that it was useful

to practice writing short texts and receiving feedback in a timely manner, which supports the

assumption that they attended to ongoing feedback as part of learning and test preparation in this

particular context.

Importantly, except the formative assessment, the feedback from the instructor generally focused on the generic aspects of writing in a second language rather than targeting students' source integration. Researchers have shown that L2 writers' challenges with integrated writing tasks mostly relate to understanding source materials, performing linguistic revision of source ideas and representation source content accurately (e.g., Keck, 2014; Plakans & Gebril, 2012, 2013; Wette, 2010). In the current study, despite her high self-efficacy, Moza identified paraphrasing and summarizing as the most challenging aspects of writing from source saying: "The hardest part is you have to rephrase it. You cannot just copy and paste it, and it's new to me". Similarly, Liu and Qian mentioned having no prior experience in source-based writing and experiencing difficulties in terms of selecting ideas from sources and linking them to their personal opinions. Although is possible that the course instructor prioritized language use and idea generation over source use at this time of the semester, it is a key for source use to be emphasized as an integral part of corrective feedback in EAP settings to prepare students for writing in their disciplines.

Turning to the use of scoring criteria as part of teacher feedback, all three students agreed that it is beneficial to receive a score based on the rubric, which is used to evaluate their integrated writing performance in the exams. For example, Qian said: "When I saw my performance on the rubric, I kind of knew why I am not "above standard". Liu and Moza shared the same perception as Qian although none of the students had used the rubric for self-assessment or exam preparation. Also, when asked to comment on rubric categories and descriptors, the students could hardly recall those without actually seeing the rubric. Their perceptions were centered upon the performance levels (i.e., below standard, standard, above standard), rather than the constructs outlined in the criteria. Thus, the rubric appeared to function

as a tool for students to reflect on their performance levels, rather than replacing or constituting corrective feedback. Because evaluation criteria promote autonomous learning and self-monitoring in L2 writing classrooms (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013), it is critical to make effective use of rubrics to facilitate students' understanding of teacher feedback.

Taken together, the students experience challenges with understanding and following test instructions, and they view test condition and feedback effects as a negative influence regarding validity of classroom-based integrated writing assessment. The following section will describe the students' use of an integrated writing rubric for self-assessment in connection with the results from textual analysis and instructor evaluation of the essays.

Self-assessment Using an Analytic Rubric

During the one-on-one meetings, the students were introduced to the new assessment criteria, which evaluated source use as a separate category, and included more detailed and explicit descriptors for content and organization compared the current EAP program rubric. They were encouraged to ask questions after the initial discussion of the rubric and while elaborating on their scoring decisions in reference to the specific descriptors. Even though they had not encountered the rubric prior to this study, the students did not seem to experience much difficulty understanding the descriptors. On the other hand, the way they treated the rubric during self-assessment showed variation. For example, Liu approached the criteria more holistically, compared to Qian and Moza and rather than highlighting specific descriptors across different score levels, she assigned a score of 2 (out of 4) for source use and language use, and 3 for content and organization. Her comments regarding her source use performance were as following:

Because source use, I think I am not perfect, not even 3. I know in this paragraph I

needed another resource to support this idea. I try to write in my words but maybe not

perfect. Also, my sentences and vocabulary should be better when summarizing.

As shown in Table 9, there was an overlap between Liu's self-assessment and the instructor

evaluation of her essay using the same criteria although she received a slightly higher score for

language use from the instructor.

Table 9: Comparison of Self-assessment and the Instructor Evaluation

	Content		**Organization**		**Source use**		**Language use**	
	Self	Instructor	Self	Instructor	Self	Instructor	Self	Instructor
Liu	3	2.5	3	3	2	2	2	2.5
Qian	3	2.5	3.5	3	2	2.5	2.5	3
Moza	3.5	3	4	3.5	3	3	3.5	3.5

Textual analysis of Liu's essay for linguistic revision of source ideas revealed instances of

verbatim copying and close paraphrasing in her essay (see Table 10). When asked about these

particular instances, she commented:

The professor knows this [unacknowledged source use] *paraphrasing is from the*

reading. So maybe I still need the author and date, right? I do not remember but I

think I changed the words in this example [verbatim copying].

Table 10: Textual Analysis of Student Essays

Source use variables	**Liu**	**Qian**	**Moza**
Essay length (N of words)	501	540	590
Citation density (N of citations)	5	4	6

Instances of verbatim copying (4+ strings)	2	0	1
Direct quotations	0	3	0
Close paraphrasing	1	0	1
Unacknowledged source use	1	0	1

As for Qian, he used the criteria more analytically, and annotated his essay based on the individual descriptors corresponding to his task performance. For example, he highlighted *overreliance on quotes* under source use and *limited range and variety of vocabulary* under language use and provided specific examples from his essay to support his scoring decisions. Qian regarded his textual organization "almost perfect", except for the use of transition words between the paragraphs. He was mostly critical of his source use, which, he thought, had a detrimental effect on content development:

> *I know I have deficiency in quoting and citing some information. So, in terms of this, I gave myself score 2. Also, there were few ideas supported with information from resources. This effects my content, as well.*

Although Qian's self-assessment did not align well with the instructor scores (Table 10), he was able to use the criteria insightfully. He considered self-assessment as an eye-opening practice as shown by his comment below:

> *This rubric is very very useful because it gave me some guidance to judge my writing from another point of view. I need to do this more often and criticize myself from another perspective. It helps me a lot.*

Turning to Moza, as a reflection of her high self-efficacy, she focused on the higher score levels (3 and 4) while applying the rubric on her cause-and-effect writing. She was the least

confident in her source use and assigned herself a score of 3. While explaining her scoring decisions, she pointed to the unacknowledged source use in the first body paragraph and commented:

> *In this paragraph, 30% of the information is from the reading, but the rest is from my background knowledge. Now looking at the criteria, I feel that I should have mentioned the source.*

Moza acknowledged the instance of close paraphrasing in her essay, and she identified fatigue as the leading factor. Regarding the source verbatim copying in the introduction paragraph, she said: "I thought that since it is the hook and it is a general idea, it is not that important to cite the source". Overall, the instructor's evaluation of Moza's essay (Table 10) was consistent with her self-assessment. She attained the highest scores among the three students for all rubric categories. In terms of her self-assessment experience, Moza commented: "It is so good to grade yourself. I should do this more often".

The fact that students were able to use the evaluation criteria for self-assessment with ease suggests they could benefit from having access to a detailed and explicit integrated writing rubric in their EAP writing courses. All students regarded rubric use for self-evaluation as useful and their self-assessment showed similarity to the instructor evaluation. In particular, Qian and Moza were precise in their use of source use criteria as they were able to identify the potential problems related to linguistic revisions of source ideas. On the other hand, Liu and Moza shared the misconception with regards to the requirement for explicit citation of source ideas. This problem could be addressed by providing students explicit feedback on their source integration patterns. In addition, previous research which focused on L2 writers' integrated writing development have provided evidence of improvement in students' citation behavior after

receiving targeted instruction (e.g., Hendricks & Quinn, 2000; Machili et al., 2019). Thus, it is important to reconsider assessment and feedback as integral components of learning and instruction and incorporate students' voices in assessment decisions, including the development and refinement of evaluation criteria.

Discussion and Conclusions

Guided by L2 writers' experiences and challenges with classroom-based integrated writing tasks, the purpose of this study was to shed light on students' conceptualizations of integrated writing assessment in an EAP context and their use of an analytic rubric for self-assessment. Utilizing multiple data sources (i.e., students' cause-and-effect integrated writing tests, a self-efficacy questionnaire, individual retrospective interviews, and course materials), and adopting a case study approach, the study centralized a focus on students' contextualized challenges with integrated writing assessment. Three major themes emerged from the findings, which pertains to the students' conceptualizations of classroom-based assessment: task requirements, test conditions, and feedback from the instructors.

First, although the students had practiced writing an integrated essay as part of pedagogical practices, they needed additional time and further instructions after the midterm examination to conceive of integrated writing task requirements. Their particular concerns were associated with the use of personal opinions along with the information from sources, and citation of the source materials. These concerns might stem from different individual characteristics, such as cultural and rhetorical background (Shi, 2004), or contextual factors, including the types of pedagogical tasks students were exposed to prior to the test, as well as the nature of task instructions by the instructor (Greene, 1995; Plakans, 2009, 2010). The fact that

86

these students were able to articulate their experiences with classroom-based assessment provides evidence that students' perspectives can enhance test processes and procedures.

Second, time pressure and lack of background knowledge about the topics prompted negative feelings, such as test anxiety, and fatigue, and affected the students' performance. As discussed by Cheng and DeLuca (2011), a decrease of test anxiety can motivate students to perform better on tests. In addition, allocating additional time for cognitive processes (e.g., planning, review) during the assessment helps students develop metacognitive strategies, which would guide them to monitor the writing process (Gebril & Plakans, 2009). Therefore, student perceptions of testing conditions are needed to ensure the validity of outcomes from test scores.

A final contextual assessment challenge that these participants encountered entailed the type of feedback that they received on their integrated essays. The students valued the feedback on the pedagogical tasks and ungraded quizzes; however, they required additional training to apply it to the integrated writing task. In some instances, they regarded the instructors' comments on their writing negatively, perhaps because of a lack of information on how to interpret the individualized feedback on their content and organization. This concern was accompanied by their approach to the current integrated writing rubric, which did not seem to facilitate their self-monitoring. These findings suggest construct-oriented rubrics need to be tailored for student use to facilitate their interpretation of test results (Alderson & Hamp-Lyons, 1996) and promote transparency in assessment (Crusan, 2015; Hudson, 2005).

Regarding the students' use of an analytic rubric for self-assessment, they were able to understand the constructs and identify the potential issues in their writing in light of the rubric descriptors. Moreover, although they were mixed in their approach to the criteria, the students deemed self-assessment as useful, and they were encouraged for purposing it as part of test

preparation. These findings do not just support the validity and practicality of the rubric, but also provide evidence for the role of evaluation criteria in promoting autonomous learning and self-monitoring (Benson, 2007; Jonsson, 2014).

Overall, findings from this study suggest that student perceptions of fundamental aspects of language assessment can contribute to best practices in EAP contexts and offer practical considerations for stakeholders, including course instructors and test developers. Although the outcomes from this study are inevitably linked to the context of the research, in which students' perceptions were influenced by the local teaching and assessment practices, such as the exam format, and the evaluation criteria, there are a few important implications for classroom-based integrated writing assessment. First of all, because transitioning from summary writing to integrated writing tasks was challenging for these students, EAP instructors and curriculum developers might consider introducing students to response writing and providing them examples of contextualized source use to help bridge the gap between summary writing and integrated writing tasks. In addition, to facilitate students' task representation, instructors may need to flesh out the verbal and written task instructions. The use of practice tests supplemented with sample responses might familiarize students with the test format and sample prompts might also help students internalize the test requirements. Given that learning to write from sources is a gradual process of development, classroom-based assessment tasks might be designed taking a processed based approach. Completing a draft before writing the final version in a timed exam condition might help with test-anxiety and fatigue associated with exam conditions. Furthermore, to develop students' autonomy and encourage students' engagement with teacher feedback, it is essential to introduce them to revision strategies and to the use of evaluation criteria as part of self- and peer-assessment. Most importantly, integrated writing rubrics need to be localized

through taking an evidence-based approach to rubric design. Evaluation criteria which represent the curricular objectives will address students' concerns regarding the interpretation of teacher feedback and centralize a focus on individual aspects of source use.

Chapter 5. General Discussion and Conclusion

This dissertation has offered a conceptual framework for assessing the construct of integration in EAP writing and contributed to our understanding of how source integration skills are evaluated in classroom-based assessment contexts from teacher and student perspectives. Although an increasing number of studies have been exploring the construct of integrated writing, particularly in second language writing domains, only a small number of them have provided a definition for source integration (Cumming et al., 2006). Prior attempts to formulate the notion of textual connections have been mostly limited to synthesis writing relying on Spivey's discourse synthesis model, which included the operations of *organizing, selecting*, and *connecting* ideas from multiple sources (Ascensión, 2005; Plakans, 2009; Yang & Plakans, 2012). Even though linguistic and rhetorical aspects of source integration have been acknowledged in the current definitions of integrated writing (e.g., Cumming et al., 2006; Knoch & Sitajalabhorn, 2013), researchers have not offered an inclusive theory and conceptual framework to inform the assessment of classroom-based integrated writing tasks. To address this gap, this dissertation has refined the construct of integration by combining major processes involved in synthesis writing (i.e., organizing, selecting, connecting) with rhetorical and linguistic resources devoted to decontextualizing source-text information.

To inform assessment practices in EAP settings and to address instructional concerns, the dissertation addressed three specific objectives: (a) designing a rubric drawing upon EAP instructors' perceptions and practices when assessing L2 students' integrated essays, (b) taking an empirically-based approach to validate an analytic rubric for assessing the construct of integration in EAP student writing, and (c) examining EAP students' perceptions of integrated writing assessment and their use of an analytic rubric for self-assessment in an EAP writing

course. Based on these objectives, the three studies were conducted, with each having its own specific goals while remaining connected with each other. In the next section, I will discuss key findings of the three studies and the usefulness of defining source integration in terms of three sub-processes: understanding, restructuring, and responding to source information. I will draw conclusions from the studies, suggest implications for contextual and situational assessment of integrated writing in EAP programs, and conclude with directions for future research.

Overview of the Key Findings

As discussed in the Introduction Chapter, a process model has been proposed in this dissertation to clarify and redefine the constructs within L2 integrated writing. Although the model accounts for the major operations involved in the process of integrated writing, this dissertation has mainly focussed on the assessment of three sub-constructs within source integration (i.e., introducing, restructuring, and responding to source information) outlined in the process model. Conceptually, defining source integration through the use of a process model appeared to be useful in Study 1 and Study 2 which focused on designing and validating an analytic rubric for assessing integrated writing. The first study identified specific constructs EAP instructors oriented to in students' argumentative integrated writing. More specifically, EAP instructors considered three constructs as key for the assessment of source integration: using source information to support personal ideas, modifying source-text language, and representing source content accurately. With regards to introducing source ideas, the instructors attended to students' source use purposes and awarded higher content and organization scores to the essays in which source ideas were introduced to elaborate on personal opinions. This finding, in particular has supported the motivation for refining the construct of integration, since source integration, which is different from synthesis writing, requires establishing a balance between

personal ideas and source-text information. Furthermore, the instructors expected students to restructure source ideas through linguistic modification of source-text language and accurate representation of source content. Overall, Study 1 has illustrated the value of utilizing a model of source integration with regards to identification of the underlying constructs within source integration that should be represented in the evaluation criteria. For example, reflecting the EAP instructors' scoring practice and perception data, one of the rubric descriptors in the content referred to supporting personal claims with information from sources. As such, since responding to source information was not salient for the instructors, source use criteria did not include a descriptor for rhetorical acts. Thus, establishing a relationship between the theory and practice in Study 1 has appeared to improve the outcomes.

The methodological focus of Study 1, which targeted EAP instructors' definition and evaluation of source integration in the light of the three sub-constructs also carried over to Study 2. However, in the case of Study 2, the attention was placed on reformulating the rubric descriptors. Adopting a model and construct definition was crucial in conducting the analysis as it helped better situate the research so that a variety of methods could be applied to establish the scoring validity. The fact that the raters (i.e., EAP instructors) showed acceptable levels of severity and awarded higher scores to more-proficient students' integrated essays has supported the validity of the constructs represented in the evaluation criteria. Furthermore, EAP instructors played an active role in the process of validation by making scoring decisions and sharing their feedback on the criteria, which helped us contextualize the criteria to reflect curriculum objectives and assessment practices in an EAP context (Crusan & Matsuda, 2018). In Study 3, the process model was again helpful for understanding particular challenges that students experience with integrating source information. For example, the students reported experiencing

difficulties with presenting source ideas along with personal opinions and modifying source-text language while integrating information from sources. These challenges were reflective of the sub-constructs discussed in the process model. Furthermore, even though the students had no prior experience with the rubric, which has been developed and validated in Study 1 and Study 2, they were able to use it effectively for self-assessment. This finding is promising in terms of the potential application of the rubric criteria in the EAP program to guide teaching and assessment practices.

Taken together, the three studies in this dissertation demonstrate that source integration, operationalized as a multifaceted construct, can be taught and assessed using a conceptual framework in educational settings. The analytic rubric, which has been designed based upon the skills and competencies included in the process model of source integration might help writing teachers reflect and act upon L2 learners' challenges while composing integrated essays. This in turn could provide guidance for the development of instructional materials to help students have the literacy skills needed to succeed in academic degrees. The following section provides overall implications of the outcomes from this dissertation for teaching and evaluating L2 students' integrated writing tasks in EAP contexts.

Overall Implications

Despite centralizing a focus on the assessment of integrated writing skills in EAP contexts, the outcomes from this dissertation study have also pedagogical and theoretical implications. While there exist differences in their design and overarching goals, the studies shared a common purpose to refine the construct of source integration in L2 writing. Considering the fact that integrated writing skills targeted in EAP courses help students get ready for their undergraduate courses in academic disciplines, the process model of source integration could be

tested and adapted for assessing disciplinary writing assignments. Despite the pivotal role that it plays in disciplinary assignments, the current conceptualizations of integrated writing as a skill remains controversial in university courses. For example, students in hard sciences, such as engineering and physics, deal with writing tasks such as short reports, proposals, case studies, lab reports and progress reports that require them to integrate facts and figures using disciplinary conventions (Hutchins, 2015). In social sciences, the instructional goals are grounded in achieving rhetorical purposes, rather than creating a structure, adopting a point of view in relation to source content, and developing convincing arguments as an outcome of the genre-based approach (Hyland & Bondi, 2006). In the field of history, for example, students typically engage in document-based tasks that require evidence-based thinking in a way that recognizes and reconciles historical perspectives from primary and secondary sources (Monte-Sano & De La Paz, 2010). To complete such tasks, students compose texts synthesizing information from multiple sources and contextualize evidence to introduce their perceptions for a greater purpose (Lent, 2016). Although reading and writing are viewed as interconnected in such disciplines, components of source-based writing, inclusive of integration, are not explicitly taught or assessed as separate skills in educational settings (Chang et al.,, 2002; De La Paz & Felton, 2010; Reynolds & Perin, 2009). Therefore, informed by the empirical work that narrowly focuses on academic subjects, introducing a process model for source integration in university courses can facilitate teaching and assessment practices.

In addition, a further step towards an understanding of the usefulness of the process model is to research the impact of the task environment (i.e., task representation and idea development) on students' integrated writing performance. The three studies in this dissertation mainly explored the sub-constructs within source integration. However, as discussed in the

Introduction chapter, L2 writers' task performance tend to be impacted by their task representation, and mental activities associated with the use of source texts (Cheng, 2009; Plakans & Liao, 2018). Although Study 3 demonstrated some of the challenges students experienced with understanding task instructions, and developing metacognitive strategies, extending research into the mediating factors within the task environment might provide a better understanding of individual writers' challenges with integrated writing tasks.

Focusing more narrowly on the assessment of classroom-based integrated writing tasks, the findings form the three studies suggest ways to improve assessment practices in EAP contexts and helps bridge the gap between curriculum and assessment. An important take-away from the three studies is to determine to what extent language proficiency scales, originally created by language testing professionals, serve the needs of the local users (instructors, and students, in particular). For example, in Study 1 and Study 2, EAP instructors, as the active users of evaluation scales, have played a key role in the process of rubric development and validation, and their perceptions have contributed to the contextualization of the evaluation criteria (Crusan & Matsuda, 2018). Eliciting instructor perceptions have also supported the approach that, teachers, one of the major stakeholders in educational settings, could provide important insights into isolating key features and refining rubric descriptors (Hudson, 2005; North, 2000). Study 3 has shifted the attention from the instructors to students to understand their beliefs, experiences, and expectations related to classroom-based integrated writing assessment. Despite the differences in their backgrounds, proficiency levels and self-efficacy ratings, all three students could articulate their perspectives about task requirements, test conditions, and instructor feedback. This has provided further evidence that incorporating students' voices can enhance teaching and assessment practices in educational contexts (Cheng & DeLuca, 2011; Malone &

Mantoee, 2014). Thus, an important implication for EAP programs is to acknowledge students' and instructors' perspectives of integrated writing assessment during the process of curriculum design and task development.

Limitations and Suggestions for Future Research

Although the limitations of each study were discussed previously, the three studies share similar limitations that need to be considered. Both Study 1 and Study 2 used sample argumentative essays from the CWEAT corpus. Even though these essays were written as a final examination in response to the same prompt and using the same source-texts, it is essential to establish the validity of the rubric on different essay types, and on a full test taker population before adopting it for more widespread use in an EAP program. Also, the source materials were assigned to the students before the examination so they had additional time to select and make notes of the source ideas they could possibly cite in their essays. Therefore, additional studies are needed to determine how varying writing conditions and different essay types targeted in EAP programs impact instructor perceptions of source integration.

Both Study 1 and Study 2 recruited EAP instructors among those who have experience in teaching integrated writing in a Canadian context. It is therefore important to replicate these results in different contexts to better understand how integrated writing assessment is contextualized in ESL/EFL programs with different curricular objectives. Furthermore, larger sample sizes are recommended to develop and validate rubrics to increase the validity of the outcomes. As such, student perceptions of the classroom-based integrated writing tasks, which were targeted in Study 3, should be further investigated using a mixed-methods design, and with more diverse populations. Due to global pandemic, only three students from an EAP writing course were willing to participate in the case study. The fact that students' perceptions were

influenced by the local teaching and assessment practices, such as course delivery format (remote), the exam conditions, and the evaluation criteria, it is unclear whether the results would be similar in different local assessment contexts.

Another agenda for future studies is to implement the analytic rubric, developed and validated in this dissertation, as part of a quasi-experimental design to determine whether it captures writing development. Previous studies of writing development have largely focused on which language features change over time in independent writing tasks (e.g., Lu, 2011; Yoon, 2017). Thus, there is an ongoing need for longitudinal and cross-sectional studies targeting L2 students' integrated writing development. Given the importance of writing development studies for understanding varying stages in the learning process, examining how development is manifested in integrated writing will have implications for curricular and assessment decisions.

Conclusion

Findings from this dissertation demonstrate the value of operationalizing a construct definition and utilizing a research-informed process model to design and validate evaluation criteria for assessing L2 integrated writing tasks. Along with supporting scoring validity of integrated writing tasks, outcomes from the three studies testify to the importance of incorporating stakeholder perspectives into assessment development. I hope that this dissertation will benefit writing instructors and assessment developers in various contexts with an interest in improving integrated writing teaching and assessment practices.

References

Abdel-Latif, M. (2015). Sources of L2 writing apprehension: A study of Egyptian university students. *Journal of Research in Reading, 38*, 194-212. https://doi.org/10.1111/j.14679817.2012.01549.x

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language testing, 13*(3), 280-297. https://doi.org/10.1177/026553229601300304

American Psychological Association (2020). *Publication Manual of the American Psychological Association* (7th ed.). https://www.apa.org/pubs/books/publication-manual-7th-edition

Applebee, A. N. (2000). Alternative models of writing development. In R. Indrisano & J. R. Squire (Eds.), *Perspectives on writing: Research, theory, and practice* (pp. 90–110). International Reading Association. https://doi.org/10.1598/0872072681.4

Ascención-Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*, 140-150. https://doi.org/10.1016/j.jeap.2008.04.001

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

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

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. In A. Kunnan (Vol. Ed.), *The companion to language assessment* (pp. 1301–1322). Oxford, United Kingdom: Wiley-Blackwell.

Belcher, D. D., & Hirvela, A. (Eds.). (2001). Linking literacies: Perspectives on L2 reading-

writing connections. Ann Arbor: University of Michigan Press.

https://doi.org/10.3998/mpub.11496

Benson, P. (2007). Autonomy in language teaching and learning. *Language teaching*, *40*(1), 21.

https://doi.org/10.1017/s0261444806003958

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the

Human Sciences. Mahwah, NJ: L. Erlbaum

Brindley, G. (2001). Outcomes-based assessment in practice: Some examples and emerging

insights. *Language Testing, 18*(4), 393-407.

https://doi.org/10.1177/026553220101800405

Bruce, E., & Hamp-Lyons, L. (2015). Opposing tensions of local and international standards for

EAP writing programmes: Who are we assessing for?. *Journal of English for Academic*

Purposes, 18, 64-77. https://doi.org/10.1016/j.jeap.2015.03.003

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In

Nikolov, M. (Ed.), *Assessing young learners of English: Global and local perspectives*

(pp. 291–315). Springer.

Butler, Y. G. (2019). Assessment of young English learners in instructional settings. In Gao, A.

(Ed.), *Second handbook of English language teaching* (pp. 477–495). Springer.

Butler, Y. G., Peng, X., & Lee, J. (2021). Young learners' voices: Towards a learner-centered

approach to understanding language assessment literacy. *Language Testing.*
https://doi.org/10.1177/0265532221992274

Campbell, C. (1990). Writing with others' words: Using background reading text in academic

compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the*

classroom (pp. 211–230). Cambridge: Cambridge University Press

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher is course.* Boston, MA: Heinle/Cengage Learning.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing writing, 26*, 20-37. https://doi.org/10.1016/j.asw.2015.07.004

Chang, K. E., Sung, Y. T., & Chen, I. D. (2002). The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education, 71*(1), 5-23. https://doi.org/10.1080/00220970209602054

Cheng, L. (2005). Changing language teaching through language testing: A washback study (Vol. 21). Cambridge University Press.

Cheng, F. W. (2009). Task Representation and Text Construction in Reading-to-Write. *Journal of Pan-Pacific Association of Applied Linguistics, 13*(2), 1-21.

Cheng, L., DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104–122. https://doi.org/10.1080/10627197.2011.584042

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Coffin, C. J. (2009). Incorporating and evaluating voices in a film studies thesis. thesis. *Writing and Pedagogy, 1*(2)163–193. https://doi.org/10.1558/wap.v1i2.163

Corbin, J., & Strauss, A. (2015). *Basics of qualitative research.* Thousand Oaks, CA: Sage.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*.

    Sage publications.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text

    cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior*

    Research Methods, 48(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Crusan, D., & Matsuda, P. K. (2018). Classroom writing assessment. *The TESOL Encyclopedia*

    of English Language Teaching. https://doi.org/10.1002/9781118784235.eelt0541

Crusan, D. (2010). Assessment in the second language writing classroom. University of

    Michigan Press.

Crusan, D. J. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing, 26,* 1.

    https://doi.org/10.1016/j.asw.2015.08.002

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and

    perils. *Language Assessment Quarterly, 10*(1), 1-8.

    https://doi.org/10.1080/15434303.2011.622016

Cumming, A., Kantor, R., Baba, K., Erdoosy, U., Eouanzoui, K., & James, M. (2006). *Analysis*

    of discourse features and verification of scoring levels for independent and integrated

    tasks for the new TOEFL. (TOEFL Monograph No. MS-30). Princeton, NJ: ETS

Cumming, A., Kantor, R., & Powers, D. E. (2002), Decision making while rating ESL/EFL

    writing tasks: A descriptive framework. *The Modern Language Journal, 86*, 67-96.

    https://doi.org/10.1111/1540-4781.00137

Davies, A, Brown, A, Elder, C, Hill, K, Lumley, T, & McNamara, T. (1999). *Dictionary of*

    language testing. Cambridge: Cambridge University Press.

De La Paz, S., & Felton, M. K. (2010). Reading and writing from multiple source documents in

history: Effects of strategy instruction with low to average high school writers. *Contemporary Educational Psychology, 35*(3), 174-192. https://doi.org/10.1016/j.cedpsych.2010.03.001

Dooey, P. (2010). Students' perspectives of an EAP pathway program. *Journal of English for Academic Purposes, 9*(3), 184-197. https://doi.org/10.1016/j.jeap.2010.02.013

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2011). Item banking for C-tests: A polytomous Rasch modeling approach. *Psychological Test and Assessment Modeling, 53*(4), 414.

Erickson, G., & Gustafsson, J. (2005). Some European students' and teachers' views on language testing and assessment. A report on a questionnaire survey. *European Association of Language Testing and Assessment*. http://www.ealta.eu.org/documents/resources/enlta%20activity%204%20report.pdf

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English test. *Canadian Modern Language Journal, 58*(4), 599–622. https://doi.org/10.3138/cmlr.58.4.599

Ewert, D., & Shin, S. Y. (2015). Examining instructors' conceptualizations and challenges in designing a data-driven rating scale for a reading-to-write task. *Assessing Writing, 26,* 38-50. https://doi.org/10.1016/j.asw.2015.06.001

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). New York: Routledge.

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition*

and Communication, 32(4), 365-387. https: //doi.org/10.2307/356600

Flowerdew, J., & Peacock, M. (Eds.) (2001). Research perspectives on English for Academic

Purposes. Cambridge: Cambridge University Press

http://dx.doi.org/10.1017/CBO9781139524766

Gass, S. M., & Mackey, A. (2000). Stimulated recall methodology in second language research.

Routledge.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing:

Rater reactions to integrated tasks. *Assessing writing, 21*, 56-73.

https://doi.org/10.1016/j.asw.2014.03.002

Gebril, A. & Plakans, L. (2016). Source-based tasks in L2 writing assessment: Lexical diversity,

textual borrowing, and proficiency. *Journal of English for Academic Purposes*, *24*, 78-88.

http://dx.doi.org/10.1016/j.jeap.2016.10.001

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in

integrated writing tests. S*paan Fellow Working Papers in Second or Foreign Language

Assessment, 7*(1), 47-84.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The

interface between discourse features and proficiency. *Language Assessment Quarterly,

10*(1), 9-27. https://doi.org/10.1080/15434303.2011.642040

Grabe, W. (2008). *Reading in a Second Language: Moving from Theory to Practice* (Cambridge

Applied Linguistics). Cambridge: Cambridge University Press.

https://doi.org/10.1017/CBO9781139150484

Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English

for academic purposes teaching and learning. *TESOL Journal, 4*(1), 9-24.

https://doi:10.1002/tesj.65

Greene, S. (1995). Making sense of my own ideas: The problems of authorship in a beginning

writing classroom. *Written Communication, 12*(2), 186-218.

https://doi.org/10.1177/0741088395012002002

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay

quality in both integrated and independent second language writing samples: A

comparison study. *Assessing Writing, 18*(3), 218-238.

https://doi.org/10.1016/j.asw.2013.05.002.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In *International handbook

of English language teaching* (pp. 487-504). Springer, Boston, MA.

Hamp-Lyons, L. (2014). Writing assessment in global context. *Research in the Teaching of

English, 48*(3), 353-362.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). *Metacognition in educational

theory and practice.* Routledge.

Harwood, N. (2009). An interview-based study of the functions of citations in academic writing

across two disciplines. *Journal of Pragmatics, 41*(3), 497-518.

https://doi.org/10.1016/j.pragma.2008.06.001

Harwood, N., & Petri, B. (2012). Performance in the citing behavior of two student writers.

Written Communication, 29, 55–103. https ://doi.org/10.1177/07410 88311 424133.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach.

Written Communication, 17, 307-352. https://doi.org/10.1177/0741088300017003001

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In M.

    C. Levy, & S. Randall (Eds.), *The science of writing: Theories, methods, individual*

    differences, and applications. Mahwah, NJ: Lawrence Erlbaum Associates.

Hendricks, M., & Quinn, L. (2000). Teaching referencing as an introduction to epistemological

    empowerment. *Teaching in Higher Education, 5*(4), 447-457.

    https://doi.org//10.1080/713699175

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?": Undergraduate ESL writers'

    engagement with source-based academic writing and reading. *Journal of English for*

    Academic Purposes, 12(2), 87-98. https://doi.org/10.1016/j.jeap.2012.11.005

Hirvela, A. (2004). Connecting reading and writing in second language writing instruction. Ann

    Arbor: University of Michigan Press.

Howard, R. M., Serviss, T., & Rodrigue, T. K. (2010). Writing from sources, writing from

    sentences. *Writing and Pedagogy, 2*(2), 177-192. https://doi.org/10.1558/wap.v2i2.177

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment.

    Annual Review of Applied Linguistics, 25, 205-227.

    https://doi.org/10.1017/S0267190505000115

Hutchins, T. D. (2015). A new approach to teaching business writing: Writing across the core-a

    document-based curriculum. American Journal of Business Education (AJBE), 8(2), 131-

    138.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse.

    Discourse studies, 7(2), 173-192. https://doi.org/10.1177/1461445605050365

Hyland, K. (2012). *Disciplinary identities: Individuality and community in academic*

    discourse. Cambridge: Cambridge University Press

Hyland, K. (2015). Genre, discipline and identity. *Journal of English for Academic Purposes, 19*, 32-43. https://doi.org/10.1016/j.jeap.2015.02.005

Hyland, K. (2019). *Second language writing.* Cambridge University Press

Hyland, K., & Bondi, M. (Eds.). (2006). *Academic discourse across disciplines*. Frankfort: Peter Lang.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for Academic Purposes, 1*(1), 1-12. https://doi.org/10.1016/S1475-1585(02)00002-4

Inoue, A. B. (2009). The technology of writing assessment and racial validity. In S. Schreiner Christopher (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 97–120). Philadelphia: Information Science Reference

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing, 26,* 51-66. https://doi.org/10.1016/j.asw.2015.07.002

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied linguistics, 11*(3), 253-271. https://doi.org/10.1093/applin/11.3.253

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment & Evaluation in Higher Education, 39*(7), 840-852. https://doi.org/10.1080/02602938.2013.875117

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (p. 131–153). Lawrence Erlbaum Associates Publishers.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261-278. https://doi.org/10.1016/j.jslw.2006.09.006

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing, 25*, 4–22. https ://doi.org/10.1016/j.jslw.2014.05.005.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304. https://doi.org/10.1177/0265532208101008

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing Writing, 12*(1), 26-43. https://doi.org/10.1016/j.asw.2007.04.001

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing, 18*(4), 300-308. https://doi.org/10.1016/j.asw.2013.09.003

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Kwan, B. S. (2008). The nexus of reading, writing and researching in the doctoral undertaking of humanities and social sciences: Implications for literature reviewing. *English for Specific Purposes, 27*(1), 42-56. https://doi.org/10.1016/j.esp.2007.05.002

Lavelle, E., & Bushrow, K. (2007). Writing approaches of graduate students. *Educational Psychology, 27*(6), 807-822. https://doi.org/10.1080/01443410701366001

Leki, I. (2017). Undergraduates in a second language: Challenges and complexities of academic literacy development. Mahwah, NJ: Lawrence Erlbaum

Leki, I., & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31*(1), 39-69. https://doi.org/10.2307/3587974

Lewkowicz, J. A. (1997). The integrated testing of a second language. *Encyclopedia of Language and Education, 7,* 121-130. https://doi.org/10.1007/978-1-4020-4489-2_12

Li, Y., & Casanave, C. P. (2012). Two first-year students' strategies for writing from sources: Patchwriting or plagiarism? *Journal of Second Language Writing, 21*(2), 165–180. https://doi.org/10.1016/j.jslw.2012.03.002.

Linacre, J. M. (2014). Winsteps Rasch Measurement computer program (Version 3.81.0). Beaverton, OR.:Winsteps.com.

Linacre, J. (2010). A user's guide to FACETS Rasch-model computer programs. Chicago, IL: Winsteps.com

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95-110

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33*(4), 623-636. https://doi.org/10.1016/j.system.2005.02.002

Llosa, L., & Malone, M. E. (2017). Student and instructor perceptions of writing tasks and performance on TOEFL iBT versus university writing courses. *Assessing Writing, 34*, 88-99. https://doi.org/10.1016/j.asw.2017.09.004

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*, 36-62. https://doi.org/10.5054/tq.2011.240859

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71. https://doi.org/10.1177/026553229501200104

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

Lunz, M. E., Wright, B. D., & & Linacre, J. M. (1990) Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345. https://doi.org/10.1207/s15324818ame0304_3

Machili, I., Papadopoulou, I., & Kantaridou, Z. (2019). Effect of strategy instruction on EFL students' video-mediated integrated writing performance. *Journal of Second Language Writing, 48.* https://doi.org/10.1016/j.jslw.2019.100708

Malone, M. E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability. *ETS research report No. RR-4-42*. Princeton, NJ: Educational Testing Service.

Mansourizadeh, K., & Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes, 10*(3), 152-161. https://doi.org/10.1016/j.jeap.2011.03.004

Mateos, M., Solé, I., Martín, E., Cuevas, I., Miras, M. & Castells, N. (2014). Writing Synthesis from multiple sources as a learning activity. In G. Rijlaarsam, P.D. Klein, P. Boscolo, L. C. Kirckpartick, & C. Gelati (Eds.), *Writing as a learning activity* (pp. 169-190). Leiden: Brill

McAlpine, L., & Amundsen, C. (2011). Doctoral education: research-based strategies for

    doctoral students, supervisors and administrators. London: Springer.

    https://doi.org/10.1007/978-94-007-0507-4

McDonough, K., Neumann, H., & Leu, S. (2018). *Concordia Written English Academic Texts*

    (CWEAT) corpus. Concordia University: Montreal, QC.

McNamara, T. (1996). *Measuring second language performance.* United Kingdom: Blackwell.

Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a

    Spanish test for academic purposes using the argument-based approach to validation.

    Assessing Writing, 35, 41-55. https://doi.org/10.1016/j.asw.2017.12.003

Monte-Sano, C., & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical

    reasoning. *Journal of Literacy Research, 44*(3), 273-299.

    https://doi.org/10.1177/1086296X12450445

Neumann, H., Leu, S., & McDonough, K. (2019). L2 writers' use of outside sources and the

    related challenges. *Journal of English for Academic Purposes, 38,* 106-120.

    https://doi.org/10.1016/j.jeap.2019.02.002

North, B. (2000). The development of a common framework scale of language proficiency. New

    York: Peter Lang.

Ohkubo, N. (2009). Validating the integrated writing task of the TOEFL internet-based test

    (iBT): Linguistic analysis of test takers' use of input material. *Melbourne Papers in*

    Language Testing, 14(1), 1-31.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment

    purposes revisited: A review. *Educational research review, 9,* 129-144.

    https://doi.org/10.1016/j.edurev.2013.01.002

Patton, M. Q. (2014). Qualitative research & evaluation methods: Integrating theory and practice (4th edition). Sage publications.

Payant, C., McDonough, K., Uludag, P., & Lindberg, R. (2019). Predicting integrated writing task performance: source comprehension, prewriting planning, and individual differences. *Journal of English for Academic Purposes, 40,* 87–97. https://doi.org/10.1016/j.jeap.2019.06.001

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of second language writing, 12*(4), 317-345. https://doi.org/10.1016/j.jslw.2003.08.004

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111-129. https://doi.org/10.1016/j.asw.2008.07.001

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561-587. https://doi.org/10.1177/0265532209340192

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *Tesol Quarterly*, *44*(1), 185-194. https://doi.org/10.5054/tq.2010.215251

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217-230. https://doi.org/10.1016/j.jslw.2013.02.003

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*(1), 18-34. https://doi.org/10.1016/j.asw.2011.09.002

Plakans, L., & Gebril, A. (2017a). An assessment perspective on argumentation in writing. *Journal of Second Language Writing*, *36*, 85–86. https://doi.org/10.1016/j.jslw.2017.05.008

Plakans, L., & Gebril, A. (2017b). Exploring the relationship of organization and connection with   scores in integrated writing assessment. *Assessing Writing, 31*, 98-112. https://doi.org/10.1016/j.asw.2016.08.005

Plakans, L. & Liao, J. T. (2018). Task Representation. In *The TESOL Encyclopedia of English Language Teaching.* https://doi.org/10.1002/9781118784235.eelt0532

Polio, C. & Shi, L. (2012). Perceptions and beliefs about textual appropriation and source use in second language writing. *Journal of Second Language Writing 21*(2), 95–101 https://doi.org/10.1016/j.jslw.2012.03.001

Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: a learning-focused interpretation 1. *International Journal of Applied Linguistics, 16*(2), 163-188. https://doi.org/10.1111/j.1473-4192.2006.00112.x

Reid, J. (2001). Advanced EAP writing and curriculum design: What do we need to know? In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 143-160). Mahwah, NJ: Lawrence Erlbaum Associates

Reynolds, G. A., & Perin, D. (2009). A comparison of text structure and self-regulated writing strategies for composing from sources by middle school students. *Reading Psychology, 30*(3), 265-300. https://doi.org/10.1080/02702710802411547

Roberts, M. J., & Erdos, G. (1993). Strategy selection and metacognition. *Educational Psychology, 13*(3-4), 259-266. https://doi.org/10.1080/0144341930130304

Rosenfeld, M., Leung, S., & Oltman, P. (2001). The reading, writing, speaking, and listening

    tasks important for academic success at the undergraduate and graduate levels.

    TOEFL monograph series 21. Princeton, NJ: Educational Testing Service.

Ruiz-Funes, M. (2011). Reading to write in a foreign language: Cognition and task

    representation. In T. Cimasko & M. Reichelt (Eds.), *Foreign language writing*

    instruction: Principles and practices (pp. 22– 43). Anderson, SC: Parlor Press.

Sato, T., Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact

    of face validity on student learning. *Language Testing in Asia, 5*(1), 1–15.

    https://doi.org/10.1186/s40468-015-0019-z

Sawaki, Y., Quinlan, T., & Lee, Y. W. (2013). Understanding learner strengths and weaknesses:

    Assessing performance on an integrated writing task. *Language Assessment Quarterly,*

    10(1), 73-95. https://doi.org/10.1080/15434303.2011.633305

Schuemann, C. M. (2008). Myth 2: Teaching citation is someone else's job. In J. Reid (Ed.),

    Writing myths (pp. 18- 40). Applying second language research to classroom teaching.

    Ann Harbour, MI: University of Michigan Press.

Shaw, P., & Pecorari, D. (2013). Source use in academic writing: An introduction to the special

    issue. *Journal of English for Academic Purposes, 12*(2).

    https://doi.org/10.1016/j.jeap.2012.11.001

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing*

    second language writing. Cambridge University Press.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*(2),

    171-200. https://doi.org//10.1177/0741088303262846

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores?. *Language Testing, 32*(2), 259-281. https://doi.org/10.1177/0265532214560257

Simons, H. (2009). *Case study research in practice.* Thousand Oaks, CA: Sage.

Spivey, N. N. (1990). Transforming texts constructive processes in reading and writing. *Written Communication, 7*(2), 256-287. https://doi.org/10.1177/0741088390007002004

Spivey, N. N. (1991). The shaping of meaning: Options in writing the comparison. *Research in the Teaching of English, 25(4), 390-418.

Stahl, A., Mann, R., & Loren, D. D. (2004). Writing for many: Interdisciplinary communication, constructionism, and the practices of writing. *Historical Archaeology, 38*(2), 83-102.

Swales, J. (1990). Genre analysis: English in academic and research settings. Cambridge University Press.

Swales, J. M. (2014). Variation in citational practice in a corpus of student biology papers: From parenthetical plonking to intertextual storytelling. *Written Communication, 31*(1), 118-141. https://doi.org/10.1177/0741088313515166

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills.* Ann Arbor: University of Michigan Press.

Thompson, G. (2001). Interaction in academic writing: Learning to argue with the reader. *Applied linguistics, 22*(1), 58-78. https://doi.org/10.1093/applin/22.1.58

Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning & Technology, 5*(3), 91-105.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment Research and Evaluation, 9*(2), 1-7. https://doi.org/10.7275/jtvt-wg68

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests?. *Language Testing, 22*(2), 174-210. https://doi.org/10.1191/0265532205lt299oa

Turley, E. D., & Gallagher, C. G. (2008). On the uses of rubrics: Reframing the great rubric debate. *English Journal, 79*(4), 87–92.

Tweedie, M. G., & Kim, M. (2015). EAP curriculum alignment and social acculturation: Student perceptions. *TESL Canada Journal*, 41-7. htp://doi.org/htp://dx.doi.org/10.18806/tesl.v33i1.1226

Uludag, P., Lindberg, R., McDonough, K., & Payant, C. (2019). Exploring l2 writers' source-text use in an integrated writing assessment. *Journal of Second Language Writing, 46*. https://doi.org/10.1016/j.jslw.2019.100670

Vlanti, S. (2012). Assessment practices in the English language classroom of Greek junior high school. *Research Papers in Language Teaching and Learning, 3*(1), 92–122. https://rpltl.eap.gr/images/2012/RPLTL-03-01-fulltext.pdf

Yilmaz, M., & Özdem Erturk, Z. (2017). A Contrastive Corpus-Based Analysis of the Use of Reporting Verbs by Native and Non-Native ELT Researchers. *Novitas-ROYAL (Research on Youth and Language), 11*(2), 112-127.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27-55. https://doi.org/10.1016/j.asw.2004.01.002

Wette, R. (2010). Evaluating student learning in a university-level EAP unit on writing using

sources. Journal of Second Language Writing, 19(3), 158-177.

https://doi.org/10.1016/j.jslw.2010.06.002

Wette, R. (2017). Using mind maps to reveal and develop genre knowledge in a graduate writing

course. *Journal of Second Language Writing, 38*, 58-71.

https://doi.org/10.1016/j.jslw.2017.09.005

Wette, R. (2018). Source-based writing in a health sciences essay: Year 1 students' perceptions,

abilities and strategies. *Journal of English for Academic Purposes, 36*, 61-75.

https://doi.org/10.1016/j.jeap.2018.09.006

Wilson, M. (2007). Why I won't be using rubrics to respond to students' writing. *English

Journal, 96*(4), 62-66. https://doi.org/10.2307/30047167

Winke, P., Lee, S., Yoon, H.-J., Ahn, J. I., Choi, I., Cui, Y. (2018). The cognitive validity of

child English-language tests: What young language learners and their native-speaking

peers can reveal. *TESOL Quarterly, 52*(2), 274–303. https://doi.org/10.1002/tesq.396

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student

learning. *Journal of Effective Teaching, 7*(1), 3-14.

Yang, H. C., & Plakans, L. (2012). Second language writers' strategy use and performance on an

integrated reading-listening-writing task. *TESOL Quarterly, 46*(1), 80-103.

https://doi.org/10.1002/tesq.6

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese

EFL learners at different proficiency levels. *Linguistics and Education, 23*(1), 31-48.

https://doi.org/10.1016/j.linged.2011.09.004

Yilmaz, M., & Erturk, Z. (2017). A Contrastive Corpus-Based Analysis of the Use of

    Reporting Verbs by Native and Non-Native ELT Researchers. *Novitas-ROYAL*

    (Research on Youth and Language), 11(2), 112-127.

Yin, R. (2014). *Case study research design and methods* (5th ed.). Thousand Oaks, CA: Sage

Yoon, H. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and

    construct multidimensionality. *System, 66,* 130-141.

    https://doi.org/10.1016/j.system.2017.03.007

Youn, S. Y. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed

    methods. *Language Testing, 32*(2), 199-225. https://doi.org/10.1177/0265532214557113

Yu, G. (2013). The use of summarization tasks: Some lexical and conceptual analyses. Language

    Assessment Quarterly, 10(1), 96-109. https://doi.org/10.1080/15434303.2012.750659

Zhu, W. (2005). Source articles as scaffolds in reading to write: The case of a Chinese student

    writing in English. *Journal of Asian Pacific Communication, 15*(1), 129-152.

    https://doi.org/10.1075/japc.15.1.09zh

Appendix A: Integrated Writing Rubric from the EAP Program

	Content & Organization	Grammar & Vocabulary	Mechanics
4	**C** thesis (explicit or implicit) and topic sentences express the writer's position on the assigned topic in a clear and sophisticated way **C** informative, convincing and relevant supporting ideas **C** information from sources are accurately interpreted and acknowledged **C** exemplary selection and integration of source material **C** acknowledgement of and response to opposing view(s) is thorough and effective **O** clarity of message enhanced by clear pattern of organization between and within paragraphs **O** effective and varied transitions between ideas	**G** variety of sentence structures used with no major sentence problems **G** rare minor language errors **V** sophisticated and precise word choice and accurate word form **V** extensive range and variety of vocabulary	**M** conventions of punctuation, capitalization, spelling, and indentation respected with rare minor errors **M** proper and complete mechanics for citing information from sources
3	**C** thesis (explicit or implicit) takes a clear position on the assigned topic; topic sentences identifiable and relevant **C** supporting ideas are mostly sufficiently developed **C** most information from sources is accurately interpreted and acknowledged **C** mostly effective selection and integration of source material **C** mostly effective acknowledgment of and response to at least one opposing view **O** mostly logical sequencing of ideas and smooth transitions between and within paragraphs	**G** variety of sentence structures used with rare major sentence problems **G** occasional minor language errors **V** mostly accurate and appropriate word choice and form **V** adequate range and variety of vocabulary	**M** conventions of punctuation, capitalization, spelling, and indentation mostly respected with occasional errors **M** mechanics for citing information from sources respected with minor errors
2	**C** thesis and topic sentences identifiable, but weak; e.g. vague, narrow, or not causal **C** some supporting ideas are inadequately developed, irrelevant, and/or repetitive **C** some information from sources is misinterpreted and/or unacknowledged **C** ineffective selection and integration of source material **C** acknowledgement of opposing view(s) is present but unclear, insufficient, or not coherently integrated **O** significant problem(s) with principles of essay organization **O** relationship between ideas is poorly established between and/or within paragraphs; transitions sometimes not smooth	**G** major sentence problems or limited sentence variety **G** frequent minor language errors **V** imprecise word choice or frequent word form errors **V** limited range and variety of vocabulary	**M** frequent problems with conventions of punctuation, capitalization, spelling, and indentation **M** frequent errors with mechanics for citing information from sources
1	**C** thesis and topic sentences missing, irrelevant, or incomprehensible **C** supporting points are inadequately developed, irrelevant, unconvincing and/or repetitive **C** little or no information from sources used **C** relies on lengthy direct quotations; sentences or sections of text copied or copied directly from resource sheets **C** no acknowledgment of opposing views in argument **O** no clear pattern of organization **O** sequence of ideas not logical or coherent; poor or absent transitions between ideas **X** insufficient information for evaluation	**G** frequent major sentence structure problems; lack of sentence variety **G** pervasive or severe errors impede comprehension of ideas **V** limited and inaccurate word choice interferes with meaning; pervasive word form errors **V** very narrow range of vocabulary	**M** pervasive or severe problems with conventions of punctuation, capitalization, spelling, and indentation **M** mechanics for citing information not properly employed or not attempted

Appendix B: Stimulated Recall Instructions

Before the Stimulated Recall Interview:

I am interested in learning what you think about as you read and evaluate this essay. To do this, I will ask you to spend 15-20 minutes to read it, take some notes on the text using the comment function, and/or underline or highlight sections that are relevant to your judgements of certain features. Once you finish, I will ask you to share your screen with me, recall and talk about what you were thinking while evaluating it.

I will record what you say using the recording function over the Zoom. Do you understand what I am asking you to do? Do you have any questions?

During the Stimulated Recall Interview:

What did you think of this essay?

What were you thinking here/at this point/right then?

Can you tell me what you were thinking at that point?

Is there anything else that comes to your mind?

Do you remember anything else about what you were thinking at that moment?

Appendix C: Interview Questions

1. What is your overall opinion of the rubric?

2. Do you think there is a category that should be excluded or included?

3. Do you think the descriptors included in the rubric were enough to assess each category?

4. Did you think that the descriptors were clear enough?

5. Did you think there was any overlap between the categories?

6. Was there a category that was difficult to use?

7. If you had difficulties in using any of the categories, what strategy or criteria did you use to grade them?

8. In your opinion, rank the rubric categories* based on their importance.

9. Can you compare your experience with using the current EAP rubric and the new rubric? What are the similarities and differences?

10. Do you think EAP students would understand the rubric? Would it help them identify how to improve their source-based writing?

11. Do you think this rubric could be used to assess student writing in the EAP program (ESL 204)?

(Adapted from Mendoza & Knoch, 2018)

Integrated Writing Rubric: Version 1

	Content	**Organization**	**Source Use**	**Language and Mechanics**
4	- The prompt is fully addressed; the ideas are explored in-depth with supporting details. - The essay has a strong presence of personal opinions which are supported with information from sources - Acknowledgement of and response to opposing view(s) is thorough and effective	- The essay has a clear, narrow thesis; introduction and conclusion are complete and effective - Topic sentences clearly stem from the thesis and body paragraphs are cohesive - Ideas between and within paragraphs are linked with smooth and effective transitions	- Information from sources are relevant, accurate and complete - Source texts are paraphrased/summarized in the student's words with structural and lexical changes - In-text references are documented properly; a range of reporting structures are used	- The essay includes a variety of sentence styles and length with no major structural problems - A wide range of sophisticated academic vocabulary is used accurately - The essay is virtually free of punctuation, capitalization, and spelling errors
3	- The prompt is addressed adequately; the ideas are explored with supporting details - The essay generally reflects the student's perspective on the topic, and personal views are mostly supported with information from sources - Acknowledgment of and response to opposing view(s) is mostly effective	- The thesis is mostly clear; introduction and conclusion are mainly complete and effective - Topic sentences generally stem from the thesis and body paragraphs are mostly cohesive - Most ideas between and within paragraphs are linked with transitions	- Information from sources are mostly relevant, accurate and complete but not quite enough/a bit too much - Source texts are mostly paraphrased/summarized in the student's words with adequate structural and lexical changes - In-text references are mostly documented properly; some reporting structures are used	- The essay includes an appropriate range of sentence styles and length with some structural problems - An adequate range of academic vocabulary is used accurately - There are occasional spelling, capitalization, and punctuation errors, which do not interfere with meaning

2	- There is some attempt to address the prompt; some ideas are supported with details - The essay somewhat reflects the student's perspective on the topic, and personal views are hard to distinguish from source information - Acknowledgment of opposing view(s) is present but unclear/insufficient.	- The thesis is present but unclear; introduction and conclusion are attempted but lack some required elements - Inappropriate topic sentences may be present; paragraphs show structure but lack cohesion - Some ideas between and within paragraphs may be linked with transitions	- Information from sources may be irrelevant, inaccurate; there is overmuch evidence from sources/some evidence but not enough - Source texts are somewhat paraphrased/summarized but too similar to original - There are a number of problems with documentation of in-text references and the use of reporting structures	- Sentences show structural errors and little or no variety - Some academic vocabulary is used appropriately - There are some errors of spelling, capitalization, and punctuation errors, which interfere with meaning in places
1	- The prompt is minimally addressed; the ideas are underdeveloped - The essay rarely presents the student's perspective on the topic; there is no sense of individuality - Opposing view(s) are not acknowledged	- There is no discernible thesis; introduction and conclusion are missing all required elements - Paragraphing is mostly inappropriate with little, if any, cohesion - Transitions are used minimally	- Most of writing is from sources/ little or poor source evidence given/reliance on quotes - Largely unchanged paraphrases/summaries; much unacknowledged evidence suggesting plagiarism - There are serious problems with the use of documentation conventions	- Sentences show multiple and serious structural errors and little or no variety - There is limited use of academic vocabulary - There are frequent errors of spelling, capitalization, and punctuation; Little, if any, proofreading is evident
0	- The essay shows no engagement with the topic	- There is no indication of paragraphing.	- No evidence is provided; no academic conventions are followed	- Intrusive and/or inaccurate language use which greatly impedes communication

Integrated Writing Rubric: Version 2

	Content	Organization	Source Use	Language and Mechanics
4	- The prompt is fully addressed; the ideas are explored in-depth with supporting details. - The student's stance/voice is clear in the essay, and arguments are always supported with evidence from sources. - Acknowledgement of and response to opposing view(s) is thorough and effective.	- The essay has a clear, focussed thesis; introduction and conclusion are complete and effective. - Topic sentences clearly stem from the thesis and body paragraphs are cohesive. - Ideas between and within paragraphs are linked with smooth and effective transitions.	- Information from sources is relevant, accurate and complete. - Source information is paraphrased/summarized in the student's words with structural and lexical changes. - APA citations are accurate.	- The essay includes a variety of sentence styles and length with no major structural and language problems. - An extensive range and variety of vocabulary is used accurately. - The essay is virtually free of punctuation, capitalization, and spelling errors.
3	- The prompt is addressed adequately; the ideas are explored with supporting details. - The student's stance/voice is generally evident in the essay, and arguments are mostly supported with evidence from sources. - Acknowledgment of and response to opposing view(s) is mostly effective.	- The thesis is mostly clear; introduction and conclusion are mainly complete and effective. - Topic sentences generally stem from the thesis and body paragraphs are mostly cohesive. - Most ideas between and within paragraphs are linked with transitions.	- Information from sources is mostly relevant, accurate and complete but not quite enough/a bit too much. - Source information is mostly paraphrased/summarized in the student's words with adequate structural and lexical changes. - APA citations are mostly accurate.	- The essay includes an appropriate range of sentence styles and length with some structural and/or language problems. - An adequate range and variety of vocabulary is used mostly accurately. - There are occasional spelling, capitalization, and punctuation errors, which do not interfere with meaning.
2	- There is some attempt to address the prompt; some ideas are supported with details. - The student's stance/voice is somewhat evident in the essay; arguments are hard to	- The thesis is present but unclear; introduction and conclusion are attempted but lack some required elements. - Inappropriate topic sentences may be present; body	- Information from sources may be irrelevant/inaccurate; there is some evidence but not enough. - Source information is somewhat	- Sentences show structural and/or language errors and little or no variety. - Limited range and variety of vocabulary is used with some accuracy.

123

	distinguish from source information/limited evidence is used to support arguments. - Acknowledgment of opposing view(s) might be present but unclear/insufficient.	paragraphs show structure but lack cohesion. - Some ideas between and within paragraphs may be linked with transitions.	paraphrased/summarized but too similar to original; there might be some unacknowledged information from sources. - There are a number of problems with APA citations.	- There are some errors of spelling, capitalization, and punctuation errors, which interfere with meaning in places.
1	- The prompt is minimally addressed; the ideas are underdeveloped. - The student's stance/voice is rarely evident in the essay; there is no sense of individuality/no evidence is used to support arguments. - Opposing view(s) are not acknowledged.	- There is no discernible thesis; introduction and conclusion are missing/ lacking all required elements. - Body paragraphs lack structure and cohesion. - Transitions are used minimally.	- Little or poor source evidence is given; the student relies mostly on quotes. - There are largely unchanged text chunks from sources/ much unacknowledged information suggesting plagiarism. - There are serious problems with citation of source information.	- Sentences show multiple and serious structural and language errors and little or no variety. - Very narrow range of vocabulary is used with pervasive errors. - There are frequent errors of spelling, capitalization, and punctuation; Little, if any, proofreading is evident.
0	- The essay shows no engagement with the topic.	- There is no indication of paragraphing.	- No evidence is provided; no academic referencing/citation conventions are followed.	- Intrusive and/or inaccurate language use which greatly impedes communication.

Integrated Writing Rubric: Final Version

	Content	Organization	Source Use	Language and Mechanics
4	- The prompt is fully addressed; the ideas are explored in-depth with supporting details. - The student's position is clear in the essay, and personal claims are always supported with evidence from sources. - Acknowledgement of and response to opposing view(s) is thorough and effective.	- The essay has a clear and focussed thesis; introduction and conclusion are complete and effective. - Topic sentences clearly stem from the thesis and body paragraphs are cohesive. - Ideas between and within paragraphs are linked with smooth and effective transitions.	- Information from sources is always relevant and accurate. - Source information is paraphrased/summarized in the student's words with structural and lexical changes. - Source information is cited properly.	- The essay includes a variety of sentence styles and length with no major structural and language problems. - An extensive range and variety of vocabulary is used accurately. - The essay is virtually free of punctuation, capitalization, and spelling errors.
3	- The prompt is addressed adequately; the ideas are explored with supporting details. - The student's position is generally evident in the essay, and personal claims are mostly supported with evidence from sources. - Acknowledgment of and response to opposing view(s) is mostly effective.	- The thesis is mostly clear; introduction and conclusion are mainly complete. - Topic sentences generally stem from the thesis and body paragraphs are mostly cohesive. - Most ideas between and within paragraphs are linked with transitions.	- Information from sources is mostly relevant, and accurate but not quite enough/a bit too much. - Source information is mostly paraphrased/summarized in the student's words with adequate structural and lexical changes. - Source information is mostly cited properly.	- The essay includes an appropriate range of sentence styles and length with occasional structural and/or language problems. - An adequate range and variety of vocabulary is used mostly accurately. - There are occasional spelling, capitalization, and punctuation errors, which do not interfere with meaning.

125

2	- There is some attempt to address the prompt; supporting details were used occasionally. - The student's position is somewhat evident in the essay; opinions are hard to distinguish from source information OR only few arguments are supported with evidence from sources. - Acknowledgment of opposing view(s) might be present but unclear/insufficient.	- The thesis is present but unclear; introduction and conclusion are attempted but lack some required elements. - Inappropriate topic sentences may be present; body paragraphs show structure but lack cohesion. - Some ideas between and within paragraphs may be linked with transitions.	- Information from sources may be irrelevant/inaccurate; there is some evidence but not enough. - Source information is somewhat paraphrased/summarized but too similar to original; there might be some unacknowledged information from sources. - There are a number of problems with citation of source information.	- Sentences show structural and/or language errors and little variety. - Limited range and variety of vocabulary is used with some accuracy. - There are several errors of spelling, capitalization, and punctuation errors, which may interfere with meaning in places.
1	- The prompt is minimally addressed; the ideas are underdeveloped. - The student's position is rarely evident in the essay; there is no sense of individuality/no evidence is used to support arguments. - Opposing view(s) are not acknowledged.	- There is no discernible thesis; introduction and conclusion are missing/ lacking required elements. - Body paragraphs lack structure and cohesion. - Transitions are used minimally.	- Little or poor source evidence is given; the student relies mostly on quotes. - There are largely unchanged text chunks from sources/ much unacknowledged information suggesting plagiarism. - There are serious problems with citation of source information.	- Sentences show multiple and serious structural and language errors and no variety. - Very narrow range of vocabulary is used with pervasive errors. - There are frequent errors of spelling, capitalization, and punctuation; Little, if any, proofreading is evident.
0	- The essay shows no engagement with the topic.	- There is no indication of paragraphing.	- No evidence is provided; no academic referencing/citation conventions are followed.	- Intrusive and/or inaccurate language use which greatly impedes communication.